Automated Assessment of and Feedback on Concept Maps During Inquiry Learning

Karel A. Kroeze, Stéphanie M. van den Berg, Bernard P. Veldkamp, and Ton de Jong

***Abstract*—A tool is presented that can automatically assess the quality of students' concept maps and provide feedback based on a reference concept map. It is shown that this tool can effectively assess the quality of concept maps, and that it can provide accurate and helpful feedback on a number of specific shortcomings often evident in students' concept maps. However, it was also found that students who had access to feedback often did not request it, or did not take full advantage of the feedback given.**

***Index Terms*—Adaptive hypermedia, automated assessment tools, automated feedback, computer-aided instruction, concept maps, educational technology, inquiry learning, personalized learning, virtual labs.**

I. Introduction

CONTEMPORARY theories of learning prescribe that students should be actively involved in their own learning process [1]–[4]. Inquiry learning environments are designed to help students perform inquiries into science topics independently, which requires students to be actively involved in the learning process [5]. In following an inquiry cycle, students engage in various processes that aid the construction of new knowledge (e.g., see [6]).

Concept maps represent relations between concepts in the form of propositions. Two concepts connected by a relationship form a proposition, or a unit of meaning. Concept maps help students to organize and represent knowledge [7] and have been successfully used as instructional material (e.g., [8]–[10]) and for assessment (e.g., [11]–[13]). In addition to helping students to organize their knowledge, the creation of concept maps can also help teachers to form a picture of the mental models students have [14].

Manuscript received July 23, 2020; revised May 10, 2021; accepted August 5, 2021. Date of publication August 12, 2021; date of current version October 15, 2021. *(Corresponding author: Karel A. Kroeze.)*

Karel A. Kroeze was with the Departments of Instructional Technology and Research Methods, Measurement, and Data Analytics, University of Twente, 7500 AE Enschede, The Netherlands. He is now with the Behavioral Data Science Incubator, University of Twente, 7500 AE Enschede, The Netherlands (e-mail: k.a.kroeze@utwente.nl).

Stéphanie M. van den Berg and Bernard P. Veldkamp are with the Cognition, Data, and Education Research Group, University of Twente, 7500 AE Enschede, The Netherlands (e-mail: stephanie.vandenberg@utwente.nl; b.p.veldkamp@utwente.nl).

Ton de Jong is with the Instructional Technology Research Group, University of Twente, 7500 AE Enschede, The Netherlands (e-mail: a.j.m.dejong@utwente.nl).

Digital Object Identifier 10.1109/TLT.2021.3103331

Supporting students in increasing their knowledge at the start of an inquiry has been linked to better performance in the later steps of the inquiry process [15], and concept maps have been shown to be effective in filling this role; see [16].

While concept maps are widely used to scaffold student inquiries, students also have serious difficulties creating concept maps themselves [17]. In this context, there has been a substantial amount of research into automated assessment and feedback on concept maps. Some examples include Betty's Brain [18], [19], COMPASS [12], HIMATT [20], Conlon's Reasonable Fallible Analyzer (RFA) [21], and work on automated extensions for the concept mapping tools in the SCY [22] and WISE [23], [24] learning environments.

The tool we present and evaluate in this study builds on previous work by combining assessment and feedback based on network theory metrics and comparison to a reference map. The tool has two main functions: 1) give an overall quality score that can be used as an indicator in a higher level overview—for example, a teacher dashboard or adaptive learner model; and 2) generate actionable feedback to students on the process of creating a concept map and guide their attention to specific deviations from the reference map.

When designing adaptive tools, there is always a tradeoff between portability, flexibility, and accuracy. Portability is the ability of a tool to be deployed in varying domains, flexibility is the amount of freedom it allows students, and accuracy is the extent to which a tool can mimic expert teachers' judgments. The guidance provided in WISE [23], [24], for example, is informed by the detailed analysis of common mistakes made by previous students and the extensive collection of feedback prompts from experts. A set of carefully crafted rules provides accurate feedback when it detects specific misconceptions. However, the tool limits students to using a preconfigured set of concepts and propositions, and adapting it to a different domain would require collecting new data and feedback. COMPASS [12], HIMATT [20], AISLE [25], and RFA [21] all use a reference map created by an expert as the basis for assessment. HIMATT provides assessments and feedback based on the structural and semantic properties of students' entire concept maps, whereas both COMPASS and RFA also provide feedback based on individual deviations from the expert model. By reducing the structural and semantic quality of concept maps to a set of mathematical equations, these tools sacrifice some interpretability for increased portability. The use of a reference map also aims at portability, but assuming the reference map to be the absolute truth limits the ability of these tools to accurately assess students' knowledge.

In SCY, Weinbrenner *et al.* [22] instead used a domain ontology as the input for a reference model for their automated version of the concept mapping tool. Their tool infers important concepts and propositions from the information enshrined in the ontology and evaluates the overlap with students' existing concept maps. The use of an ontology allows students more freedom in the creation of their concept maps, but the creation of the ontology is a laborious process involving both domain experts and computer scientists.

As is the case for WISE, for the automated assessment and feedback in both COMPASS and HIMATT to work, students are limited to using only a preconfigured set of concepts and propositions, as these tools cannot disambiguate synonyms and spelling mistakes.

Overcoming this limitation is a main feature of RFA, which was specifically designed to let students argue for why their concepts and propositions are equivalent to those used in the reference map. This feature was also implemented for SCY.

The tool we present here is inspired by, builds upon, and combines what we consider to be the best features of the tools described above. We use a reference map as the basis of assessment and feedback to allow teachers to adapt the tool to different domains. An overall quality score is calculated using criteria derived from Ifenthaler's work on HIMATT. Our tool also identifies and provides feedback for individual deviations from the reference map. Unlike RFA, which provides a list of directive feedback, we take an approach similar to COMPASS and provide feedback prompts that encourage students to rethink (parts of) their concept map. Students are able to freely create concepts and propositions, as our tool is designed to recognize synonyms and spelling mistakes.

In designing the tool we present here, we let students create their own concepts and relations freely. This allows students full expression, and gives a rich representation of their mental models. By providing students with suggestions instead of directions, we encourage students to create their own concept maps while considering the expert suggestions given to them. The use of a single reference map for generating assessments and feedback, as is predominantly done in this research field [26], makes the tool easy to adapt to different domains and easier to deploy in educational practice. A reference map can easily be generated by teachers. In addition, feedback is immediate and integrated into the concept mapping tool, minimizing the distance between students' work and feedback in both time and space. Feedback is available at the student's request, and is then presented by an avatar who suggests possible areas of improvement (see Fig. 1).

The tool is embedded in the Go-Lab ecosystem [27], [28], an online sharing and learning platform for inquiry learning. In Go-Lab, teachers, researchers, and developers can create online inquiry learning environments (inquiry learning spaces, or ILSs) covering various domains related to science, technology, engineering, and math (STEM). Typically, inquiry learning spaces are built around a virtual or remote laboratory in which students can experiment. These labs are embedded in an environment that provides rich multimedia information (text, video) on the domain and software scaffolds that guide

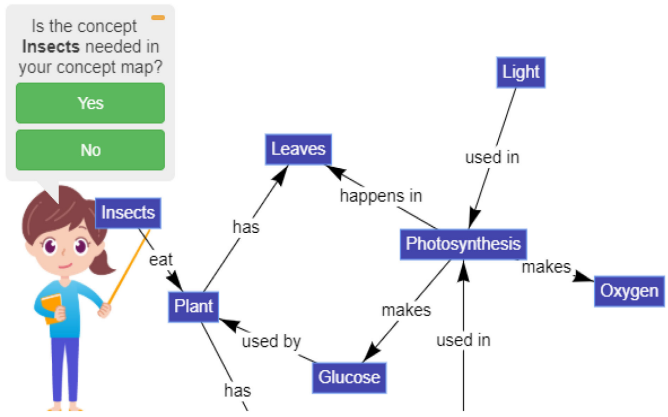

Fig. 1. Screenshot of feedback for a superfluous concept.

students in each step of the inquiry cycle: orientation, conceptualization (e.g., concept mapping and hypothesis creation), investigation (experimentation and observation), conclusion, and discussion [6]. Software scaffolds (or apps) are small tools that support a specific inquiry process, such as hypothesis generation or experiment design.

Creating concept maps is typically considered to be part of the orientation and conceptualization phases of the inquiry cycle, and therefore usually takes place at the start of an inquiry. By creating a concept map, students reactivate their prior knowledge and after receiving feedback, they can also create a solid basis for the following inquiry process. Experimental work and PISA 2015 results confirm that such a prior knowledge base is essential for effective inquiry learning [15], [29], [30].

If an improvement in concept map quality is accompanied by an improvement in students' initial mental models of the domain, we expect to see a corresponding improvement in the quality of students' inquiry processes.

To assess the effectiveness of our automated feedback, we first establish assessment criteria and to what extent automated assessments correspond with (human) experts' assessments. We then investigate whether and how students use feedback, and finally we assess if the quality of the inquiry process is affected. This leads to three main research questions.

1) Can the tool accurately assess the overall quality of a concept map?
2) Can the tool provide effective feedback to individual shortcomings?
3) Does using the tool improve the quality of students' inquiry learning?

The first question assesses the validity of automated quality scores. The tool provides a quality score based on network theoretical metrics as well as comparison to a reference map. For this score to have value, it should correspond to human assessments of the same concept map.

The next set of questions investigates the accuracy of the feedback, and students' interactions with the feedback.

2a) Can the automated tool accurately identify individual shortcomings?
2b) Can the automated tool provide accurate feedback?

2c) Can the automated tool provide helpful feedback?

From the students' perspective:

2a) How often do students request feedback?

2b) Do students respond to the feedback they are given?

If automated feedback improves students' conceptual understanding, that should improve the quality of inquiry artifacts at all steps in the inquiry cycle. This leads to the final set of research questions.

3a) Does automated feedback on concept maps improve the quality of concept maps?

3b) Does automated feedback on concept maps improve the quality of hypotheses?

3c) Does automated feedback on concept maps improve the quality of experimentation?

These questions were addressed in a field experiment. Students were randomly assigned to a learning environment with automated feedback or to a control condition without automated feedback.

II. Methodology

A. Participants

Five biology and physics teachers from three schools in the Twente region of the Netherlands participated. The experiment was performed in 12 classes (258 students) taught by the participating teachers. All participating students were in the first year of Dutch secondary schools, aged 12.3 to 13.8 years. The sample included students from the entire spectrum of Dutch educational levels, from the lowest to the highest (in Dutch): *kader* ($n = 34$), *VMBO* ($n = 53$), *HAVO* ($n = 128$), and *VWO* ($n = 21$). Students' parents had given prior active consent for research projects, and students were informed about the use of their data and given the opportunity to opt out at any point during the study. The research protocol was approved by the humanities and social sciences ethical review board of the University of Twente (#190073), and is fully GDPR compliant. No students opted out or objected to the use of their data. Data for 22 students who were not present during the experiments and four students who were present but whose login credentials could not be matched with background data were excluded from analysis. Analysis was performed on the data for the remaining 236 participating students. Their basic background information can be seen in Table I.

The experiment was designed and performed with the Dutch educational system in mind, and all textual and audiovisual materials were Dutch. The textual materials presented here have all been translated into English by the authors.

B. Learning Environment

The intervention consisted of an online learning environment that fit within students' regular curriculum. The intervention was performed in a single lesson which lasted 45–50 mins, depending on the school. Two inquiry learning spaces were created in close cooperation with teachers, one for the photosynthesis domain in biology, and one for the heat transfer domain in physics. The domains used were chosen in cooperation with the teachers, and selected because they fit the curriculum for students at all educational levels. Reference maps were created by the teachers involved, and deliberately tailored for a mid-level curriculum. Both spaces were identical in structure, presenting a short inquiry cycle: a basic introduction in text and video (orientation); information on concept maps and creating a concept map (conceptualization); creating a hypothesis, experimenting in a virtual lab about the domain, and writing down observations (investigation); writing down conclusions (conclusion); and finally a short questionnaire about students' experiences with concept maps in general, and the concept mapping tool developed in the current study in particular. Each student worked in only one of the learning environments, depending on the subject taught by the participating teacher.

The learning environment included several scaffolds to support students in each phase of their inquiry process: a modified version of the concept mapping tool (the subject of this study), a modified version of the hypothesis scratchpad, and a tool to write down observations [28], [31]. The photosynthesis learning environment included a virtual laboratory that allowed students to manipulate a light source and ambient temperature, and measure the amount of oxygen produced [32]. The heat transfer learning environment included a virtual laboratory that modeled convection in a closed room, and allowed students to move heater, walls, and windows around to see the effects on convection patterns [33].

TABLE I
Background Characteristics of Participating Students, by Experimental Condition, and Feedback Use

	Control	Experimental	
Feedback used?	No	No	Yes
Participants (count)	92	78	66
Female (%)	47.8%	43.6%	50.0%
Age [mean (SD)]	13 (0.75)	13 (0.44)	13 (0.42)
Grade[a] [mean (SD)]	6.9 (0.78)	6.9 (0.74)	7.1 (0.71)
Level			
Kader, $n = 34$	18 (20%)	12 (15%)	4 (6%)
VMBO, $n = 53$	19 (21%)	18 (23%)	16 (24%)
HAVO, $n = 128$	50 (54%)	39 (50%)	39 (59%)
VWO, $n = 21$	5 (5%)	9 (12%)	7 (11%)

[a]The Dutch educational system uses grades ranging from 1–10, where one is the lowest and ten the highest. A five-point-five is typically the threshold for a passing grade, and an eight is considered excellent. Scores over eight are exceptional and a perfect ten is rarely given.

C. Procedure

The experiment was performed in students' regular classrooms, under the supervision of their class teacher and the first author. After a brief verbal introduction covering the study and the use of students' data, students were instructed to visit a preset URL. From this URL, they were randomly assigned to the control or experimental condition (using [34]). Students were thus randomly assigned within classes. The learning environments for the control and experimental condition differed only in the availability of feedback. The control condition included the concept mapping tool without feedback,

whereas the experimental condition included a version of the concept mapping tool with feedback. In earlier work [35], we found that roughly half of students who had feedback available chose not to use it. We therefore assigned roughly two-thirds of students to the experimental condition, with the expectation that about half of all students in the experimental condition would actually use feedback. This theoretically would leave us with three roughly equal-sized groups: students who could not use feedback, students who could ask for feedback but did not use it, and students who could ask for feedback and chose to do so. The roughly equal distribution of students across these three groups shown in Table I shows that this approach was successful.

In previous work [35], we found that letting students explore in a similar learning environment on their own took too much time, and some instruction on how to use the environment was needed. Students were therefore given an oral description of the apps, labs, and resources available to them in the learning environment at the beginning of each phase, and directed to wait for further instructions after completion of each phase. However, this structuring was not present in the learning environment itself, and students were free to move backward and forward between phases as they pleased. Students were instructed to work alone. Where possible, seats and tables were moved to "exam positions," creating a setup where students could not see each other's screens. Both the teacher and the primary author were present to stop students cooperating or distracting each other, and to answer process-related questions.

D. Assessment

Two forms of assessment of concept maps took place: an overall quality score and criteria for a number of specific shortcomings. While there is some overlap between the two forms of assessment, they serve different purposes.

The overall score is calculated to provide a "high-level" indication of the quality of students' concept maps. This score can then be used in teacher dashboards, or as an indicator in learner models for adaptive learning environments.

Specific criteria are used to generate feedback to be presented to students, with the aim of directly improving specific aspects of students' concept maps and the quality of their inquiry in general.

The following two sections give further details on both forms of automated assessment.

1) Quality Score: Concept maps are a form of directed graph, and there is a rich literature on their analysis. It is not our aim to expand upon that literature, but we will introduce some of the mathematical terminology used, in the process of explaining the assessment criteria used.

In our study, concept maps are scored on five criteria derived from Ifenthaler's work [36] and commonly used graph and network theory measures. The first three criteria provide an overview of the complexity of concept maps without depending on domain knowledge or semantic content, while the final two give an estimate of the extent of overlap between student and expert concept maps. Crucially, all of these criteria can be automated.

TABLE II
FEEDBACK PROMPTS

Shortcoming	Priority	Feedback prompt	Response options
Typo	1	Did you mean [label]?	Oops, yes [a] / That's the same thing [b] / No
Superfluous concept	1.5	Is [concept] really necessary in your concept map?	Yes / No
Superfluous proposition	1.5	Does this relation add anything to your concept map?	Yes / No
		Are [concept] and [concept] related?	Yes / No
Concept density	2	Start by adding the most important concepts	Ok / Which concepts? [c]
Missing concept	2/3[c]	Did you forget a concept?	Maybe? [d] / No
	2/3[c]	Could [concept] be important in your concept map?	Yes / No
Proposition density	2	Now try to add the relations between these concepts.	Ok / I don't understand [d]
Missing proposition	2/3[c]	Is there a relation between [concept] and [concept]?	Yes / No
Reversed proposition	3	Are you sure [link] is in the correct direction?	Yes / No
Mislabeled proposition	3	Are you sure [link] is the correct label for this relation?	Yes / No
		Could this relation be called [link]?	Oops, yes [a] / That's the same thing [b] / No
Shortcut	3	Is there another concept in between this relation?	Maybe? [d] / No
		Is [concept] important in this relation?	Yes / No

English translations of the Dutch feedback prompts. Words surrounded by square brackets are placeholders and replaced with relevant content when the feedback prompt is presented.
[a]If chosen, the students' label is automatically corrected.
[b]If chosen, the current label is added as a synonym.
[c]Feedback can be given as a follow-up to density feedback with priority two, or as separate feedback with priority three.
[d]If chosen, the next feedback prompt for this criterion is immediately shown.

A concept map M is an ordered pair $M = (V, E)$, where V is a set of nodes and E a set of edges. Each node v_i represents a concept and each edge e_{ij} a proposition from node v_i to v_j. Each node v_i and proposition e_{ij} in a concept map has a label l_i or l_{ij} associated with it. M_s, V_s, and E_s are used to denote a student concept map and the sets of concepts and propositions in it. M_x, V_x, and E_x denote the reference map, concepts, and propositions.

Before we can compare concept maps, we use labels to match students' concepts and propositions to those in the reference map. We dealt with alternate labels and common misspellings of the concepts and propositions used in the reference map by creating an *a priori* list of synonyms, and by suggesting corrections to the student when student labels were similar to known labels or synonyms. The student then had the option to correct the misspelling, or to add her version as a synonym (see also the "mislabeled link" prompt in Table II).

Density of concepts is calculated as the proportion of concepts in the reference map that are present in the student map

$$D_{concepts} = \frac{|V_s \cap V_x|}{|Vx|} \ .$$

Density of propositions is the number of propositions in the student map as a proportion of the total number of propositions that were possible in the student map compared to the same measure in the reference map

$$D_{propositions} = \frac{|E_s|}{|E_s^*|} \ / \frac{|E_x|}{|E_x^*|}$$

where E^* is the set of all possible propositions in M

$$E^* = \left\{ e_{ij} i \in \mathbb{N}, j \in \mathbb{N}, i \leq |V|, \ j \leq |V|, i \neq j \right\} \ .$$

Matching structure is the length of the longest continuous chain of propositions in the student map, relative to the same measure in the reference map. If $d(i, j)$ is the distance between nodes v_i and v_j, then the length of the longest continuous chain is given as

$$\mu \ = \max_{ij} \ \ d(i, j).$$

The matching structure is simply

$$\frac{\mu_s}{\mu_x}.$$

The loose and strict deep structure criteria give an estimate of the overlap between student and reference maps, including correct, missing, and superfluous propositions. Specifically, they are a Tversky index

$$\sigma_{sx} = \frac{|E_x \cap E_s|}{|E_x \cap E_s| + \alpha |E_x - E_s| + \beta |E_s - E_x|}$$

where $|E_s \cap E_x|$ is the number of propositions present in both the student and reference maps, $E_s - E_x$ denotes the relative complement of E_x in E_s, that is, the propositions present in the reference map that are missing in the student map, and $E_x - E_s$ is the relative complement of E_s in E_x. Setting values for α and β allows for assigning more weight to missing propositions, or conversely, to superfluous propositions. When $\alpha = \beta = 0.5$, as we have done, equal weight is assigned to both types of imperfections, and the Tversky index is equal to the Sørensen–Dice coefficient.

The loose and strict variants differ in what propositions they consider equal. In the strict variant, a proposition e_{ij} is considered equal to proposition e_{kl} if and only if they connect the same nodes, in the same direction, and have the same label

$$e_{ij} \equiv e_{mn} \text{ if, and only if } i = m, j = n \text{ and } \ l_{ij} = l_{mn}.$$

In contrast, the loose variant ignores the label and direction of a proposition, and simply asserts that two concepts are related. That is

$$e_{ij} \equiv e_{mn} \text{ if } i = m \text{ and } j = n, \text{ or } i = n \text{ and } j = m.$$

A score on a 0–1 scale was calculated for each criterion. Principal component analysis (PCA) showed that much of the variance in scores on these five criteria could be explained by a single underlying "quality" component. With the PCA loadings, a single "quality score" was calculated for each concept map. As an automatically calculated score largely based on network metrics, this score is by definition a "black box" score. The extent to which the score corresponds with human experts' assessment of concept map quality was evaluated by comparing the scores to those given by human experts.

2) Specific Shortcomings: Criteria were developed to assess specific shortcomings: typos, missing concepts, superfluous concepts, mislabeled propositions, reversed propositions, missing propositions, superfluous propositions, and shortcuts. Most of these are self-explanatory. Missing and superfluous concepts and propositions are determined by comparing students' concept maps to the reference map. Typos are determined by comparing student concepts to similar concepts (and their synonyms) in the reference map. Shortcuts are defined as propositions that are missing intermediate concepts. For example, the proposition "plants have photosynthesis" should include the concept of leaves: "plants have leaves" and "photosynthesis takes place in leaves." Mislabeled and reversed propositions are determined by comparing labels for student propositions with those for propositions in the reference map, and identifying student propositions that contain identical concepts to those in the reference map, but either have a different label, or reverse the direction of the proposition.

Using these criteria, the automated tool can identify any number of shortcomings for any concept map. Feedback prompts were developed to give feedback to students based on these individual shortcoming criteria and the concept and proposition density criteria described in the previous section.

Given that several hundred shortcomings could be identified when the student requests feedback, a feedback selection mechanism is required so as not to overwhelm the student. Each prompt provides feedback with a fixed priority, and a variable weight based on the importance of the affected concepts in the concept map (see Table II). Any time a change is made to the concept map, all possible feedback is created, sorted by priority and then weight, and stored for later use. When the student requests feedback, the relevant prompt with the highest priority and weight that has not yet been used is presented to the student. Concept importance can either be set by the teacher or estimated based on the concept's centrality in the reference map.

Each feedback prompt presented to the student presents one or more response options. These options usually include an "agree" and a "disagree" option, and sometimes a "more information" option. Making a significant change to the con-

cept map or clicking one of these options will dismiss the feedback and suppress further feedback for a set period. Once feedback has been dismissed, that specific feedback prompt will never reappear. The orange button in the top right of the feedback prompt (see Fig. 1) will suppress feedback for 30 seconds, but not dismiss the current feedback (it will remain a valid feedback option, and may reappear). See Table II for an overview of the possible feedback prompts for each criterion.

Given that each criterion targets a specific discrepancy between student and reference concept maps, repeated requests for feedback would eventually allow students to make perfect copies of the reference map, regardless of their knowledge of the domain or ability to make concept maps. To prevent students from using feedback as a step-by-step tutorial and never actually reflecting on the concept map created, no further feedback could be requested for 30 seconds after receiving feedback.

E. Data

The Go-Lab ecosystem allows teachers and researchers to log and store students' interactions with the learning environment. The resulting log files provide a wealth of information about students' learning processes and products. For this research, we collected students' interactions with the concept mapper and their feedback requests and prompts, as well as the contents of the hypothesis scratchpad and the observation tool. Students' age, gender, and current grade in the class (biology grade for students in the *photosynthesis* groups, and physics for students in the *heat transfer* group) were collected as possible control and interaction variables.

Students were asked to identify themselves by their student numbers, a quasi-anonymous means of identification. Some students used personally identifiable credentials (e.g., names, e-mail addresses), but these were removed soon after gathering the data. Students' responses were linked to their background information based on their student numbers; at no point was sensitive personal data collected or stored. Data were stored in compliance with the University of Twente ethical and data management policies, which are fully compliant with the European Union's General Data Protection Regulation (GDPR).

F. Measures

Several outcome measures were used to answer the research questions. Many of these measures rely on human coding of the processes and artifacts created by students and extracted from the log files. Given the amount of materials to be coded, coding was done by the authors and other experts. Interrater reliability between coders, and whenever possible between humans and the automated tool, was calculated and reported with either Cohens' κ or where appropriate, a correlation (Pearson's r). The scoring procedures used in relation to each research question are explained below. Two sets of measures are distinguished, the first relating to the accuracy and quality of feedback (measures 1–7), and the second to the effect of feedback on students' learning process (measures 8–9).

The tool assigns scores and calculates feedback each time a change is made to the content of a concept map, leading to 13 269 parsed concept map "snapshots" in our dataset. Only those snapshots taken at crucial moments were scored: those for the final concept maps ($n = 236$), and before and after requesting feedback ($n = 688$, from which 287 unique snapshots were scored—snapshots that were identical to a snapshot that was already scored were not scored again).

TABLE III
CODING CRITERIA FOR CONCEPT MAPS

Criterion	Category	Explanation
Comprehensiveness [a]	Weak	The concept map contains three or fewer of the concepts in the reference map
	Average	The concept map contains four to six of the concepts in the reference map
	Strong	The concept map contains seven or more of the concepts in the reference map
Correctness [b]	Weak	The concept map contains three or fewer of the propositions in the reference map
	Average	The concept map contains four to six of the propositions in the reference map
	Strong	The concept map contains seven or more of the propositions in the reference map
Understanding [c]	1–10[d]	Based on this concept map, how would you grade the students' understanding of the domain? Score as you would a test for this domain.

[a]A list of core concepts present in the reference map was provided; coders were instructed to interpret all concepts not on this list as incorrect.
[b]A domain-specific list of key misconceptions and core relations present in the reference map was provided, coders were instructed to interpret all propositions not in the reference map as incorrect.
[c]A domain-specific list of core competencies was provided. Coders were instructed to ignore additional content in student maps that was not present in the reference map.
[d]Where one is the minimum and ten is the maximum score. See also the footnotes for Table I.

To identify changes before and after feedback, the most recent snapshot before receiving feedback was compared with the next available snapshot at least 30 seconds and at most 90 seconds after having received feedback. These boundaries were set to ensure that any changes were likely made as a result of the feedback, and so that students had some time to actually implement the feedback. Simply taking the next available snapshot would have included partial responses to the feedback (e.g., moving a concept before adding a new proposition). If no snapshot existed within these boundaries, the student was assumed not to have reacted to the feedback.

1) Can the Automated Tool Accurately Assess the Overall Quality of a Concept Map?: The quality score generated by the automated tool is meant to provide an indication of concept map quality, but is based in large part on graph-theoretical indicators that represent the mathematical structure of a concept map, not the mental model it represents. In order to assess to what extent the abstract "black box" overall quality score corresponds with expert impressions, concept maps were scored by experts on a 1–10 scale.

In addition, to provide some context for the indicators underlying the quality score, concept maps were also scored by experts on their comprehensiveness and correctness (Table III). The number of concepts and propositions required for *weak*, *average,* and *strong* codes using these criteria was

determined based on teacher input and discussions between coders during iterative coding of early results. There reference maps used in both domains were crafted to use the same number of concepts and propositions and the same coding scheme was used for both domains.

The association between expert and automated scores was determined using Pearson's r. Associations between expert assessments of the completeness and correctness and automated scores was determined using Spearman's ρ.

2) Can the Automated Tool Accurately Identify Individual Shortcomings?: The accuracy of identifying shortcomings was established based on simple agreement between the automated parser and human coders. The parser often identified dozens of shortcomings each time a change was made to the concept map, leading to thousands of identified shortcomings for each student. Students were only presented with feedback for the shortcoming with the highest priority and weight whenever they requested feedback, and thus were only exposed to a tiny fraction of the total number of identified shortcomings. Only those shortcomings that were the basis of feedback presented to the student were considered in the reliability analysis. This set by definition only includes shortcomings identified by the parser, and omits any "true" shortcomings that were not identified by the parser. Consequently, the outcome measure here is a simple agreement rate.

3) Can the Automated Tool Provide Accurate Feedback?: The accuracy of feedback was determined based on human coders' assessments, scoring each feedback prompt presented to students as correct or incorrect. Note that by only looking at shortcomings identified by the parser, it is not possible to obtain a false negative rate. Consequently, the outcome measure here is a simple agreement rate.

4) Can the Automated Tool Provide Helpful Feedback?: Even though feedback is technically accurate, that does not necessarily mean that it is helpful. Feedback helpfulness was determined by human coders scoring each feedback prompt as helpful or not helpful, where helpful feedback is defined as feedback that would make a concept map meaningfully better if acted upon. As above, the outcome measure is a simple agreement rate.

5) How Often Do Students Request Feedback?: The number of feedback requests was determined by the number of times a student opened the feedback dialog by clicking on the avatar. Feedback requests in the first 30 seconds of opening the tool were dismissed as exploration of the tool by students, as were feedback requests made after students had already completed their conclusion. The outcome measure is then the number of feedback requests made while students were, presumably, meaningfully engaged with the concept mapping tool.

6) Do Students Respond to the Feedback They Are Given?: Students' responses to receiving feedback were coded by comparing their concept map at the time of requesting feedback, to the first snapshot taken at least 30 seconds and at most one minute later. If a change was made that affected the concepts and/or propositions mentioned in the feedback, it was coded as having acted as a result of that feedback. As with criteria one through five, the outcome measure is a simple agreement rate.

TABLE IV
CODING CRITERIA FOR EXPERIMENTATION AND CONCLUSION

Criterion	Category	Explanation
Observation	Weak	student has failed to report observations in a way that provides evidence towards answering the research question
	Average	student reports observations that provide evidence towards the research question, but made no attempt to discover alternative explanations
	Strong	student reports observations that allow conclusively answering the research question, and rule out alternatives
Measurement	Weak	student reports vague or imprecise measurements
	Average	student reports some effort to make precise measurements
	Strong	student reports precise and specific measurements
Interpretation	None	student has failed to draw any conclusions
	Invalid	student has drawn conclusions for which there was no evidence in the observations made
	Missing	student has failed to draw conclusions for which there was evidence in the observations
	Partial	student has drawn conclusions for which there was only partial evidence in the observations made
	Strong	student has drawn conclusions that are fully supported by the observations, and fully explain the observations.

7) Does Automated Feedback on Concept Maps Improve the Quality of Concept Maps?: The effect of the presence of the automated feedback tool on concept map quality was determined by comparing the grades given by human coders to the final concept maps of students in the control and experimental groups. Given that a large proportion of the experimental group did not request (or therefore use) feedback, two further contrast analyses were done. First, to investigate the effect of using feedback, students who used feedback were contrasted with students who did not use feedback (those in both the control and experimental conditions). Second, students in the experimental condition who did not use feedback were contrasted with students in the control condition (who could not use feedback). If there was a significant difference between these groups, which would indicate that there was a confounding variable that might explain why these students did not use feedback. Perhaps students who chose not to use feedback were simply less able than their peers.

To assess the effectiveness of feedback based on individual criteria, we compared the automatically generated quality score before and after receiving feedback based on each criterion. A multilevel model was used to model the difference in quality of concept maps before and after receiving feedback. Class and domain were included as random effects to account for interclass correlation effects. Students' gender, age, and grade for the domain were controlled for by including them as fixed effects. Marginal means were then estimated to give an overview of the effectiveness of each criterion.

TABLE V
LINEAR REGRESSION OF UNDERSTANDING SCORES ON AUTOMATED QUALITY CRITERIA INDIVIDUALLY, AND AGGREGATED QUALITY SCORE

Predictor	β	[95% CI]	r
Full model			
(Intercept)	-3.98**	[-5.60, -2.35]	
Loose deep structure	1.20**	[0.64, 1.77]	.63**
Strict deep structure	0.86*	[0.02, 1.71]	.44**
Matching structure	1.88**	[1.47, 2.28]	.66**
Density of concepts	0.65**	[0.23, 1.07]	.56**
Density of links	7.67**	[5.78, 9.56]	.67**
R^2	.679**	[0.62,0.72]	
Aggregated model			
(Intercept)	1.39**	[1.03, 1.74]	
Score	3.23**	[2.95, 3.51]	.78**
R^2	.607**	[0.55,0.66]	

Regression coefficients for a full and aggregated model. β indicates unstandardized regression weights. A 95% confidence interval for the coefficients is given between brackets. r represents the zero-order correlation of the predictor with understanding scores, R^2 the proportion of variance in understanding scores explained by the model. Significant results are marked * if $p < 0.05$, and ** if $p < 0.01$.

TABLE VI
COUNT AND PROPORTION CORRECT, HELPFUL, FOLLOWED, AND IMPROVED FOR EACH FEEDBACK TYPE

Feedback	N	Correct	Helpful	Followed	Improved
Concept density	25	80.0%	80.0%	52.00%	48.00%
Proposition density	18	100.0%	100.0%	27.78%	27.78%
Missing proposition	63	96.8%	96.8%	41.27%	34.92%
Missing node	96	97.9%	97.9%	57.29%	55.21%
Reversed proposition	15	46.7%	46.7%	40.00%	33.33%
Shortcut proposition	15	20.0%	20.0%	40.00%	20.00%
Superfluous proposition	39	66.7%	64.1%	35.90%	30.77%
Superfluous concept	33	72.7%	69.7%	24.24%	24.24%
Typo	4	75.0%	75.0%	25.00%	25.00%

8) Does Automated Feedback on Concept Maps Improve the Quality of Hypotheses?: Hypothesis quality was determined with the criteria described in [35], using a newly created grammar for each of the two domains involved [37]. As above, quality of hypotheses was compared between the control and experimental groups, and further contrasts were made between students who used feedback and those who did not, and between students who did not use feedback when it was available to them, and those for whom feedback was not available. A binomial multilevel model was used to model the syntactic correctness of hypotheses, with class and domain as random effects and gender, age, and grade as fixed effects.

9) Does Automated Feedback on Concept Maps Improve the Quality of Experimentation?: The quality of observations and measurements made during experimentation was determined using the criteria described in Table IV. These criteria were again compared between the control and experimental groups, as well as contrasts for the (lack of) use of feedback. An ordinal multilevel model was used to model the quality of observations and measurements. A multinomial multilevel model was used to assess students' interpretations. In both cases, class and domain were added as random effects, and gender, age, and grade as fixed effects.

III. RESULTS

A. Overall Quality of Feedback

A subset of 57 concept maps was scored by an additional expert. Experts reached an agreement of 93% (Cohen's $\kappa = 0.88$) on the comprehensiveness criterion, and 74% ($\kappa = 0.56$) on the correctness criterion. There was a strong correlation (Pearson's $r = 0.89$) between the grades given for the understanding criterion by both experts.

Table V shows the outcomes of fitting a linear regression model to predict understanding scores by the five indicators provided by the automated quality assessment. This model was able to explain 68% of the variability in understanding scores, but it was based on the relatively small sample of snapshots that were manually scored.

To incorporate information from all 13 268 concept map snapshots, PCA was performed on the five indicators provided by the automated quality assessment. The PCA revealed that a single underlying component explained 58% of the variance present in the indicator scores across all snapshots. Using the loadings on this component (0.51, 0.4, 0.46, 0.42, and 0.44 for the loose deep structure, strict deep structure, matching structure, density of propositions, and density of concepts criteria, respectively), a single aggregated quality score was calculated.

As shown in the bottom part of Table V, this aggregated quality score had a strong and significant correlation (Pearson's $r = 0.78$) with the understanding grades given by human experts. There was also a significant correlation between overall quality score and experts' assessments of comprehensiveness (Spearman's $\rho = 0.76$) and correctness ($\rho = 0.75$).

B. Accuracy of Identified Individual Shortcomings

In total, 83% of feedback prompts were considered accurate by human coders, who reached 89% agreement, with a Cohen's κ of 0.47. For most of the nine criteria used (Table II), human coders agreed with the algorithm when it detected a shortcoming (see the "Correct" column in Table VI). The main exceptions are the reversed and shortcut proposition criteria, where human coders agreed with the algorithm 47% and 20% of the time, respectively.

C. Helpfulness of Feedback Prompts

Overall, 82% of feedback prompts were considered helpful by human coders, who had 84% agreement (Cohen's $\kappa = 0.52$). As might be expected, feedback that was incorrect was almost never considered helpful.

D. Quantity of Student Feedback Requests

Of the 144 students able to request feedback, 92 (64%) requested feedback at least once and 66 (46%) requested

feedback multiple times. In total, 308 requests for feedback were made.

E. Student Responses to Feedback

In a total of 134 cases (44%), students were judged to have made a change to their concept map after receiving feedback. Coders reached 83% agreement with a Cohen's κ of 0.65.

Students' responses seem to be at least in part related to the correctness and helpfulness of the feedback received. When feedback was accurate, it was overwhelmingly coded as helpful, and when feedback was inaccurate, it was coded as unhelpful. When feedback was helpful, students made changes to their concept maps in 45% of cases. However, unhelpful, and incorrect feedback still encouraged students to make changes to their concept map in 37% of cases. In particular, this was the case for the reversed-proposition and superfluous-concept criteria, where 50% of the unhelpful feedback still led students to respond.

F. Effect of Feedback on Concept Map Quality

There was no statistically significant difference between experimental and control conditions in a multilevel model for overall quality scores of students' final concept maps ($\beta_{condition} = -0.07, SE_{\beta} = 0.04, CI = [-0.15, 0.02], p = 0.135$). Furthermore, neither the difference between students who used feedback and those who did not ($\beta_{feedbackUsed} = -0.09, SE_{\beta} = 0.06, CI = [-0.20.\ 02], p = 0.125$) nor that between students who did not have feedback available and those that had feedback available but did not use it ($\beta_{feedbackAvailable} = -0.05, SE_{\beta} = 0.05,\ CI = [-0.15, 0.04], p = 0.284$) was significant.

Human experts judged improvements to have been made after 39% of feedback prompts. Improvements followed a pattern similar to students' responses, with helpful and unhelpful feedback leading to improvements being made in 41% and 33% of cases, respectively. Feedback based on the concept density criterion led to the concept map being improved in 48% of cases, whereas that based on the shortcut and superfluous concept criteria only led to improvements in 20% and 24% of cases, respectively. Incorrect feedback based on the reversed proposition or superfluous concept criteria led to improvements being made more often (47%) than correct feedback based on these criteria (16%).

When comparing quality scores directly before and shortly after feedback based on the different criteria, only the missing concept criterion led to a significant change in quality, which increased by 0.05 to 0.13 after each feedback prompt. Fig. 2 shows the estimated marginal means for quality gain for each criterion. The horizontal bar represents a 95% confidence interval, and the grey dots mark the underlying data.

G. Effect of Feedback on Hypothesis Quality

The quality of hypotheses was scored using an automated tool developed in previous research [35]. Neither the effect of being in the experimental group ($\beta_{condition} = -0.09, SE_{\beta} = 0.3, CI_{OR} = [0.51, 1.64], p = 0.767$), nor that of using feedback ($\beta_{feedbackUse} = -0.14, SE_{\beta} = 0.4, CI_{OR} = [0.41.91], p = 0.734$) was significant. We found no effect of concept map quality on the quality of hypotheses ($\beta_{CM_{quality}} = -0.16, SE_{\beta} = 0.48, CI_{OR} = [0.33,\ 2.17], p = 0.733$).

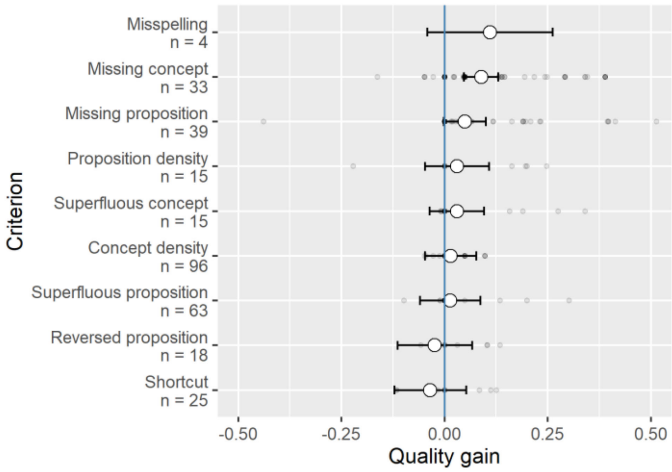

Fig. 2. Estimated marginal mean quality gain for concept maps after receiving feedback, per criterion.

H. Effect of Feedback on Experimentation

We found no significant effect of experimental condition, feedback use or concept map quality on the quality of observation, accuracy of measurement and validity of interpretations. Students in the highest level of education did perform significantly better than their peers in lower levels of education on the observation ($\beta_{VWO} = 3.56, SE_{\beta} = 1.61, CI_{OR} = [1.51821.41],\ p = 0.027$) and interpretation ($\beta_{VWO} = 2.88, SE_{\beta} = 1.27, CI_{OR} = [1.47213.87], p = 0.023$) criteria, but educational level was not associated with the accuracy of measurements.

IV. Discussion and Conclusion

We set out to help students create better concept maps, with the assumption that students who created better concept maps would have a clearer overview of the domain, and therefore be better able to conduct inquiries independently. We created an automated tool that can parse concept maps, assess their quality, find individual shortcomings, and provide feedback to students.

We answered three main questions, centered on the validity of automated quality scores, the quality of feedback, and the effectiveness of providing feedback for the quality of students' inquiry process. We performed an experiment in which students were assigned to two conditions: two thirds were assigned to the experimental condition, with the opportunity to ask for automated feedback; the remaining third was assigned to the control condition where there was no automated feedback available. We expected about half of the students in the experimental condition to use the feedback, leading to three groups: control, experimental with feedback

use, and experimental without feedback use. Roughly half of the students in the experimental condition used the automated feedback during their inquiry, forming three roughly equal groups aligned with our expectation.

Using PCA, we found that a single component explained about half of the variance in the individual quality criteria calculated by the automated tool. This quality component was highly correlated with grades given by human coders, over the entire process of creating concept maps. This gives strong evidence that the tool presented here could be a valuable addition to student models and/or teacher dashboards.

The comparison between machine and human assessments was made to validate the automated products, as well as to seek meaning in those assessments. We did not expect to find perfect correlations between human and machine measures, as they may take somewhat different approaches toward understanding a concept map. We were therefore less interested in the specific scores obtained, and more interested in the rankings of scores. If both human and automated assessment can reliably identify lower quality concept maps, that would allow tutors (whether they are human or automated) to focus their attention where it is most needed.

This information can be part of dashboards providing overviews of students' activities and products. Overview studies [38] have shown that these dashboards often display many indicators, but often at a more superficial level (e.g., screen activities) and, for that reason, use of the dashboards in practice is relatively low (e.g., [39]). A very recent overview study of research on learning analytics dashboards [40] established this conclusion: learning dashboards have thus far been reported to have low impact. Interviews with users confirmed this low usage and indicated that users expect a more detailed, cognitive assessment of the learning process to be displayed [41], in which the concept map quality indicator we have developed here could be included.

For the detailed diagnosis of the concept maps, the overwhelming majority (83%) of shortcomings in students' maps identified by the automated tool were confirmed by human coders. This level of accuracy is comparable to that reported by other automated tools [21]. The only exceptions were the reversed proposition and shortcut criteria. The low accuracy of the reversed proposition criterion can be explained by the fact that with an appropriate change of label, any proposition in a concept map can be reversed. The shortcut criterion was meant to detect missing intermediate concepts in a proposition, but did not function as expected. With the data collected here, this criterion could be improved for future work.

Incorrect or unhelpful feedback may have had a negative effect on students' willingness to incorporate feedback and request further feedback. This is particularly true for the reversed proposition and shortcut criteria, which produced feedback that was incorrect more often than it was correct. However, we also found that even when students received incorrect and/or unhelpful feedback, they occasionally still made changes to their concept maps. When coding, we noticed that in these cases, students often clarified their point of view by improving the labels for relations, and/or moving concepts so that they had a more logical structure. This may also explain why incorrect feedback sometimes still led to improvements being made.

Fig. 3 shows two examples of feedback events, with the students' initial concept map and the feedback presented in the left panels, and their concept map 30 seconds to one minute later in the right two panels. The top panels show feedback that was both correct and helpful, and to which the student responded by adding the suggested proposition. The bottom two panels show an incorrect feedback prompt, suggesting that the student add "glucose" to an existing proposition. While the prompt itself was incorrect, the student did make changes related to the prompt that improved the overall quality of the concept map.

We often found that students added all relevant concepts to their concept maps at the start of creating the concept map. This behavior was seen in students in both experimental conditions. We suspect that a scaffolding function in the concept mapping tool is responsible, whereby students get a list of suggested concepts to add whenever they add a concept. Given that this scaffold appears to be quite effective and was present in both experimental conditions, it most probably reduced the effectiveness of feedback based on the concept density quality criterion. Indeed, a ceiling effect is clearly visible for scores on this criterion, with almost two-thirds of students obtaining a perfect score. This ceiling effect disappears when we look at the combined score based on all five quality criteria. Use of this concept-adding scaffold may also partly explain the quality gain following feedback based on the concept density criterion shown in Fig. 2. It is possible that the gains we detected are because feedback on this criterion was typically given early in the concept map creation process when students were more likely to subsequently make improvements to their concept maps. For future research, interesting avenues concern the differential effects of scaffolding, feedback, and the use of nudging techniques [42] for students with different backgrounds.

Although students used the tool, and the tool was able to give accurate quality scores, accurately identified shortcomings, and provided helpful feedback, we were unable to detect any effect of the availability or the use of feedback on any subsequent inquiry activities—concept map quality, hypothesis quality or the quality of experimental designs. We also found no relation between concept map quality and the quality of hypotheses or experimental designs, products developed later in the inquiry cycle. Given the overwhelming amount of evidence to support the efficacy of inquiry learning, we suspect that a single session of 45–50 mins was simply not enough time for a measurable effect on the inquiry process to occur. In studies in which the learning time was considerably longer (e.g., 12 hours, see [43]), an effect on learning was found from providing students with automated feedback on a concept map. Our study also has some limitations. A practical limitation of our study that we noticed concerned cooperation between students. While we made efforts to limit students' cooperating, placing them in "exam conditions" was not always feasible, for example, because classes took place in a

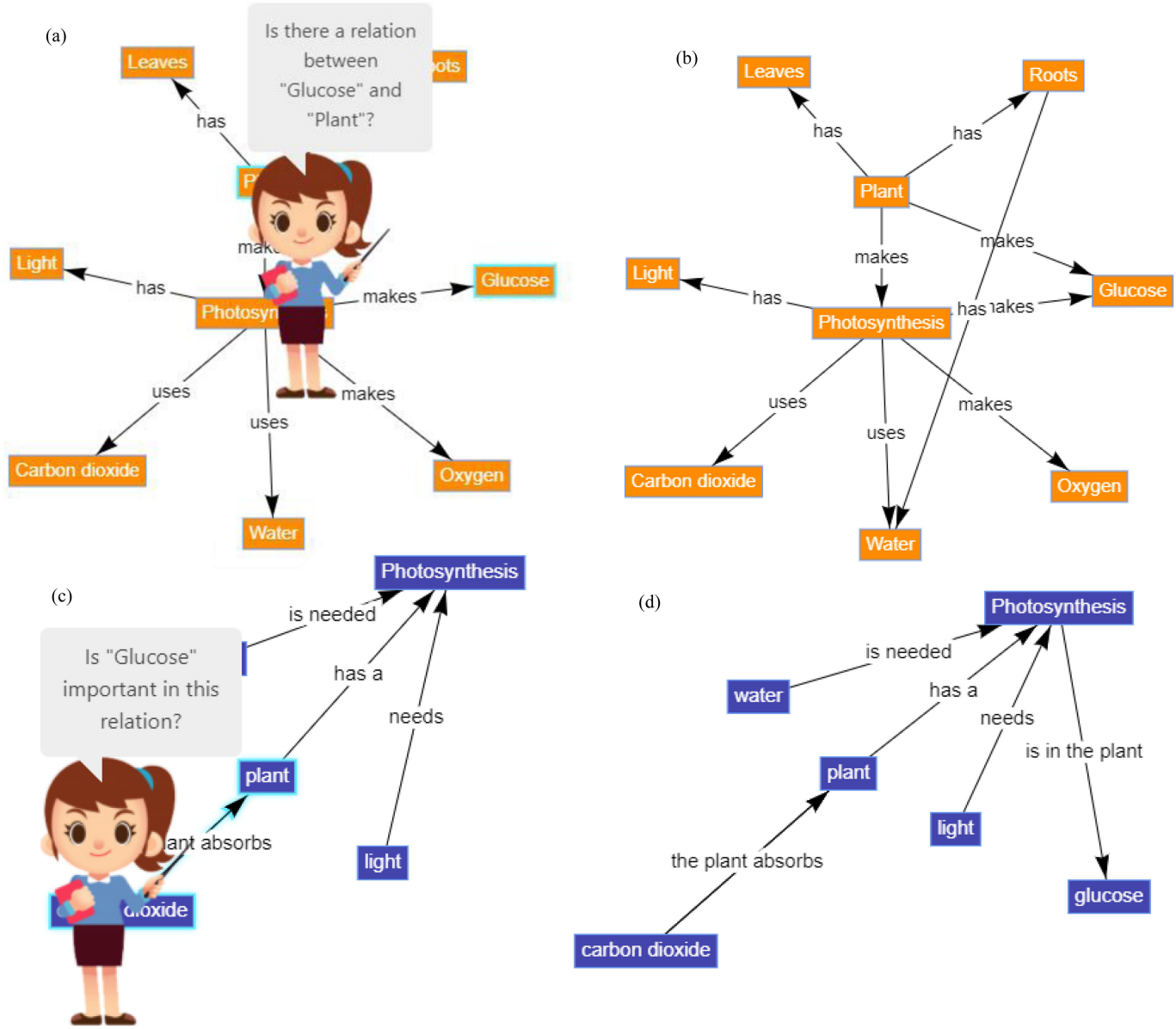

Fig. 3. Two examples of feedback events triggering a student response. (a) Suggestion to add a proposition between two concepts. (b) Student response adding the suggested proposition, as well as another relevant proposition. (c) Erroneous feedback suggesting an intermediate concept. (d) Student responds by adding the suggested concept, as part of aa different correct proposition. When referring to a specific concept or proposition, the avatar will be positioned such that her pointer indicates the target of the feedback. The concepts in the top two concept maps were colored orange by the student. Screenshots were taken in the teacher interface, which adds highlighting around the concepts and propositions targeted by the feedback. This highlighting was not present for students.

physics classroom with fixed table positions. Students were instructed—and if necessary reminded—to work individually. Nevertheless, evidence of cooperation was visible both in the classroom and in the logs of actions taken. For example, two students in the control group seemed to change their concept map in a manner that was very similar to suggestions in the feedback they would have received from the automated tool. On closer inspection, it appears they had almost identical results to a peer assigned to the experimental group; the likely conclusion is that they worked together, using the feedback made available to the peer in the experimental condition.

A second limitation was that approximately half the students in the experimental group did not use feedback when it was available to them. This number is comparable to what has been reported in other studies (e.g., [44]). We chose to analyze these students as a separate group, but this group was self-selected. While we did not find any evidence that students who used feedback were meaningfully different from students who did not, self-selection may undermine the validity of that aspect of the experiment. However, the alternative would be to somehow force all students in the experimental group to use feedback. We felt this would create an experimental setup that would be incompatible with the practice of inquiry learning, where students are encouraged to perform largely self-guided inquiries. The third limitation is that feedback is constrained by a single reference map for each domain. Valid

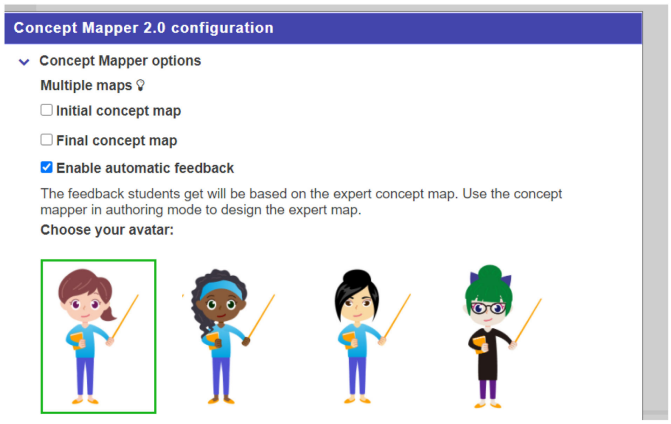

Fig. 4. Authoring view of activating automated feedback on concept maps in the Go-Lab ecosystem.

alternative explanations or approaches to modeling a domain are regarded as wrong by the algorithm, as is additional information. Students who add concepts and propositions covering alternative interpretations or additional topics that are correct but not represented in the reference map are presented with feedback that encourages the removal of this extra information. This may be criticized, as a concept map is meant for students to freely express their knowledge and there may even be different expert views on the same domain. However, we do have good reasons to follow this approach. First of all, as indicated in the introduction, we used this type of feedback because it is a practical and portable approach, which we have already actually implemented in the current version of the concept mapper app of the Go-Lab ecosystem, based on the prototype evaluated in this article. In this app, teachers can draw their own expert concept map and choose an avatar, which is all they need to do to make the automatic feedback work (see Fig. 4). Second, in order to respect students' choice to make their own concept map, the feedback given in this study has a nondirective character, most of the time asking students to make additions to the concept map on their own (see Table II). A third consideration is that it is very common to base feedback on expert information. In a very recent overview, Deeva *et al.* [26] analyzed 109 automated feedback systems, distinguishing data-driven and expert-driven feedback generation models, and found that almost half of the existing systems were expert based, 32% were data driven, and 19% combined both approaches. According to these authors, mixed approaches offer better opportunities than pure expert models. We agree with that conclusion and see our system as a first step on that path. Finally, and perhaps most importantly, empirical studies have shown that expert models may provide students with a useful structure or feedback, in the context of creating concept maps as well as in other contexts (e.g., see [25], [45]–[47]).

The use of a single reference map has important implications for considering what level of reference map to use. The students taking part in this study were taught the topics with varying levels of complexity, depending on their educational level. If we were to use the lowest level scoring rubric for all students, we would have to disregard significant portions of the higher level students' concept maps as wrong. Using the highest level scoring, rubric would mean giving low-level students feedback about concepts and propositions they may have never heard about. In either case, some feedback would be at best meaningless, and at worst detrimental.

We chose to use a single reference map based on an intermediate-level curriculum for students from all educational levels. This caused a dilemma when validating the algorithm for how human coders should score concept maps that included extra information. We instructed coders to base their coding on the intermediate-level reference map, and therefore to disregard extra information. This choice was made for two reasons: we felt that making coders aware of the students' educational level might bias their judgements, and using the same reference map would allow validating the algorithm on equal terms. The effect of educational level seems to have been minor; the correlation between scores from the algorithm and human coders was marginally lower for low-level students ($r = 0.88$) and high-level students ($r = 0.82$) than it was for intermediate-level students ($r = 0.9$).

Students of different ability levels benefit most from different types of concept maps, with low-ability students learning little from complex concept maps [8]–[10]. Given these results, it may be counterproductive to help students create concept maps that are (too far) above their ability level. It would be interesting to see whether, and at what point, additional feedback becomes detrimental. Similarly, we found that even incorrect or unhelpful feedback can motivate some students to *explain* their concept map better. Further analysis of the effect of correct and incorrect feedback on students of different ability levels might be fruitful.

Based on our findings, we theorize that positive feedback, in which students are encouraged to add information, is more likely to be beneficial than negative feedback, in which students are encouraged to remove incorrect information. A possible solution would be the use of a relatedness matrix (using weights for the degree of relation between each pair of concepts) rather than an adjacency matrix (using a binary system denoting if concepts are connected or not). Such a relatedness matrix would allow for larger reference maps and more nuanced feedback. Students creating propositions that are correct but deviate from the reference map would then no longer receive negative feedback for their creativity. We plan to investigate this approach further in our future research.

This study was performed with students who were 12–13 years old, but inquiry learning is used from kindergarten to college and in professional education. Our highly portable tool would make it possible to directly compare and contrast the effects of automated feedback for students of different ages and abilities, which may help bridge gaps in the existing literature.

We contend that the tool presented here gives ample avenues for future research. It provides an automated, valid, and accurate measure of concept map quality that can be employed easily in any domain. We were able to accurately identify a large number of shortcomings in students' concept maps, and

although we were unable to find an effect of feedback on the quality of concept maps in this study, we are confident that with improvements to the algorithm and the feedback given, the tool will be a valuable addition to any inquiry learning activity.

REFERENCES

[1] J. D. Bransford, A. L. Brown, and R. R. Cocking, Eds. *How People Learn: Brain, Mind, Experience, and School*. Washington, DC, USA: Nat. Academies Press, 2000, doi: 10.17226/9853.

[2] C. C. Kuhlthau, L. K. Maniotes, and A. K. Caspari, *Guided Inquiry: Learning in the 21st Century*. Westport, CT, USA: Libraries Unlimited, 2015.

[3] S. Freeman *et al.*, "Active learning increases student performance in science, engineering, and mathematics," *Proc. Nat. Acad. Sci.*, vol. 111, no. 23, pp. 8410–8415, Jun. 2014, doi: 10.1073/pnas.1319030111.

[4] M. T. H. Chi and R. Wylie, "The ICAP framework: Linking cognitive engagement to active learning outcomes," *Educ. Psychol.*, vol. 49, no. 4, pp. 219–243, Oct. 2014, doi: 10.1080/00461520.2014.965823.

[5] T. de Jong, "Moving towards engaged learning in STEM domains; there is no simple answer, but clearly a road ahead," *J. Comput. Assist. Learn.*, vol. 35, no. 2, pp. 153–167, Apr. 2019, doi: 10.1111/jcal.12337.

[6] M. Pedaste *et al.*, "Phases of inquiry-based learning: Definitions and the inquiry cycle," *Educ. Res. Rev.*, vol. 14, pp. 47–61, Feb. 2015, doi: 10.1016/j.edurev.2015.02.003.

[7] J. D. Novak and A. J. Cañas, "The theory underlying concept maps and how to construct and use them," *Inst. Human Mach. Cogn.*, Pensacola, FL: CMapTools, Jan. 2008. [Online]. Available: https://cmap.ihmc.us/docs/pdf/TheoryUnderlyingConceptMaps.pdf

[8] F. Amadieu, A. Tricot, and C. Mariné, "Prior knowledge in learning from a non-linear electronic document: Disorientation and coherence of the reading sequences," *Comput. Hum. Behav.*, vol. 25, no. 2, pp. 381–388, Mar. 2009, doi: 10.1016/j.chb.2008.12.017.

[9] F. Amadieu, T. van Gog, F. Paas, A. Tricot, and C. Mariné, "Effects of prior knowledge and concept-map structure on disorientation, cognitive load, and learning," *Learn. Instruct.*, vol. 19, no. 5, pp. 376–386, Oct. 2009, doi: 10.1016/j.learninstruc.2009.02.005.

[10] S. Gerstner and F. X. Bogner, "Concept map structure, gender and teaching methods: An investigation of students' science learning," *Educ. Res.*, vol. 51, no. 4, pp. 425–438, 2009, doi: 10.1080/00131880903354758.

[11] I. M. Kinchin, D. B. Hay, and A. Adams, "How a qualitative approach to concept map analysis can be used to aid learning by illustrating patterns of conceptual development," *Educ. Res.*, vol. 42, no. 1, pp. 43–57, Jan. 2000, doi: 10.1080/001318800363908.

[12] E. Gouli, A. Gogoulou, K. Papanikolaou, and M. Grigoriadou, "COMPASS: An adaptive web-based concept map assessment tool," in *Proc. 1st Int. Conf. Concept Mapping*, Pamplona, Spain, Sep. 14–17, 2004, pp. 295–302.

[13] A. M. van Dijk and A. W. Lazonder, "Scaffolding students' use of learner-generated content in a technology-enhanced inquiry learning environment," *Interact. Learn. Environ.*, vol. 24, no. 1, pp. 194–204, Feb. 2016, doi: 10.1080/10494820.2013.834828.

[14] K. M. Edmondson, "Assessing science understanding through concept maps," in *Assessing Science Understanding*, J. J. Mintzes, J. H. Wandersee, and J. D. Novak, Eds. Amsterdam, The Netherlands: Elsevier, 2005, pp. 15–40. doi: 10.1016/B978-012498365-6/50004-4.

[15] J. A. C. Hattie and G. M. Donoghue, "Learning strategies: A synthesis and conceptual model," *NPJ Sci. Learn.*, vol. 1, Dec. 2016, Art. no. 16013, doi: 10.1038/npjscilearn.2016.13.

[16] J. C. Nesbit and O. O. Adesope, "Learning with concept and knowledge maps: A meta-analysis," *Rev. Educ. Res.*, vol. 76, no. 3, pp. 413–448, Sep. 2006, doi: 10.3102/00346543076003413.

[17] S. Löhner, W. R. van Joolingen, E. R. Savelsbergh, and B. van Hout-Wolters, "Students' reasoning during modeling in an inquiry learning environment," *Comput. Hum. Behav.*, vol. 21, no. 3, pp. 441–461, May 2005, doi: 10.1016/j.chb.2004.10.037.

[18] K. Leelawong and G. Biswas, "Designing learning by teaching agents: The Betty's brain system," *Int. J. Artif. Intell. Educ.*, vol. 18, no. 3, pp. 181–208, Oct. 2008.

[19] G. Biswas, J. R. Segedy, and K. Bunchongchit, "From design to implementation to practice a learning by teaching system: Betty's brain," *Int. J. Artif. Intell. Educ.*, vol. 26, no. 1, pp. 350–364, Mar. 2016, doi: 10.1007/s40593-015-0057-9.

[20] D. Ifenthaler, "Bridging the gap between expert-novice differences," *J. Res. Technol. Educ.*, vol. 43, no. 2, pp. 103–117, Dec. 2010, doi: 10.1080/15391523.2010.10782564.

[21] T. Conlon, "But is our concept map any good?: Classroom experience with the reasonable fallible analyser," in *Proc. 1st Int. Conf. Concept Mapping*, Pamplona, Spain, Sep. 2004, pp. 159–166.

[22] S. Weinbrenner, J. Engler, and H. U. Hoppe, "Ontology-supported scaffolding of concept maps," in *Proc. 15th Int. Conf. Artif. Intell. Educ.*, Jun./Jul. 2011, vol. 6738, pp. 582–584. doi: 10.1007/978-3-642-21869-9_108.

[23] L. F. Gerard, K. Ryoo, K. W. McElhaney, O. L. Liu, A. N. Rafferty, and M. C. Linn, "Automated guidance for student inquiry," *J. Educ. Psychol.*, vol. 108, no. 1, pp. 60–81, Jan. 2016, doi: 10.1037/edu0000052.

[24] K. Ryoo and M. C. Linn, "Designing automated guidance for concept diagrams in inquiry instruction," *J. Res. Sci. Teach.*, vol. 53, no. 7, pp. 1003–1035, 2016, doi: 10.1002/tea.21321.

[25] G. P. Jain, V. P. Gurupur, J. L. Schroeder, and E. D. Faulkenberry, "Artificial intelligence-based student learning evaluation: A concept map-based approach for analyzing a student's understanding of a topic," *IEEE Trans. Learn. Technol.*, vol. 7, no. 3, pp. 267–279, Jul.–Sep. 2014, doi: 10.1109/TLT.2014.2330297.

[26] G. Deeva, D. Bogdanova, E. Serral, M. Snoeck, and J. De Weerdt, "A review of automated feedback systems for learners: Classification framework, challenges and opportunities," *Comput. Educ.*, vol. 162, Mar. 2021, Art. no. 104094, doi: 10.1016/j.compedu.2020.104094.

[27] T. de Jong, S. Sotiriou, and D. Gillet, "Innovations in STEM education: The go-lab federation of online labs," *Smart Learn. Environ.*, vol. 1, no. 3, pp. 1–16, Dec. 2014, doi: 10.1186/s40561-014-0003-6.

[28] T. de Jong *et al.*, "Understanding teacher design practices for digital inquiry–based science learning: The case of go-lab," *Educ. Technol. Res. Develop.*, vol. 69, no. 2, pp. 417–444, Apr. 2021, doi: 10.1007/s11423-020-09904-z.

[29] M. Schneider and F. Preckel, "Variables associated with achievement in higher education: A systematic review of meta-analyses," *Psychol. Bull.*, vol. 143, no. 6, pp. 565–600, Jun. 2017, doi: 10.1037/bul0000098.

[30] L.-K. Chen, E. Dorn, M. Krawitz, C. S. H. Lim, and M. Mourshed, "Drivers of student performance: Asia insights," McKinsey, Jan. 2018. [Online]. Available: https://www.mckinsey.com/industries/public-and-social-sector/our-insights/drivers-of-student-performance-asia-insights

[31] Z. C. Zacharia *et al.*, "Identifying potential types of guidance for supporting student inquiry when using virtual and remote labs in science: A literature review," *Educ. Technol. Res. Dev.*, vol. 63, no. 2, pp. 257–302, Apr. 2015, doi: 10.1007/s11423-015-9370-0.

[32] L. Siiman, "Rate of photosynthesis lab," 2020. [Online]. Available: https://www.golabz.eu/lab/rate-of-photosynthesis-lab-html5

[33] Concord Consortium, "Convection: The stack effect." [Online]. Available: https://learn.concord.org/resources/758/convection-the-stack-effect

[34] K. A. Kroeze, "Redirect balancer," 2019. [Online]. Available: https://dx.doi.org/10.5281/zenodo.5155411, doi: 10.5281/zenodo.5155411.

[35] K. A. Kroeze, S. M. van den Berg, A. W. Lazonder, B. P. Veldkamp, and T. de Jong, "Automated feedback can improve hypothesis quality," *Front. Educ.*, vol. 3, pp. 1–14, Jan. 2019, doi: 10.3389/feduc.2018.00116.

[36] D. Ifenthaler, "Relational, structural, and semantic analysis of graphical representations and concept maps," *Educ. Technol. Res. Develop.*, vol. 58, no. 1, pp. 81–97, Feb. 2010, doi: 10.1007/s11423-008-9087-4.

[37] K. A. Kroeze, "Adaptive hypothesis grammars," 2020. [Online]. Available: https://dx.doi.org/10.5281/zenodo.3739003, doi: 10.5281/zenodo.3739003.

[38] K. Verbert *et al.*, "Learning dashboards: An overview and future research opportunities," *Pers. Ubiquitous Comput.*, vol. 18, no. 6, pp. 1499–1514, Nov. 2013, doi: 10.1007/s00779-013-0751-2.

[39] R. Bodily and K. Verbert, "Review of research on student-facing learning analytics dashboards and educational recommender systems," *IEEE Trans. Learn. Technol.*, vol. 10, no. 4, pp. 405–418, Oct.–Dec. 2017, doi: 10.1109/TLT.2017.2740172.

[40] W. Matcha, N. A. Uzir, D. Gasevic, and A. Pardo, "A systematic review of empirical studies on learning analytics dashboards: A self-regulated learning perspective," *IEEE Trans. Learn. Technol.*, vol. 13, no. 2, pp. 226–245, Apr. 2020, doi: 10.1109/TLT.2019.2916802.

[41] C. Schumacher and D. Ifenthaler, "Features students really expect from learning analytics," *Comput. Hum. Behav.*, vol. 78, pp. 397–407, Jan. 2018, doi: 10.1016/j.chb.2017.06.030.

[42] R. J. Weijers, B. B. de Koning, and F. Paas, "Nudging in education: From theory towards guidelines for successful implementation," *Eur. J. Psychol. Educ.*, vol. 36, pp. 883–902, Sep. 2021, doi: 10.1007/s10212-020-00495-0.

[43] P.-H. Wu, G.-J. Hwang, M. Milrad, H.-R. Ke, and Y.-M. Huang, "An innovative concept map approach for improving students' learning performance with an instant feedback mechanism," *Brit. J. Educ. Technol.*, vol. 43, no. 2, pp. 217–232, Mar. 2012, doi: 10.1111/j.1467-8535.2010.01167.x.

[44] H. K. Sinclair and J. A. Cleland, "Undergraduate medical students: Who seeks formative feedback?," *Med. Educ.*, vol. 41, no. 6, pp. 580–582, Jun. 2007, doi: 10.1111/j.1365-2923.2007.02768.x.

[45] K. E. Chang, Y. T. Sung, and S. F. Chen, "Learning through computer-based concept mapping with scaffolding aid," *J. Comput. Assist. Learn.*, vol. 17, no. 1, pp. 21–33, Mar. 2001, doi: 10.1046/j.1365-2729.2001.00156.x.

[46] R. Azevedo, J. G. Cromley, and D. Seibert, "Does adaptive scaffolding facilitate students' ability to regulate their learning with hypermedia?," *Contemporary Educ. Psychol.*, vol. 29, no. 3, pp. 344–370, Jul. 2004, doi: 10.1016/j.cedpsych.2003.09.002.

[47] Z. Wang, O. Adesope, N. Sundararajan, and P. Buckley, "Effects of different concept map activities on chemistry learning," *Educ. Psychol.*, vol. 41, no. 2, pp. 245–260, Feb. 2021, doi: 10.1080/01443410.2020.1749567.

Karel A. Kroeze was born in Noordwijkerhout, The Netherlands. He received the B.Sc. degree in public governance from the University of Twente, Enschede, The Netherlands, in 2013, and the M.Sc. degree in methodology and statistics for the behavioral, biomedical, and social sciences from the University of Utrecht, University Medical Centre Utrecht, Utrecht, The Netherlands, and the University of Twente, in 2015.

He is currently a Researcher and Data Scientist with the University of Twente, while finishing his doctoral dissertation. His research interests include adaptive learning environments, computerized adaptive testing, computational statistics, statistical learning, and data visualization.

Stéphanie M. van den Berg received the M.Sc. degree in psychology from Leiden University, Leiden, The Netherlands, in 1998, and the Ph.D. degree in psychology from the University of Amsterdam, Amsterdam, The Netherlands, in 2002.

She is currently an Associate Professor with the University of Twente, Enschede, The Netherlands. She is the Head of the Cognition, Data, and Education section of the Learning, Data Analytics, and Technology Department, University of Twente. She also leads the Behavioral Data Science Incubator, which aims to make advanced measurement, collection, and analytics available for all researchers at the faculty of Behavioral and Management Sciences. Her research interests include machine learning, psychometrics, Bayesian statistics, and modeling time-intensive data. She teaches courses in data science and statistics, and is currently writing a book on linear methods for social sciences students.

Bernard P. Veldkamp received the M.Sc. degree in applied mathematics and the Ph.D. degree in behavioral sciences from the University of Twente, Enschede, The Netherlands, in 1996 and 2001, respectively.

He is currently the Head of the Department of Learning, Data Analytics, and Technology with the University of Twente. He specializes in research methodology and data analytics. His research interests include psychometrics, operations research, data and text mining, optimization, and computer-based assessment. His work spans a range of issues in educational, psychological, and health sciences, from the development of new methods/models for the design and construction of (adaptive) psychological and educational tests, to the development of data mining models for analyzing verbal data and large datasets in fraud detection.

Ton de Jong received the M.Sc. degree in educational and cognitive psychology from the University of Amsterdam, Amsterdam, The Netherlands, in 1981, and the Ph.D. degree in technological science from the Eindhoven University of Technology, Eindhoven, The Netherlands, in 1986.

He is currently a Full Professor of Instructional Technology with the University of Twente, Enschede, The Netherlands, with the Faculty of Behavioral Sciences where he acts as the Department Head of the Department Instructional Technology. He is the author or co-author of more than 100 journal papers and more than 90 book chapters, and was the Editor or Co-Editor of a number of books. He has three publications in *Science*. He was the Coordinator of eight EU projects including the 7th framework Go-Lab project on learning with online laboratories in science and its Horizon 2020 follow-up project Next-Lab, and is currently on the editorial board of eight journals.

Dr. Jong is a Fellow of the AERA and ISLS, and was an Elected Member of the Academia Europaea in 2014.