frontiers in Education

ORIGINAL RESEARCH
published: 04 January 2019
doi: 10.3389/feduc.2018.00116

Automated Feedback Can Improve Hypothesis Quality

Karel A. Kroeze[1,2]*, *Stéphanie M. van den Berg*[2], *Ard W. Lazonder*[3], *Bernard P. Veldkamp*[2] *and Ton de Jong*[1]

[1] Department of Instructional Technology, University of Twente, Enschede, Netherlands, [2] Department of Research Methodology, Measurement and Data Analysis, University of Twente, Enschede, Netherlands, [3] Behavioural Science Institute, Radboud University, Nijmegen, Netherlands

Stating a hypothesis is one of the central processes in inquiry learning, and often forms the starting point of the inquiry process. We designed, implemented, and evaluated an automated parsing and feedback system that informed students about the quality of hypotheses they had created in an online tool, the hypothesis scratchpad. In two pilot studies in different domains ("supply and demand" from economics and "electrical circuits" from physics) we determined the parser's accuracy by comparing its judgments with those of human experts. A satisfactory to high accuracy was reached. In the main study (in the "electrical circuits" domain), students were assigned to one of two conditions: no feedback (control) and automated feedback. We found that the subset of students in the experimental condition who asked for automated feedback on their hypotheses were much more likely to create a syntactically correct hypothesis than students in either condition who did not ask for feedback.

Keywords: automated feedback, hypotheses, inquiry learning, context-free grammars, online learning environment

OPEN ACCESS

Edited by:
Samuel Greiff, University of Luxembourg, Luxembourg

Reviewed by:
Trude Nilsen, University of Oslo, Norway
Jessica Andrews-Todd, Educational Testing Service, United States

***Correspondence:**
Karel A. Kroeze
k.a.kroeze@utwente.nl

Specialty section:
This article was submitted to Educational Psychology, a section of the journal Frontiers in Education

Received: *14 September 2018*
Accepted: *11 December 2018*
Published: *04 January 2019*

Citation:
Kroeze KA, van den Berg SM, Lazonder AW, Veldkamp BP and de Jong T (2019) Automated Feedback Can Improve Hypothesis Quality. Front. Educ. 3:116. doi: 10.3389/feduc.2018.00116

INTRODUCTION

Active forms of learning are seen as key to acquiring deep conceptual knowledge, especially in science domains (Hake, 1998; Freeman et al., 2014). One of the active forms of learning is inquiry learning. Inquiry learning has been defined in many different ways with as its kernel that the method starts from questions for which students need to find answers [see e.g., (Prince and Felder, 2007)]. In the current work, we focus on one of the ways inquiry is used in instruction, namely "learning science by doing science": students are expected to form and test hypotheses by performing experiments and analyzing data. In following an inquiry cycle, students learn both science content and the scientific method. In this study, we focus on the practice of the scientific method, and in particular on the creation of hypotheses.

Most models of inquiry-based learning encompass an orientation and conceptualization phase that enables students to familiarize themselves with the topic of investigation. Common activities during orientation are studying background information and conducting a few explorative experiments with the equipment at hand. The intended outcome of these initial explorations is the formation of theories and ideas, formalized in hypotheses (Pedaste et al., 2015). Hypotheses are integral to the inquiry cycle: they direct students' attention to specific aspects of the research problem and, hence, facilitate experimental design and data interpretation (Klahr and Dunbar, 1988; Zimmerman, 2007). In a classic study, Tschirgi (1980) found that both children and adults design more conclusive experiments when trying to test a hypothesis that contradicts prior

evidence. Hypothesis testing also increases the amount of domain knowledge students gain from an inquiry (Burns and Vollmeyer, 2002; Brod et al., 2018), which is probably due to the fact that hypotheses, regardless of their specificity and truth value, provide direction to students' inquiry process (Lazonder et al., 2009).

The importance of hypothesizing nevertheless stands in marked contrast with its occurrence in high school science classes. Research has consistently shown that inquiry is a complex process in which students make mistakes (Mulder et al., 2010). Specifically, students of all ages have problems in formulating hypotheses, particularly when they are unfamiliar with the topic of inquiry (Gijlers and de Jong, 2005; Mulder et al., 2010), and when experimental data is anomalous (Lazonder, 2014). As a consequence, few students generate hypotheses on their own account, and when they do, they often stick to a single hypothesis that is known to be true (i.e., confirmation bias) or formulate imprecise statements that cannot be tested in research. These natural tendencies demonstrate that unguided inquiry learning is likely to be ineffective (Mayer, 2004; Kirschner et al., 2006; de Jong and Lazonder, 2014). However, *guided* inquiry learning has been shown to compare favorably to both direct instruction (D'Angelo et al., 2014) and unguided inquiry learning (Furtak et al., 2012), and helps foster a deeper conceptual understanding (Alfieri et al., 2011).

Inspired by these positive findings we set out to design and evaluate a software scaffold that presented students with automatically generated feedback on the quality of their hypotheses.

THEORETICAL FRAMEWORK

Adaptive and Automated Scaffolding

Inquiry learning often takes place in virtual or remote laboratories and, to be successful, should be supplemented with guidance (de Jong and Lazonder, 2014). Furthermore, de Jong and Lazonder (2014) postulated that different types of students require different types of guidance. Recent work on differentiated guidance lends credence to this argument, finding a moderating effect of students' age (Lazonder and Harmsen, 2016) and prior knowledge (van Riesen et al., 2018) on learning activities and knowledge gains. Moreover, Furtak et al. (2012) showed teacher-led inquiry activities to be more effective than student-led inquiry, implying that teachers are effective suppliers of guidance. However, given that teachers' time is an increasingly valuable resource, several adaptive software agents have recently been developed to support teachers on specific tasks and that adapt the guidance to students' characteristics. While Belland et al. (2016) found no added effect of limited adaptive scaffolding over static scaffolding, intelligent tutoring systems (Nye et al., 2014), adaptive environments (Durlach and Ray, 2011; Vandewaetere et al., 2011), and automated feedback (Gerard et al., 2015, 2016) have all shown promising results. The common-sense conclusion appears to be that the more guidance is adapted to the individual student, the better the guidance—and thus the student—performs. Indeed, Pedaste et al. (2015) recently identified the development of "*virtual teacher assistants that analyse and respond to individual learners to create meaningful learning activities*" as one of the main challenges in the field.

Although adaptive and automated elements are increasingly common in online learning environments (e.g., Aleven et al., 2010; Lukasenko et al., 2010; Vandewaetere et al., 2011; Gerard et al., 2015, 2016; Ryoo and Linn, 2016), they have typically been designed and implemented for a single learning activity in a specific domain. The reason for this is simple; even adaptive guidance for a single well-defined learning task generally requires years of research and development. Data must be gathered and coded, models have to be trained and fitted, appropriate feedback has to be fine-tuned and a digital environment has to be developed. Each of these steps involves the input of experts from different fields; teachers, statisticians, educational researchers, and computer scientists. As a result, scaffolds in multi-domain environments such as Go-Lab (de Jong et al., 2014) and WISE (Linn et al., 2003) generally do not adapt to the individual student, nor can they automatically assess products or provide context-sensitive feedback. The hypothesis scaffold we describe and test in this paper aims to fill this gap.

We have been unable to find any existing literature on the automated scoring of and feedback on free-text hypotheses. In contrast, a variety of increasingly sophisticated natural language processing (NLP) techniques have been employed for automated essay scoring. However, the techniques applied to scoring essays typically require a large amount of training data, and even when training data is available they are unlikely to provide the level of detail on the underlying structure of hypotheses required to give meaningful feedback. Training data is not readily available for hypotheses, and would be expensive to gather (Shermis and Burstein, 2013).

Anjewierden et al. (2015) noted that the "language" of hypotheses is a subset of natural language with a specific structure. They suggested using a domain-specific list of variables and categorical values (the *lexicon*), in conjunction with a *grammar* of hypotheses. Together, the lexicon and grammar could be used to create a hypothesis parses that is robust, and can be adapted to different domains with relative ease. The work reported here attempts to implement such a context-free grammar.

Feedback

The *informative tutoring feedback* model [ITF, (Narciss, 2006, 2008)] distinguishes between *internal* feedback and *external* feedback, and a wide variety of feedback *types*. Internal feedback is provided by individual cognitive monitoring processes (Ifenthaler, 2011), external feedback can be provided by for example; teachers, peers, or automated scaffolds. Both types of feedback may conflict with or reinforce an *internal reference value*. Careful feedback design can help students regulate their learning process, particularly when internal and external feedback conflict (Narciss, 2008).

The function of feedback may be *cognitive*, *meta-cognitive,* or *motivational*, and a distinction can be made between *simple* (e.g., knowledge of performance, correct result) and *elaborated* (e.g., knowledge about task constraints, mistakes, and concepts) forms of feedback. These components broadly overlap with *outcome*,

corrective and *explanatory* feedback types (e.g., Johnson and Priest, 2014). In a second-order meta-analysis on the effects of feedback, Hattie and Timperley (2007) prescribed that good feedback should set clear goals (*feed up*), inform the student of their progress (*feed back*), and provide steps to improve (*feed forward*). Finally, *immediate* feedback has been shown to give larger benefits than *delayed* feedback (Van der Kleij et al., 2015).

Research Goal and Context

This project is performed in the Go-Lab ecosystem (de Jong et al., 2014). Go-Lab is an online environment where teachers and authors can share online and remote laboratories (Labs) and scaffolding applications (Apps). Apps and Labs can, together with multimedia material, be combined to create Inquiry Learning Spaces (ILS), which can also be shared on the Go-Lab environment. **Figure 1** shows a screenshot of a typical ILS. This ILS is organized in six phases that follow an inquiry cycle (in this case; Orientation, Conceptualization, Investigation, Interpretation, Conclusion, and Discussion), and can be navigated freely.

The hypothesis scratchpad app [**Figure 2**; (Bollen and Sikken, 2018)] is used to support students with hypothesis generation. This study aimed to create an adaptive version of the hypothesis scratchpad that can scaffold the individual student in hypothesizing in any domain, with a minimum of set-up time for teachers. This new version will need to (1) identify mistakes in students' hypotheses, and (2) provide students with appropriate feedback to correct these mistakes. If the app achieves both of these goals, it will be a considerable step toward "*empowering science teachers using technology-enhanced scaffolding to improve inquiry learning*" (Pedaste et al., 2015).

DESIGN

For this project the hypothesis scratchpad currently available in Go-Lab has been extended. An automated feedback system was developed that can identify flaws in students' hypotheses and provide tailored feedback that enables students to correct their mistakes. The aim is to improve the quality of students' hypotheses.

The following sections will (1) describe the main components of hypotheses and the criteria used to assess them, (2) introduce the process of parsing hypotheses and applying criteria, (3) present the feedback given to students, and (4) formalize the outcome measures and statistical analyses used.

Criteria

Quinn and George (1975) were the first to formally define a set of criteria for evaluating hypotheses: (1) *it makes sense*; (2) *it is empirical,* a (partial) scientific relation; (3) *it is adequate*, a scientific relation between at least two variables; (4) *it is precise*—a qualified and/or quantified relation; and (5) *it states a test*, an explicit statement of a test. Subsequent research on hypothesis generation has broadly followed the same criteria, or a subset thereof. Van Joolingen and De Jong (1991, 1993) used a "syntax" and a "precision" measure, that correspond roughly with the "it makes sense" and "precise' criteria of Quinn and George. Mulder et al. (2010) used a "specificity" scale, using criteria comparable to those of Quinn and George.

Based on the criteria used by Quinn and George, and the measures used by Van Joolingen and de Jong, we developed a set of criteria that could be implemented in automated feedback. **Table 1** lists these criteria, providing a short explanation and examples from the electrical circuits domain for each criterion. In the automated feedback, the first two criteria are straightforward in that they rely on the presence of certain words. The remaining criteria are established using a context-free grammar parser, which is described in the next section.

Parser

To detect mistakes, the automated system needs to interpret hypotheses on the criteria listed in **Table 1**. Given the observation that hypotheses are a relatively structured subset of natural language (Anjewierden et al., 2015), we can define a *context-free grammar* [CFG, (Chomsky, 1956)] that covers all well-structured hypotheses.

CFGs can be used to define natural languages, and are ideally suited to define heavily structured languages [e.g., programming languages, (Chomsky, 1956)]. A CFG is comprised of a set of *production rules*. All the sentences that can be produced by the repeated application of these rules are the *formal language* of that grammar.

The grammar that defines hypotheses looks something like the following[1];

```
HYPOTHESIS -> if ACTION then ACTION
HYPOTHESIS -> ACTION if ACTION
ACTION -> VAR INTERACTOR VAR
ACTION -> VAR MODIFIER
ACTION -> MODIFIER VAR
ACTION -> ACTION and ACTION
VAR -> PROPERTY VAR
VAR -> bulbs
VAR -> voltage
VAR -> brightness
INTERACTOR -> is greater than
INTERACTOR -> is smaller than
INTERACTOR -> is equal to
MODIFIER -> increases
MODIFIER -> decreases
QUALIFIER -> series circuit
QUALIFIER -> parallel circuit
```

Each line is a production rule, the left-hand side of the rule can be replaced by the right-hand side. Uppercase words refer to further rules (they are *non-terminal*) and lowercase words refer to tokens (they are *terminal*). A token can be anything, but in our case, they are (sets of) words, e.g., "voltage" or "is greater than."

Consider the following hypothesis; *"if the number of bulbs in a series circuit increases, the brightness of the bulbs decreases."* If we were to apply our grammar, we can decompose this hypothesis

[1]For the complete grammar, see https://github.com/Karel-Kroeze/adaptive-hypothesis-grammars.

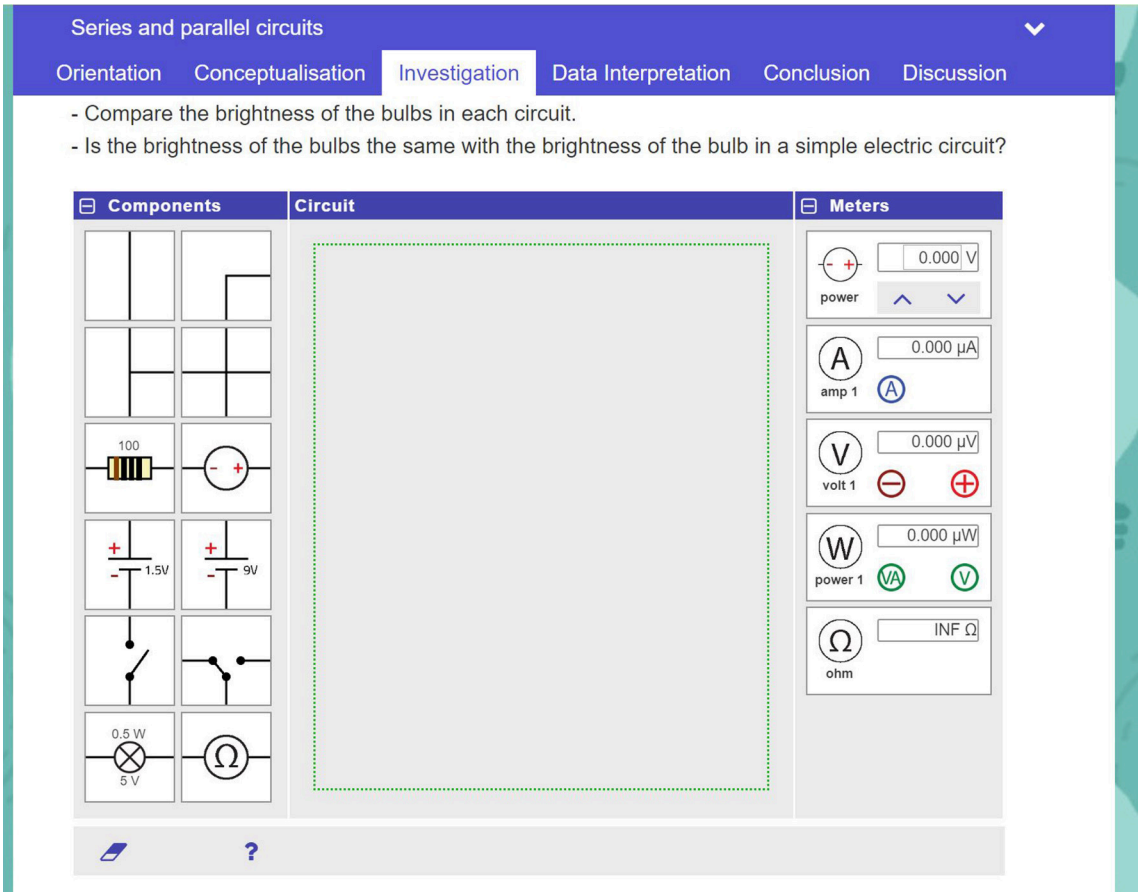

FIGURE 1 | Screenshot of a typical inquiry learning space on the Go-Lab environment.

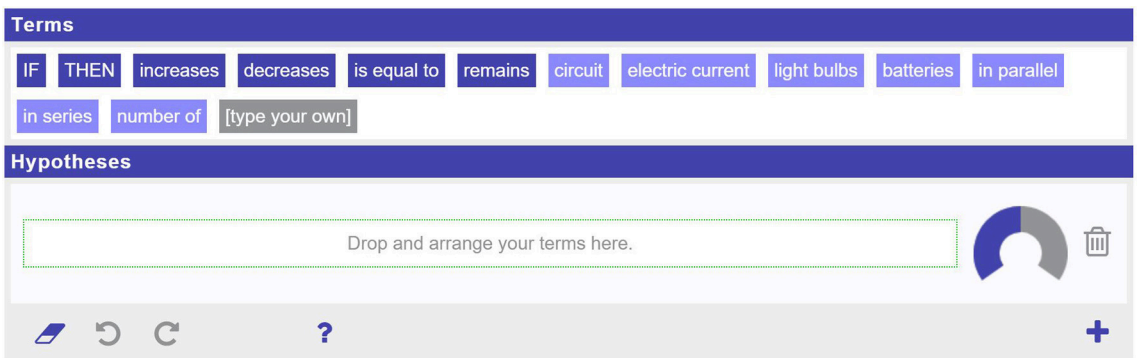

FIGURE 2 | Screenshot of the hypothesis scratchpad.

as per **Figure 3**. Although this decomposition provides the structure of the hypothesis, it still does not contain the *semantic* information necessary to evaluate the criteria.

If we add semantic information to each of the tokens, and rules on how to *unify* this information to each of the production rules, we can extract all relevant information from the hypothesis (Knuth, 1968; Shieber, 2003). **Figure 4** shows an example of the final parse result[2] which contains all the information needed to evaluate the criteria discussed.

[2]The parser was created using the Nearley.js package (Hardmath123., 2017), which implements the Earley context-free parsing algorithm (Earley, 1970). The source code of the parser is available on GitHub; https://github.com/Karel-Kroeze/adaptive-hypothesis-utils/.

TABLE 1 | Scoring criteria.

Criterion	Name	Description	Examples
1	Contains at least two variables	The hypothesis should contain at least two variables. Without two variables, the hypothesis can at best be an observation, and is likely to be nonsense.	✗"the current increases" ✓ "the current increases and the brightness increases"
2	Contains a modifier	The hypothesis should contain at least one modifier (e.g., "increases," "floats," but not "remains the same"). Without a modifier, the hypothesis can at best describe a static situation, and is likely to be nonsense.	✗ "the current remains the same" ✓ "the current increases"
3	Is a syntactically correct sentence	The hypothesis should be a correct sentence. Not only is the hypothesis likely to be nonsense if it is not a sentence, but moreover the automated system can only parse syntactically correct sentences.	✗ "the current increases decreases" ✓ "the current increases"
4	Manipulates exactly one independent variable	In order to test an effect of *x* on *y*, *x* should change, and no other variable should change.	✗ "if the current remains the same, the brightness increases" ✗ "if the number of bulbs increases and the current increases, the brightness remains the same" ✓ "if the number of bulbs increases, the brightness decreases"
5	Qualifies the variables	For some variables, it is their context that defines them. e.g., for buoyancy, density is defined by mass *and* volume, and in electrical circuits the *type* of circuit is crucial.	✗ "if the number of bulbs increases, the brightness remains the same" ✗ "if the mass of the object is larger than the volume of the fluid, the object sinks" ✓ "if the number of bulbs in a parallel circuit increases, the brightness remains the same"
6	Specifies interactions between variables	In some domains, it is the interaction between variables that is important. In our dataset this refers mainly to buoyancy, the relevant variable is the density of an object, as related to the density of the fluid.	✗ "if the density of the object increases, the object sinks" ✓ "if the density of the object is larger than the density of the fluid, the object sinks"

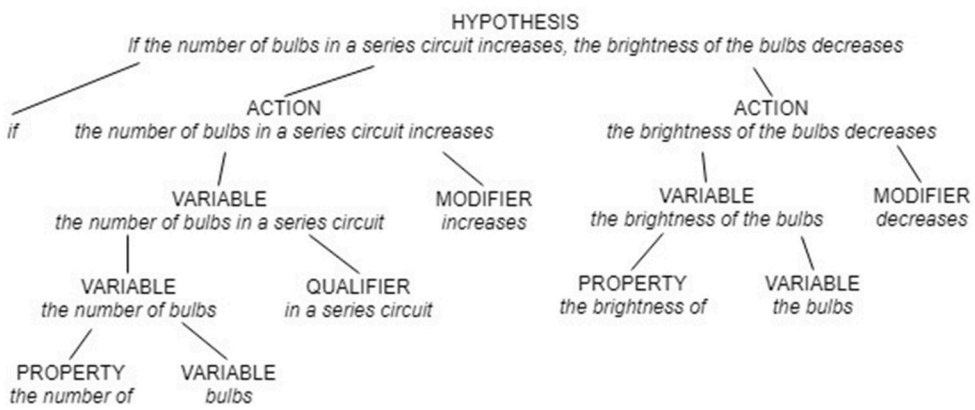

FIGURE 3 | Example of a hypothesis parse tree.

Feedback

The automated hypothesis scratchpad gives students the opportunity to request feedback. **Figure 5** shows an example of the automated hypothesis scratchpad, with the feedback button highlighted (the highlight is not part of the interface).

Table 2 gives an overview of the feedback used. The feedback follows the guidelines set by Hattie and Timperley (2007) in that it informs students of their progress, is specific about the mistakes made, and—where relevant—suggests modes of improvement. The first three criteria from **Table 2** are required conditions; if a hypothesis does not have variables, a modifier or cannot be parsed, the other criteria are not shown. Conversely, if these criteria are met, feedback is presented only on the other relevant criteria.

Feedback was presented to the student in textual form in a pop-up window and was shown immediately after a student

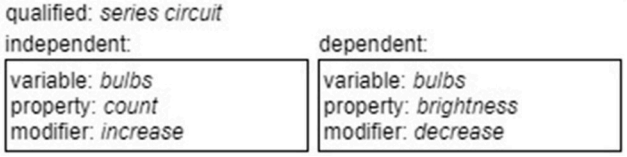

FIGURE 4 | Parse result with semantic information.

requested it by clicking the feedback button. Feedback was never presented automatically. After receiving feedback, students could revise their hypothesis, and ask for feedback again. No explicit limits were placed on the amount of times students could ask for feedback.

Measures

Three outcome measures are of interest; (1) do students use the feedback tool, (2) does the parser correctly classify mistakes, and (3) do students' hypotheses improve after receiving feedback.

All student actions within a Go-Lab inquiry learning space are logged to a database. Specifically, the history of all hypotheses is tracked, including requests for feedback, and the feedback received. Feedback counts can thus be readily determined from the log files. A snapshot of a hypothesis is made whenever a student asks for feedback, and of the final state of the hypothesis. The collection of snapshots for a hypothesis creates a "story" for that hypothesis, tracking it over time.

The validity of classifications made by the parser is evaluated by calculating an inter-rater reliability between the results of the parser and human coders. The human coders were instructed to code as a teacher, ignoring small mistakes in spelling and syntax if the intention of a hypothesis was clear. To train the human coders, a sample of snapshots was coded, and any disagreements were discussed. After reaching agreement, each coder independently coded the remaining snapshots. Agreement is calculated using Cohens' κ, and interpreted using rules of thumb Landis and Koch (1977) .

Each snapshot is given a score based on the number of criteria passed, resulting in a score in a $0-k$ range, where k is the number of criteria used (three in the first pilot, six in the second pilot and final experiment). Improvement of hypotheses is evaluated by comparing the score for a snapshot to the score for the previous snapshot. The quality of a hypothesis is the quality of the final snapshot of that hypothesis.

If feedback is effective, we expect to see that students who have feedback available create higher quality hypotheses, and that hypothesis quality increases after students ask for feedback: each consecutive snapshot should have a higher quality than the last.

During the study, it became apparent that the aggregate score does not follow a parametric distribution, and therefore could not be used as an outcome measure. The variables and modifier criteria were satisfied by almost all students in our samples. The syntax criterion was often indicative for success on the manipulation, CVS and qualified criteria. Thus, even though the variables, modifier and CVS criteria might be important from a science education perspective, the syntactically correct criterion was used as an indicator for hypothesis quality.

Multilevel logistic models (i.e., generalized linear mixed models) were used to account for the inherent group structure in the data, controlling for student and class effects where appropriate. The models used were comprised of two levels, students and classes. All reported effects are on the student level. To perform the models, we used R (R Core Team, 2018) and the package "lme4" (Bates et al., 2015). The scripts used in analyses are deposited along with the raw and generated datasets at DANS (Kroeze, 2018).

FIELD STUDIES

Three field studies were conducted. An initial pilot study was conducted with an early version of the hypothesis parser to assess the feasibility of automated parsing of hypotheses using a context-free grammar. Following that, a second pilot study was conducted with the complete version of the parser to identify any remaining issues with the parser and ILS before moving on to the final experiment. The final experiment used a quasi-experimental design to assess the benefit of the tool in improving students' hypotheses. Each of these studies is described in more detail in the following sections.

First Pilot Study

Participants

Four classes of 13- to 14-year-old secondary education students ($n = 99$), spread over three HAVO classes (preparing for a university of applied science, $n = 76$) and one VMBO class (preparing for vocational education, $n = 23$) at a local high school participated in the pilot. Students had already studied the subject matter (supply and demand) as part of their regular curriculum and had previously participated in studies using Go-Lab ILSs and a version of the hypothesis scratchpad that did not provide feedback.

Materials and Procedure

The pilot revolved around a short ILS set in the *supply & demand* domain, where students were introduced to the interactions between price, supply, and demand. The ILS was created in collaboration with a participating economics teacher. Each class performed the study in a single 50-min session. At the beginning of a session, students were given an oral introduction detailing how to use the environment and refreshing them on what a hypothesis is. They were then asked to open the inquiry learning space, where they were first presented with information on the domain. They were then asked to create as many hypotheses about this domain as possible in the automated hypothesis scratchpad, and to use the feedback mechanism when they were stuck or wanted to check their hypothesis. An initial version of the parser was used that could detect the first three criteria: *it has* two *variables*, *it has a modifier*, and *it is a syntactically correct sentence*. Students were regularly encouraged and reminded to create as many hypotheses as

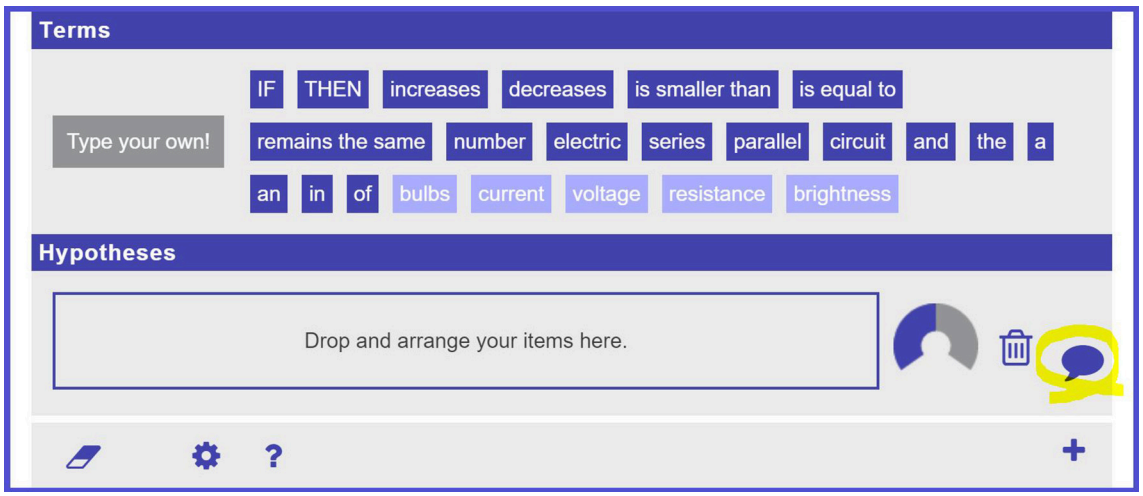

FIGURE 5 | Automated hypothesis scratchpad. The feedback button is highlighted.

TABLE 2 | Feedback for each criterion.

Criterion	Feedback	
	Wrong	**Correct**
Variables	Not enough variables, A hypothesis should always have at least two variables.	–
Modifier	You can only test a hypothesis if something changes. Without change, you cannot test the hypothesis.	–
Syntax	It appears you've entered an incomplete hypothesis. I can only give feedback on full hypotheses[a].	–
	I don't understand your hypothesis. Are you sure this is a correct hypothesis?; "[HYPOTHESIS]"[b]	–
CVS	If you don't change the value of [INDEPENDENT], you won't be able to test if this has an effect on [DEPENDENT].[c]	You're changing the value of [INDEPENDENT] to see if that has an effect on [DEPENDENT].[c]
	You're changing [INDEPENDENT] at the same time. You can't be sure which of these changes has an effect on [DEPENDENT].[d]	You're changing only the value of [INDEPENDENT], so you can be certain that any change in [DEPENDENT] is caused by [INDEPENDENT].[d]
Qualified	You did not describe the conditions in which your hypothesis applies.	You specified that your hypothesis only applies in a [QUALIFIER].

[HYPOTHESIS], [INDEPENDENT], [DEPENDENT], and [VARIABLE] will be dynamically replaced with the actual hypothesis and variables used by the student and recognized by the parser. The feedback has been translated from the Dutch original used in the experiments.
[a] Used when a hypothesis starts valid but is incomplete (partial parse).
[b] Used when a hypothesis cannot be parsed (nonsense, or syntax error).
[c] Used when the independent variable is not manipulated.
[d] Used when multiple independent variables are manipulated.

possible[3], but no attempt was made to force the creation of hypotheses or the use of the feedback tool. The session was concluded with a small user satisfaction questionnaire. During each session, the researcher and the classroom teacher monitored the class, answering process-related questions, and eliciting feedback if any out of the ordinary situations or interactions were encountered.

[3]Unfortunately, during one of the HAVO sessions the teacher instructed students to create 'at least 4' hypotheses, which was immediately interpreted as 'create 4 hypotheses'.

Results

A total of 979 hypotheses were collected from 96 students. Most students created three to five hypotheses and asked for feedback multiple times over the course of the experiment. One student asked for feedback 84 times and was removed as an outlier.

Inter-rater reliability between the parser and two human experts was almost perfect on all three criteria (Cohen's $\kappa = 0.81 - 0.96$), showing high parser accuracy. Hypotheses for which students requested and received feedback at least once were more likely to be correct on all criteria. This relation is visible in **Figure 6**, and statistically significant using a multilevel logistic model estimating the probability of a syntactically correct

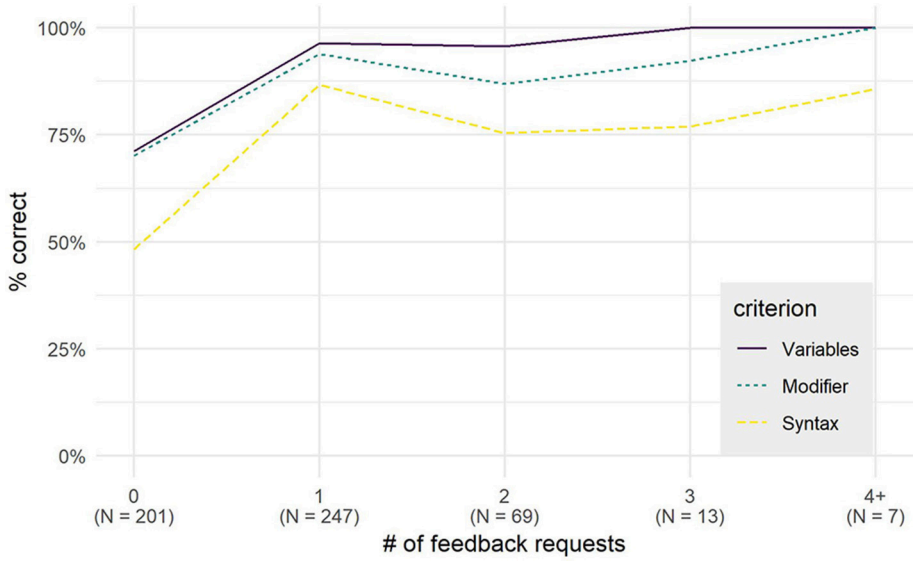

FIGURE 6 | Average performance on each criterion, by number of feedback requests.

hypothesis by the number of feedback requests, corrected for student and class effects, gender, and age ($\beta_{feedbackCount} = 1.00$, $SE_\beta = 0.17$, $CI_{OR} = 1.93 - 3.83$, $p < 0.001$), where $\beta_{feedbackCount}$ is the effect of each additional feedback request, and CI_{OR} the confidence interval of the Odds Ratio.

Discussion

The first pilot took place under test conditions; students were told to create as many hypotheses as possible, and the learning space was only there to provide a setting for hypotheses to be created. Such conditions are different from usual educational practice. Nevertheless, high parser accuracy and significantly increased quality of hypotheses showed that a parser is feasible, and that a hypothesis scratchpad enhanced with automated scoring and feedback is promising.

Therefore, a second pilot study was conducted using an expanded version of the context-free grammar that included all criteria listed in **Table 1**. In addition, the automated scratchpad was embedded in a full ILS, aligning much closer to how the tool is likely to be used in practice.

Second Pilot Study

Participants

Participants came from one HAVO class of 13 to 14-year-old secondary educations students ($n = 27$), at a local high school. The students had recently been introduced to electrical circuits as part of their regular curriculum but were familiar with neither Go-Lab environments nor the hypothesis scratchpad prior to the experiment.

Materials and Procedure

A short ILS in the *electrical circuits* domain that could be completed in a single 50-min session was created in collaboration with participating teachers. At the beginning of a session, students were given an oral introduction detailing how to use the tools in the ILS and refreshing them on what a hypothesis is. They were then asked to open the ILS, where they were presented with a short pre-test, followed by some information on the domain. To guide students' hypothesis construction, they were asked to enter two predictions about the change in brightness of lightbulbs in series and parallel circuits after adding another bulb. In the next steps, students were asked to turn these predictions into hypotheses in the automated hypothesis scratchpad, and design an experiment in the *Experiment Design* app [see e.g., (van Riesen et al., 2018)] to test their hypotheses. Finally, students were given time to create an experimental setup in the *Circuit Lab* virtual laboratory, test their hypotheses, and enter their conclusions.

All student actions took place in the ILS, which encompassed a full inquiry cycle, from orientation to conclusion. This created an environment more likely to occur in real educational settings. An expanded version of the automated hypothesis scratchpad was used, designed to be able to classify and give feedback to all the relevant criteria.

During the session, the researcher and the classroom teacher monitored the class, answering process-related questions and eliciting feedback if any out of the ordinary situations or interactions were encountered.

Results

Both the researcher and the classroom teacher noticed that students had problems working with the ILS and staying on-task. These problems were process related (e.g., students got distracted, skipped steps) and tool related (i.e., students did not know how to work with the tool). Attempts to provide instructions during the experiment were largely ineffective because students were at different stages of the ILS (making group instructions difficult), and there were too many students to provide individual instructions.

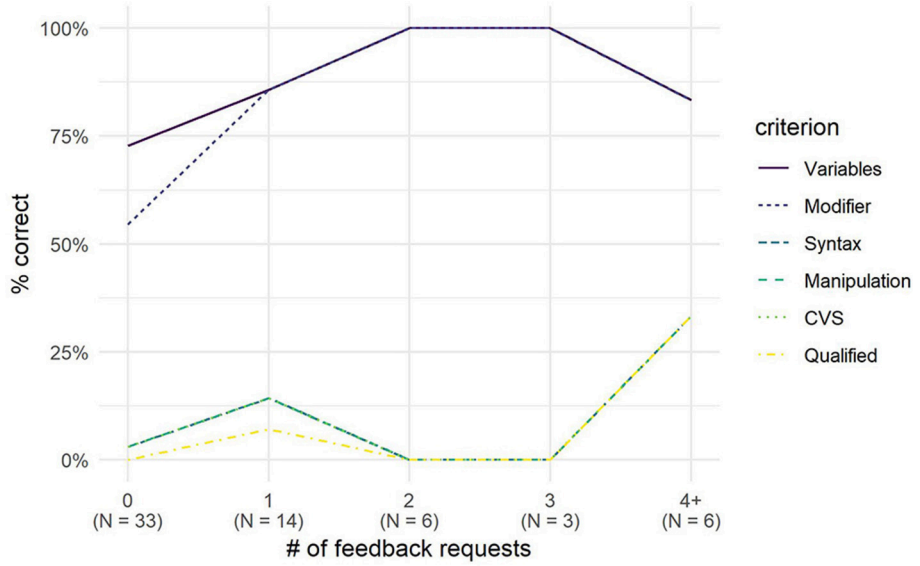

FIGURE 7 | Average performance on each criterion, by number of feedback requests. Note that the poor performance is at least partially due to low parser accuracy, and that the scores for the Syntax, manipulation, and qualified criteria overlap.

In addition, some of the written instructions were too long. For example, upon seeing the instructions, one student immediately uttered: "*too long, won't read.*" It seems likely that his sentiments were shared by other students, highlighting the need for verbal (or at least more interactive) instructions.

A total of 50 hypotheses were collected from 27 students. The plurality (13) of students created two hypotheses each, 7 students did not create any hypotheses. Most (16) students asked for feedback at least once, 11 students did not ask for feedback. One student asked for feedback 23 times and was removed as an outlier.

Parser accuracy was below expectations, achieving a Cohens' κ of 0.91, 0.90, and 0.40 on the *contains at least* two *variables*, *contains a modifier*, and *is a syntactically correct sentence* criterion, respectively. Accuracy for the *manipulates exactly* one *variable* and *is qualified* criteria is not reported, as the parser failed to recognize 30 out of 46 syntactically correct snapshots, leaving only 16 parsed snapshots.

Although there does appear to be a positive effect of feedback on hypothesis quality (see **Figure 7**), this effect was not statistically significant, as shown by a multilevel logistic model estimating the probability of a syntactically correct hypothesis by the number of feedback requests, correcting for student effects, gender and age ($\beta_{feedbackCount} = 0.46$, $SE_\beta = 0.24$, $CI_{OR} = 0.98 - 2.57$, $p = .058$).

Discussion

The number of collected hypotheses per student was lower than in the first pilot. In part, that was by design: the first pilot was specifically set up to encourage students to create as many hypotheses as possible, whereas, in this pilot students were guided to create two hypotheses. The participants in this pilot also had less experience working in an ILS, which caused several process-related issues during the session that likely influenced the number of hypotheses created. A more structured lesson plan where students start and end each step in the inquiry cycle at the same time will allow for verbal instructions to be given before starting each section.

Many students failed to distinguish between series and parallel circuits in their hypotheses, even when their predictions did show they understood the differences between the types of circuits. This does seem to indicate the need for supporting the creation of hypotheses while at the same time highlighting that the currently implemented support is insufficient.

Poor parser accuracy can be attributed to students' difficulties in working with the ILS, additional criteria introducing more complexity to the grammar, and a lack of training data for the *Electrical Circuits* domain in the target language (Dutch) to calibrate the parser. Using the data gathered in the pilot, we were able to make improvements to the grammar used by the parser. When applying this new grammar to the gathered hypotheses, inter-rater agreement on the syntax criterion was raised to moderate (Cohens' $\kappa = 0.53$).

Main Study

Participants

Six classes of 13- to 15-year-old secondary education students ($n = 132$), from two local high schools participated in the study. Six students used incorrect login credentials and were left out of the analyses. The remaining participants came from 4 HAVO classes ($n = 78$), and 2 VWO classes ($n = 48$). Students were randomly assigned to one of two conditions. Students in the experimental condition ($n = 68$) used the automated hypothesis scratchpad, while those in the control condition ($n = 58$) used a version of the hypothesis scratchpad that did not

provide feedback. No significant differences were present in the distribution of age, gender, and current physics grade across conditions (**Table 3**).

Materials and procedure

A single 50-min session was used, covering the same material as that of the second pilot study. The ILS used in the second pilot study was used again, with some minor changes to ameliorate some of the process-related issues students encountered. In particular, written descriptions and instructions were shortened. Instead, at the outset of the session and each phase, students were given a short oral introduction.

Students received a link to a randomizer[4] that assigned each student to one of two conditions and redirected them to the corresponding ILS. Students were instructed not to move to the next phase until told to do so.

At pre-set intervals during the sessions, the researcher gave an oral introduction to the next phase of the inquiry cycle, and the corresponding tools in the ILS. Students where then encouraged to start with that phase. In each session, the researcher and the class teacher monitored the students, answering process-related questions, and eliciting feedback if any extra-ordinary situations or interactions were encountered.

Results

Most students were already familiar with the GoLab environment and its tools and encountered no significant difficulties. Based on observations during the sessions, oral introductions prior to each phase of the ILS appeared to keep most students on task, most of the time.

Students in the experimental condition created 201 hypotheses, for 56 of which feedback was requested. Of the 68 students in the experimental condition, exactly half never asked for feedback.

Parser accuracy was moderate to almost perfect, achieving a Cohens' κ of 0.84, 0.70, and 0.59 on the *contains at least* two *variables*, *contains a modifier*, and *is a syntactically correct sentence* criterion, respectively, and $>$ 0.80 for the *manipulates exactly* one *variable* and *is qualified* criteria.

Figure 8 appears to show that on average the hypotheses generated in the experimental condition scored higher on all criteria. In addition, **Figure 9** suggests a positive relation between the number of feedback requests and the quality of hypotheses. In particular, hypotheses for which feedback was requested at least once appear to be of higher quality.

To test the effect of our tool on hypothesis quality, we fitted a multilevel logistic model, controlling for student and class effects, as well as gender, age, physics grade, and academic level. We found no significant effect from being assigned to the experimental condition ($\beta_{condition} = 0.25$, $SE_\beta = 0.34$, $CI_{OR} = 0.66 - 2.50$, $p = 0.472$). Given that half of all participants in the experimental group never requested feedback, this outcome was not unexpected.

However, when we split the experimental group in two, based on whether students requested feedback or not ($n = 34$ in both groups, **Figure 10**), and contrast those who requested feedback against those who did not or could not, controlling for student and class effects, as well as gender, age, physics grade and academic level, the effect of requesting feedback is significant ($\beta_{feedbackCount} = 1.47$, $SE_\beta = 0.42, CI_{OR} = 1.92 - 9.89$, $p < 0.001$).

It could be argued that students who did not request feedback when it was made available to them are less proficient students. However, a contrast analysis comparing students in the control condition (who could not ask for feedback) and those in the experimental condition who did not request feedback found no significant difference between the two groups on the syntactically correct criterion ($\beta_{condition} = -0.30$, $SE = 0.39$, $CI_{OR} = 0.34 - 1.60$, $p = 0.445$). We thus found no evidence to suggest that there was a difference between students who could have asked for feedback but did not do so, and students who did not have the option to ask for feedback.

GENERAL DISCUSSION

The creation of hypotheses is a critical step in the inquiry cycle (Zimmerman, 2007), yet students of all ages experience difficulties creating informative hypotheses (Mulder et al., 2010). Automated scaffolds can help students create informative hypotheses, but their implementation in the regular curriculum is often cost-prohibitive, especially since they can typically only be used in one specific domain and language. This study set out to create a hypothesis scratchpad that can automatically evaluate and score hypotheses and provide students with immediate feedback. We use a flexible Context-Free Grammar approach that can relatively easily be adapted and extended for other languages and domains. We described the development process of this tool over two pilot studies and evaluated its instructional effectiveness in a controlled experiment.

Across three studies, we showed that a hypothesis parser based on a context-free-grammar is feasible, attaining moderate to almost perfect levels of agreement with human coders. The required complexity of the parser is directly linked to the syntactical complexity of the domain. For example, the electrical circuits domain requires a more complex parser than the supply and demand domain. Further development of the context-free-grammar used in the parser will contribute to higher reliability and may extend it to other languages and domains.

The second pilot study illustrated that a lack of familiarity of students with the online environment and the tools used can have a negative effect on their performance. Students were distracted by technical and process related issues, and had difficulty remaining on-task. In the final experiment, we used a largely identical learning environment, but students were verbally introduced to each phase. These introductions allowed students to focus on the content of the learning environment, rather than on how to use the learning environment itself.

Nevertheless, when using the automated hypothesis scratchpad in a "typical" ILS, students often did not request

[4]A separate ILS was created for each condition. The randomizer forwarded the students browser to one of these conditions. Randomization was weighted to ensure a roughly equal distribution across conditions in each session.

TABLE 3 | Participant characteristics, by condition.

		Overall	Control	Experimental	Test statistic (df)	*P*-value
		126	58	68		
Gender (%)	Female	57 (45.2)	27 (46.6)	30 (44.1)	$\chi^2(1) = 0.01$	0.925
	Male	69 (54.8)	31 (53.4)	38 (55.9)		
Level (%)	HAVO	78 (61.9)	35 (60.3)	43 (63.2)	$\chi^2(1) = 0.02$	0.882
	VWO	48 (38.1)	23 (39.7)	25 (36.8)		
Mean age (SD)		13.96 (0.64)	13.98 (0.66)	13.95 (0.63)	$t(119.09) = 0.34$	0.736
Mean grade (SD)		6.50 (0.86)	6.52 (0.86)	6.49 (0.86)	$t(120.51) = 0.21$	0.836

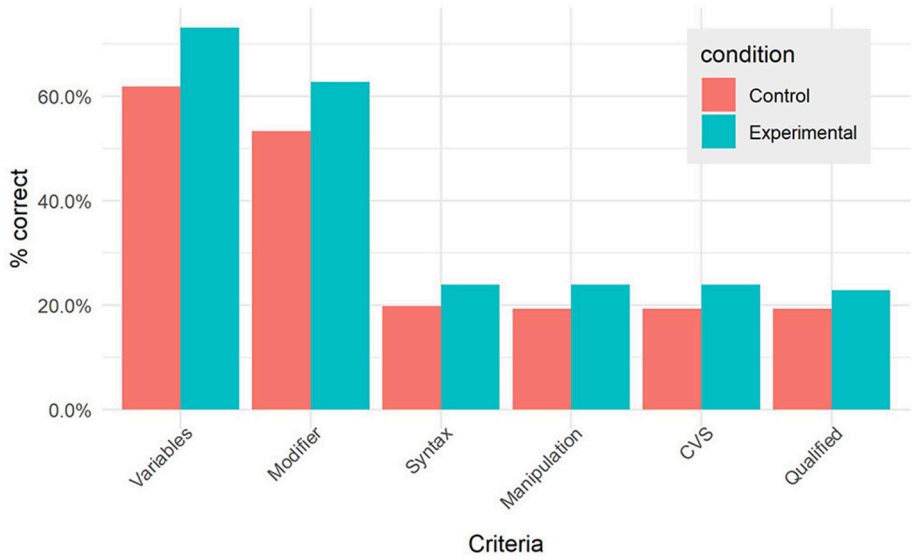

FIGURE 8 | Average performance by criterion, by condition.

feedback. Timmers et al. (2015) found a relation between gender and the willingness to ask for feedback, but such a relation was not present in our sample. In fact, none of the background variables collected (age, gender, physics grade and educational level) were significantly related to feedback requests or the quality of hypotheses.

If the goal was to obtain as many hypotheses as possible and assess the performance of the parser alone, we would have been better off following the approach taken in the first pilot. However, we deliberately chose to embed the automated hypothesis scratchpad in a typical ILS in the second pilot and main study, with the aim of replicating "real-world" conditions. In doing so, we can draw conclusions that are likely to be applicable to educational practice, rather than in laboratory conditions alone.

In the first pilot, the number of feedback requests was significantly related to the quality of hypotheses. This result was confirmed in a controlled experiment, where students who requested feedback were significantly more likely to create syntactically valid hypotheses than those who did not. The effects of feedback were immediate; hypotheses for which feedback was requested once where more likely to be correct.

To the best of our knowledge, no other tool exists that can reliably score hypotheses, can easily be adapted to different domains, and that allows students to create free-text hypotheses. The automated hypothesis scratchpad we present here can provide a clear and immediate benefit in science learning, provided students request feedback. By increasing the quality of students' hypotheses, we may assume that students are able to engage in more targeted inquiries, positively impacting their learning outcomes. How students can best be encouraged to request (and use) feedback is an open problem, and out of scope for this project. The automated hypothesis scratchpad could also be adapted to be a monitoring tool, highlighting students that may have difficulties creating hypotheses, allowing teachers to intervene directly.

The ability to reliably score hypotheses presents possibilities besides giving feedback. For example, hypothesis scores could serve as an indicator of inquiry skill. As such, they can be part of student models in adaptive inquiry learning environments. Crucially, obtaining an estimate from students' inquiry products

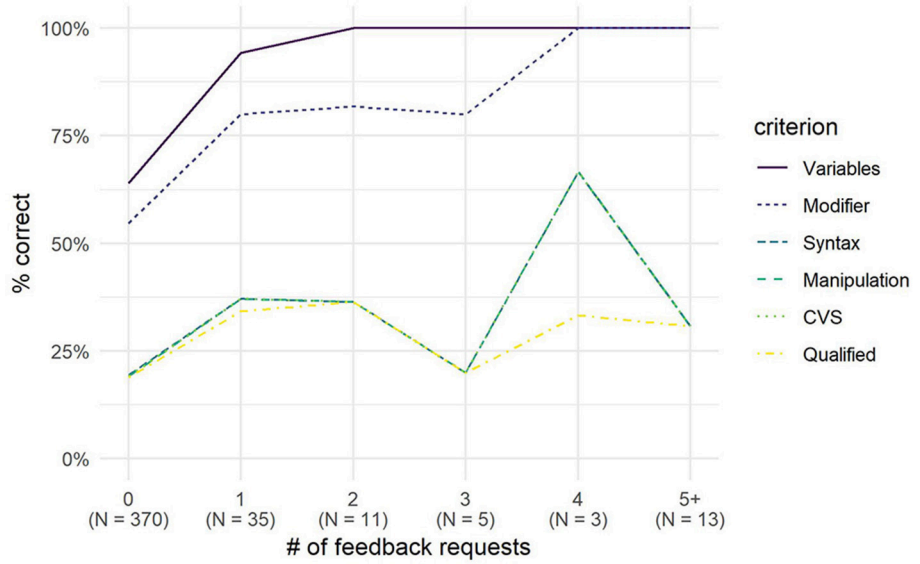

FIGURE 9 | Average performance by criterion for the experimental group, by number of feedback requests.

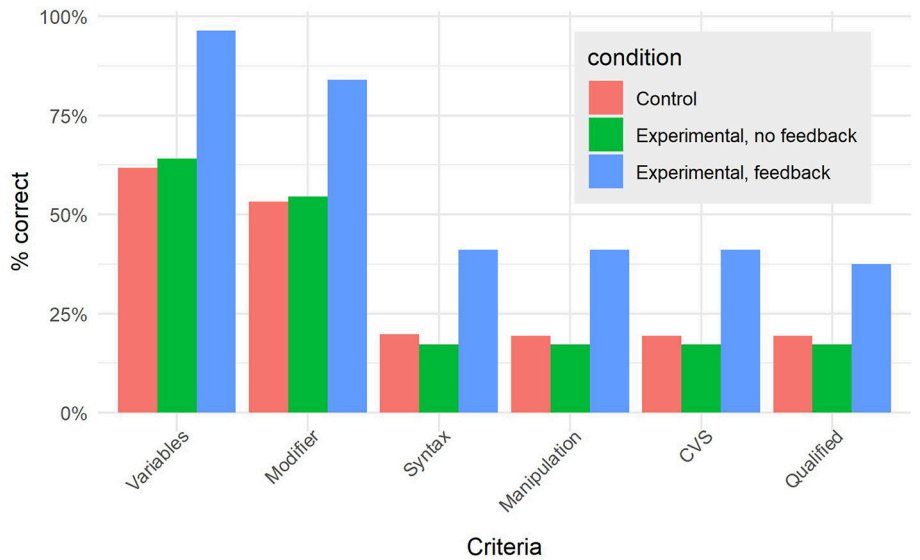

FIGURE 10 | Average performance on each criterion, by condition and feedback use.

is less obtrusive than doing so with a pre-test, and likely to be more reliable than estimates obtained from students' inquiry processes.

The aggregate hypothesis score computed for students did not have a known parametric distribution. This represents a serious limitation, as the score could not be used in statistical analyses. As a result, we chose to only test statistical significance based on the syntax criterion. Investigating alternative modeling techniques to arrive at a statistically valid conclusion based on multiple interdependent criteria will be part of our future work.

An automated hypothesis scratchpad providing students with immediate feedback on the quality of their hypotheses was implemented using context-free grammars. The automated scratchpad was shown to be effective; students who used its feedback function created better hypotheses than those who did not. The use of context-free grammars makes it relatively straightforward to separate the basic syntax of hypotheses, language specific constructs, and domain specific implementations. This separation allows for the quick adaptation of the tool to new languages and domains, allowing

configuration by teachers, and inclusion in a broad range of inquiry environments.

ETHICAL STATEMENT

All participating schools have obtained written and informed consent from students' parents to perform research activities that fall within the regular curriculum. Parents were not asked to give consent for this study specifically. The experiments we performed were embedded in the students' curriculum, and the collected data was limited to learning processes and outcomes. Students were briefed that their activities in the online learning environment would be logged, and that this data would be used in anonymized form. Both the research protocol and consent procedures followed were approved by the ethical board of the faculty of Behavioural, Management and Social Sciences of the University of Twente (ref # 17029).

AUTHOR CONTRIBUTIONS

KK, TdJ, AL, and SvdB designed the experiment. KK and AL designed the intervention. KK, SvdB, and BV performed statistical analyses, TdJ and AL helped put experimental results into context. KK wrote the manuscript, aided by TdJ, SvdB, AL, and BV.

REFERENCES

Aleven, V., Roll, I., Mclaren, B. M., and Koedinger, K. R. (2010). Automated, unobtrusive, action-by-action assessment of self-regulation during learning with an intelligent tutoring system. *Educ. Psychol.* 45, 224–233. doi: 10.1080/00461520.2010.517740

Alfieri, L., Brooks, P., Aldrich, N. J., and Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? A meta-analysis. *J. Educ. Psychol.* 103, 1–18. doi: 10.1037/a0021017

Anjewierden, A., Kamp, E. T., and Bollen, L. (2015). *Analysis of Hypotheses in Go-Lab*. Enschede.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using {lme4}. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Belland, B. R., Walker, A. E., Kim, N. J., and Lefler, M. (2016). Synthesizing results from empirical research on computer-based scaffolding in STEM education: a meta-analysis. *Rev. Educ. Res.* 87, 309–344. doi: 10.3102/0034654316670999

Bollen, L., and Sikken, J. (2018). *Hypothesis Scratchpad*. Available online at: https://www.golabz.eu/app/hypothesis-scratchpad (Accessed November 28, 2018).

Brod, G., Hasselhorn, M., and Bunge, S. A. (2018). When generating a prediction boosts learning: the element of surprise. *Learn. Instr.* 55, 22–31. doi: 10.1016/j.learninstruc.2018.01.013

Burns, B. D., and Vollmeyer, R. (2002). Goal specificity effects on hypothesis testing in problem solving. *Q. J. Exp. Psychol. Sec. A* 55, 241–261. doi: 10.1080/02724980143000262

Chomsky, N. (1956). Three models for the description of language. *IRE Transac. Inform. Theory* 2, 113–124. doi: 10.1109/TIT.1956.1056813

D'Angelo, C., Rutstein, D., Harris, C., Bernard, R., Borokhovski, E., and Haertel, G. (2014). *Simulations for STEM Learning: Systematic Review and Meta-Analysis Executive Summary*. Menlo Park, CA: SRI International.

de Jong, T., and Lazonder, A. W. (2014). "The guided discovery learning principle in multimedia learning," in *The Cambridge Handbook of Multimedia Learning,* ed R. Mayer (Cambridge: Cambridge University Press), 371–390. doi: 10.1017/CBO9781139547369.019

de Jong, T., Sotiriou, S., and Gillet, D. (2014). Innovations in STEM education: the Go-Lab federation of online labs. *Smart Learn. Environ.* 1:3. doi: 10.1186/s40561-014-0003-6

Durlach, P. J., and Ray, J. M. (2011). *Designing Adaptive Instructional Environments: Insights From Empirical Evidence (Technical Report 1297)*. Available online at: http://www.adlnet.gov/wp-content/uploads/2011/11/TR-1297.pdf

Earley, J. (1970). An efficient context-free parsing algorithm. *Commun. ACM* 13, 94–102. doi: 10.1145/362007.362035

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., et al. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proc. Natl. Acad. Sci.U.S.A.* 111, 8410–8415. doi: 10.1073/pnas.1319030111

Furtak, E. M., Seidel, T., Iverson, H., and Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: a meta-analysis. *Rev. Educ. Res.* 82, 300–329. doi: 10.3102/0034654312457206

Gerard, L. F., Matuk, C., McElhaney, K., and Linn, M. C. (2015). Automated, adaptive guidance for K-12 education. *Educ. Res. Rev.* 15, 41–58. doi: 10.1016/j.edurev.2015.04.001

Gerard, L. F., Ryoo, K., McElhaney, K. W., Liu, O. L., Rafferty, A. N., and Linn, M. C. (2016). Automated guidance for student inquiry. *J. Educ. Psychol.* 108, 60–81. doi: 10.1037/edu0000052

Gijlers, H., and de Jong, T. (2005). The relation between prior knowledge and students' collaborative discovery learning processes. *J. Res. Sci. Teach.* 42, 264–282. doi: 10.1002/tea.20056

Hake, R. R. (1998). Interactive-engagement versus traditional methods: a six-thousand-student survey of mechanics test data for introductory physics courses. *Am. J. Phys.* 66, 64–74. doi: 10.1119/1.18809

Hardmath123. (2017). *Nearley.js. GitHub*. Available online at: https://nearley.js.org

Hattie, J., and Timperley, H. (2007). The power of feedback. *Rev. Educ. Res.* 77, 81–112. doi: 10.3102/003465430298487

Ifenthaler, D. (2011). Bridging the gap between expert-novice differences: the model-based feedback approach. *J. Res. Technol. Educ.* 43, 103–117. doi: 10.1080/15391523.2010.10782564

Johnson, C. I., and Priest, H. A. (2014). "The feedback principle in multimedia learning," in *The Cambridge Handbook of Multimedia Learning*, ed R. Mayer (Cambridge: Cambridge University Press), 449–463. doi: 10.1017/CBO9781139547369.023

Kirschner, P. A., Sweller, J., and Clark, R. E. (2006). Why minimal guidance during instruction does not work: an analysis of the failure of constructivist, discovery, problem-based, experiential and inquiry-based teaching. *Educ. Psychol.* 41, 75–86. doi: 10.1207/s15326985ep4102_1

Klahr, D., and Dunbar, K. (1988). Dual space search during scientific reasoning. *Cogn. Sci.* 12, 1–48. doi: 10.1207/s15516709cog1201_1

Knuth, D. E. (1968). Semantics of context-free languages. *Math. Syst. Theory* 2, 127–145. doi: 10.1007/BF01692511

Kroeze, K. A. (2018). *Automated Hypothesis Scratchpad*. DANS. doi: 10.17026/dans-znq-knky

Landis, J. R., and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics* 33, 159–174. doi: 10.2307/2529310

Lazonder, A. W. (2014). "Inquiry learning," in *Handbook of Research on Educational Communications and Technology,* eds J. M. Spector, M. D. Merrill, J. Elen, and M. J. Bishop (New York, NY: Springer New York), 1–34. doi: 10.1007/978-1-4614-3185-5_36

Lazonder, A. W., and Harmsen, R. (2016). Meta-analysis of inquiry-based learning: effects of guidance. *Rev. Educ. Res.* 86, 681–718. doi: 10.3102/0034654315627366

Lazonder, A. W., Wilhelm, P., and van Lieburg, E. (2009). Unraveling the influence of domain knowledge during simulation-based inquiry learning. *Instr. Sci.* 37, 437–451. doi: 10.1007/s11251-008-9055-8

Linn, M. C., Clark, D., and Slotta, J. D. (2003). WISE design for knowledge integration. *Sci. Educ.* 87, 517–538. doi: 10.1002/sce.10086

Lukasenko, R., Anohina-Naumeca, A., Vilkelis, M., and Grundspenkis, J. (2010). Feedback in the concept map based intelligent knowledge assessment system. *Sci. J. Riga Technic. Univ. Comp. Sci.* 41, 8–15. doi: 10.2478/v10143-010-0020-z

Mayer, R. E. (2004). Should there be a three-strikes rule against pure discovery learning? The case for guided methods of instruction. *Am. Psychol.* 59, 14–19. doi: 10.1037/0003-066X.59.1.14

Mulder, Y. G., Lazonder, A. W., and de Jong, T. (2010). Finding out how they find it out: an empirical analysis of inquiry learners' need for support. *Int. J. Sci. Educ.* 32, 2033–2053. doi: 10.1080/09500690903289993

Narciss, S. (2006). "Informatives tutorielles Feedback," in *Entwicklungs-Und Evaluationsprinzipien Auf Der Basis Instruktionspsychologischer Erkenntnisse,* Münster:Waxmann.

Narciss, S. (2008). "Feedback strategies for interactive learning tasks," in *Handbook of Research on Educational Communications and Technology*, 3rd Edn., eds J. M. Spector, M. D. Merrill, J. van Merrienboer, and M. P. Driscoll (New York, NY: Lawrence Erlbaum Associates), 125–144. doi: 10.4324/9780203880869.ch11

Nye, B. D., Graesser, A. C., and Hu, X. (2014). "Multimedia learning with intelligent tutoring systems," in *The Cambridge Handbook of Multimedia Learning*, ed R. Mayer (Cambridge: Cambridge University Press), 705–728. doi: 10.1017/CBO9781139547369.035

Pedaste, M., Lazonder, A. W., Raes, A., Wajeman, C., Moore, E., and Girault, I. (2015). "Grand challenge problem 3: empowering science teachers using technology-enhanced scaffolding to improve inquiry learning," in *Grand Challenge Problems in Technology Enhanced Learning II: MOOCS and Beyond: Perspectives for Research, Practice, and Policy Making Developed at the Alpine Rendez- Vous in Villard-de-Lans,* eds K. Lund, P. Tchounikine, and F. Fischer (Cham: Springer Briefs in Education), 18–20.

Pedaste, M., Mäeots, M., Siiman, L. A., de Jong, T., van Riesen, S. A. N., Kamp, E. T., et al. (2015). Phases of inquiry-based learning: definitions and the inquiry cycle. *Educ. Res. Rev.* 14, 47–61. doi: 10.1016/j.edurev.2015.02.003

Prince, M. J., and Felder, R. M. (2007). The many faces of inductive teaching and learning. *J. Coll. Sci. Teach.* 36, 14–20. Available online at: www.jstor.org/stable/42992681

Quinn, M. E., and George, K. D. (1975). Teaching hypothesis formation. *Sci. Educ.* 59, 289–296. doi: 10.1002/sce.3730590303

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. Vienna. Available online at: https://www.r-project.org/

Ryoo, K., and Linn, M. C. (2016). Designing automated guidance for concept diagrams in inquiry instruction. *J. Res. Sci. Teach*. 53, 1003–1035. doi: 10.1002/tea.21321

Shermis, M. D., and Burstein, J. (2013). *Handbook of Automated Essay Evaluation: Current Applications and New Directions*. New York, NY: Routledge. doi: 10.4324/9780203122761

Shieber, S. M. (2003). *An Introduction to Unification-Based Approaches to Grammar*. Brookline, MA: Microtome Publishing.

Timmers, C. F., Walraven, A., and Veldkamp, B. P. (2015). The effect of regulation feedback in a computer-based formative assessment on information problem solving. *Comp. Educ.* 87, 1–9. doi: 10.1016/j.compedu.2015.03.012

Tschirgi, J. E. (1980). Sensible reasoning: a hypothesis about hypotheses. *Child Dev.* 51, 1–10. doi: 10.2307/1129583

Van der Kleij, F. M., Feskens, R. C. W., and Eggen, T. J. H. M. (2015). Effects of feedback in a computer-based learning environment on students' learning outcomes: a meta-analysis. *Rev. Educ. Res.* 85, 475–511. doi: 10.3102/0034654314564881

Van Joolingen, W. R., and De Jong, T. (1991). Supporting hypothesis generation by learners exploring an interactive computer simulation. *Instruct. Sci.* 20, 389–404. doi: 10.1007/BF00116355

Van Joolingen, W. R., and De Jong, T. (1993). Exploring a domain with a computer simulation: traversing variable and relation space with the help of a hypothesis scratchpad. *Simulat. Based Exp. Learn.* 191–206. doi: 10.1007/978-3-642-78539-9_14

van Riesen, S. A. N., Gijlers, H., Anjewierden, A., and de Jong, T. (2018). The influence of prior knowledge on experiment design guidance in a science inquiry context. *Int. J. Sci. Educ.* 40, 1327–1344. doi: 10.1080/09500693.2018.1477263

Vandewaetere, M., Desmet, P., and Clarebout, G. (2011). The contribution of learner characteristics in the development of computer-based adaptive learning environments. *Comput. Human Behav.* 27, 118–130. doi: 10.1016/j.chb.2010.07.038

Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. *Dev. Rev.* 27, 172–223. doi: 10.1016/j.dr.2006.12.001

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.