Computers & Education 82 (2015) 460–485

Contents lists available at ScienceDirect

Computers & Education

journal homepage: www.elsevier.com/locate/compedu

Connection making between multiple graphical representations: A multi-methods approach for domain-specific grounding of an intelligent tutoring system for chemistry

Martina A. Rau*, Joseph E. Michaelis, Natalie Fay

Department of Educational Psychology, University of Wisconsin–Madison, USA

ARTICLE INFO

Article history:
Received 21 August 2014
Received in revised form
4 December 2014
Accepted 5 December 2014
Available online 18 December 2014

Keywords:
Connection making
Multiple representations
Perceptual learning
Intelligent tutoring systems
Chemistry

ABSTRACT

Making connections between graphical representations is integral to learning in science, technology, engineering, and mathematical (STEM) fields. However, students often fail to make these connections spontaneously. Intelligent tutoring systems (ITSs) are suitable educational technologies to support connection making. Yet, when designing an ITS for connection making, we need to investigate what concepts and learning processes play a role within the specific domain. We describe a multi-methods approach for grounding ITS design in the specific requirements of the target domain. Specifically, we applied this approach to an ITS for connection making in chemistry. We used a theoretical framework that describes potential target learning processes and conducted a series of four empirical studies to investigate what role graphical representations play in chemistry knowledge and to investigate which learning processes related to connection making play a role in students' learning about chemistry. These studies combined multiple methods, including knowledge testing, eye tracking, interviews, and log data analysis. We illustrate how our findings inform the design of an ITS for chemistry: Chem Tutor. Results from two pilot studies done in the lab and in the field with altogether 99 undergraduates suggest that Chem Tutor leads to significant and large learning gains on chemistry knowledge.

1. Introduction

Multiple graphical representations are ubiquitous in science, technology, engineering, and mathematics (STEM) domains. For instance, line graphs, coordinate systems, pie and bar charts, and sets are used in mathematics (Arcavi, 2003; Cheng, 1999; Noss, Healy, & Hoyles, 1997); Lewis dot structures, electrostatic potential maps, and ball-and-stick figures are used in chemistry (Kozma, Chin, Russell, & Marx, 2000; Stieff, Hegarty, & Deslongchamps, 2011; Zhang & Linn, 2011); diagrams, charts, and graphs are used in physics (Larkin & Simon, 1987; Lewalter, 2003; Urban-Woldron, 2009). In all of these domains, learning of the domain knowledge depends on students' ability to make connections between representations (Ainsworth, 2006; Gobert et al., 2011; de Jong et al., 1998), and many students struggle doing so (Ainsworth, Bibby, & Wood, 2002; Rau, Rummel, Aleven, Pa cilio, & Tunc-Pekkan, 2012). Multiple graphical representations can enhance learning of the domain knowledge because different representations emphasize complementary conceptual aspects of the learning material and have different effects on mental processing (Kozma et al., 2000; Larkin & Simon, 1987; Schnotz & Bannert, 2003). However, students' benefit from multiple representations depends on their ability to make connections between them (Ainsworth, 2006; Bodemer & Faust, 2006; Bodemer, Ploetzner, Bruchmüller, & Häcker, 2005; Bodemer, Ploetzner, Feuerlein, & Spada, 2004; Brünken, Seufert, & Zander, 2005; Butcher & Aleven, 2008; Gutwill, Frederiksen, & White, 1999; van der Meij & de Jong, 2006; Seufert & Brünken, 2006; Taber, 2001). For instance, to learn about chemical bonding, students need to make connections between Lewis structures, ball-and-stick figures, space-filling models, and electrostatic potential maps (EPMs; see Fig. 1). Connection making is a difficult task that students often do not engage in spontaneously, even though it is critical to their learning (Ainsworth et al., 2002; Rau, Rummel, et al., 2012). Hence, they need

* Corresponding author. Department of Educational Psychology, University of Wisconsin–Madison, 1025 W Johnson St, Madison WI 53706, USA.
E-mail address: marau@wisc.edu (M.A. Rau).

http://dx.doi.org/10.1016/j.compedu.2014.12.009
0360-1315/

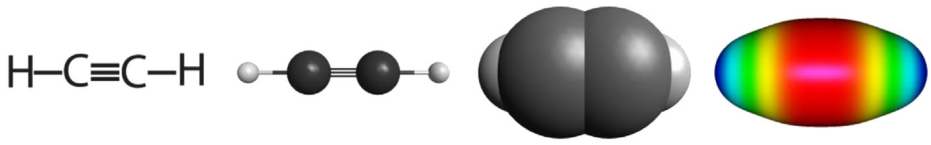


Fig. 1. Graphical representations of ethyne: Lewis structure, ball-and-stick figure, space-filling model, electrostatic potential map (EPM).

support to make these connections. Prior research shows that connection-making support can enhance students' learning outcomes in STEM domains (Bodemer & Faust, 2006; van der Meij & de Jong, 2006; Seufert, 2003).

Recent research indicates that intelligent tutoring systems (ITSs) can be effective in supporting connection making (Rau, Aleven, Rummel, & Rohrbach, 2012). ITSs support step-by-step problem solving (VanLehn, 2011) and provide adaptive instructional support (Corbett, Koedinger, & Hadley, 2001; Koedinger & Corbett, 2006). Adaptive support in ITSs typically includes feedback upon the diagnosis of a student's misconception (e.g., based on certain errors he/she makes while solving a problem), hints on demand (e.g., the student requests help on solving a step), and problem selection (e.g., based on the student's diagnosed knowledge level, the tutor selects a new problem that is considered to be of appropriate difficulty).

A key open question we face when designing connection making support is how to identify what specific learning processes play a role within the given target domain; that is, how to *ground* the design of support in the domain-specific requirements. The goal of this paper is to describe a multi-methods approach for grounding the design of an ITS in a particular domain. We describe how we applied our approach to the design of an ITS for connection making in chemistry: Chem Tutor. We conducted four empirical studies. Studies 1 and 2 focused on the role connection making plays in how chemistry knowledge is structured. Studies 3 and 4 focused on the role of connections between graphical representations in how students learn about chemistry. Across these studies, we pursued the following research goals:

1. Identify learning processes that are important for connection making between multiple graphical representations in chemistry;
2. Identify visual attention behaviors that indicate productive learning processes as students make connections between multiple graphical representations in chemistry;
3. Improve students' learning of important concepts in chemistry.

We conclude this paper by arguing that, even though we address these goals within the chemistry domain, our approach is applicable to other STEM domains than chemistry and to other educational technologies than ITSs. Furthermore, we believe that our approach can fundamentally improve STEM education by helping students take better advantage of multiple graphical representations that are ubiquitous in their learning materials.

2. Theoretical background

As mentioned, prior research shows that learning of domain knowledge critically depends on the students' ability to make connections between multiple representations (Ainsworth, 2006; Ainsworth et al., 2002; Cook, Wiebe, & Carter, 2007; Eilam & Poyas, 2008; Gutwill et al., 1999; de Jong et al., 1998; Özgün-Koca, 2008; Schnotz & Bannert, 2003; Schwonke, Ertelt, & Renkl, 2008; Schwonke & Renkl, 2010; Taber, 2001). We distinguish between the broader category of external representations (which includes symbolic and graphical representations), and the more specific category of graphical representations. Symbolic representations, such as text, are composed of features that have arbitrary relation to the real-world aspects they describe. Symbolic representations are interpreted based on their semantic meaning that we encode based on previously learned conventions (e.g., "1" stands for a quantity of one of something). Graphical representations are composed of perceptual features that have identifiable correspondence to the real-world aspects they depict. Therefore, graphical representations can be encoded based on their perceptual meaning (e.g., the ball-and-stick figure for ethane in Fig. 1 shows four spheres because ethane is composed of four atoms). To use graphical representations to learn about domain content, students have to learn which perceptual features of the graphical representations to attend to, how to interpret these features, and how to map these features to other representations (i.e., in the case of multiple external representations to symbolic representations, or in the case of multiple graphical representations to other graphical representations). For example, to understand the graphical representations shown in Fig. 1, students need to learn that the color in ball-and-stick figures and space-filling models denotes the identity of the atom, whereas that the color (in web version) in EPMs denotes regions of high electron density. Thus, learning with graphical representations involves a considerable amount of perceptual learning (Kellman & Massey, 2013; Kellman, Massey, & Son, 2009).

Most research on representations has focused on the broader category of learning with multiple *external* representations (Ainsworth & Loizou, 2003; Bodemer et al., 2005; Butcher & Aleven, 2007; Magner, Schwonke, Aleven, Popescu, & Renkl, 2014; Rasch & Schnotz, 2009), but only few studies have focused on learning with multiple *graphical* representations. Yet, multiple graphical representations are ubiquitous in STEM domains (Arcavi, 2003; Cook et al., 2007; Kordaki, 2010; Kozma et al., 2000; Lewalter, 2003; Nathan, Walkington, Srisurichan, & Alibali, 2011) and can significantly enhance learning outcomes compared to text and one additional graphical representation (Rau, Aleven, & Rummel, 2014).

A critical difference between multiple external and multiple graphical representations is that multiple external representations typically involve text as a dominant representation. One may assume that students are highly fluent in processing text. When students make connections between multiple external representation, the text guides their visual attention as they process the graphical representation (Rayner, Rotello, Stewart, Keir, & Duffy, 2001; Schmidt-Weigand, Kohnert, & Glowalla, 2010). By contrast, in the case of multiple graphical representations, there is no dominant text representation that we can assume students to be highly fluent with because they may not yet be fluent in processing graphical representations. To make connections between multiple graphical representations, students need to map relevant perceptual features across different representations. This task is not straightforward, because different graphical representations

tend to share both critical and incidental perceptual features. For example, when making connections between the graphical representations shown in Fig. 1, students need to learn that corresponding colors (in the web version) in EPMs and ball-and-stick figures denote different information and thus are not critical for connection making between these two particular graphical representations. Thus, the main difference to multiple external representations is that students have to engage in these perceptual learning tasks without the guidance of a dominant text representation. Because they cannot rely on the guidance of text, students may need additional support for the perceptual learning processes involved in connection making when they learn with multiple graphical representations. Therefore, to help students learn with multiple graphical representations, we may need to take a stronger focus on perceptual learning processes than has been common in research on multiple external representations. To address this need, we draw on a theoretical framework that specifically focuses on the case of learning with multiple graphical representations: the MGR-framework (Rau, under review) to inform the design of Chem Tutor.

2.1. Theoretical framework for learning with multiple graphical representations

The MGR-framework (Rau, under review) proposes that two types of connection-making abilities play a role in domain expertise. In specifying these abilities, the MGR-framework draws on the Knowledge-Learning and Instruction (KLI) framework (Koedinger, Corbett, & Perfetti, 2012), which distinguishes different learning processes that lead to the acquisition of different types of knowledge. In most domains, learning involves the ability to make sense of domain-relevant concepts (Koedinger et al., 2012). With respect to learning with multiple graphical representations, *sense-making ability* is defined as principled understanding of concepts depicted in graphical representations based on their knowledge components. Knowledge components are "acquired units of cognitive function or structure that can be inferred from performance on a set of related tasks" (Koedinger et al., 2012, p. 764). Sense-making ability means that a student is able to establish relations between corresponding knowledge components of different graphical representations. As previously mentioned, prior research shows that the ability to make connections between representations is a crucial prerequisite to students' learning of the domain knowledge (Ainsworth, 2006; Ainsworth et al., 2002; Cook et al., 2007; Eilam & Poyas, 2008; Gutwill et al., 1999; de Jong et al., 1998; Özgün-Koca, 2008; Schnotz & Bannert, 2003; Schwonke et al., 2008; Schwonke & Renkl, 2010; Taber, 2001). A student with high sense-making ability can relate aspects that correspond to one another across representations because they depict the same concept (e.g., in the example shown in Fig. 1, relating the local negative charge that results from the triple bond shown in the Lewis structure to the region of high electron density depicted by the red color (in the web version) in the EPM). In the following, we will refer to learning processes that result in the students' acquisition of sense-making ability as sense-making processes, and to learning interventions that aim at helping students acquire sense-making ability as sense-making support.

However, expertise does not only involve the ability to make sense of concepts; knowledge is only useful if it is readily accessible whenever needed. A learner who has readily accessible knowledge is said to have fluency in that knowledge (Koedinger et al., 2012). Often, fluency is considered as the ability to retrieve facts from memory (Arroyo, Royer, & Woolf, 2011). By contrast, we focus on *perceptual fluency*, which has been described as the ability to "extract information […] as the result of experience and practice" (Gibson, 1969, p.3). Kellman and Garrigan (2009) describe perceptual fluency as the ability to quickly and effortlessly pick up "relevant features and structural relations that define important classifications" (p. 55), and as the ability to "[extract] information more quickly and automatically with practice" (Kellman et al., 2009, p. 287). This type of fluency is an important aspect of domain expertise (Kellman et al., 2008; Kellman et al., 2009). It is acquired via unconscious forms of learning (Fahle & Poggio, 2002), and is neither conceptual or procedural knowledge (Kellman & Garrigan, 2009). Perceptual fluency results from experience with the perceptual properties of graphical representations and is characterized by readily accessible perceptual knowledge about graphical representations. A student who is perceptually fluent can rapidly and effortlessly find representations that depict the same concept, by relying on perceptual characteristics (Kellman et al., 2008, 2009; e.g., by rapidly seeing that the representations in Fig. 1 likely show the same molecule based on their linear geometry), rather than by reasoning about their knowledge components. In other words, perceptually fluent students treat one graphical representation as a single perceptual chunk, which allows them to perform domain-relevant tasks quickly and effortlessly. In the following, we will refer to learning processes that result in the acquisition of perceptual fluency as fluency-building processes, and to learning interventions that help students become perceptually fluent as fluency-building support.

2.2. Connection making in chemistry

Chemistry is a suitable domain to investigate how to support connection making. Connection making between multiple graphical representations is an important educational problem in chemistry. Chemistry instruction heavily relies on the use of graphical representations (Bodner & Domin, 2000; Coll & Treagust, 2003a), because many key concepts cannot be observed with the regular eye (Davidowitz & Chittleborough, 2009), and because many concepts are inherently abstract (Justi & Gilbert, 2002). Different representations provide complementary views on these concepts (Coll & Treagust, 2003a; Kozma & Russell, 2005b). Relying on only one representation can severely interfere with students' learning (Furio, Calatayud, Barcenas, & Padilla, 2000; Gabel & Bunce, 1994). Thus, connection making is key to students' learning of chemistry concepts (Kozma & Russell, 2005b; Wu, Krajcik, & Soloway, 2001). Indeed, the chemistry education literature widely acknowledges that students' conceptual difficulties are related to their difficulties in making connections between graphical representations (Dori & Barak, 2001; Gilbert & Treagust, 2009; Talanquer, 2013; Wu & Shah, 2004). Furthermore, the chemistry education literature widely acknowledge that both sense-making abilities (Linenberger & Bretz, 2012; Wu et al., 2001) and perceptual fluency (Justi, Gilbert, & Ferreira, 2009; Kozma & Russell, 2005a; Wu et al., 2001) are important aspects of connection making. Both abilities are considered to be important prerequisites to students' learning of chemistry concepts. Conceptually making sense of why different representations show the same phenomenon (Stieff et al., 2011; Strickland, Kraft, & Bhattacharyya, 2010) and how they provide complementary information (Talanquer, 2013; Williamson, 2014) is critical to understanding chemistry concepts. In addition, perceptual fluency—that is, the ability to fluently use multiple representations and to translate among them rapidly and with ease—is an important prerequisite for students' ability to learn about chemistry (Taber, 2013, 2014).

Given the importance of connection making in chemistry, it is not surprising that several educational technologies exist that support connection making between graphical representations of chemical phenomena. However, existing educational technologies focus on *sense-making ability* in connection making while disregarding *perceptual fluency*. For example, Connected Chemistry (Stieff, 2005) presents students with multiple graphical representations of phase changes and allows students to manipulate one representation while observing changes in another one. This intervention targets students' ability to conceptually understand how different representations depict the same concept, which is an important aspect of sense-making ability. SVM:Chem (Kozma & Russell, 2005a) is designed to facilitate class discussions and homework questions by exposing them to graphical and symbolic representations. This intervention helps students make sense of how graphical and symbolic representations of chemistry concepts relate to one another, which is also an important aspect of sense-making ability. ChemSense (Michalchik, Rosenquist, Kozma, Kreikemeier, & Schank, 2008) presents a variety of graphical and symbolic representations. This technology is designed to enhance students' sense-making ability through collaborative learning. eChem (Wu et al., 2001) supports students in making sense of connections between representations by providing feedback and hints.

Recently, a few interventions have focused on perceptual fluency. Eastwood (2013) describes a case study of a game that helps students to become fluent in translating from symbolic representations to physical ball-and-stick figures. Moreira (2013) describes an observational study in which students learn to rapidly name a molecule presented visually—in other words, to translate a graphical representation into a symbolic representation. Both studies show positive effects on engagement and reproduction, but they did not assess students' learning of chemistry concepts. Furthermore, these interventions were not technology based.

The lack of focus on perceptual fluency in educational technologies may be surprising, given that the chemistry education literature suggests that both abilities are important aspects of chemistry expertise (Cheng & Gilbert, 2009; Gilbert & Treagust, 2009). One reason for this tendency may be the aforementioned fact that prior research on connection making has mostly focused on the case of learning with multiple *external* representations, which corresponds to the format used in text books.

In summary, our goal to develop an ITS for connection making in chemistry targets an important educational problem and addresses the fact that existing educational technologies do not integrate support for sense-making and fluency-building processes.

3. Domain-specific grounding in chemistry knowledge structures

Our first goal in grounding the design of an ITS for connection making in the chemistry domain was to focus on how chemistry knowledge is structured. To this end, we conducted two empirical studies that instantiate the MRG-framework for the specific domain of chemistry. Our approach was to combine complementary data sources that yield insights into process-level and performance-level aspects of chemistry knowledge. Findings from these studies provided the basis for the development of a first version of Chem Tutor and for the tests we use to evaluate the effectiveness of Chem Tutor.

3.1. Study 1: Assessment of sense-making ability and perceptual fluency

The chemistry education literature documents the importance of both sense-making ability and perceptual fluency in connection making (Cheng & Gilbert, 2009; Gilbert & Treagust, 2009). Confirming the claim that these are indeed distinct abilities is a prerequisite for the design of separate ITS activities to support each of these abilities. Thus, Study 1 investigated:

Research question 1: Are sense-making ability and perceptual fluency separate connection-making abilities in chemistry?

To investigate the hypothesis that sense-making ability and perceptual fluency are distinct abilities, we conducted an *a priori* factor analysis on a chemistry knowledge test.

3.1.1. Methods

3.1.1.1. Participants. Undergraduate and graduate students from a large Midwestern university were recruited to take a 30–40 min test online. The institution's undergraduate population for the relevant semester was composed of 44.9% men and 55.1% women, 2.8% African American, 6.2% Asian, 1.2% Native American, 4.3% Hispanic, 11.9% international, and 72% Caucasian students.

To target students with varying levels of chemistry expertise, we advertised the study via mailing lists of the chemistry department and via posters that were hung at various locations in the chemistry department. One-hundred eighteen students started the test, but six of them terminated before answering questions about their prior chemistry courses, so a sample of 112 students was used for further analyses. One-hundred eleven students reported having taken at least one introductory undergraduate course. Forty-two students reported having taken at least one intermediate undergraduate course. Six students reported having taken advanced undergraduate courses. The student who did not report having taken any introductory undergraduate courses reported having taken five intermediate and one advanced undergraduate courses.

3.1.1.2. Procedure. Students took the test online, and their participation was anonymous. They had a chance of winning a $25 cash prize for their participation and were told that the chance of winning is 1:25. Before starting the test, students agreed to a consent form and selected the chemistry courses they had taken from a list of introductory, intermediate, and advanced chemistry courses that are offered at the study institution.

3.1.1.3. Materials. The test contained 16 items designed to assess sense-making ability, and 9 items to assess fluency. We included two types of sense-making items. Sense-similarities items asked students to reason about similarities between pairs of graphical representations of the same molecule (8 items) and sense-differences items asked students to reason about differences between pairs of representations of the same molecule (8 items). Students solved these items via a multiple-choice selection. Fig. 2 shows an example of a sense-differences item.

The perceptual fluency items required students to match pairs of graphical representations that showed the same molecule. In each item, students were given six graphical representations of one type (e.g., six space-filling models of different molecules) and had to map them to one of six graphical representations of a different type (e.g., six EPMs of different molecules). Students had the option of answering that none of the choices applied (however, this choice was always incorrect). Fig. 3 shows an example of a perceptual fluency item.

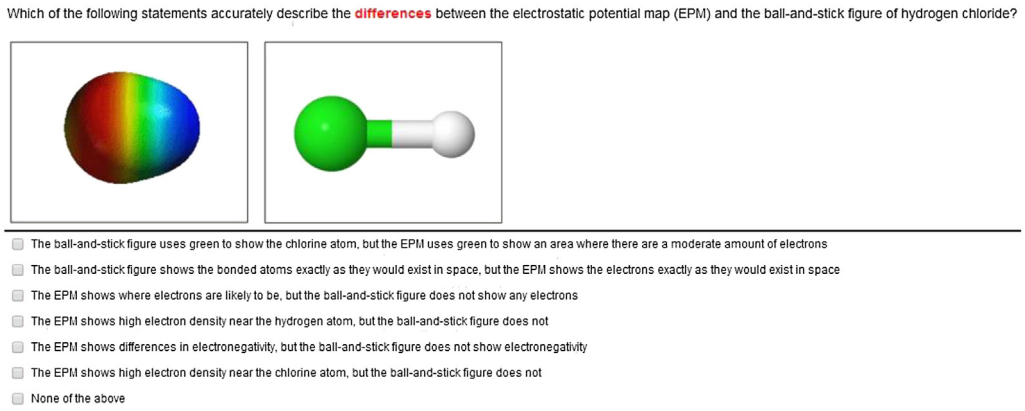


Fig. 2. Sample item for sense-making differences test used in Study 1.

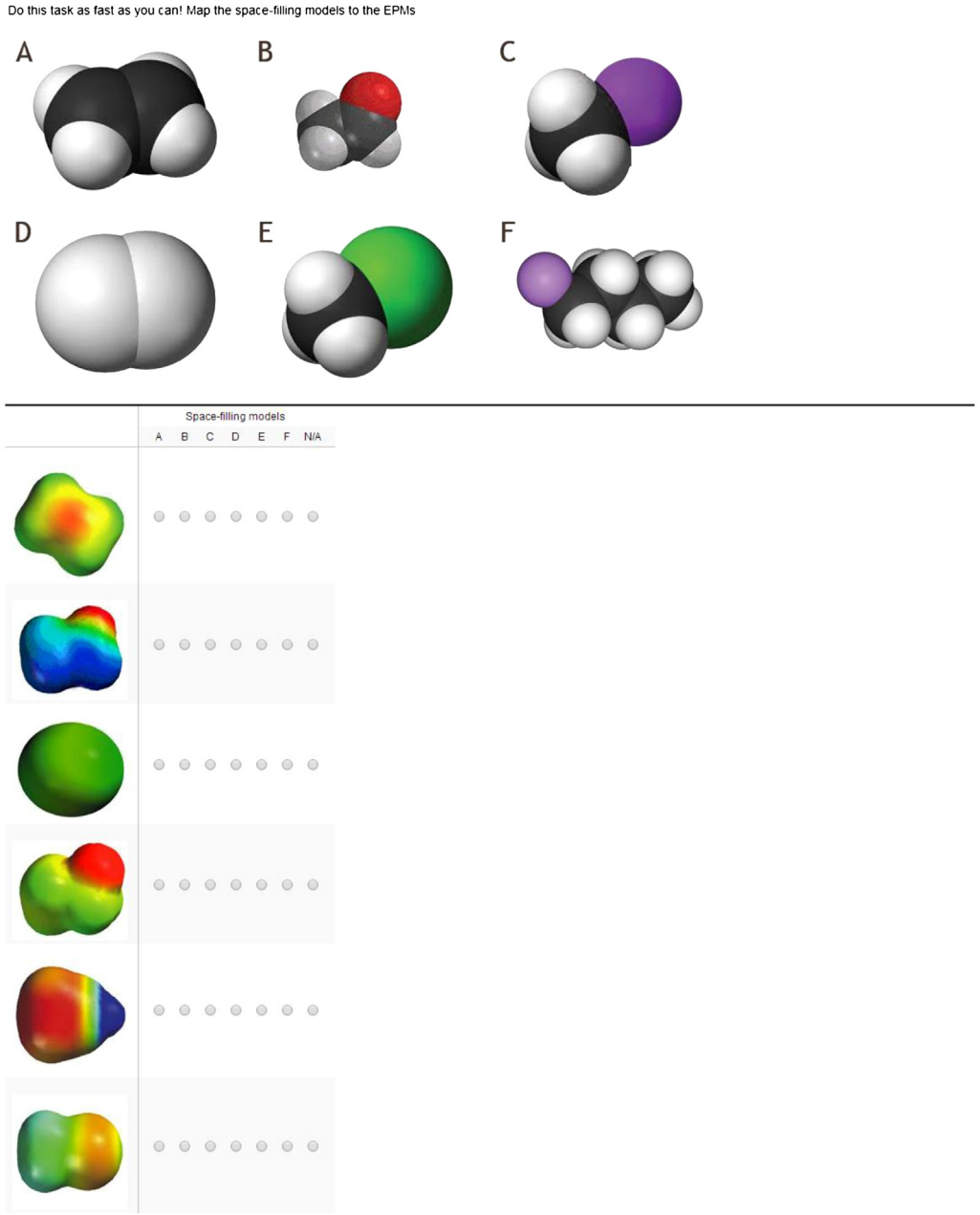


Fig. 3. Sample item for perceptual fluency test used in Study 1.

3.1.1.4. Analysis. We used the SPSS AMOS software to compare several factor models: a single-factor model (that does not distinguish between sense-making ability and fluency), a two-factor model (that distinguishes sense-making ability and fluency), and a three-factor model (that distinguishes sense-similarities, sense-differences, and fluency).

3.1.2. Results

Out of the 112 students who started working on the test items, 44 students completed all test items. Test items were presented in random order so that the missing data resulting from incomplete tests can be assumed to be at random. Thus, we excluded missing values on an item-by-item basis. To compare the fit of the tested models, we used root mean squared error (RMSE). The results show that the 3-factor model (RMSE = .072) and the 2-factor model (RMSE = .082) both yielded a better fit than the 1-factor model (RMSE = .088). Because the sense-differences and sense-similarities factors in the 3-factor model correlated highly with $r = .93$, we choose the 2-factor model for further analyses. The resulting two factors, sense-making ability and perceptual fluency, correlated moderately with $r = .62$. Both scales had good reliability, with a Cronbach's Alpha of .80 for the sense-making scale and a Cronbach's Alpha of .90 for the perceptual-fluency scale.

A repeated measures ANOVA showed that students performed significantly better on the sense-making scale ($M = .75$; $SD = .12$) than on the perceptual-fluency scale ($M = .62$; $SD = .24$; $p < .01$). To investigate the relation of these two abilities with chemistry proficiency, we computed correlations with the number of chemistry courses taken. The number of courses taken was associated with marginally higher sense-making ability ($r = .22$, $p < .10$), and with significantly higher perceptual fluency ($r = .45$, $p < .01$).

3.1.3. Discussion

Research question 1 asked: Are sense-making ability and perceptual fluency separate connection-making abilities in chemistry? To address this question, we designed test items that assessed students' ability to make sense of differences and similarities between representations (i.e., aspects of sense-making ability), and test items that assessed their ability to quickly find corresponding representations based on perceptual aspects. Based on the chemistry education literature (Cheng & Gilbert, 2009; Gilbert & Treagust, 2009), we hypothesized that sense-making ability and perceptual fluency are separate skills in chemistry. Our findings are in line with this hypothesis: we found a better model fit when distinguishing between sense-making items and the perceptual-fluency items.

The finding that students have higher sense-making ability than fluency is not surprising: it mimics the previously-mentioned current trend in educational practice that connection making focuses solely on sense-making processes. Thus, the findings from Study 1 encourage the goal to develop problems for Chem Tutor that specifically target perceptual fluency. By contrast, the finding that chemistry proficiency (approximated by the number of courses taken) is more strongly associated with perceptual fluency than with sense-making ability is surprising. It seems that chemistry instruction does not sufficiently target the ability to make sense of connections between graphical representations. Given that students' performance on the sense-making scale is still relatively low ($M = .75$; $SD = .12$), these findings indicate that there is an instructional need to support students' sense-making ability in connection making.

Study 1 has several limitations. First, we did not collect data on how long it took students to solve sense-making and perceptual-fluency items. Therefore, we cannot investigate how efficiently students made connections between representations. Efficiency in connection making would be an interesting compliment to the accuracy measures we derived from students' performance on the sense-making and perceptual-fluency scales. Second, Study 1 may have suffered from a selection bias. Because students were self-selected, our findings might be more representative of students with relatively high motivation. Students' self-reported chemistry courses indicate that students who participated in the survey had mostly taken introductory and intermediate courses. Few students had taken advanced chemistry courses. Thus, our findings might be most representative of students at the introductory and intermediate levels. Third, Study 1 suffered from attrition. Over half of the students who started working on the test items did not finish all test items. The problem of attrition in survey studies is well documented in the social sciences literature (e.g., Little & Rubin, 1989; Means, Toyama, Murphy, Bakia, & Jones, 2009; Shih & Fan, 2008). Since students participated anonymously online, it is impossible to inquire about why they chose not to finish the test. It is possible that they found the test too hard or that the incentive of winning a cash prize was not motivating enough. The fact that test items were presented in random order allowed us to include the data we had from all students. However, overall, attrition may have affected which student populations our results may be most representative of. It is likely that motivated students with high interest in chemistry were more likely to complete the test than students with low motivation and low interest. For these reasons, it would be worthwhile to repeat Study 1 in a setting where students have to complete the test, for example as part of a chemistry course.

Finally, the results from Study 1 were not conclusive as to the role that students' reasoning about similarities and differences between graphical representations plays in their ability to make sense of connections. Thus, an open question remains as to how sense-similarities and sense-differences abilities relate to students' domain knowledge. Study 2 focused on this question.

3.2. Study 2: Eye-tracking and interview study on sense-making ability

One goal of Study 2 was to investigate how students' ability to make sense of similarities and differences between representations relates to their ability to reason about domain-relevant concepts. In addition, a second goal was to identify which visual attention behaviors indicate low and high quality reasoning about chemistry. As mentioned, learning with graphical representations requires students to understand the meaning of perceptual features, and students need to visually attend to these features to make sense of these graphical representations. Investigating how students direct their visual attention to graphical representations as they reason about chemistry may thus yield interesting insights into what constitutes productive or unproductive processing. We anticipated that we could use these insights in future studies to evaluate how Chem Tutor supports productive learning processes. To achieve this goal, Study 2 combined eye-tracking and interview data. By combining these two different methods, we were able to investigate how students' visual attention behaviors relate to their sense-making abilities and their reasoning about chemistry.

Finally, a third goal was to identify specific concepts that Chem Tutor should target because helping students make connections between representations with respect to these concepts might lead to significant learning gains. Specifically, we were interested in identifying concepts that are important but difficult for undergraduate students to notice when making connections. To this end, Study 2 compared undergraduate students to graduate students in the following fashion. We assumed that concepts that graduate students are likely to mention

What are the similarities and differences between the space-filling model and the Bohr model of Ammonia?

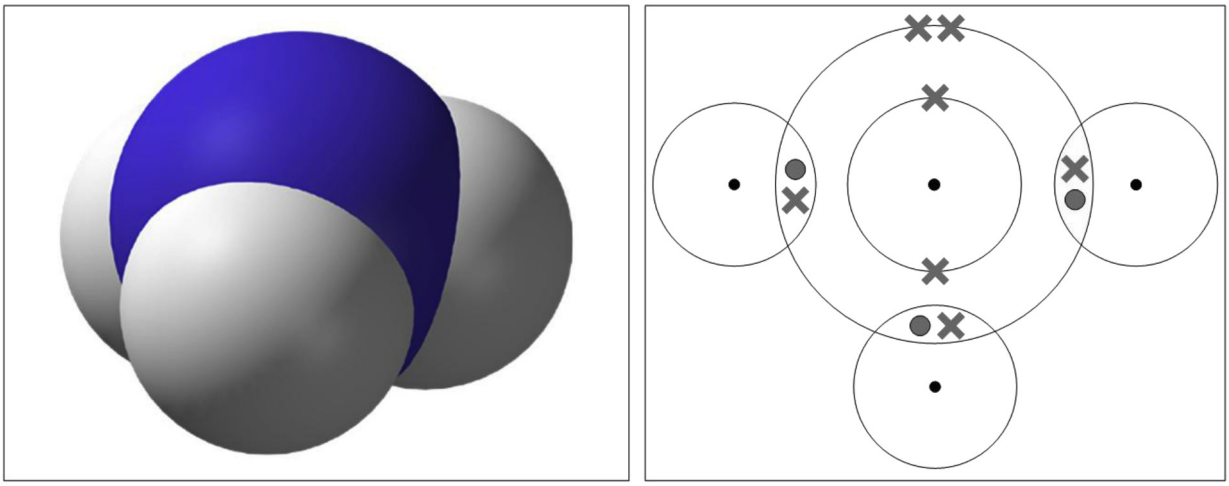


Fig. 4. Example of a sense-making problem used in Study 2.

are important. We further assumed that if undergraduate students mention these concepts infrequently, they have difficulty "seeing" these concepts in the representations. In other words, we were interested in which concepts were mentioned frequently by graduate students but infrequently by undergraduate students. This approach allowed us to compare knowledge structures in our target population (i.e., undergraduate students) to the desired knowledge structure that is typical of highly successful students (i.e., graduate students). Thus, our approach was to identify the "knowledge gap" (between undergraduate and graduate students) that Chem Tutor would seek to close.

Specifically, Study 2 investigated the following research questions:

Research question 2.1: What is the relation between sense-similarities and sense-differences connections and students' reasoning about chemistry concepts?

Research question 2.2: Which visual attention behaviors are associated with connection making and reasoning about chemistry concepts?

Research question 2.3: What specific concepts are important but difficult for undergraduate students to notice when making connections?

With respect to research question 2.1, we hypothesized that the number of sense-similarities-connections and the number of sense-differences connections students made would be positively associated with making inferences about chemistry concepts, because the ability to make sense of the connections between representations involves understanding similarities and differences between different graphical representations. Further, we hypothesized that the number of surface features students noticed would be negatively associated with inferences about chemistry concepts, but positively associated with misconceptions about chemistry. Because research questions 2.2 and 2.3 were exploratory questions we did not have specific hypotheses.

3.2.1. Methods

3.2.1.1. Participants. Undergraduate and graduate students were recruited from the same institution as in Study 1. Study 2 took place in the same semester as Study 1 (see Section 3.1.1 for the population's demographic information). To target students with varying levels of chemistry expertise, we advertised the study via mailing lists of the chemistry department and via posters that were hung at various locations in the chemistry department. Twenty-six students participated in Study 2 (21 undergraduate students and 5 graduate chemistry students). Three of the undergraduate students had taken only high-school chemistry. All other eighteen undergraduate students had taken at least one introductory chemistry course, eight had taken at least one intermediate undergraduate course, and two had taken at least one advanced undergraduate course. All graduate students had taken at least ten undergraduate classes and had experience as teaching assistants in undergraduate courses.

3.2.1.2. Procedure. Sessions took place in the laboratory and lasted 30–45 min. Students were paid for their participation. They worked through a series of sense-making problems on an SMI RED250 eye tracker. Students were asked to first think about the answer to the problem. Once they were ready, they alerted the experimenter, who annotated the eye-tracking data with a note that the student would start talking. This procedure was necessary because jaw movements that result from talking interfere with the quality of the eye-tracking data. After the experimenter had annotated the eye-tracking data, students provided their answer to the problem verbally.

3.2.1.3. Materials. The problems were identical to the ones in Study 1 with the following three exceptions. First, only sense-making problems were included. Second, students were asked about similarities and differences on the same problem. Third, students responded verbally, not via multiple choice. Specifically, the sense-making problems asked students to describe similarities and differences between two graphical representations of the same molecule. Fig. 4 shows an example of a sense-making problem used in Study 2. All verbal responses were audiotaped.

Table 1
First-level codes for verbal responses to interviews in Study 2.

Code	Definition	Example
Surface	Student makes a connection between representations, based on some conceptu-ally irrelevant feature	"um so they're both like red on the top"
Similarities	Student refers to a structural feature of representations that depict the same concept	"the space-filling model and the EPM both in shape are very similar cause they show the electron cloud")
Differences	Student refers to a structural feature of two representations that differs between representations or to information that differs between representations	"the space filling model gives you a better of idea of what the actual atoms of the molecule are […], but the electrostatic potential map gives you a much better idea of the electronic structure of the molecule"
Inference	Student explains a concept that goes beyond what is depicted	"this [the EPM] just shows that on the oxygen it's more reactive because there's lone pairs"

3.2.1.4. Analyses. To analyze the eye-tracking data, we created areas of interest (AOIs) that corresponded to the graphical representations shown on the screen, one AOI for each representation. We considered two measures. First, we considered frequency of switching between AOIs, because switching between conceptually relevant parts of the instructional materials is often used to indicate that students attempt to conceptually integrate these parts (Holsanova & Holmberg, 2009; Johnson & Mayer, 2012; Mason, Pluchino, Tornatora, & Ariasi, 2013). We computed AOI switches as the number of times a fixation on one AOI was followed by another. Second, we considered first-inspection and second-inspection durations. First inspections of an AOI are often considered to indicate initial processing of material (Hyönä, Lorch, & Rinck, 2003; Hyönä & Nurminen, 2006; Mason, Pluchino, & Tornatora, 2013). Second inspections (i.e., when a student re-inspects an AOI after the first inspection) are considered to reflect intentional processing to integrate the information with other information (Hyönä et al., 2003; Hyönä & Nurminen, 2006; Mason, Pluchino, & Tornatora, 2013; Schlag & Ploetzner, 2011). We computed first-inspection durations as the sum of durations of students' first fixation on a given AOI. We computed second-inspection durations as the sum of all fixation durations that occurred after the first fixation on a given AOI (in other words, second-inspection durations include all fixations except the first).

To analyze the verbal responses, we used a two-level coding scheme. The first-level codes were adapted from prior research on connection making (Rau, Rummel, et al., 2012). Specifically, we distinguished connections based on surface features, similarities, or differences, and whether students made inferences about concepts not explicitly shown in the representations. Table 1 provides descriptions and examples for first-level codes. We constructed the second-level codes in a bottom-up fashion: we first collected all concepts that students mentioned during the interview and then coded for their occurrence across all participants. Table A1 in the Appendix provides descriptions and examples for second-level codes. Interrater reliability on two randomly selected students was good with 85% agreement for first-level codes and 72.9% for second-level codes.

3.2.2. Results

Table 2 provides the means, standard deviations, and range for the variables we derived from the eye-tracking data and the interview data. Fig. 5 provides a summary of our findings on research questions 2.1 and 2.2. To address research question 2.1 (what is the relation between students' ability to identify similarities and differences between graphical representations and their reasoning about domain-relevant concepts?), we computed correlations among first-level interview codes (see Table 1). Specifically, we investigated how surface-connections, similarities-connections, and differences-connections relate to inferences and misconceptions that students uttered during the interview. We found that difference-connections were associated with significantly more inferences ($r = .56$, $p < .01$). There were no associations between surface-connections and inferences ($p = .13$) or between similarities-connections and inferences ($p = .22$). There were also no associations between surface-connections and misconceptions ($p = .87$), between similarities-connections and misconceptions ($p = .53$), or between differences-connections and misconceptions ($p = .30$).

To address research question 2.2 (which visual attention behaviors are associated with connection making and reasoning about domain-relevant concepts?), we computed correlations between eye-tracking variables (i.e., frequency of switching, first-inspection durations, and second-inspection durations) and first-level interview codes. We found that frequency of switching, first-inspection durations, and second-inspection durations were associated with significantly more surface-connections ($r = .56$, $p < .01$ for switching; $r = .54$, $p < .01$ for first-inspection durations; $r = .60$, $p < .01$ for second-inspection durations). We found no significant correlations between eye-tracking variables with similarity-connections. We found that second-inspection durations were associated with marginally more difference-connections ($r = .39$, $p < .10$), and inferences ($r = .36$, $p < .10$). We found no significant correlations between eye-tracking variables and misconception utterances.

Table 2
Means, standard deviations, and range of eye-tracking and interview data in Study 2.

	Means	Standard deviations	Range
Eye-tracking			
Frequency of AOI switches	1081.35	371.14	350–1905
Duration of 1st-inspection fixations (in ms)	16,965.96	2416.75	13,604–23,658
Duration of 2nd-inspection fixations (in ms)	723,191.42	373,609.59	145,101–1,711,988
Interviews			
Frequency of surface utterances	3.89	7.70	0–23
Frequency of similarity utterances	9.15	7.89	0–26
Frequency of differences utterances	49.27	15.30	25–110
Frequency of misconception utterances	4.31	3.74	0–16
Frequency of inference utterances	3.08	3.94	0–15

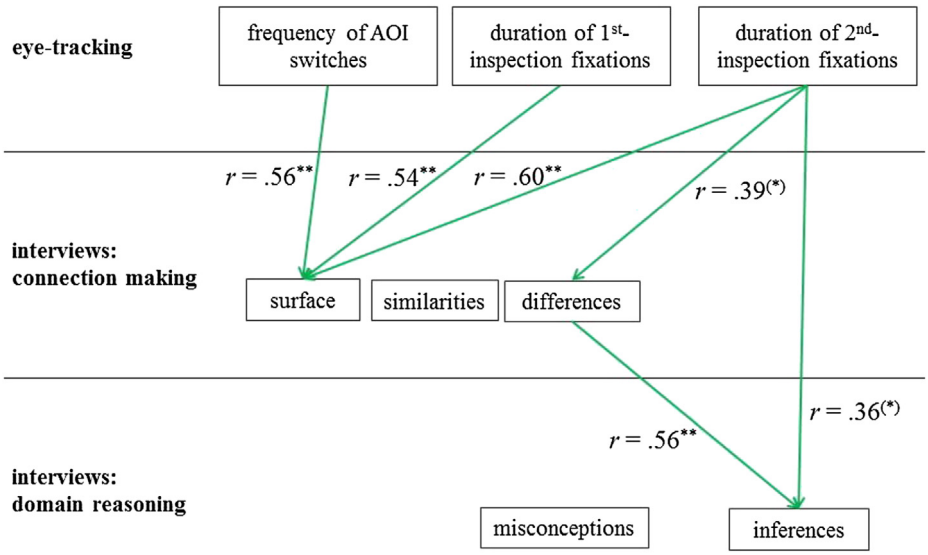


Fig. 5. Overview of correlation analyses in Study 2.

To investigate research question 2.3 (what specific concepts are important but difficult for undergraduate students to notice when making connections?), we analyzed the second-level interview codes. Using the bottom-up approach described above, we identified concepts related to the topics of atom identity (symbol, number of electrons, CPK color coding, general identity information), molecule structure (bond angle, bond length, conformation, geometry, atomic radii, electron cloud), energy (steric interactions, relative energy), electrons (core, valence, shared, lone), atomic structure (shells, orbitals, hybridization potential, spin states), and bonding (type, electronegativity, charge distribution). To get insights into which of these concepts are important but particularly difficult for undergraduates, we compared the relative frequency of a concept being discussed by graduate versus undergraduate students. We used differences larger than 1 SD to indicate that undergraduates had difficulties seeing this concept (i.e., they mentioned it infrequently), even though it is an important concept (i.e., graduate students mentioned it frequently). We found that graduate students were more likely than undergraduates to mention CPK color coding, bond angle, atomic radii, relative energy, bonding type, and reactivity. In addition, graduate students were more likely use these concepts to make inferences about the behavior of electrons, atoms, and molecules to explain bonding than undergraduate students. We consider these concepts to be important target concepts for Chem Tutor.

Furthermore, we considered the complexity of students' reasoning as the number of concepts mentioned per problem. Across all concepts, there were no differences between undergraduate and graduate students: undergraduate students mentioned on average 1.47 concepts per problem, and graduate students mentioned on average 1.48 concepts per problem. However, when considering only utterances that contained a reference to concepts we identified as target concepts, we found that graduate students' utterances with respect to target concepts were more complex than those made by undergraduate students': graduate students' utterances that referred to target concepts contained on average 3.56 concepts, whereas undergraduate students' utterances that referred to target concepts contained on average 2.55 concepts.

To investigate whether these differences in frequencies are indicative of different quality in undergraduate and graduate students' reasoning, we rated the quality of students' utterances with respect to these concepts. Using the bottom-up approach described above, three levels of quality emerged: whether students merely *describe* the differences (e.g., about CPK color coding "… the atoms are labeled with carbon or hydrogen, whereas in the ball-and-stick figure they just use the basic, black is for carbon and white is for hydrogen"), whether they comment on how different representations *complement* one another in terms of what information they show or which representation makes it easier to see certain information (e.g., about bond angles, "… it's more difficult to see in the space filling model of ethane, um what atoms there are in the molecule, and it's not very easy um to sort of, imagine it as a structure … um … but you can more clearly see what the dihedral situation is, uh about the cc bonds."), and whether they relate concepts to *conceptually relevant aspects* that are not explicitly shown in the representations (e.g., reasoning about bonding when comparing a ball-and-stick figure and an EPM, "You can clearly see what the polarity is based on. Um. The color. And you can clearly see that the carbon atom of the formaldehyde is deficient, as well as the hydrogen atoms. Oxygen atoms are quite rich in comparison.").

Table 3 shows a comparison of the quality ratings for undergraduate and graduate students. We found that undergraduate students were more likely than graduate students to *describe* differences in how the representations show information without going beyond what was shown in the representations (i.e., in 60.19% of all cases for undergraduate students, compared to 17.35% for graduate students). By contrast, graduate students were more likely than undergraduate students to describe *complementary* functions of the representations (i.e., in 75.51% of all cases for graduate students, compared to 36.11% for undergraduate students). Furthermore, graduate students were more likely than undergraduate students to relate what was shown in the representations to *conceptual* aspects of the domain that were not explicitly depicted in the representations (i.e., in 39.80% of all cases for graduate students, compared to 9.72% for undergraduate students).

3.2.3. Discussion

Research question 2.1 asked: What is the relation between students' ability to identify similarities and differences between graphical representations and their reasoning about domain-relevant concepts? In line with our hypothesis, our findings show that difference-

Table 3
Absolute frequences and relative frequences (as percentages, in parentheses) of quality ratings for undergraduate and graduate students.

	Undergraduate students	Graduate students
Descriptive	130 (60.19%)	17 (17.35%)
Complementary	78 (36.11%)	74 (75.51%)
Conceptual	21 (9.72%)	39 (39.80%)
Total	*216*	*98*

connections were associated with making more inferences about domain-relevant concepts. However, counter to our hypothesis, we found no positive associations between similarity-connections and inferences about chemistry concepts. It may be that expertise in chemistry relies on the use of different graphical representations for different purposes because they provide complementary information, rather than in using them interchangeably because they provide similar information. Also counter to our hypothesis, we did not find associations of surface-connections with inferences or misconceptions. It may be that the number of surface-connections is not predictive of reasoning about chemistry because noticing surface features is not necessarily unproductive behavior, as long as students relate surface features to differences in what they communicate about chemistry concepts. Taken together, these findings lead to a new hypothesis, namely that sense-making support might be more effective if it focuses on how different graphical representations depict *complementary* information than if it focuses on how they depict *similar* concepts.

Research question 2.2 asked: Which visual attention behaviors are associated with connection making and reasoning about domain-relevant concepts? We found that frequency of switching and first-inspection durations were mostly associated with low-quality processes (i.e., with more surface-connections and, for first-inspection durations, fewer inferences about chemistry concepts). It seems that switching between representations might not indicate that students successfully integrate conceptually relevant aspects of the material, but might rather indicate superficial processing or confusion about what to focus on. Second-inspection durations were the only measure that was associated with more reflective learning processes; namely with differences-connections and inferences. We also found that second-inspection durations were associated with surface-connections—however, since surface-connections were not associated with students making fewer inferences, this association does not imply that second-fixation durations indicate unproductive processes.

Research question 2.3 asked: What specific concepts are important but difficult for undergraduate students to notice when making connections? To address this question, we sought to identify the "knowledge gap" between graduate students' reasoning about connections with respect to conceptual aspects of chemistry, and undergraduate students' reasoning about connections. In comparing undergraduate and graduate students, we combined measures of how often students mentioned chemistry concepts when making connections and the quality of the reasoning process. Our findings suggest that Chem Tutor should target the concepts of CPK color coding, bond angle, atomic radii, relative energy, bonding type, and reactivity. These concepts may be difficult because they are complex: they are typically used to reason about bonding phenomena that involve the interaction of one molecule with additional atoms and molecules rather than about the structure of individual atoms and molecules. The fact that graduate students' utterances included more concepts than undergraduate students' utterances when they reasoned about these particular concepts (but not when they reasoned about other concepts) supports the notion that these concepts are difficult because they are complex. Another striking finding was that undergraduate students tend to not go beyond merely describing differences between representations: they rarely reason about complementary aspects of representations and they rarely relate the representations to conceptual aspects that go beyond what the representations depict explicitly. Taken together with the finding that difference-connections seem to be particularly important (research question 2.1), this finding further supports the notion that learning how different representations complement one another may be particularly important in chemistry. Thus, an important goal of Chem Tutor should be to help undergraduate students notice differences in how representations depict important target concepts, and how they provide complementary information that can be used to make inferences about conceptual aspects that are not explicitly depicted in the representations.

Study 2 has several limitations. First, as in Study 1, students were self-selected. It is possible that both undergraduate and graduate students in our study were more motivated to work on chemistry problems than students who did not participate. Second, the number of graduate students in our sample was rather small. It is possible that concepts graduate students mentioned were influenced by concepts that are relevant to their own research projects. Third, the setting of the study (i.e., a laboratory) might have influenced how students thought about chemistry concepts. It is possible that if we had interviewed students while they were solving chemistry homework problems, or even while conducting experiments in a wet-lab setting, we might have obtained different results. In particular, the eye-tracking procedure, which required students to first think about the problem and then talk about it, might have felt artificial to students and might have influenced how they thought about the chemistry concepts. To address these limitations, it would be necessary to validate the findings from Study 2 using a larger sample of graduate students and by conducting the study in a variety of different study settings, possibly without eye-tracking.

A fourth limitation of Study 2 is that it focused on knowledge structures and not on learning processes: Study 2 did not involve an instructional intervention. Thus, our findings do not necessarily allow for conclusions as to whether these visual attention patterns relate to productive or unproductive *learning processes.* Rather, our goal was to form new hypotheses as to which visual attention behaviors relate to productive or unproductive processes that we expect to result in low or high learning gains. We hypothesize that frequency of switching and first-fixation durations are indicators of superficial processing, which should result in low learning gains. Furthermore, second-inspection durations seem to indicate conceptual processing, which should result in high learning gains.

Finally, it is important to note that Study 2 contained only sense-making problems, not fluency-building problems. Thus, our hypotheses about which visual attention behaviors indicate productive (or unproductive) processes do not necessarily generalize to instructional activities such as fluency-building problems, which serve an entirely different educational purpose, namely to help students become fluent in using graphical representations. Studies 3 and 4 address some of these limitations.

4. Design of a chemistry tutor for connection making

Study 1 leads to the hypothesis that an ITS for chemistry that targets sense-making ability and perceptual fluency through separate activities might be effective. Study 2 leads to the hypothesis that sense-making activities that focus on differences between representations might be more effective than sense-making activities that focus on similarities—especially if students receive guidance in relating differences to inferences about how the representations complement one another and in relating these differences to difficult domain-relevant concepts. Here, we first describe the general features of Chem Tutor, which are typical of ITSs. Then, we describe how the findings from Studies 1 and 2 informed the design of Chem Tutor.

4.1. An intelligent tutoring system for chemistry

Chem Tutor is an ITS: a type of educational technology that is grounded in cognitive theories of learning and artificial intelligence. ITSs pose complex problem-solving activities and provide individualized step-by-step guidance at any point during the problem-solving process (VanLehn, 2011). At the heart of ITSs lies a cognitive model of the students' problem-solving steps. This model allows ITSs to detect multiple strategies a student might use to solve a problem (Aleven, McLaren, Sewall, & Koedinger, 2009), and to provide detailed feedback and (on the student's request) hints on how to solve a step in the tutor problem (Corbett et al., 2001). Traditional ITSs use a rule-based cognitive model that is based on production-rule theories of learning, such as ACT-R (Anderson, Corbett, Koedinger, & Pelletier, 1995; Corbett, 2001; Ritter, Anderson, Koedinger, & Corbett, 2007). Chem Tutor is a newer type of ITSs, called example-tracing tutors (Aleven et al., 2009). Example-tracing tutors use a cognitive model that is not rule based, but instead relies on generalized examples of correct and incorrect solution paths. Chem Tutor was created using Cognitive Tutor Authoring Tools (CTAT; Aleven et al., 2009), which allows for rapid iterations of prototyping and pilot-testing.

Typical ITSs have several adaptive features. First, they provide hints on demand in an adaptive fashion (Corbett et al., 2001; VanLehn, 2011). The cognitive model allows the ITS to infer which step the student is currently working on and to provide hints for how to solve this step. Chem Tutor assumes that the student is working on steps in the order in which they are presented in the interface, unless the student previously attempted to solve a different step. In this case, Chem Tutor assumes that the student will continue to solve this step. When the student asks for a hint, Chem Tutor provides a sequence of messages that provide increasingly specific guidance for solving the current step. The first level of hints provides a clarification of what the student is asked to do (e.g., “Why is chlorine more electronegative than hydrogen?”). The second level of hints provides additional conceptual information that may help the student solve the step (e.g., “Chlorine is lower in the periodic table than hydrogen, but it is also further right than hydrogen. Elements that are lower in the periodic table have a greater atomic radius, which increases the distance between the valence electron and the nucleus. Elements that are further right in the periodic table have more valence electrons, which increases the amount of energy it would take for the element to lose an electron. Which of the trends in the periodic table explain why chlorine is more electronegative than hydrogen?”). The third level of hints provides the reason for the correct answer—without explicitly saying what the right answer is (e.g., “The electronegativity of an element increases the further right it is in the periodic table, and the higher it is in the periodic table.”). The fourth hint level gives students the correct answer (e.g., “Chlorine is more electronegative than hydrogen because it is further right in the periodic table.”). The final hint level tells students exactly what to do (e.g., “Please select ‘further right’ from the highlighted menu.”).

Second, ITSs typically provide adaptive error feedback. When a student makes a mistake that is indicative of a common misconception, the ITS detects this misconception and provides a feedback designed to challenge it. Chem Tutor detects misconceptions that have been described in the chemistry education literature about bonding. For example, if a student knows about trends in the periodic table but gets the direction of the trend wrong, Chem Tutor might provide the following message: “You're close! You're right that chlorine is lower in the periodic table than hydrogen. But electronegativity decreases for lower elements in the periodic table, because these elements have a greater atomic radius, which decreases the distance between the valence electron and the nucleus. So, even though you're right about the location of hydrogen and chlorine in the periodic table, you got the trend of electronegativity wrong.”

Third, many ITSs adapt the selection of tutor problems to the individual student's learning progress. To do so, they use a cognitive model that determines which concepts the student has already mastered and hence does not need more practice on, and which concepts the student has not yet mastered but that are within his/her reach because he/she has acquired the prerequisite concepts and skills. Based on this assessment, ITSs can select appropriate problems for students to work on—appropriate in the sense that these problems provide opportunities to learn concepts and skills that are within reach (i.e., to select problems of appropriate difficulty), but not yet mastered (i.e., to prevent selecting problems that would over-practice already-learned skills or concepts). Chem Tutor does not yet provide adaptive problem selection. Rather, we view the research presented in this paper as a first step towards the goal of developing a cognitive model of how students make connections between representations and relate these connections to chemistry concepts. Based on this cognitive model, we will then be in a position to make new hypotheses about how best to adapt Chem Tutor's problems to the individual student's learning progress.

4.2. Chemistry content and representations

The current version of Chem Tutor focuses on chemical bonding, because this was one of the major topics that we identified to be difficult but important in Study 2. Our own findings align with the chemistry education literature, which documents that bonding is a particularly difficult topic (Coll & Treagust, 2003a, 2003b; Furio et al., 2000) in which students develop a variety of misconceptions (Taber, 2009; Taber & Coll, 2002). Research across high school, undergraduate, and graduate levels shows that students struggle with these crucial concepts (Coll & Treagust, 2003a, 2003b; Gabel & Bunce, 1994; Nicoll, 2001). The chemistry education literature provides an abundance of examples of students' misconceptions about bonding (Kind, 2004). For instance, even senior chemistry majors believe that electrons attract one another and cannot accurately explain how ionic, covalent, and polar bonds form (Nicoll, 2001). A survey with second-year chemistry undergraduates (Nakiboglu, 2003) showed that more than 75% held misconceptions about hybridization, believing, for example, that hybridization is a process by which an atom completes the number of its valence electrons to a full shell.

Table 4
Topics covered by the current version of Chem Tutor.

a. Introduction: Types of bonds and graphical representations	Electronegativity; continuum of ionic, polar covalent, and covalent bonds
b. Covalent bonding	Covalent bonds as shared electron pairs; σ and π bonds; orbital hybridization; shared electron distributions
c. Ionic bonding	Ionic bonds as electron transfer; lattice structures; electron distributions; lattice structures; formal charge
d. Polar covalent bonding	Dipole moment; electron distributions; resonance structure

Table 4 lists the topics and concepts covered by the Chem Tutor curriculum. The curriculum is based on the content typically covered in introductory-level courses on general chemistry and in the first weeks of organic chemistry courses. It aligns with the American Chemistry Society (ACS) standards for undergraduate programs[1] and with some of the advanced 9–12th grade content standards described in the National Science Education Standards (NSES)[2] and in the Next Generation Science Standards (NGES).[3] It also aligns with the ACS recommendations for development of student skills to enable students to use representations and technologies to communicate about chemical reactions.[4] Finally, the focus on connection-making abilities aligns with the NSES to understand the complementary functions of different models to explain scientific phenomena.

Table 5 shows examples for each of the graphical representations used in Chem Tutor and describes the conceptual aspects of chemical bonding each of them emphasizes. The sequence of activities in Chem Tutor is organized as follows. Unit 1 is an introductory unit on bonding types and graphical representations. Unit 2 covers each type of bond with three pairs of graphical representations: Bohr models with orbital diagrams, Lewis structures with EPMs, and Bohr models with EPMs. Unit 3 covers each type of bond with three different pairs of graphical representations: Lewis structures with ball-and-stick figures, Bohr models with space-filling models, and Lewis structures with space-filling models. Finally, unit 4 covers each type of bond, again with three different pairs of graphical representations: ball-and-stick figures with EPMs, space-filling models with orbital diagrams, and ball-and-stick figures with orbital diagrams. This sequence ensures that students cannot anticipate which type of bond they will encounter, which is crucial because part of their task is to identity which type of bond forms between two given atoms. Furthermore, the sequence is set up so that students encounter the maximum number of possible combinations of pairs of graphical representations.

4.3. Tutor design

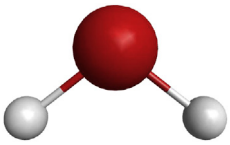
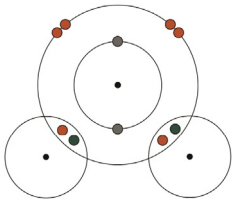
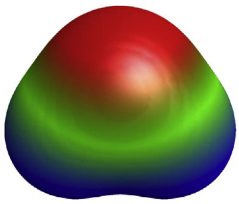
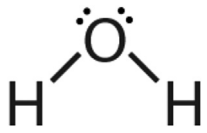
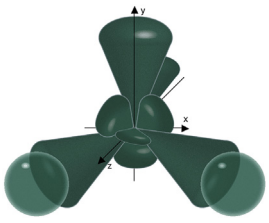
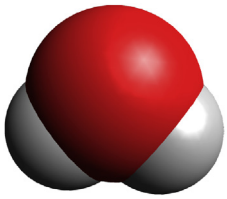
Before students can make connections between different graphical representations, they have to acquire some basic understanding of each individual graphical representation (Ainsworth, 2006; Eilam, 2013). Therefore, we designed *individual-representation problems* to help students understand how a given graphical representation depicts information about atoms, molecules, and bonds. Consider a problem that targets one of the concepts that we found to be particularly difficult in Study 2: bonding type and electron behavior (Fig. 6). Students identify the type of bond between atoms and make inferences about how electrons are distributed across the molecule. First, they solve this problem with one representation (e.g., a Lewis structure, see Fig. 6A). Second, they solve a corresponding problem with another representation (e.g., an EPM, see Fig. 6B). Students receive two individual-representations problems per representation. The individual-representations problems are sequenced so that consecutive problems provide different graphical representations, and so that students have encountered at least one individual-representations problem for each representation they will encounter in subsequent sense-making problems and fluency-building problems.

In line with prior research (Bodemer & Faust, 2006; van der Meij & de Jong, 2006; Seufert, 2003), *sense-making problems* are designed to help students in relating conceptually relevant aspects of different graphical representations (Fig. 7). In order to investigate the effects of sense-making problems that focus on similarities or differences, we designed two types of sense-making problems. In sense-similarities problems, students are prompted to explain similarities between representations (e.g., both representations show local negative charges; Fig. 7A). In sense-differences problems, students are prompted to explain differences between representations (e.g., the local negative charge is shown by a larger number of electron-dots shown in Lewis structures, and by red color (in the web version) in EPMs; Fig. 7B). Students receive one sense-similarities and one sense-differences problem for each pair of representations, presented in random order.

The design of the *fluency-building problems* is based on Kellman and colleagues' (2009) perceptual learning paradigm and the notion that students who are perceptually fluent will have more cognitive resources available to engage in higher-order learning tasks. Therefore, rather than focusing on why or how different representations correspond to one another, fluency-building support aims at helping students become faster and more efficient at extracting relevant information from graphical representations based on repeated experience with a large variety of problems. Thus, fluency-building activities provide numerous practice opportunities to find corresponding graphical representations based on their perceptual properties. Fig. 8 shows two sample problems in which students have to choose a representation that shows the same molecule. Choices are designed to contrast which perceptual aspects provide relevant information. For instance, to solve the example in Fig. 8A, students have to attend to how EPMs depict the geometry of the molecule. To solve the example in Fig. 8B, students need to attend to the lone pair in Lewis structures, which has implications for electronegativity that the EPM depicts as color (in the web version). To encourage perceptual rather than conceptual problem solving, students are encouraged to solve them fast, by using perceptual properties and without overthinking the problem. Students receive a series of ten fluency-building problems for each pair of representations, presented in random order.

[1] See the ACS guidelines for chemistry in 2-year college programs and the ACS Inorganic Chemistry Supplement.
[2] See National Science Education Standards: http://www.nap.edu/catalog/4962.html.
[3] See Next Generation Science Standards: http://www.nextgenscience.org/next-generation-science-standards.
[4] See ACS Development of Student Skills in a Chemistry Curriculum.

Table 5
Overview of graphical representations used in Chem Tutor.

Representation name	Example: Water molecule	Conceptual foci
Ball-and-stick figure		Atom identity, explicit bonds, geometry (bond angles and lengths), conformation
Bohr model		Shells, core electrons, valence electrons, shared electrons, explicit bonds
Electrostatic potential map (EPM)		Electronic distribution, molecule size
Lewis structure		Atom identity, valence electrons, lone pairs, explicit bonds
Orbital diagram		Orbital differentiation, geometry (bond angles)
Space-filling model		Atom identity, geometry (bond angles and lengths), conformation, atomic radii

5. Domain-specific grounding in chemistry learning processes

Studies 1 and 2 had focused on the role of connection making between multiple graphical representations in how students think about knowledge in chemistry. However, when designing a learning intervention, it is not only important to know how the target knowledge is structured, but it is also important to know how this knowledge is acquired; that is, to know how learning takes place. Therefore, we conducted two empirical studies that focus on learning processes. At the same time, these two studies served to pilot test Chem Tutor. We conducted one study in the lab and one in the field. In the lab study, Study 3, we combined learning outcome measures (i.e., pretests and posttests) with a number of learning process measures; namely, eye tracking, interviews, and data on problem-solving behaviors that the ITS collects automatically. In the field study, Study 4, undergraduate students used Chem Tutor for a voluntary online homework assignment as part of an introductory chemistry course.

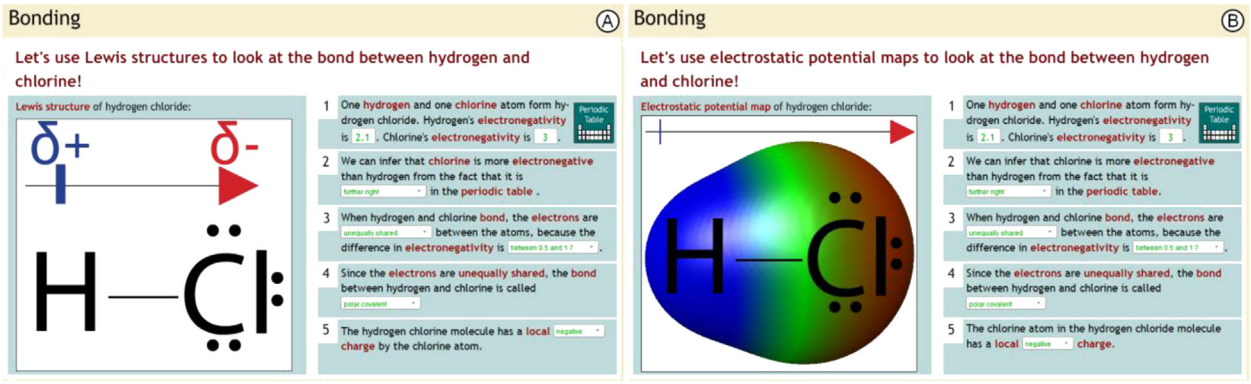


Fig. 6. Individual-representation problems with Lewis structure (A) and with electrostatic potential map (EPM, B).

5.1. Study 3: Pilot test in the lab

One goal of Study 3 was to pilot test Chem Tutor in a controlled setting. We investigated:

Research question 3.1: Does Chem Tutor help students learn about chemistry?

We hypothesized that undergraduate students would perform better on a test of chemistry knowledge after having worked with Chem Tutor than before.

In addition, a second goal was to further explore the relation of students' visual attention behaviors with problem-solving behaviors and with learning gains. We investigated:

Research question 3.2: Which visual attention behaviors are associated with students' performance on tutor problems and with their learning gains?

Based on Study 2, we expected that for sense-making problems, second-inspection durations would be associated with higher performance on tutor problems and higher learning gains. Further, we expected that for sense-making problems, frequency of switching and first-inspection durations would be associated with lower performance on tutor problems and lower learning gains. Since Study 2 included only sense-making problems, we did not have specific hypotheses with respect to individual-representation problems and fluency-building problems.

A further goal of Study 3 was to investigate how the different components of Chem Tutor relate to students' learning gains, specifically regarding individual-representation problems, sense-similarities problems, sense-differences problems, and fluency-building problems. We investigated:

Research question 3.3: What is the relation between students' performance on the different components of Chem Tutor and their learning gains?

Based on prior research that indicates that students' understanding of individual representations is a prerequisite for their ability to make connections between different representations (Ainsworth, 2006; Eilam, 2013), we hypothesized that higher performance on individual-representation problems would be associated with higher learning gains. Based on the findings from Study 1, we hypothesized that higher performance on sense-making and fluency-building problems would both be associated with higher learning gains. Based on the finding from Study 2, we hypothesized that higher performance on sense-differences problems would be more strongly associated with learning gains than performance on sense-similarities problems.

Finally, a third goal was to explore students' reactions to Chem Tutor and suggestions that they might have for improvement.

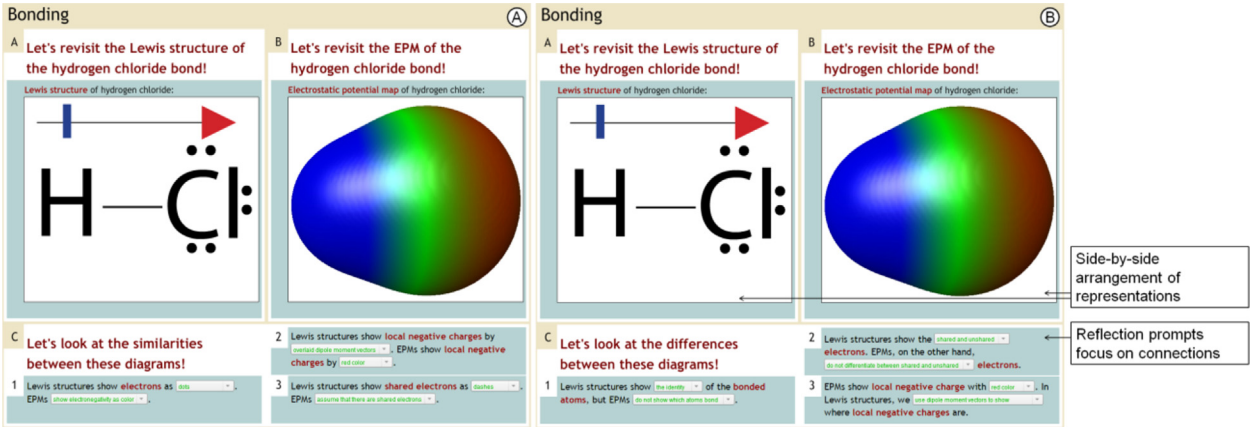


Fig. 7. Sense-making problems for connection making: Sense-similarities problem with Lewis structure and EPM (A) and sense-differences problem with Lewis structure and EPM (B).

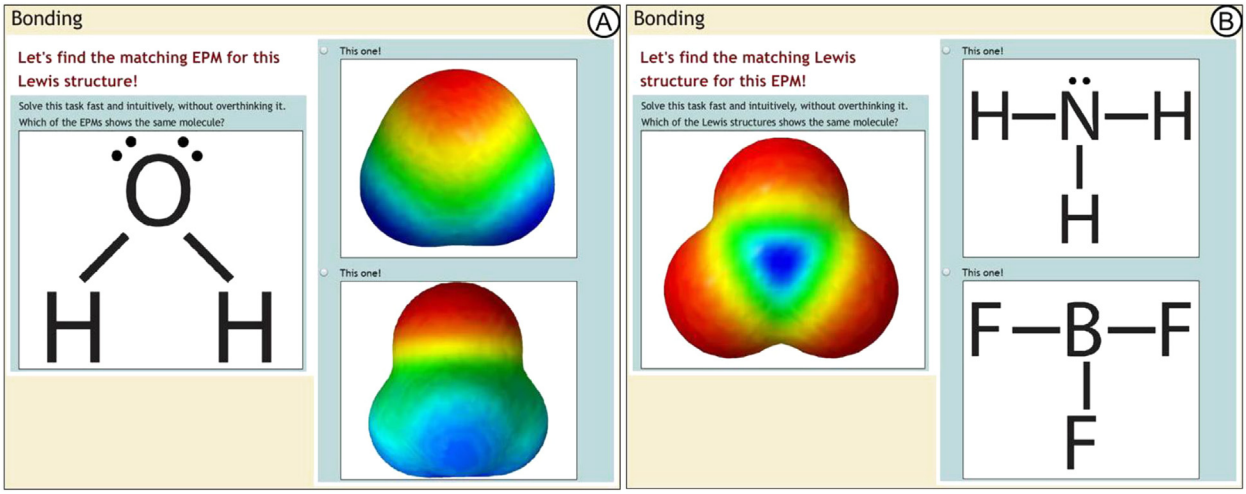


Fig. 8. Fluency-building problems: Students have to attend to particular perceptual features to solve these problems, for instance to the molecule's geometry (A), or to the lone pairs (B).

5.1.1. Methods

5.1.1.1. Participants. Twenty-five undergraduate students from a large Midwestern university participated in the study. The institution's undergraduate population for the relevant semester was composed of 49.1% men and 50.9% women, 2.9% African American, 6.3% Asian, 1.2% Native American, 4.4% Hispanic, 12.2% international, and 73.2% Caucasian students.

Because Chem Tutor is designed for undergraduate students in introductory chemistry courses, we recruited students from an introductory chemistry lecture. Study 3 took place in the middle of the semester, so that students from the chemistry lecture likely had a considerable amount of prior chemistry knowledge. An instructor who uses Chem Tutor in a lecture is likely to do so from the beginning of the semester onwards, so that undergraduate users of Chem Tutor might have very little prior knowledge when they first start using Chem Tutor. Therefore, we also recruited students from an introductory educational psychology lecture: most of these students had not taken chemistry since high school and may thus be similar (at least in terms of prior chemistry knowledge) to students who just started their first introductory chemistry course at the undergraduate level. On average, 68% of the students in the study sample had previously taken or were currently enrolled in at least one introductory undergraduate chemistry course. None of the students had taken or were currently taking advanced undergraduate or graduate level chemistry courses. Students received extra course credit for participating in the study.

5.1.1.2. Procedure. Students participated in two sessions that took place on different days but were no more than three days apart. Session 1 lasted at most 1.5 h. First, students took a pretest and worked through unit 1 of Chem Tutor, which introduces the different bond types and the graphical representations used in Chem Tutor. Then, they worked on units 2 and 3 of Chem Tutor. Next, they took an intermediate test. Session 2 lasted at most 1 h. Students first worked on unit 4 of Chem Tutor. Then they took a final posttest. Students performed all tasks (i.e., the tests and Chem Tutor) on an SMI RED 250 eye tracker. On average, students spent 01 h:26 m:08 s on Chem Tutor (including all three tests). At the end of the study, students were interviewed about their opinions about Chem Tutor, specifically with respect to what they liked and what they suggest we change about the system.

5.1.1.3. Measures. Students took three tests: a pretest, an intermediate test, and a final posttest. For this purpose, we created three equivalent test forms that included the same questions but used different chemicals. Each test comprised 14 multiple-choice items and three open-ended items. Multiple-choice items assessed students' knowledge about chemical bonding and about graphical representations. In designing the multiple-choice options for the tests, we drew on the verbal data we collected in Study 2. That is, the options reflect misconceptions and correct reasoning that undergraduate and graduate students engage in. To analyze students' performance on the multiple-choice items, we used effectiveness and efficiency measures. The effectiveness measures were computed as the mean scores on the tests. To analyze students' efficiency on the tests, we used a measure of efficiency described by Van Gog and Paas (2008) and by Lewis and Barron (2009). Efficiency measures assess whether students got faster at achieving a good test score, which is of interest because most tests at the college level are timed. We computed efficiency measures by combining Z-standardized average scores and the Z-standardized average time spent on the tests:

$$Efficiency = \frac{Z(score\ \ on\ \ test) + Z(time\ \ spent\ \ on\ \ test)}{\sqrt{2}}$$

The open-ended items asked students to describe in their own words how the behavior of electrons and of electronegative forces affects the type of bond that forms between atoms. We assigned quality rankings from 0 (guess or inadequate) to 3 (thorough) to each student response to each open ended question, with point values assigned to each ranking. Table A2 in the Appendix provides an overview of the coding scheme we used to assign these quality rankings. Since the students' response to these questions was not timed, we did not compute efficiency measures for the open-ended items.

To assess students' performance on the tutor problems, we used errors students made on the tutor problems. Specifically, we considered an attempt at solving a step in the tutor problems as incorrect if the first attempt at solving the step was either a hint request or an incorrect

Table 6
Means[a] and standard deviations (in parentheses) by test time and measure of Study 3.

	Pretest	Intermediate test	Final posttest
Effectiveness	.37 (.11)	.54 (.13)	.51 (.20)
Efficiency	−.87 (.70)	.41 (.57)	.37 (.78)
Open-ended	.38 (.30)	.59 (.23)	.72 (.17)

[a] Effectiveness and open-ended scores are on a scale between 0 and 1, efficiency scores are on a Z-score scale with extreme values (in this study) of −2.48 and 1.47.

response. This way of computing error rates is common in ITS research (Koedinger et al., 2010). To create our measures of interest, we computed average error rates for individual-representation problems, sense-similarities problems, sense-differences problems, and fluency-building problems. These averages were weighted by the number of steps each of these components involved.

To assess students' visual attention behaviors, we created areas of interest (AOIs) for each graphical representation. We computed frequency of switches between different representations as the number of times a fixation on an AOI was followed by a fixation on a graphical representation on which students had not fixated before, relative to the total number of switches a student made. Thus, this measure included switches between different graphical representation and switches between any other area on the screen and a graphical representation, but it did not include switches within a given graphical representation. We computed first-inspection durations as the sum of the durations of students' first fixation on a graphical representation. If there was more than one graphical representation on the screen (i.e., in the sense-making and fluency-building problems), the first fixation on each graphical representation was considered a first fixation. We computed second-inspection durations as the sum of fixation durations that occurred after the first fixation on a graphical representation.

To analyze the results from the interviews, we used a bottom-up approach. Specifically, for each interview question, we listened to the recordings, while noting categories of responses that were mentioned by several students. We then listened to the recordings again, while consolidating the original categories into a coherent set of distinct codes that describe qualitatively different sets of responses. The final codes are listed in Table A3 in the Appendix.

5.1.1.4. Analyses. We used ANOVAs and post-hoc comparisons to analyze the test results. Reported p-values for post-hoc comparisons were adjusted using the Bonferroni correction. We report partial η^2 for effect sizes on effects including more than two conditions, and Cohen's d for effect sizes of pairwise comparisons. According to Cohen (1988), an effect size partial η^2 of .01 corresponds to a small effect, .06 to a medium effect, and .14 to a large effect. An effect size d of .20 corresponds to a small effect, .50 to a medium effect, and .80 to a large effect.

5.1.2. Results

Table 6 shows means and standard deviations for students' effectiveness and efficiency scores on the multiple-choice items and their open-ended scores by test (pretest, intermediate test, final posttest).

To investigate research question 3.1 (does Chem Tutor help students learn about chemistry?), we conducted a repeated measures ANOVA with test (pretest, intermediate test, final posttest) as the independent variable and students' *effectiveness scores* as the dependent variable. We found a significant effect of test, $F(2,48) = 14.66$, $p < .01$, partial $\eta^2 = .38$. Post-hoc comparisons showed that students performed significantly better at the intermediate test than at the pretest, $t(24) = 5.31$, $p < .01$, $d = 1.43$, and that they performed significantly better at the final posttest than at the pretest, $t(24) = 3.59$, $p < .01$, $d = .94$. We conducted the same ANOVA with students' *efficiency scores* as the dependent variable. We found a significant effect of test, $F(2,48) = 32.72$, $p < .01$, partial $\eta^2 = .58$. Post-hoc comparisons showed that students performed significantly better at the intermediate test than at the pretest, $t(24) = 7.63$, $p < .01$, $d = 2.03$, and that they performed significantly better at the final posttest than at the pretest, $t(24) = 5.84$, $p < .01$, $d = 1.67$. We conducted the same ANOVA with students' *open-ended scores* as the dependent variable. We found a significant effect of test, $F(2,48) = 27.64$, $p < .01$, partial $\eta^2 = .54$. Post-hoc comparisons showed that students performed significantly better at the intermediate test than at the pretest, $t(24) = 3.91$, $p < .05$, $d = .80$, and that they performed significantly better at the final posttest than at the pretest, $t(24) = 6.65$, $p < .01$, $d = 1.44$.

To investigate research question 3.2 (which visual attention behaviors are associated with students' performance on tutor problems and with their learning outcomes?) and research question 3.3 (what is the relation between students' performance on the different components of Chem Tutor and their learning outcomes?), we conducted correlation analyses. Table 7 provides the means, standard deviations, and

Table 7
Means, standard deviations, and range of eye-tracking variables in Study 3, organized by tutor problems types.

	Means	Standard deviations	Range
Individual-representation problems			
Frequency of AOI switches into graphical representation	124.96	36.62	56–215
Duration of 1st-inspection fixations on graphical representation (in ms)	1645.12	754.38	734–3369
Duration of 2nd-inspection fixations on graphical representation (in ms)	67,781.56	28,314.54	28,874–167,101
Sense-making problems			
Frequency of AOI switches into graphical representations	158.32	62.21	30–292
Duration of 1st-inspection fixations on graphical representations (in ms)	3022.40	1295.33	1195–5773
Duration of 2nd-inspection fixations on graphical representations (in ms)	72,926.84	29,862.88	16,001–139,844
Fluency-building problems			
Frequency of AOI switches into graphical representations	706.88	143.58	446–1040
Duration of 1st-inspection fixations on graphical representations (in ms)	4218.44	1236.48	2076–6577
Duration of 2nd-inspection fixations on graphical representations (in ms)	290,414.32	61,849.16	138,701–409,966

Table 8
Means, standard deviations, and range of error rates derived from the tutor log data in Study 3, organized by tutor problems types.

	Means	Standard deviations	Range
Individual-representation problems	.26	.08	.13–.47
Sense-similarities problems	.66	.18	.42–1.04
Sense-differences problems	.36	.11	.16–.61
Fluency-building problems	.30	.06	.18–.42

range for the variables we derived from the eye-tracking data. Table 8 provides the means, standard deviations, and range for the error rates we derived from the tutor log data. Fig. 9 provides a summary of the findings.

First, we investigated the relation between students' prior knowledge with visual attention behaviors and problem-solving performance, using correlations of pretest measures with visual attention measures and with students' error rates on the different tutor components as dependent variables. With respect to visual attention behaviors, we found that higher pretest effectiveness scores were associated with marginally longer first-inspection durations on graphical representations in individual-representation problems ($r = .37$, $p < .10$), but no other correlations were significant. Higher pretest efficiency scores were associated with marginally shorter second-inspection durations in fluency-building problems ($r = -.38$, $p < .10$), but no other correlations were significant. We found no associations between pretest open-ended scores and visual attention behaviors. With respect to problem-solving performance, we found that higher pretest effectiveness and efficiency scores, as well as higher open-ended scores were associated with significantly lower error rates on individual-representation problems ($r = -.59$, $p < .01$ for effectiveness scores; $r = -.44$, $p < .05$ for efficiency scores; $r = -.48$, $p < .05$ for open-ended scores), on sense-similarities problems ($r = -.52$, $p < .01$; $r = -.41$, $p < .05$; $r = -.48$, $p < .05$), and on sense-differences problems ($r = -.58$, $p < .01$; $r = -.47$, $p < .05$; $r = -.50$, $p < .05$). There were no associations between any pretest measure and error rates on fluency-building problems.

To investigate the relation between visual attention behaviors and students' learning gains, we computed partial correlations of visual attention measures with final posttest measures, while controlling for pretest measures. Longer first-inspection durations on individual problems were associated with marginally higher posttest effectiveness scores ($r = 35$, $p < .10$) and efficiency scores ($r = .40$, $p < .10$). Longer second-inspection durations on sense-making problems were associated with marginally higher posttest efficiency scores ($r = .35$, $p < .10$). There were no other significant correlations between visual attention measures and posttest measures.

To investigate the relation between students' visual attention behaviors with their performance on the tutor problems, we computed partial correlations of visual attention behaviors with students' error rates, while controlling for pretest measures. We found that first-inspection durations in individual-representation problems were associated with significantly lower error rates on individual-representation problems ($r = -.47$, $p < .05$) and sense-similarities problems ($r = -.47$, $p < .05$). Longer second-inspection durations on fluency-building problems were associated with marginally lower error rates on individual-representation problems ($r = -.34$, $p < .10$) and sense-differences problems ($r = -.37$, $p < .10$). There were no other significant correlations between frequency of switching and error rates.

To investigate the relation between students' performance on the tutor problems and their learning gains, we computed partial correlations of error rates with final posttest measures, while controlling for pretest. We found that higher error rates on individual-representation problems was associated with significantly lower posttest effectiveness scores ($r = -.77$, $p < .01$), efficiency scores

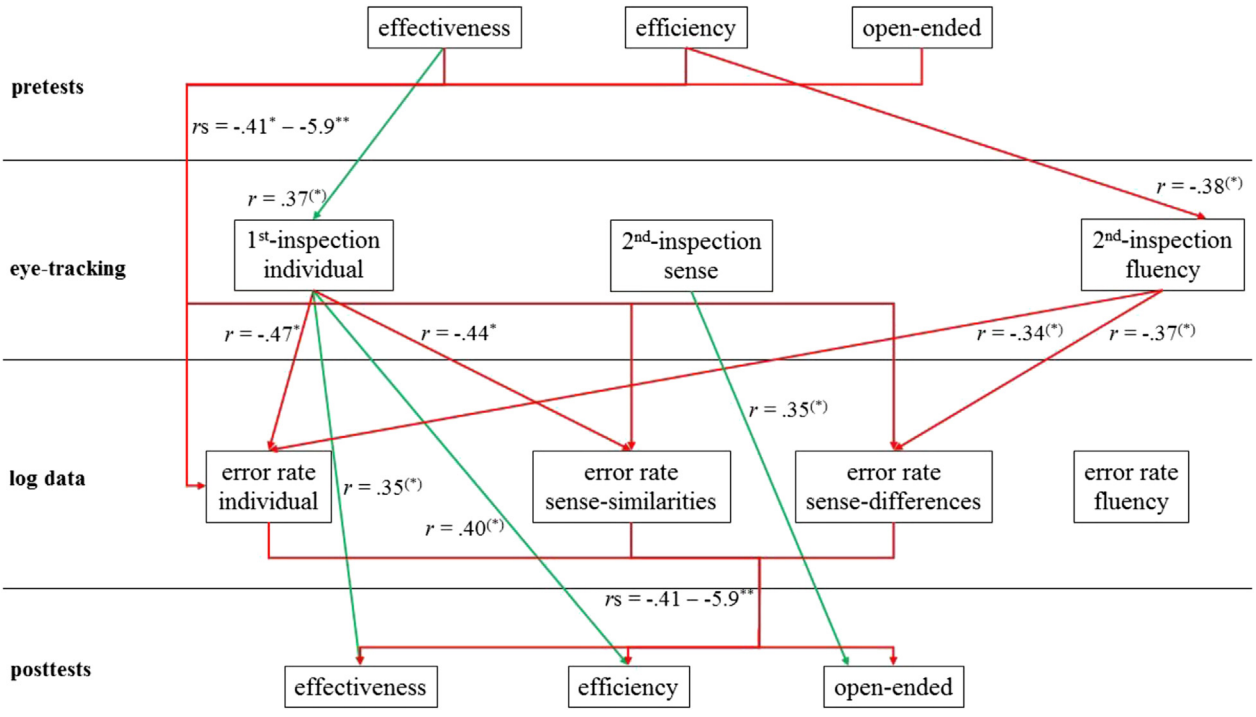


Fig. 9. Overview of correlation analyses in Study 3.

($r = -.64$, $p < .01$), and open-ended scores ($r = -.51$, $p < .05$). Likewise, higher error rates on sense-similarities problems were associated with significantly lower posttest effectiveness scores ($r = -.69$, $p < .01$), efficiency scores ($r = -.59$, $p < .01$), and open-ended scores ($r = -.47$, $p < .05$). Similarly, higher error rates on sense-differences problems were associated with significantly lower posttest effectiveness scores ($r = -.69$, $p < .01$), efficiency scores ($r = -.58$, $p < .01$), and open-ended scores ($r = -.51$, $p < .05$). There were no significant correlations between error rates on fluency-building problems and posttest measures when controlling for pretest. To investigate whether there is a correlation between students' performance on fluency-building problems if we do not control for pretest performance, we computed regular correlations between error-rates on fluency-building problems and learning outcomes. We found a marginal correlation between error rates on fluency-building problems and posttest effectiveness scores ($r = -.38$, $p < .10$).

Finally, we explored students' opinions about Chem Tutor based on the interviews. Table 9 shows the coded results from the interviews. First, we asked students to state their general opinion about Chem Tutor ("What did you think about the tutoring system?"). Overall, students mentioned 38 aspects, 11 of which were negative, 27 of which were positive. The majority of negative aspects mentioned concerned confusion about how to solve the tutor problems. The majority of positive aspects mentioned concerned repetition of questions, and the fact that the system provided tutoring (i.e., error feedback and hints on demand). Second, we asked students to state what they liked the most about Chem Tutor ("Was there anything that stood out to you that you particularly liked?"). Overall, repetition and the use of graphical representations stood out positively to students. Finally, we asked students what they would like for us to change about Chem Tutor ("Was there anything that you think we should improve?"). Overall, the majority of students did not suggest specific changes. Most students who did make suggestions commented on the fact that there was very little instruction and recommended to add instructional content. In addition, some students disliked repetition. Some students mentioned a specific problem that was difficult for them to solve. Since it was striking that some students explicitly liked repetition whereas others explicitly disliked repetition, we compared the average pretest scores of students who liked or disliked repetition. We found that students who disliked repetition had significantly lower pretest effectiveness scores, $t(15) = 2.23$, $p < .05$, $d = 1.73$, but did not differ with respect to their error rates or learning gains ($ps > .10$).

5.1.3. Discussion

Research question 3.1 asked: Does Chem Tutor help students learn about chemistry? As hypothesized, our findings show that students perform significantly better on a test of chemistry knowledge after having worked with Chem Tutor than before. We found large effect sizes in learning gains on effectiveness and efficiency measures that assessed students' knowledge about chemistry and about graphical representations in a multiple-choice format. We also found large and significant learning gains on measures that assessed students' reasoning about chemical bonding in an open-ended format. Surprisingly, the learning gains on the multiple-choice items were smaller at the final posttest than at the intermediate posttest. By contrast, the learning gains on the open-ended items were larger at the final posttest than at the intermediate posttest. It is possible that students experienced test fatigue when they were asked to take the third test. Even though we used different test forms for the pretest, intermediate, and final posttest, the fact that the tests were created to be highly similar may have decreased students' willingness to read test questions carefully. The fact that students' performance on open-ended items increased from the intermediate posttest (given at the end of session 1) to the final posttest (given at the end of session 2) lends credibility to the notion that there were indeed additional learning gains in the second session of Study 3. In summary, Study 3 indicates that Chem Tutor leads to large learning gains in a controlled setting, on tests that assess their conceptual reasoning about important chemistry concepts.

Research question 3.2 asked: Which visual attention behaviors are associated with students' performance on tutor problems and with their learning gains? Based on Study 2, we had hypothesized that second-inspection durations would indicate productive learning processes, whereas first-inspection durations and frequency of switches would indicate unproductive learning processes, at least with respect to sense-making problems. The results from Study 3 provide partial support for this interpretation. In line with Study 2, we found that longer second-inspection durations on sense-making and fluency-building problems were associated with higher learning gains or with lower error rates, suggesting that this measure indicates productive learning processes. However, we did not find evidence that first-inspection durations indicated unproductive learning processes. To the contrary—first-inspection durations on individual problems were associated with lower error rates and higher learning outcomes and thus seemed to indicate productive learning processes. Furthermore, we did not find evidence that frequency of switching indicated unproductive learning processes, because we did not find any significant correlations between frequency of switching and problem-solving performance or learning outcomes. Thus, we cannot draw the conclusion that first-inspection durations and frequency of switching reflect superficial or unproductive processing of graphical representations. It is possible

Table 9
Students' responses to interview questions.

	Question 1: What did you think about the tutoring system?		Question 2: Was there anything that stood out to you that you particularly liked?		Question 3: Was there anything that you think we should improve?	
Dislike: Unspecific	1	11				26
Dislike: Confusion	7					
Dislike: Instruction	1				6	
Dislike: Specific problem	1				4	
Dislike: Technical issue					3	
Dislike: Repetition					4	
Dislike: Other	1				1	
Dislike: Nothing					8	
Approval: Unspecific	2	27		23		
Approval: Repetition	9		7			
Approval: Graphical representations	3		5			
Approval: Tutoring	7					
Approval: Other	6		9			
Approval: Nothing			2			

that general first-inspection durations and general frequency of switching are not fine-grained enough assessments of what aspect of the graphical representations students attend to. For example, it may matter what aspects of the graphical representations students fixate on when they first inspect a graphical representation, and it may also matter what aspects of the graphical representations they switch between. It is also possible that a students' prior knowledge determines how long a "productive" first inspection is. Thus, further investigation of how to interpret first-inspection durations and switching between graphical representations is needed.

Research question 3.3 asked: What is the relation between students' performance on the different components of Chem Tutor and their learning gains? As hypothesized, we found that higher performance on individual-representation problems and sense-making problems (indicated by lower error rates) was associated with higher learning outcomes. Counter to our hypothesis, students' performance on fluency-building problems was not associated with their learning outcomes. This lack of association is surprising because Study 1 had indicated, in line with the chemistry education literature (Cheng & Gilbert, 2009; Gilbert & Treagust, 2009), that perceptual fluency is an important aspect of chemistry expertise. We would thus expect that students who do better on fluency-building problems have better learning outcomes. Students were encouraged to solve the fluency-building problems quickly and without over-thinking them. Therefore, students who used the fluency-building problems as intended were encouraged to make many mistakes. Yet, we would expect that, as students acquire perceptual fluency, their error rates decrease, which would result in the expected association between error rates and learning outcomes. Especially because students were encouraged to make many mistakes, it is possible that ten fluency-building problems per unit were not enough to yield sufficient decrease in error rates to yield that association.

Based on Study 2, we had further expected that students' performance on sense-differences problems would be more strongly associated with their learning outcomes than their performance on sense-similarities problems. Study 3 does not support this hypothesis, but shows instead that performance on sense-similarities and performance on sense-differences problems are equally associated with students' learning outcomes. Thus, it seems that both the ability to make sense of similarities between graphical representations and the ability to make sense of differences between graphical representations are important components of chemistry learning. The differences between the findings in Studies 2 and 3 may be due to the fact that Study 2 focused on knowledge structures and not on learning processes. It is possible that knowing about both similarities and differences between graphical representations play an important role as students *acquire* knowledge about chemistry (Study 3), but that differences are more salient when students *report* on their knowledge about chemistry (Study 2). This finding is interesting because it might suggest that students tend to pay less attention to similarities between graphical representations than to differences, even though knowing about similarities is an important component of learning. Thus, this finding supports our choice of including components that help students reason about similarities between graphical representations, because they might otherwise attend only to differences.

A limitation of Study 3 is that it was conducted in a laboratory setting. Working with Chem Tutor in a laboratory (i.e., with an experimenter close by, knowing that one's eye-gaze behaviors are being assessed) likely yields an entirely different learning experience than working with Chem Tutor as a regular homework system would. Furthermore, as in Studies 1 and 2, students were self-selected. Since they received extra course credit for their participation, it is possible that lower-performing students who needed extra credit were more likely to participate than stronger students. Thus, our findings might generalize to lower-performing students more so than to higher-performing students. Finally, Study 3 contained a diverse population of students, some of whom were not enrolled in a chemistry course. Even though, students who were not currently enrolled in a chemistry were similar to the target population in terms of their prior chemistry knowledge (e.g., students at the beginning of their first introductory chemistry course at the college level), these students may have been different from the target population in terms of other factors, such as their interest in learning chemistry. Thus, it is unclear whether the findings from Study 3 generalize to the setting and population that Chem Tutor was designed for. Study 4 aims to address these shortcomings.

5.2. Study 4: Pilot test in the field

The goal of Study 4 was to validate the findings from Study 3 in a realistic educational setting. We investigated:

Research question 4.1: Does Chem Tutor help students learn about chemistry?

Research question 4.2: What is the relation between students' performance on the different components of Chem Tutor and their learning gains?

5.2.1. Methods

5.2.1.1. Participants. Eighty-five undergraduate students from a large Midwestern university participated in the study. All students were recruited from the same introductory chemistry lecture as students who participated in Study 3, but from a different section, taught by the same instructor. Study 4 took place a few weeks after Study 3 in the same semester (see Section 5.1.1 for the population's demographic information). Five students started the tutor but did not finish. On average, students who did not finish completed 15.6% of the problems. None of the remaining students were currently taking advanced undergraduate chemistry courses. One student reported to have previously taken or to be currently taking two advanced graduate chemistry courses. On average, students reported a GPA of 3.30. 62.5% of the students had already declared a major that was related to chemistry (including, for instance, bio-chemistry, geo-science, and industrial engineering). All students received extra course credit for participating in the study.

5.2.1.2. Procedure. Students were given a personal account for Chem Tutor and could work with the system whenever they wanted. The sequence of activities was identical to Study 3, except that it did not involve eye tracking or interviews. On average, students spent 02 h:01 m:18 s on Chem Tutor (including all three tests).

5.2.1.3. Measures and analyses. Measures and analyses were identical to those in Study 3, except that we did not collect eye-tracking data or interview data.

Table 10
Means[a] and standard deviations (in parentheses) by test time and measure of Study 4.

	Pretest	Intermediate test	Final posttest
Effectiveness	.41 (.13)	.50 (.13)	.54 (.15)
Efficiency	−.56 (.79)	.25 (.64)	.57 (.63)
Open-ended	.60 (.21)	.66 (.19)	.68 (.20)

[a] Effectiveness and open-ended scores are on a scale between 0 and 1, efficiency scores are on a Z-score scale with extreme values (in this study) of −3.14 and 2.00.

5.2.2. Results

Six students were excluded from the analysis because their pretest scores were statistical outliers (i.e., 2 standard deviations lower or higher than average), resulting in a sample of $N = 74$. Table 10 shows the means and standard deviations for students' scores on the effectiveness, efficiency, and open-ended measures at the pretest, the intermediate test, and the final posttest. Table 11 provides the means, standard deviations, and range for the error rates we derived from the tutor log data.

To investigate research question 4.1 (does Chem Tutor help students learn about chemistry?), we conducted a repeated measures ANOVA with test (pretest, intermediate test, final posttest) as the independent variable and students' *effectiveness scores* as the dependent variable. We found a significant effect of test, $F(2,146) = 20.96$, $p < .01$, partial $\eta^2 = .22$. Post-hoc comparisons showed that students performed significantly better at the intermediate test than at the pretest, $t(73) = 4.44$, $p < .01$, $d = .73$, and that they performed significantly better at the final posttest than at the pretest, $t(73) = 5.92$, $p < .01$, $d = .93$. We conducted the same ANOVA with students' *efficiency scores* as the dependent variable. We found a significant effect of test, $F(2,146) = 61.46$, $p < .01$, partial $\eta^2 = .46$. Post-hoc comparisons showed that students performed significantly better at the intermediate posttest than at the pretest, $t(73) = 7.17$, $p < .01$, $d = 1.12$, and that they performed significantly better at the final posttest than at the pretest, $t(73) = 10.36$, $p < .01$, $d = 1.58$. We conducted the same ANOVA with students' *open-ended scores* as the dependent variable. We found a significant effect of test, $F(2,146) = 5.23$, $p < .01$, partial $\eta^2 = .07$. Post-hoc comparisons showed that students performed significantly better at the intermediate posttest than at the pretest, $t(73) = 2.17$, $p < .05$, $d = .30$, and that they performed significantly better at the final posttest than at the pretest, $t(73) = 2.91$, $p < .01$, $d = .38$.

To investigate research question 3.2 (what is the relation between students' performance on the different components of Chem Tutor and their learning outcomes?), we conducted correlation analyses. Fig. 10 provides a summary of the findings.

First, we investigated the relation between students' prior knowledge on problem-solving performance. To do so, we computed correlations of pretest measures (i.e., effectiveness scores, efficiency scores, and open-ended scores) with students' error rates on the different tutor components as dependent variables (i.e., individual-representations problems, sense-making problems with respect to both similarities and differences, and fluency-building problems). We found that higher pretest effectiveness scores were associated with lower error rates on individual-representation problems ($r = -.22$, $p < .10$), on sense-similarities problems ($r = -.30$, $p < .01$), on sense-differences problems ($r = -.30$, $p < .05$), and on fluency-building problems ($r = -.28$, $p < .05$). Furthermore, higher pretest efficiency scores were associated with lower error rates on individual-representation problems ($r = -.34$, $p < .01$), on sense-similarities problems ($r = -.38$, $p < .01$), on sense-differences problems ($r = -.33$, $p < .01$), and on fluency-building problems ($r = -.22$, $p < .10$). Finally, higher pretest open-ended scores were associated with lower error rates on individual-representation problems ($r = -.24$, $p < .05$), on sense-similarities problems ($r = -.28$, $p < .01$), and on fluency-building problems ($r = -.22$, $p < .10$).

To investigate the relation between students' performance on the tutor problems and their learning gains, we computed partial correlations of error rates with final posttest measures, while controlling for pretest. We found that higher error rates on individual-representation problems were associated with significantly lower posttest effectiveness scores ($r = -.31$, $p < .01$), and with significantly lower posttest efficiency scores ($r = -.25$, $p < .05$). Higher error rates on sense-similarities problems were associated with significantly lower posttest effectiveness scores ($r = -.32$, $p < .01$), and with significantly lower posttest efficiency scores ($r = -.28$, $p < .01$). Higher error rates on fluency-building problems were associated with significantly lower posttest effectiveness scores ($r = -.23$, $p < .05$). There were no significant correlations of error rates on sense-differences problems with any posttest measure when controlling for pretest. When we computed regular correlations between error rates on sense-differences problems with posttest measures, we found a marginally significant correlation between error rates on sense-differences problems with posttest effectiveness scores ($r = -.20$, $p < .10$). There were no significant correlations of any error rates with posttest open-ended scores when controlling for pretest. When not controlling for pretest, we found a significant correlation between error rates on sense-similarities problems with posttest open-ended scores ($r = -.23$, $p < .05$).

5.2.3. Discussion

The goal of Study 4 was to validate our findings from Study 3 in a realistic educational setting. Research question 4.1 asked: Does Chem Tutor help students learn about chemistry? In line with Study 3, we found significant learning gains on all posttest measures. The effect sizes on students' learning gains were smaller in Study 4 than in Study 3, as is typical for a less controlled setting. The difference in effect sizes was most noticeable on open-ended scores, where we found only small effect sizes in Study 4, even though we had found large effect sizes in Study 3. It is tempting to assume that students who participated in the lab study (where they had an experimenter observing them while they were taking the test) took more care of writing careful answers. However, the data do not support this interpretation: Tables 6 and 10 show that students' performance on the open-ended posttests of Studies 3 and 4 were comparable. The difference in learning gains seems to

Table 11
Means, standard deviations, and range of error rates derived from the tutor log data in Study 4, organized by tutor problems types.

	Means	Standard deviations	Range
Individual-representation problems	.29	.08	.15–.50
Sense-similarities problems	.48	.14	.18–.81
Sense-differences problems	.49	.12	.23–.73
Fluency-building problems	.29	.08	.12–.55

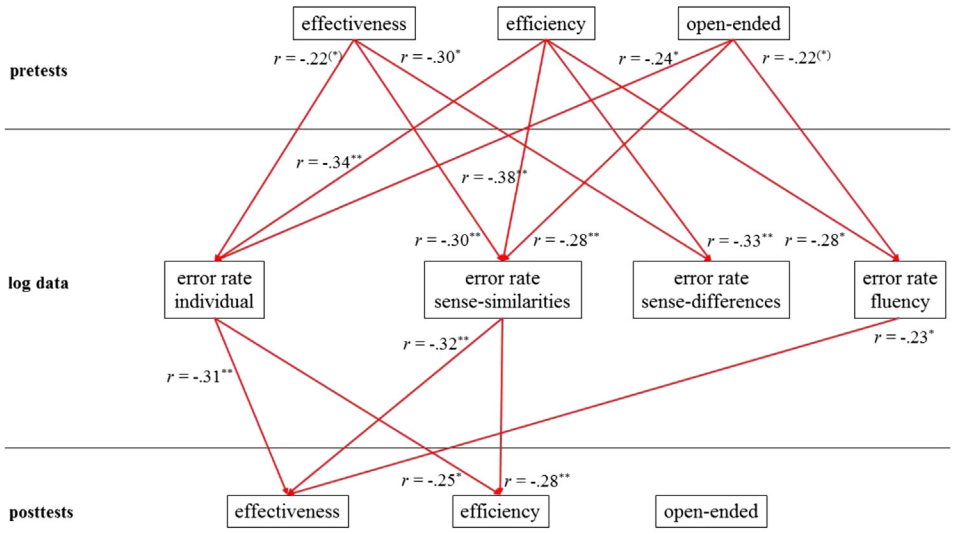


Fig. 10. Overview of correlation analyses in Study 4.

be due to the fact that students in Study 4 had higher pretest open-ended scores than students in Study 3. In other words, there was more room for improvement for students in Study 3 than for students in Study 4, with respect to the open-ended scores. There are many factors that might have contributed to the differences between students' pretest open-ended scores in Studies 3 and 4. Students were drawn from different sections of the same chemistry lecture. Students' schedules of other university courses may have determined which lecture section they chose, so that students in Studies 3 and 4 may have differed with respect to majors or interest levels. It is also possible that students in Study 4 had higher pretest open-ended scores because Study 4 took place later in the semester than Study 3 and they had already learned more. These explanation attempts are highly speculative. In any case, it seems that Chem Tutor is effective in helping students perform better on tests that require them to explain in their own words how bonding occurs, even if the effect sizes might range between small and large, possibly depending on students' prior knowledge about chemistry.

With respect to effectiveness and efficiency measures, Study 4 shows—as did Study 3—large learning gains. On efficiency measures, the learning gains exceeded one standard deviation, which is promising for a 2 h long intervention. Thus, Study 4 replicates the finding that Chem Tutor is effective in a realistic, natural setting, in which students use the system as a homework assignment.

Research question 4.2 asked: What is the relation between students' performance on the different components of Chem Tutor and their learning gains? As hypothesized, we found that higher performance on individual-representation problems, sense-making problems, and fluency-building problems (indicated by lower error rates) was associated with higher learning outcomes. There were three major differences in the findings of Study 4 compared to those of Study 3. First, Study 3 did not show associations between performance on fluency-building problems and learning outcomes. In Study 4, by contrast, we found that students' performance on fluency-building problems correlates with students' learning outcomes (i.e., with posttest effectiveness scores). This finding is in line with our original hypothesis that fluency-building processes are an important aspect of connection making.

Second, counter to our hypothesis, we did not find evidence that students' performance on the sense-differences components of Chem Tutor was associated with learning outcomes when controlling for pretest performance, but only when not controlling for pretest performance. Apparently, students' performance on sense-differences problems did not have an impact on their learning outcomes beyond what is expected given their prior knowledge. This finding is surprising because it stands in contrast to the results of Studies 2 and 3, where we found such an association independent of students' prior knowledge.

Third, counter to our hypothesis, we found no associations between problem-solving performance with students' open-ended scores. The lack of an association of problems-solving performance with open-ended scores might be due to the lower learning gains on these measures. Lower learning gains mean that the variance of students' performance on the final posttest may be mostly explained by students' prior knowledge and only to a lesser extent by their interactions with Chem Tutor that lead to an increase in their knowledge.

An important limitation of Study 4 is that it was conducted in the middle of the semester, so that students likely had considerable knowledge about the concepts Chem Tutor is designed to target. As mentioned above, if an instructor decides to use Chem Tutor, he/she is likely to do so during the entire semester, from the beginning of the course. Thus, even though Chem Tutor was conducted in the field, as part of a chemistry course, it remains to be investigated whether Chem Tutor is also effective when used from the beginning of a chemistry course throughout the semester. A further limitation of Study 4 is that students were self-selected. As mentioned before, the extra-credit opportunity might have been more attractive to lower-performing students than to higher-performing students. Self-selection may have an impact on the generalizability of our findings to a broad range of students. To address these shortcomings, we are planning a study in which all students enrolled in an introductory chemistry course use Chem Tutor as a homework system at the beginning of the semester.

6. General discussion

We described a multi-methods approach to *ground* the design of an ITS for connection making in the requirements specific to the target domain; that is, to investigate what specific learning processes play a role within the given target domain. Our empirical approach was guided by several research goals, discussed in the following.

6.1. Goal 1: Identify learning processes

Our first research goal was to identify which learning processes are important for connection making between multiple graphical representations in chemistry and should be supported by Chem Tutor. We addressed this goal by building on a theoretical framework that proposed two separate processes for connection making: sense-making processes and fluency-building processes. These learning processes result in two separate abilities: sense-making ability and perceptual fluency.

We then conducted two studies with the goal to better understand the role these hypothesized connection-making abilities play in students' knowledge about chemistry. Study 1 supports the hypothesis that sense-making abilities and fluency-building abilities are separate aspects of connection-making competence. Study 2 suggested that the ability to make sense of differences between graphical representations is more important than the ability to make sense of similarities between representations. These two studies on knowledge structures led to the hypothesis investigated in Studies 3 and 4—that both sense-making and fluency-building processes are critical aspects of learning in chemistry and that sense-making support should focus on differences between representations. Furthermore, Study 2 indicates that sense-making support should help students move beyond describing differences between representations, by reasoning about the complementary function of these representations and by relating what students can see in the representations to domain-relevant concepts that are not explicitly shown.

We then conducted two studies that focused on learning processes. Results from both Studies 3 and 4 are in line with the hypothesis that *sense-making support* for connection making has an impact on students' learning. Study 3 indicated that both aspects of sense-making processes are important, whereas Study 4 indicated that sense-similarities processes have a larger impact on chemistry learning than sense-differences processes. The different results in Studies 3 and 4 might be due to the fact that students in Study 4 had higher prior knowledge than students in Study 3. Another potential explanation may be that making sense of differences between representations is more important earlier in the learning process, whereas making sense of similarities is more important later in the learning process. Given that this interpretation is post hoc and relies on the comparison of different study populations, it is highly speculative, and can merely serve as a hypothesis for future empirical research.

The results from Study 4 are in line with the hypothesis that *fluency-building support* has an impact on chemistry learning, whereas the results from Study 3 are not. We consider differences in students' prior knowledge between Studies 3 and 4 for potential explanations of these disparate findings. It may be that fluency-building processes play a larger role later in the learning process because they build on students' ability to make sense of connections. We argued above that sense-making abilities typically receive more attention in traditional instruction than perceptual fluency. Therefore, students in Study 4 might have received more instruction that focused on sense-making abilities in their chemistry introduction lecture than students in Study 3 did. However, as noted before, any attempt at explaining differences between Studies 3 and 4 is highly speculative because many other unknown facts may have contributed to differences between Studies 3 and 4.

In summary, our studies provide support for the overall hypothesis that sense-making and fluency-building processes are important aspects of connection making between multiple graphical representations in chemistry. Open questions remain as to how different aspects of sense-making ability and perceptual fluency relate to one another, and how the learning processes that result in the acquisition of these connection-making abilities build on one another. These questions are interesting not only from a theoretical point of view, but they would also have practical implications as to how best to sequence support for sense-making and fluency-building processes so that they optimally meet the students' instructional needs.

6.2. Goal 2: Identify visual attention behaviors that indicate productive learning processes

Our second research goal was to investigate which visual attention behaviors indicate productive learning processes as students make connections between multiple graphical representations in chemistry. To address this goal, we combined eye-tracking measures that assessed visual attention behaviors with interviews that assessed students' conceptual reasoning behaviors and with tutor log data that assessed problem-solving behaviors within Chem Tutor. Our approach to combine multiple measures allowed us to disambiguate whether certain visual attention behaviors indicate productive or unproductive learning processes.

We had expected that frequency of switching between graphical representations would be associated with high-quality learning processes, based on prior research that has investigated conceptual integration of text and graphic (Holsanova & Holmberg, 2009; Johnson & Mayer, 2012; Mason, Pluchino, & Tornatora, 2013; Mason, Pluchino, Tornatora, et al., 2013). Our findings were not in line with this prior research. One key difference between our studies and earlier research is that the latter focused on instructional materials that contain text and one additional graphical representation, rather than two different graphical representations. Text is a type of representation that students have high fluency with, and it is often the dominant representation that guides processing of the graphical representations (Rayner et al., 2001; Schmidt-Weigand et al., 2010). It may be that switching between text and a graphical representation indicates deep processing because students refer to a graphical representation to make better sense of the text. In our studies, there was no dominant textual representation, so that students' visual attention was not guided by text. Switching between two graphical representations may thus indicate an entirely different process than switching between text and graphical representation. Even so, we are not the first to find that switching may not always correspond to conceptual integration: other studies on learning with text and graphic found that switching between representations is not always indicative of productive learning processes (Schmidt-Weigand et al., 2010). It has also been noted that switches may correspond to failed attempts to integrate information from different representations (Holsanova & Holmberg, 2009). Apparently, frequency of switching as a global measure may not provide useful information about the quality of visual attention behaviors. Instead, it may matter at what time during the problem-solving process switches occur (e.g., when the student first sees the problem, frequent switches may indicate confusion, whereas later in the problem-solving process, frequent switches may indicate conceptual integration). In addition, what constitutes frequent switching may differ from student to student (e.g., high prior knowledge students may require fewer switches to conceptually integrate than low prior knowledge students). Thus, further, more fine-grained analyses are needed to investigate how best to use frequencies of switching to analyze the quality of learning processes.

With respect to first-inspection durations, we had expected that they would be associated with superficial processing because prior research suggests that first inspections serve the function of initial, somewhat involuntary processing (Hyönä et al., 2003; Hyönä & Nurminen, 2006; Mason, Pluchino, & Tornatora, 2013). Studies 2 and 3 yielded conflicting results about the role of first-inspection durations. Thus, more research is needed to investigate how to interpret first-inspection durations.

With respect to second-inspection durations, we had hypothesized that longer second-inspection durations would indicate high-quality learning processes, because prior research indicates that they reflect intentional processing to integrate the information with other information (Hyönä et al., 2003; Hyönä & Nurminen, 2006; Mason, Pluchino, & Tornatora, 2013; Schlag & Ploetzner, 2011). The results from Studies 2 and 3 are in line with this prior research. Thus, with respect to the interpretation of second-fixation durations, our research on multiple graphical representations replicates prior research on multiple external representations.

In summary, our approach illustrates that the combination of multiple measures, such as visual attention measures, problem-solving performance, and learning outcomes, can yield insights into what measures of visual attention mean. We also illustrated how the use of multiple measures can prevent jumping to false conclusions, for instance, by interpreting switching between representations as indicators of productive learning processes—an interpretation that our results do not support. The combination of eye-tracking data, interview data, and log data might thus also lend useful insights into more fine-grained analyses of visual attention patterns that we believe are needed to better understand how to use switching and first-inspection durations for the analysis of learning processes. Furthermore, we found evidence that second-inspection durations are a useful global measure of productive learning processes that can help us evaluate Chem Tutor in future research.

6.3. Goal 3: Improve students' learning of important concepts in chemistry

Our third research goal was to improve students' learning of important chemistry concepts. To address this goal, we first investigated which concepts Chem Tutor should target. Based on interviews, Study 2 identified a “knowledge gap” between undergraduate and graduate students. Based on these findings, Chem Tutor focused on helping students reason about how bonding involves interactions between molecules and atoms, and about how these interactions are facilitated through the behavior of electrons. We also included these aspects in the pretests and posttests we used to pilot test Chem Tutor in Studies 3 and 4.

We then investigated whether Chem Tutor improves students' learning outcomes in chemistry. We conducted two pilot studies in which students used Chem Tutor to learn about chemistry. The learning outcome measures we considered included items that assessed conceptual knowledge about bonding, knowledge about graphical representations, and included open-ended items that required students to explain their reasoning in their own words. We assessed both students' effectiveness and efficiency in completing these tests. We found highly significant and large learning gains on these measures—not only in the artificial laboratory environment of Study 3, but also in the naturalistic, realistic setting of Study 4. These results are promising, given that students spent only 1.5–2 h with Chem Tutor, and given that these were pilot studies on the first version of Chem Tutor. Further, the interviews from Study 3 show that students had an overall positive opinion about Chem Tutor.

6.4. Limitations and future directions

There are several limitations that lead to open research questions that we will address in future research. Addressing these open questions will further improve Chem Tutor. First, even though we found large learning gains, there is considerable room for improvement. Students' scores at the final posttest were still relatively low (see Tables 6 and 10). To some extent, the low scores may reflect the fact that the tests we created for Studies 3 and 4 set a very high bar for undergraduate students. As mentioned above (see Section 5.1.1), we created the test based on the verbal utterances from undergraduate and graduate students that we collected in Study 2. Thus, the tests reflect misconceptions that undergraduate students were likely to mention and aim for reasoning that is common among graduate students. In other words, the tests were likely harder than tests that students usually encounter in their undergraduate courses. Therefore, the learning gains we found are meaningful even if posttest scores are rather low. In future studies, we will also include items drawn from exams that the students are likely to encounter in introductory chemistry courses. An additional reason why Chem Tutor did not achieve better learning outcomes may be that a 2 h long intervention is too short to result in high performance on a conceptual posttest on a complex topic. To address this issue, we will expand the number of activities covered by Chem Tutor. Another reason may be that Chem Tutor does not address all misconceptions that students have developed about bonding. To address this possibility, we will conduct think-alouds with students, combined with follow-up interviews to investigate how Chem Tutor affects their reasoning about bonding concepts.

Second, the results from Studies 2, 3, and 4 were somewhat contradictory with respect to the relative importance of sense-differences, sense-similarities, and fluency-building problems for students' learning. We argued that the effectiveness of these components may depend on students' prior knowledge. The fact that students in Studies 2, 3, and 4 were drawn from different populations that likely differed with respect to their prior chemistry knowledge might therefore explain some of the differences in the importance of sense-making problems and fluency-building problems. Furthermore, it is possible that these different learning processes build on one another. Put differently, comparing the results from Studies 2, 3, and 4 leads to the new hypothesis that fluency-building processes build on sense-making ability. This hypothesis is in line with results from earlier research in math learning (Rau, Aleven, & Rummel, 2013; Rau, Scheines, Aleven, & Rummel, 2013). However, with respect to chemistry learning, this hypothesis was formed post hoc and is based on highly speculative interpretations of differences between our study populations. Therefore, further research is needed to investigate how sense-making and fluency-building processes interact. If the hypotheses that fluency-building processes depend on the student having (at least some level of) sense-making ability holds true, Chem Tutor might be most effective if it adapts to the individual student's learning rate by providing fluency-building support only when the student has acquired the prerequisite level of sense-making ability. Such adaptive connection-making support might yield more effective instruction.

An additional open question regarding adaptive capabilities relates to our findings on the likability of the system. Even though they were overall positive, the interviews in Study 3 show that there is room for improvement with respect to how much students enjoyed working with Chem Tutor. What stands out in particular is that there were disagreements among students as to which aspects they liked or disliked.

For instance, some students liked the fact that Chem Tutor included repetition, whereas other students disliked this very fact. The finding that students who disliked repetitiveness had lower pretest scores might indicate that lower-performing students tend to prefer variety over repetitiveness. In addition, students commented that Chem Tutor involved little instruction, which resulted in confusion about how to solve the problems for some students. However, only a minority of students felt a need for additional instruction. Thus, students' needs for additional instruction appear to differ—they likely depend on their prior knowledge level. Including adaptive features into Chem Tutor might address these issues. We are planning to include adaptive features into Chem Tutor so that it adapts instruction to the needs of lower-performing students by including more variability and more instruction, at least at the beginning of the intervention. However, we cannot necessarily draw the conclusion from Study 3 that this modification would increase students' enjoyment or their learning gains. Thus, an empirical evaluation of such adaptive features would be needed.

A further limitation is that our studies were correlational—they do not provide conclusive evidence that Chem Tutor *caused* students' learning gains. Thus, even though Studies 3 and 4 evaluated Chem Tutor as a learning intervention, it is possible that students' learning gains were caused by something other than their work with Chem Tutor. To address this issue, we are planning a formal experimental study that compares students who are randomly assigned to using Chem Tutor to a control group that will not work with Chem Tutor.

7. Conclusion: Domain-specific grounding of connection-making support

We illustrated our approach to ground the design of an ITS for connection making between multiple graphical representations in the specific requirements of the target domain.

We conclude by proposing that our approach can be applied to other domains than chemistry and to other educational technologies than ITSs.

The first step in our grounding approach is to investigate what role graphical representations play in how students' domain knowledge is structured. The second step is to investigate how different types of abilities in using and interpreting graphical representations relate to the target knowledge, such as making inferences about domain-relevant concepts. The third step is to use this understanding of knowledge structures to guide the design of an initial educational technology. The fourth step is to pilot test the initial educational technology while focusing on learning processes. In our view, the combination of multiple process measures with learning outcome measures is critical to helping us understand how graphical representations shape students' acquisition of the domain knowledge. Our results illustrate that this approach can yield a highly effective first version of the technology, and will likely yield new lines of research that will improve further iterations.

Our approach has the potential to impact the development of new, effective educational technologies in the STEM domains for the following reasons. First, there is an educational need to help students learn with multiple graphical representations: they are ubiquitous in STEM domains, and they provide stumbling blocks for students' learning in many domains, including chemistry, biology, engineering, statistics, and many more (Arcavi, 2003; Cheng, 1999; Kozma et al., 2000; Larkin & Simon, 1987; Lewalter, 2003; Stieff et al., 2011; Urban-Woldron, 2009; Zhang & Linn, 2011).

Second, there is a scientific need to study learning with multiple graphical representations: prior research has mostly focused on multiple external representations (i.e., text and one graphic; Ainsworth & Loizou, 2003; Bodemer et al., 2005; Butcher & Aleven, 2007; Magner et al., 2014; Rasch & Schnotz, 2009). Yet, learning with multiple graphical representations relies more strongly on perceptual aspects because there is no guiding dominant representation that we can expect students to be highly fluent with (e.g., text). Perceptual aspects play a critical role in students' learning with multiple graphical representations because different graphical representations tend to share both critical and incidental perceptual aspects, and identifying them requires learning at the perceptual level (e.g., reading color in EPMs and ball-and-stick figures correctly, or interpreting the geometrical arrangement in Lewis structures and space-filling models correctly). We still know very little about what role these learning processes play for students' acquisition of domain knowledge, or how best to support them.

Third, there is a need to develop educational technologies that help students learn with multiple graphical representations: currently available educational technologies reflect the lack of focus on perceptual fluency in educational research and focus solely on sense-making processes (Kozma & Russell, 2005a; Michalchik et al., 2008; Stieff, 2005; Wu et al., 2001). Even though sense-making abilities are crucial for students' learning, the fact that non-technology based interventions that target perceptual fluency (Eastwood, 2013; Moreira, 2013) have recently gained attention in the STEM disciplines suggests that there is indeed a need to provide instructional support in addition to sense-making support.

In conclusion, our studies show that our approach for domain-specific grounding for connection making between multiple graphical representations is successful. We argued that our approach is applicable to a broad range of domains and educational technologies. Thus, we believe that it can fundamentally improve educational technologies for STEM learning.

Acknowledgments

This work was supported by the UW Madison Graduate School and the Wisconsin Center for Education Research. We thank Amanda Evenstone, Abigail Dreps, Brady Cleveland, William Keesler, Taryn Gordon, and Theresa Shim for their contributions.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.compedu.2014.12.009.

References

Ainsworth, S. (2006). DeFT: a conceptual framework for considering learning with multiple representations. *Learning and Instruction, 16*(3), 183–198.

Ainsworth, S., Bibby, P., & Wood, D. (2002). Examining the effects of different multiple representational systems in learning primary mathematics. *Journal of the Learning Sciences, 11*(1), 25–61.
Ainsworth, S., & Loizou, A. T. (2003). The effects of self-explaining when learning with text or diagrams. *Cognitive Science: A Multidisciplinary Journal, 27*(4), 669–681.
Aleven, V., McLaren, B. M., Sewall, J., & Koedinger, K. R. (2009). A new paradigm for intelligent tutoring systems: example-tracing tutors. *International Journal of Artificial Intelligence in Education, 19*(2), 105–154.
Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive tutors: lessons learned. *Journal of the Learning Sciences, 4*(2), 167–207.
Arcavi, A. (2003). The role of visual representations in the learning of mathematics. *Educational Studies in Mathematics, 52*(3), 215–241.
Arroyo, I., Royer, J. M., & Woolf, B. P. (2011). Using an intelligent tutor and math fluency training to improve math performance. *International Journal of Artificial Intelligence in Education, 21*(1), 135–152.
Bodemer, D., & Faust, U. (2006). External and mental referencing of multiple representations. *Computers in Human Behavior, 22*(1), 27–42.
Bodemer, D., Ploetzner, R., Bruchmüller, K., & Häcker, S. (2005). Supporting learning with interactive multimedia through active integration of representations. *Instructional Science, 33*(1), 73–95.
Bodemer, D., Ploetzner, R., Feuerlein, I., & Spada, H. (2004). The active integration of information during learning with dynamic and interactive visualisations. *Learning and Instruction, 14*(3), 325–341.
Bodner, G. M., & Domin, D. S. (2000). Mental models: the role of representations in problem solving in chemistry. *University Chemistry Education, 4*(1), 24–30.
Brünken, R., Seufert, T., & Zander, S. (2005). Förderung der Kohärenzbildung beim Lernen mit multiplen Repräsentationen. *Zeitschrift für Pädagogische Psychologie, 19*(1), 61–75.
Butcher, K., & Aleven, V. (2007). *Integrating visual and verbal knowledge during classroom learning with computer tutors*. Paper presented at the 29th Annual Conference of the Cognitive Science Society, Austin, TX.
Butcher, K., & Aleven, V. (2008). Diagram interaction during intelligent tutoring in geometry: support for knowledge retention and deep transfer. In *Proceedings of the 30th Annual Meeting of the Cognitive Science Society, CogSci 2008*. New York, NY: Lawrence Erlbaum.
Cheng, P. (1999). Unlocking conceptual learning in mathematics and science with effective representational systems. *Computers & Education, 33*, 109–130.
Cheng, M., & Gilbert, J. K. (2009). Towards a better utilization of diagrams in research into the use of representative levels in chemical education. In J. K. Gilbert, & D. F. Treagust (Eds.), *Multiple representations in chemical education* (pp. 191–208). Berlin/Heidelberg: Springer.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2 ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.
Coll, R. K., & Treagust, D. F. (2003a). Investigation of secondary school, undergraduate, and graduate learners' mental models of ionic bonding. *Journal of Research in Science Teaching, 40*(5), 464–486.
Coll, R. K., & Treagust, D. F. (2003b). Learners' mental models of metallic bonding: a cross-age study. *Science Education, 87*(5), 685–707.
Cook, M., Wiebe, E. N., & Carter, G. (2007). The influence of prior knowledge on viewing and interpreting graphics with macroscopic and molecular representations. *Science Education, 92*(5), 848–867.
Corbett, A. (2001). Cognitive computer tutors: solving the two-sigma problem. In *User Modeling: Proceedings of the 8th International Conference* (pp. 137–147). Berlin Heidelberg: Springer.
Corbett, A. T., Koedinger, K., & Hadley, W. S. (2001). Cognitive tutors: from the research classroom to all classrooms. In P. S. Goodman (Ed.), *Technology enhanced learning: Opportunities for change* (pp. 235–263). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.
Davidowitz, B., & Chittleborough, G. (2009). Linking the macroscopic and sub-microscopic levels: diagrams. In J. K. Gilbert, & D. F. Treagust (Eds.), *Multiple representations in chemical education* (pp. 169–191). Netherlands: Springer.
Dori, Y. J., & Barak, M. (2001). Virtual and physical molecular modeling: fostering model perception and spatial understanding. *Educational Technology & Society, 4*(1), 61–74.
Eastwood, M. L. (2013). Fastest fingers: a molecule-building game for teaching organic chemistry. *Journal of Chemical Education, 90*(8), 1038–1041.
Eilam, B. (2013). Possible constraints of visualization in biology: Challenges in learning with multiple representations. In *Multiple representations in biological education* (pp. 55–73). Netherlands: Springer.
Eilam, B., & Poyas, Y. (2008). Learning with multiple representations: extending multimedia learning beyond the lab. *Learning and Instruction, 18*(4), 368–378.
Fahle, M., & Poggio, T. (2002). *Perceptual learning*. The MIT Press.
Furio, C., Calatayud, M. L., Barcenas, S. L., & Padilla, O. M. (2000). Functional fixedness and function reduction as common sense reasonings in chemical equilibrium and in geometry and polarity of molecules. *Science Education, 84*, 545–565.
Gabel, D. L., & Bunce, D. M. (1994). *Research on problem solving: Chemistry handbook of research on science teaching and learning* (pp. 301–326).
Gibson, E. J. (1969). *Principles of perceptual learning and development*. New York: Prentice Hall.
Gilbert, J. K., & Treagust, D. F. (2009). Towards a coherent model for macro, submicro and symbolic representations in chemical education. In J. K. Gilbert, & D. F. Treagust (Eds.), *Multiple representations in chemical education* (pp. 333–350). Netherlands: Springer.
Gobert, J. D., O'Dwyer, L., Horwitz, P., Buckley, B. C., Levy, S. T., & Wilensky, U. (2011). Examining the relationship between students' understanding of the nature of models and conceptual learning in biology, physics, and chemistry. *International Journal of Science Education, 33*(5), 653–684.
Gutwill, J. P., Frederiksen, J. R., & White, B. Y. (1999). Making their own connections: students' understanding of multiple models in basic electricity. *Cognition and Instruction, 17*(3), 249–282.
Holsanova, J., Holmberg, N., & Holmqvist, K. (2009). Reading information graphics: the role of spatial contiguity and dual attentional guidance. *Applied Cognitive Psychology, 23*, 1215–1226. http://dx.doi.org/10.1002/acp.1525.
Hyönä, J., Lorch, R. F., & Rinck, M. (2003). *Eye movement measures to study global text processing*. The mind's eye: Cognitive and applied aspects of eye movement research (pp. 313–334).
Hyönä, J., & Nurminen, A. M. (2006). Do adult readers know how they read? Evidence from eye movement patterns and verbal reports. *British Journal of Psychology, 97*(1), 31–50. http://dx.doi.org/10.1348/000712605X53678.
Johnson, C. I., & Mayer, R. E. (2012). An eye movement analysis of the spatial contiguity effect in multimedia learning. *Journal of Experimental Psychology: Applied, 18*(2), 178–191.
de Jong, T., Ainsworth, S. E., Dobson, M., Van der Meij, J., Levonen, J., Reimann, P., et al. (1998). Acquiring knowledge in science and mathematics: the use of multiple representations in technology-based learning environments. In M. W. Van Someren, W. Reimers, H. P. A. Boshuizen, & T. de Jong (Eds.), *Learning with multiple representations*. Bingley, UK: Emerald Group Publishing Limited.
Justi, R., & Gilbert, J. K. (2002). *Models and modelling in chemical education*. Chemical education: Towards research-based practice (pp. 47–68). Kluwer Academic Publishers.
Justi, R., Gilbert, J. K., & Ferreira, P. F. (2009). The application of a "model of modelling" to illustrate the importance of metavisualisation in respect of the three types of representation. In J. K. Gilbert, & D. F. Treagust (Eds.), *Multiple representations in chemical education* (pp. 285–307). Berlin/Heidelberg: Springer.
Kellman, P. J., & Garrigan, P. B. (2009). Perceptual learning and human expertise. *Physics of Life Reviews, 6*(2), 53–84.
Kellman, P. J., & Massey, C. M. (2013). Perceptual Learning, cognition, and expertise. *The Psychology of Learning and Motivation, 558*, 117–165.
Kellman, P. J., Massey, C. M., Roth, Z., Burke, T., Zucker, J., Saw, A., et al. (2008). Perceptual learning and the technology of expertise: studies in fraction learning and algebra. *Pragmatics & Cognition, 16*(2), 356–405.
Kellman, P. J., Massey, C. M., & Son, J. Y. (2009). Perceptual learning modules in mathematics: Enhancing students' pattern recognition, structure extraction, and fluency. In *Topics in cognitive science* (Vol. 2, pp. 285–305).
Kind, V. (2004). *Beyond Appearances: Students' misconceptions about basic chemical ideas*. Durham: School of Education, Durham University.
Koedinger, K. R., Baker, R., Cunningham, K., Skogsholm, A., Leber, B., & Stamper, J. (2010). A data repository for the EDM community: the PSLC Data-Shop. In C. Romero (Ed.), *Handbook of educational data mining* (pp. 10–12). Boca Raton, FL: CRC Press.
Koedinger, K. R., & Corbett, A. (2006). *Cognitive Tutors: Technology bringing learning sciences to the classroom*. New York, NY: Cambridge University Press.
Koedinger, K. R., Corbett, A. T., & Perfetti, C. (2012). The knowledge-learning-instruction framework: bridging the science-practice chasm to enhance Robust student learning. *Cognitive Science, 36*(5), 757–798.
Kordaki, M. (2010). A drawing and multi-representational computer environment for beginners' learning of programming using C: design and pilot formative evaluation. *Computers & Education, 54*, 69–87.
Kozma, R., Chin, E., Russell, J., & Marx, N. (2000). The roles of representations and tools in the chemistry laboratory and their implications for chemistry learning. *The Journal of the Learning Sciences, 9*(2), 105–143.
Kozma, R., & Russell, J. (2005a). *Multimedia learning of chemistry Cambridge handbook of multimedia learning* (pp. 409–428).
Kozma, R., & Russell, J. (2005b). *Students becoming chemists: Developing representationl competence*. Visualization in science education (pp. 121–145). Amsterdam: Springer Netherlands.

Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science: A Multidisciplinary Journal, 11*(1), 65–100.
Lewalter, D. (2003). Cognitive strategies for learning from static and dynamic visuals. *Learning and Instruction, 13*(2), 177–189.
Lewis, D., & Barron, A. (2009). Animated demonstrations: evidence of improved performance efficiency and the worked example effect. *Human Centered Design, Lecture Notes in Computer Science, 5619*, 247–255.
Linenberger, K. J., & Bretz, S. L. (2012). Generating cognitive dissonance in student interviews through multiple representations. *Chemistry Education Research and Practice, 13*(3), 172–178.
Little, R. J., & Rubin, D. B. (1989). The analysis of social science data with missing values. *Sociological Methods & Research, 18*(2–3), 292–326.
Magner, U. I., Schwonke, R., Aleven, V., Popescu, O., & Renkl, A. (2014). Triggering situational interest by decorative illustrations both fosters and hinders learning in computer-based learning environments. *Learning and Instruction, 29*, 141–152.
Mason, L., Pluchino, P., & Tornatora, M. C. (2013). Effects of picture labeling on science text processing and learning: evidence from eye movements. *Reading Research Quarterly, 48*(2), 199–214.
Mason, L., Pluchino, P., Tornatora, M. C., & Ariasi, N. (2013). An eye-tracking study of learning from science text with concrete and abstract illustrations. *The Journal of Experimental Education, 81*(3), 356–384.
Means, B., Toyama, Y., Murphy, R., Bakia, M., & Jones, K. (2009). *Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies*. US Department of Education.
van der Meij, J., & de Jong, T. (2006). Supporting students' learning with multiple representations in a dynamic simulation-based learning environment. *Learning and Instruction, 16*(3), 199–212.
Michalchik, V., Rosenquist, A., Kozma, R., Kreikemeier, P., & Schank, P. (2008). Representational resources for constructing shared understandings in the high school chemistry classroom. In J. K. Gilbert, M. Reiner, & M. B. Nakhleh (Eds.), *Visualization: Theory and practice in science education* (pp. 233–282). Springer Netherlands.
Moreira, R. F. (2013). A game for the early and rapid assimilation of organic nomenclature. *Journal of Chemical Education*, 1035–1037.
Nakiboglu, C. (2003). Instructional misconceptions of Turkish prospective chemistry teachers about atomic orbitals and hybridization. *Chemistry Education Research and Practice, 4*(2), 171–188.
Nathan, M. J., Walkington, C. A., Srisurichan, R., & Alibali, M. W. (2011). *Modal engagements in precollege engineering: Tracking math and science concepts across symbols, sketches, software, silicone and wood*. American Society for Engineering Education.
Nicoll, G. (2001). A report of undergraduates' bonding misconceptions. *International Journal of Science Education, 23*(7), 707–730.
Noss, R. R., Healy, L., & Hoyles, C. (1997). The construction of mathematical meanings: connecting the visual with the symbolic. *Educational Studies in Mathematics, 33*, 203–233.
Özgün-Koca, S. A. (2008). Ninth grade students studying the movement of fish to learn about linear relationships: the use of video-based analysis software in mathematics classrooms. *The Mathematics Educator, 18*(1), 15–25.
Rasch, T., & Schnotz, W. (2009). Interactive and non-interactive pictures in multimedia learning environments: effects on learning outcomes and learning efficiency. *Learning and Instruction, 19*(5), 411–422.
Rau, M. A., Aleven, V., & Rummel, N. (2013). Complementary effects of sense-making and fluency-building support for connection making: a matter of sequence? In H. C. Lane, K. Yacef, J. Mostow, & P. Pavlik (Eds.), *Artificial intelligence in education* (pp. 329–338) Berlin Heidelberg: Springer.
Rau, M. A., Aleven, V., & Rummel, N. (2014). Successful learning with multiple graphical representations and self-explanation prompts. *Journal of Educational Psychology*. http://dx.doi.org/10.1037/a0037211.
Rau, M. A., Aleven, V., Rummel, N., & Rohrbach, S. (2012). Sense making alone doesn't do it: fluency matters too! ITS support for robust learning with multiple representations. In S. Cerri, W. Clancey, G. Papadourakis, & K. Panourgia (Eds.), *Intelligent tutoring systems* (Vol. 7315, pp. 174–184). Berlin/Heidelberg: Springer.
Rau, M. A., Rummel, N., Aleven, V., Pacilio, L., & Tunc-Pekkan, Z. (2012). How to schedule multiple graphical representations? A classroom experiment with an intelligent tutoring system for fractions. In J. V. Aalst, K. Thompson, M. J. Jacobson, & P. Reimann (Eds.), *The future of learning: Proceedings of the 10th international conference of the learning sciences (ICLS 2012)* (Vol. 1, pp. 64–71). Sydney, Australia: ISLS.
Rau, M. A., Scheines, R., Aleven, V., & Rummel, N. (2013). Does representational understanding enhance fluency or vice versa? Searching for mediation models. In S. K. D'Mello, R. A. Calvo, & A. Olney (Eds.), *Proceedings of the 6th International Conference on Educational Data Mining (EDM 2013)* (pp. 161–169). International Educational Data Mining Society.
Rau, M. A. (under review). Habitually closing the loop between theory and practice: How educational technologies and data mining can shape educational psychology research about learning with multiple graphical representations.
Rayner, K., Rotello, C. M., Stewart, A. J., Keir, J., & Duffy, S. A. (2001). Integrating text and pictorial information: eye movements when looking at print advertisements. *Journal of Experimental Psychology: Applied, 7*(3), 219–226.
Ritter, S., Anderson, J. R., Koedinger, K. R., & Corbett, A. T. (2007). Cognitive Tutor: applied research in mathematics education. *Psychonomic Bulletin & Review, 14*(2), 249–255.
Schlag, S., & Ploetzner, R. (2011). Supporting learning from illustrated texts: conceptualizing and evaluating a learning strategy. *Instructional Science, 39*(6), 921–937. http://dx.doi.org/10.1007/s11251-010-9160-3.
Schmidt-Weigand, F., Kohnert, A., & Glowalla, U. (2010). Explaining the modality and contiguity effects: new insights from investigating students' viewing behaviour. *Applied Cognitive Psychology, 24*(2), 226–237.
Schnotz, W., & Bannert, M. (2003). Construction and interference in learning from multiple representation. *Learning and Instruction, 13*(2), 141–156.
Schwonke, R., Ertelt, A., & Renkl, A. (2008). Fostering the translation between external representations. Does it enhance learning with an intelligent tutoring program? In J. Zumbach, N. Schwartz, T. Seufert, & L. Kester (Eds.), *Beyond knowledge: The legacy of competence* (pp. 117–119) Springer Netherlands.
Schwonke, R., & Renkl, A. (2010). *How do proficient learners construct mental representations of different but related external representations?*. Paper presented at the EARLI SIG 2, Tuebingen.
Seufert, T. (2003). Supporting coherence formation in learning from multiple representations. *Learning and Instruction, 13*(2), 227–237.
Seufert, T., & Brünken, R. (2006). Cognitive load and the format of instructional aids for coherence formation. *Applied cognitive psychology, 20*, 321–331.
Shih, T. H., & Fan, X. (2008). Comparing response rates from web and mail surveys: a meta-analysis. *Field methods, 20*(3), 249–271.
Stieff, M. (2005). Connected chemistry-A novel modeling environment for the chemistry classroom. *Journal of Chemical Education, 82*(3), 489–493.
Stieff, M., Hegarty, M., & Deslongchamps, G. (2011). Identifying representational competence with multi-representational displays. *Cognition and Instruction, 29*(1), 123–145.
Strickland, A. M., Kraft, A., & Bhattacharyya, G. (2010). What happens when representations fail to represent? Graduate students' mental models of organic chemistry diagrams. *Chemistry Education Research and Practice, 11*(4), 293–301.
Taber, S. B. (2001). *Making connections among different representations: The case of multiplication of fractions*.
Taber, K. S. (2009). Learning at the symbolic level. In J. K. Gilbert, & D. F. Treagust (Eds.), *Multiple representations in chemical education* (pp. 75–105). Berlin/Heidelberg: Springer.
Taber, K. S. (2013). Revisiting the chemistry triplet: drawing upon the nature of chemical knowledge and the psychology of learning to inform chemistry education. *Chemistry Education Research and Practice, 14*(2), 156–168.
Taber, K. S. (2014). The significance of implicit knowledge for learning and teaching chemistry. *Chemistry Education Research and Practice, 15*, 447–461.
Taber, K. S., & Coll, R. K. (2002). Bonding. In J. K. Gilbert, O. De Jong, R. Justi, D. F. Treajust, & J. H. Van Driel (Eds.), *Chemical education: Towards research-based practice* (pp. 213–234). Dordrecht Boston London: Kluwer Academic Publishers.
Talanquer, V. (2013). Chemistry education: ten facets to shape us. *Journal for Research in Mathematics Education, 90*, 832–838.
Urban-Woldron, J. (2009). Interactive simulations for the effective learning of physics. *Journal of Computers in Mathematics and Science Teaching, 28*(2), 163–176.
Van Gog, T., & Paas, F. (2008). Instructional efficiency: Revisiting the original construct in educational research. *Educational Psychologist, 43*(1), 16–26.
VanLehn, K. (2011). The relative effectiveness of human tutoring, intelligent tutoring systems and other tutoring systems. *Educational Psychologist, 46*(4), 197–221. http://dx.doi.org/10.1080/00461520.2011.611369.
Williamson, V. M. (2014). Teaching chemistry conceptually. In I. Devetak, & S. Aleksij (Eds.), *Learning with understanding in the chemistry classroom* (pp. 193–208). Dordrecht, Netherlands: Springer.
Wu, H. K., Krajcik, J. S., & Soloway, E. (2001). Promoting understanding of chemical representations: students' use of a visualization tool in the classroom. *Journal of Research in Science Teaching, 38*(7), 821–842.
Wu, H. K., & Shah, P. (2004). Exploring visuospatial thinking in chemistry learning. *Science Education, 88*(3), 465–492.
Zhang, Z. H., & Linn, M. C. (2011). Can generating representations enhance learning with dynamic visualizations? *Journal of Research in Science Teaching, 48*(10), 1177–1198.