Chemistry Education Research and Practice RSCPublishing

Paper

Cite this: DOI: 10.1039/x0xx00000x

Self-explaining effect in general chemistry instruction: Eliciting overt categorical behaviours by design

Adrian Villalta-Cerdas[a] and Santiago Sandi-Urena*[a]

Received 00th January 2012,
Accepted 00th January 2012

DOI: 10.1039/x0xx00000x

www.rsc.org/

Self-explaining refers to the generation of inferences about causal connections between objects and events. In science, this may be summarised as making sense of how and why actual or hypothetical phenomena take place. Research findings in educational psychology show that implementing activities that elicit self-explaining improves learning in general and specifically enhances authentic learning in the sciences. Research also suggests that self-explaining influences many aspects of cognition, including acquisition of problem-solving skills and conceptual understanding. Although the evidence that links self-explaining and learning is substantial, most of the research has been conducted in experimental settings. There remains a need for research conducted in the context of real college science learning environments. Working to address that need, the larger project in which this work is embedded studied the following: (a) the effect of different self-explaining tasks on self-explaining behaviour and (b) the effect of engaging in different levels of self-explaining on learning chemistry concepts. The present study used a multi-condition, mixed-method approach to categorise student self-explaining behaviours in response to learning tasks. Students were randomly assigned to conditions that included the following: explaining correct and incorrect answers, explaining agreement with another's answer, and explaining one's own answer for others to use. Textual, individual data was gathered in the classroom ecology of a university, large-enrolment general chemistry course. Findings support an association between the self-explaining tasks and students' self-explaining behaviours. Thoughtful design of learning tasks can effectively elicit engagement in sophisticated self-explaining in natural, large-enrolment college chemistry classroom environments.

Introduction

Regardless of correctness, generation of authentic explanations is a core characteristic of scientific behaviour and central to scientific and technological development (Deutsch, 2011). This centrality is reflected by the *US Next Generation Science Standards* (National Research Council, 2013) that posit the construction of explanations as one of eight practices of science essential for all students to learn. Furthermore, *A Science Framework for K-12 Science Education* (National Research Council, 2012) asserts that when students demonstrate their understanding of the implications of a scientific idea by developing their own explanations of phenomena, they take part in an essential activity by which conceptual change can occur. That is, in addition to being a desirable learning outcome in itself, the ability to generate one's own explanations supports conceptual learning. The process of generating scientific explanations requires analysis and reflection of current models and theories, thereby influencing conceptual understanding. In the process of formulating explanations, the generation of inferences relies on the application of skills associated with scientific behaviour (e.g., analytical reasoning and critical thinking). In this study we explore student engagement in the process of generating authentic explanations, *by and for themselves,* through a General Chemistry in-class activity that prompts them to self-explain.

Self-explaining is a domain-independent learning strategy whose effect has been widely replicated, and it refers to student's generation of inferences of causality (Siegler and Lin, 2009). Its effectiveness compared to other learning activities is explained by the passive-active-constructive-interactive, ICAP, theoretical framework first introduced by Chi (2009) and to which we ascribe in our work. In the ICAP framework, a learning activity is characterized by observable, overt actions undertaken by the learner. These overt actions are assumed to be an adequate proxy for the covert cognitive processes that support the manifested behaviours. Although, understandably specific overt behaviours are not a requisite for the learners to resort to specific cognitive processes, Chi (2009) argues that learners are more likely to engage in certain cognitive processes when they display certain overt actions. These actions (and their products) are in turn manipulable by the instructor or researcher and allow their use as evidence of learning, that is, they can be assessed, coded, and analysed (Fonseca and Chi, 2010). It is this quality what renders possible the classification of learning activities from the learner's perspective. Passive, active, constructive, and interactive refer to and are defined by what the learner does when in contact with the learning materials, the overt behaviours. A passive learning activity is characterised by lack of actions on the learner's part. For example, listening to a lecture or reading a text without engaging in any additional activity such as note taking or underlining. In an active learning activity, the learner does something physical in support of learning. Highlighting while reading falls under the active category as does repeating rules that describe periodic trends to increase retention. The production of "*some additional output that contains information beyond that provided in the original material*" (Fonseca and Chi, 2010, p. 301) is the hallmark of constructive learning activities. Constructing a concept map and comparing and contrasting chemical reactivity are two examples. An interactive learning activity is one in which the learner establishes a dialogue with a peer, expert, or intelligent system that includes substantive contributions from all parts and where no part's contribution is ignored. Therefore, if one interlocutor dominates the interaction or participants simply take turns speaking and ignore each other's contributions, the activity is not considered interactive. Thorough analysis of published research has contributed evidence supporting the hypothesis that learning activities produce greater learning outcomes when they are interactive compared to constructive (Chi, 2009). Likewise, constructive activities are more efficient than active and active than passive.

Self-explaining is a constructive learning activity. It requires the learner to elaborate upon the presented information by relating it to prior knowledge and integrating with other pieces of information to generate inferences beyond the learning materials. Therefore, individuals build new knowledge as they uniquely appraise their own mental model during the process of solving a given task and elaborate their self-explanations—the outcomes of self-explaining (Chi, 2000).

Despite the prolific research literature on self-explaining in specialised journals, very little research has appeared in publications that are typically within the scope of chemistry educators (Villalta-Cerdas and Sandi-Urena, 2013). This single fact may account for the widespread absence of self-explaining in chemistry instruction in contrast with the prevalence of approaches that teach chemistry as a collection of facts, which Schwab (1962) referred to as rhetoric of conclusions. Evidently, this disconnect is not exclusive to self-explaining or chemistry education. It is part of a bigger picture where *"the research communities that study and enact change are largely isolated from one another"* (Henderson et al., 2011).

We identified such a void in domain-specific research pertaining to chemical education (Villalta-Cerdas and Sandi-Urena, 2013): Only two articles have explored self-explaining in chemistry and both addressed computer-assisted learning (Crippen and Earl, 2004; Crippen and Earl, 2007). In addition, even when focused on STEM knowledge domains, research has rarely focused on STEM majors (Villalta-Cerdas and Sandi-Urena, 2013). To date, the research has been largely theoretical in nature and not applied, and it has been conducted in educational research *laboratory* settings (e.g., Chi et al., 1989; Bielaczyc et al., 1995; Schworm and Renkl, 2006; Gadgil et al., 2012; Villalta-Cerdas and Sandi-Urena, 2013). In this sense, a laboratory is a space where individuals are abstracted from their natural learning environment and function as study participants, not necessarily as students. Although this trend is changing (Villalta-Cerdas and Sandi-Urena, 2013), the need for applied research in naturalistic classroom environments persists in order to gather ecological evidence to support novel pedagogical strategies.

The 2013 National Survey of Student Engagement (NSSE) report showed that half of the respondents who majored in physical sciences, math, and computer science, *never* or only *sometimes* "prepared for exams discussing or working through course material with other students" (NSSE, 2013). Likewise, preliminary results of study habits at our own institution suggest that only a small segment of General Chemistry 1 students engage in group study outside the classroom. In the fall of 2013 only 13% reported to study in a group for up to one quarter of their study time. For the remaining students nearly all of their unsupervised learning occurred individually. Although we strongly support interactive learning in its multiple expressions, it seems reasonable to think that students do not have the opportunities to maximise use of collaborative skills they may learn in the classroom. On the other hand, in-class constructive learning activities can reinforce learning strategies students can eventually use spontaneously while studying individually. Added to the robust research evidence that supports the self-explaining effect (Chi et al., 1989; Villalta-Cerdas and Sandi-Urena, 2013), our interest in this particular constructive learning activity stems from its being an essential and desirable scientific competence (National Research Council, 2013).

In our research group, we endeavour to develop studies that address the void in domain-specific, self-explaining research pertaining to chemical education. As an initial approach, we are investigating whether framing of learning tasks may modify student self-explaining behaviour in large-enrolment General Chemistry courses. Ultimately, we are interested in assessing the impact that modifying self-explaining practices may have on conceptual learning in chemistry.

Research Goals

This study is embedded in a larger research program that focuses on the following: (a) ways to promote self-explaining during chemistry instruction and (b) the assessment of how different levels of self-explaining influence learning of specific chemistry content. This investigation of the self-explaining effect is different from other work in the field in the following regards: (a) Participants take part in this study in their normal *student* function; therefore we refer to them exclusively as students to differentiate from laboratory approaches; (b) We use a real problem situation that resembles the process of *doing* science to evoke self-explaining; (c) Prompting to self-explain occurs at various demand levels instead of relying on spontaneous production of self-explanations; (d) We focus on conceptual understanding of chemistry (as assessed by a transfer task) rather than learning declarative or procedural knowledge (e.g., using worked-out examples, reviewing an expert explanation); (e) Data collection happens within the undisturbed ecology of a college level large-enrolment chemistry classroom.

Herein we report the findings from a study within this research program that specifically addressed the following research question:

> Do tasks that require different levels of self-explaining effectively induce observable, categorical differences in self-explaining behaviour in the context of a General Chemistry classroom?

Our stance is that an association between self-explaining tasks and overt self-explaining behaviour strongly suggests that appropriate instruction in the naturalistic classroom setting can effectively modify self-explaining practices.

Methodology

Study design

The study followed a multi-condition comparison design that gathered student generated textual data during a learning event. We designed and implemented a pilot study to test logistics and gain insight about the efficacy of materials and procedures and the data analysis (Ross, 2005; van Teijlingen and Hundley, 2001).

We developed the materials specifically for use in this study (Appendix 1). The domain includes entropy and the Second Law of Thermodynamics, which we treat as individual *knowledge components* (VanLehn, 2006) for the purpose of this work.

The naturalistic classroom setting we chose to use carried the intricate complexities of a live learning environment that, in chemical terms, we liken to a complex matrix. We argue this complexity translates into enhanced ecological validity (Brewer, 2000). The complex matrix presents a series of challenges in the design, data collection and analysis, and condition comparisons. For instance, a simple comparison between self-explaining and non-self-explaining conditions was not warranted in this setting. Much like use of standard additions in the chemical analysis lab counteracts the effects of a complex matrix, we believe our approach isolates the effect of self-explaining in the complexity of the study setting. We created four conditions (Table 1), each calibrated to promote different levels of self-explaining engagement. We adhere to Chi's (2011) conceptualization of engagement as *that* what learners do with learning materials. We understand self-explaining engagement as the level of purposeful allocation of cognitive resources and strategies, time, and effort to generate explanations by and for oneself to address a particular phenomenon. We gradually increased the self-explaining demand for the conditions by modifying the prompts describing the task. We based the calibration of the conditions on literature reports (Fonseca and Chi, 2010), especially multi-condition comparison studies (Siegler and Lin, 2009) and tested them through cognitive interviews as described below. The fundamental assumption was that since the matrix was the same for all conditions, variations in the outcome or dependent variable—self-explaining behaviour—would be associated with condition membership.

The learning event consisted of a textbook passage with a general description of the Second Law of Thermodynamics and common to all the participants. A self-explaining task, SE-Task, followed this passage. There were four different SE-Tasks, each defining one of the study conditions described in Table 1. Students completed the learning event within fifteen minutes.

Table 1 Description of self-explaining tasks (SE-Task)

SE-Task	Description
SEA	Explaining own answer.
EADA	Considering others' answers and explaining one's agreement/disagreement.
SEO	Explaining answer for others to use in their studying.
SEIA	Explaining others' incorrect answer.

Unlike most of the research in the field, this learning task does not focus on advancing procedural knowledge through self-explanation of *examples* (complete or incomplete worked-out problems[1]) (e.g. Atkinson et al., 2003; Schworm and Renkl, 2006) or conceptual understanding through self-explanation of expository text, such as explaining the logic underlying statements in textbooks (e.g., Chi et al., 1994; Ainsworth and Loizou, 2003; Butcher, 2006; Ainsworth and Burcham, 2007). Neither did we utilise a conventional training study design to show or demonstrate a skill or strategy that students would perform at a later stage (e.g., Bielaczyc et al., 1995; Schworm and Renkl, 2007). Our purpose was to create an experience that was closer to *doing science* than to the procedural aspects of solving exercises or learning *about* science (Chamizo, 2012; Talanquer and Pollard, 2010).

We presented an otherwise familiar phenomenon to the students (water freezes spontaneously below 0 ºC) and a fact that would potentially induce cognitive imbalance (the change in entropy for the system in this process is negative) to prompt them to self-explain. Although not instructed to do so, we anticipated that students would be prone to use the concept introduced in the same document—Second Law of Thermodynamics—in their self-explanations. We intended to affect the engagement in self-explaining by creating different levels of encouragement to explain (Table 1) (Siegler, 2002). For this purpose, we combined two mechanisms: the effect of social engagement (e.g. explaining for others) and the depth of explaining (i.e. to explain answers that are described as correct or incorrect; Siegler, 2002).

It is reasonable to consider that the cognitive processes associated with self-explaining may take place covertly. However, our premise is that students are more likely to engage in self-explaining when an overt behaviour is required (Fonseca and Chi, 2010). Therefore, we collected written responses from students as indicators of their self-explaining behaviour. Although informative, think-aloud protocols were not an option given our desire to use large cohorts and to gather data in the most naturalistic environment possible.

We reviewed the materials after the pilot study (Table 2) and no major changes resulted from this process. We also conducted cognitive interview checks to assess interpretability of the materials. The protocol for the cognitive interviews is included in Appendix 2. Our interviewees were two second-year chemistry students who had taken General Chemistry 2 within the past year. They were recruited from a pool of chemistry undergraduate researchers, and they received no compensation for their interviews. Interviews lasted around 35 minutes, in which students completed the tasks and then discussed them in depth with the interviewer. This procedure provided evidence that supported interpretability and face validity in general. In addition, we consulted and held separate meetings with three doctoral candidates in chemical education and two experienced general chemistry instructors at the authors' home institution, who offered general advice and completed assessment rubrics to evaluate content validity of the materials. Finally, two chemical education researchers, who were external to the authors' institution and not associated with the research study, assessed the content and construct validity of the materials independently and provided feedback. No modifications were necessary upon the assessment by experts.

Context and participants

This research used a naturalistic setting and gathered data from students enrolled in General Chemistry 2 at a large, urban, public, research university in the US serving over 31 000 undergraduate students. Diverse ethnic minority students make up 39% of the undergraduate student body. Typically majors in General Chemistry 2 are distributed as follows: Pre-professional (pre-Medicine, pre-Pharmacy and Health Sciences), 61%; Chemistry, 6%; other sciences (Physics, Biology, Geology, etc.) or Math, 23%; Engineering, 8%. The remaining students are non-science/non-engineering majors. The tasks in this study were embedded within the normal requirements of the course; therefore they were simply part of normal assignments from the students' perspectives. The activity was conducted before students were formally introduced to the chemistry concepts (i.e., entropy and the Second Law of Thermodynamics). Grading guidelines for this activity were the same as those for similar assignments throughout the semester. Credit was received for the satisfactory completion of the activity and not based on performance. This study only used data from students who had previously granted informed consent. The gender distribution in the main study was representative of the university demographics (42% males, 58% females).

Data collection.

Data gathering occurred during the tenth week of the course and came from two independent cohorts of students enrolled in different semesters (Table 2). In the pilot study, we distributed alternate forms of the four SE-Tasks (Table 1) to participants. In this pseudo-randomised procedure the probability of assignment to a given SE-Task was not independent for each individual. To meet conditions for true randomisation for the main study (i.e. same probability of being assigned to any of the four conditions), we used random number generation (Microsoft Excel, 2010) to assign students to the SE-Tasks (Shadish et al., 2002; Ravid, 2010). The number of students in each condition was: SEA, 29; EADA, 31; SEO, 35; SEIA, 33.

Materials were printed, used individually without student interactions, administered during regular class schedule, and timed. Written explanations were collected, photocopied, assigned an alphanumeric code (student identifiers were removed from the photocopied materials), and later transcribed to electronic support. File names used the alphanumeric code. Drawings, diagrams, and equations were scanned and integrated to the corresponding electronic files.

The Hawthorne Effect describes how in behavioural studies participants may behave in ways different from the normality if they realise they are being observed (Franke and Kaul, 1978; Jones, 1992). Therefore, we took measures to minimise any potential risk of evoking such behaviours. This included following procedures such as distribution of materials and delivery of instructions that were not different from procedures typically used for other in-class assignments. We assumed familiarity with these procedures prevented predisposition of any kind.

Table 2 Data Collection by Study Phase

Study Phase	Sample size (n)	Dataset
Pilot Study	103	Fall 2011
Main Study	134	Fall 2012

Data analysis

The analytical methods we describe here are the final product of several iterations of the analysis of the pilot and main study datasets. For sake of simplicity and space, we omit the lengthy process of refinement of methods.

Textual analysis of learning event data. The learning event produced written explanations, which we refer to as *responses*. In preparation for textual analysis, the prompts were removed so that coders had access to the responses only. Unavoidably, in many cases the structure of the response could be associated with a specific prompt.

We used the sentences as constructed by the students as unit of response segmentation. For this purpose, independently of their syntactic accuracy, the use of a period indicated the closing of a sentence. Although the systematic analysis required segmentation, it is important to underscore that we did not intend to de-contextualise the analysis: We considered each unit of response segmentation in the light of the entire response, i.e., explanation. For the pilot study, a single researcher coded the textual data (103 responses) using a sequence of coding schemes reported in the literature (Durst, 1987; McNamara, 2004; Best, et al., 2005; McNamara and Magliano, 2009; Ford and Wargo, 2012). This preliminary analysis allowed us to ascertain the feasibility of the study; however, as an analytical tool, it was too involved and impractical. For the main study, we streamlined coding to a single scheme that was more robust and easier to apply to large cohorts. This scheme preserved fundamental codes from the literature (McNamara, 2004; McNamara and Magliano, 2009) that we modified slightly in consideration of emergent categories and subcategories and refined it through consensus coding of a subset of 50 responses by three coders. Table 3 shows the final coding scheme and a brief description of each code type. Codes BI, DI, E, and P in Table 3 derived from research reports (McNamara, 2004; McNamara and Magliano, 2009). During coding we identified two types of paraphrasing: repetition of information from the learning materials and repetition of information already in the response itself. From the total database, only three sentences were unclassifiable, U, and given the small count we dropped them from further analysis as we did with the statements deemed non-relevant, NR, since they did not provide information regarding the sophistication of the explanations (e.g., *"They are on the right track but just need to pushed* [*sic*] *in the right direction"*).

Table 3 Final coding scheme for written responses

Code Type	Description	Example from responses
BI-Bridging Inference	Relational inference linking the problem (i.e., water freezing below 0 ºC) with entropy change, and/or Second Law of Thermodynamics.	Even though ΔS_sys<0, the ΔS_univ is still positive because when the water freezes the surroundings have a sharp increase in entropy (ΔS_surr>0).
DI-Deductive Inference	Inference that uses specific content knowledge (i.e., water freezing below 0 ºC, entropy, or Second Law of Thermodynamics), but does not link to other information.	Just because the ΔS_system is negative doesn't mean that the process must all be negative.
E-Elaboration	Use of information not provided in the materials	When water begins to freeze at 0˚C, water (unlike other liquid) expands which make this less dense than when water is above 0˚C.

P-Paraphrasing	(Pa) Recount of entropy concept, or Second Law of Thermodynamics. (Pb) Repetition of previously used information within response.	If the process is indeed spontaneous, that means the ΔS_univ must be positive.
U-Unclassifiable	Statement of concepts without drawing relational inference.	Plus, although the change in entropy of the surrounding may change some in a resulting reaction that leaves ΔS_univ negative, there is still $H_2O(g)$ in the air (Earth's atmosphere). (R48, F12)
NR-Non-relevant	Comments and observations unrelated to the task.	He did not look at the big picture.

Once we had established the coding scheme, the same coders analysed 50 responses separately. Subsequently, these coded responses were team reviewed and disagreements were discussed and resolved. One researcher coded the remaining responses; the other two coders verified a different subset (42 responses each) and solved any discrepancies with the main coder. We assigned an individual code to each sentence and then tallied the codes by response. The ratio of frequency of a given code type count (n) to the total sentences in the response (N)—hereafter the code-ratio (n/N)—became the observed variable for the subsequent Latent Profile Analysis (described below). From the main study dataset we eliminated six responses that where unintelligible and 128 remained. Once the main study data was coded, we re-coded the dataset from the pilot study to investigate other potential changes to the coding scheme. Two coders worked independently on a subset of the dataset and later discussed the coding. All discrepancies were resolved, and no changes were made to the coding system. Figure 1 shows the coding of an example response.

Response	Textual Analysis						
	Value	BI	DI	E	P	U	NR
The change in entropy for the system is less than zero therefore the ΔS_surr must have a larger increase of entropy than the negative decrease of the entropy of the system for the process to be spontaneous, meaning ΔS_univ will still be greater than 0. **(DI)** When water freezes the system loses entropy, but the outside surroundings gain more entropy than what was lost by the system. **(BI)**	Count (n)	1	1	0	0	0	0
	Code-ratio (n/N)	0.50	0.50	0	0	0	0

Fig. 1 Coding example.

Latent profile analysis, LPA. LPA is a model-based statistical technique to find profile classes in continuous data (Pastor et al., 2007). It is a latent variable model, where non-observable latent constructs are inferred through mathematical modelling using observed variables (Collins and Lanza, 2010). LPA assumes that different profiles can be explained by the existence of frequency patterns in the observed variables (Marsh et al., 2009; Pastor et al., 2007). During the analysis, several profile-model solutions are generated and compared. The comparison is evaluated to select the best fitting model for the data. A number of techniques have been devised to guide selection of the best model fit (e.g., Model based hypothesis tests, Log likelihood, Akaike Information Criterion, Bayesian Information Criterion, Sample-size adjusted Bayesian Information Criterion, Entropy value; Collins and Lanza, 2010; Marsh et al., 2009; Pastor et al., 2007)[2].

Although manual inspection of data could result in the identification of patterns of response, the process would be limited to small datasets and be tedious, time-consuming and seemingly prone to researcher bias. Moreover, traits could be overlooked easily, and the process would be inherently unreliable. In our study, we used LPA to elicit otherwise undetectable trends and to minimise bias in the categorisation of student responses in explanatory behaviours. We performed LPA using the code-ratios from the textual analysis as the observed variables (i.e., four observed variables). The output of LPA was the categorisation of students into distinct profiles based on the nature of their explanations, the self-explaining profiles (SE-Profiles). LPA was performed using MPlus Version 6 (Muthén and Muthén, 2010). Figure 1 shows a coded response along with the corresponding code ratios used as observed variables in the latent profile analysis.

Analysis of the association between self-explaining tasks and self-explaining profiles. We used Chi-square tests to determine the association between self-explaining profile membership, SE-Profile, and self-explaining task, SE-Task. For the interpretation of the Chi-square test results we selected a 95% confidence level. We used IBM SPSS Statistics (Version 21.0.0.0) for the Chi-square tests.

Results and Discussion

Pilot study

The purpose of the pilot study in the initial stage of this project was to test the study design and instruments and to identify potential methodological gaps. In summary, the pilot test results suggested that tasks of different self-explaining demand elicited different self-explaining behaviours (Villalta-Cerdas, 2014). Although this evidence supports the association between the SE-Profiles and SE-Tasks, the statistical analysis was not conclusive. The pilot test supported the appropriateness of the study design, materials, and analysis procedures; it did not reveal deficiencies that required modifications prior to the implementation of the main study. Nonetheless, to enhance the design we decided to utilise true randomisation for the main study instead of pseudo-randomisation.

Main study

Code type distribution. The total count of codes showed that the deductive inference code, DI, had the highest frequency of all (Table 4). The combined count of bridging inference code, BI, and DI was 169 (44% of the total count); thereby suggesting that the generation of inferences was a considerable component of the responses. Research findings have shown that *"in the absence of specific instructions or supports, most students either do not generate self-explanations or generate superficial ones only"* (Woloshyn and Gallagher, 2009). Thus, this observation, in and of itself, suggests the tasks effectively elicited self-explaining behaviour.

In the case of the codes for elaboration, E, and paraphrasing, P, their abundance in the students' responses may reflect what Taber (2000) described as a social imperative to produce an answer in acknowledgement to a question, in this case, the SE-Task prompt. These two codes, E and P, are associated with less sophisticated explanatory behaviours as they reflect recounting of information rather than generation of causal inferences. Moreover, when students are continually exposed to instruction as rhetoric of conclusions (Schwab, 1962), one could imagine that paraphrasing may become a habitual substitute for explaining. Therefore, in the case of paraphrasing, it might be that students intended to explain but lacked the ability to construct responses beyond re-statement of information. Undeniably, some students may default to paraphrasing even when prompted otherwise.

For our research purposes, the codes unclassifiable, U, and non-relevant, NR, did not contribute valuable insight to elucidate the explanatory behaviour of the students. Therefore, we did not consider them in subsequent analyses.

Table 4 Main study learning event code type distribution by SE-Task

Code Type	Total count	SE-Task			
		%SEA	%SEIA	%EADA	%SEO
BI	47	43	21	9	28
DI	122	19	33	23	25
E	88	22	27	32	19
P	100	23	19	28	30
U	3	33	33	0	33
NR	28	7	21	64	7

χ^2 (9, N = 357) = 22.50, p < .05, codes U and NR excluded.

In a first analysis, we studied the association between the code type (e.g. bridging inference, BI, deductive inference, DI, etc.) and the self-explaining task, SE-Task (Table 4). The Chi-square test showed a statistically significant association between the code type and the SE-Task at a 95% confidence level, χ^2 (9, N = 357) = 22.50, p < .05. Cells shaded dark grey in Table 4 indicate the highest occurrence for each code type and the overall trend in the association. In the case of the bridging inference code, BI, the highest percentage of occurrences originated from the self-explaining-own-answer task, SEA, which effectively prompted students to connect chemistry concepts (i.e., entropy and the Second Law of Thermodynamics) in their effort to make sense of the phenomenon. For deductive inference, DI, the predominant source was the self-explaining-incorrect-answer task, SEIA. Encouraging students to explain the possible reasoning that led their peers to incorrect solutions generated more deductive inferences. The EADA (self-explain-agreement/disagreement) and SEO (self-explain-for-others) tasks had moderately high percentages for the DI code, thus students in these conditions engaged in the generation of deductive inferences as well. In the case of the elaboration code, E, the results showed the highest percentage in the self-explain-agreement/disagreement task, EADA. This SE-Task seemed to favour a more summative approach to self-explaining where participants brought in external information that was not provided in the materials. Despite their elaborative effort, students did not use the external information to draw deductions or bridge with other concepts; instead, they essentially recounted it in their responses. Lastly, the paraphrasing code, P, showed similar high percentages for two of the SE-Tasks: EADA and SEO. Again, we maintain this behaviour reflected the social imperative to answer a question as described by Taber (2000) even when students operated under the illusion of producing an explanation.

The code type distribution addressed the research question guiding this work: Do tasks of different self-explaining demand induce observable, categorical differences in self-explaining behaviour? Evidence supports an association between the code types in the student responses and the SE-task prompts assigned to them. This association suggests that the prompts, which we designed with differential self-explaining demand, effectively produced an observable effect on the students' behaviour as they composed their written responses. The variance of code types across SE-Tasks is indicative of the effect of individuals' characteristics. That is, students within a SE-Task still produced explanations of different sophistication. This variability is congruent with reports that have associated quantity and quality of explanations with intrinsic properties of students (Roy and Chi, 2005). This occurrence underscores the significance of randomisation of students in the conditions since otherwise the effect of task membership may be obscured by this natural variability.

Latent profile analysis: Self-explaining profiles. The results in Table 4 show quantifiable evidence for the differences in the total number of code types per SE-Task. This analysis focused at the variable-oriented level (i.e., using the code types as observed variables) and not at the person-oriented level (i.e., using the student's behaviour as observed variable). In our attempt to identify categorical explanatory behaviours at the student level, we advanced our interpretation by performing a person-oriented approach. To this end, we used latent profile analysis (LPA), a mixture model that seeks to find qualitative differences among participants based on observed variables of continuous nature (Ruscio and Ruscio, 2008). In our analysis, the code-ratios in each student's response functioned as observed variables.

To better contrast these approaches it is noteworthy to mention that the simpler association described above assumes the occurrence of all codes as independent when single students might have contributed more than one code (actually, 79% did). In addition, those who contributed more than one code did not necessarily contribute the same codes; in other words, multiple patterns of response were possible. The analysis at the person-oriented level takes these considerations into account and focuses on each individual's behaviour by integrating the number and type of codes into the categorisation of patterns. This transformative analysis allowed us to investigate whether the behaviours, and not only the codes, were linked to the SE-tasks.

Using Latent Profile Analysis (LPA) we identified patterns in code-ratios (i.e., number of code type divided by total codes in response) in student responses. These analyses required the selection of the best model for the data. We include the selection procedure and other pertinent data handling in Appendix 3. The analysis and interpretation of the models led us to select the seven-profile model solution for the main study data.

Table 5 shows the profiles in the seven-profile model solution along the number of students in each profile and the respective mean values for the four code-ratios. In the case of Profiles 1-3 and 5-7, the mean code-ratios within profiles showed a single predominant value (dark grey). Therefore, self-explaining within each of these six profiles was strongly characterised by the single class predominant code; that is, the pattern of behaviour of members within each of these profiles was homogeneous. Profile separation refers to the uniqueness of each profile; in our case that implies comparison of predominant mean code-ratio *between* profiles. Ideally, all profiles would have a maximum mean code-ratio for different codes; however, in the case at hand, there were more profiles than mean code-ratios (or code-types), which unavoidably led to profiles sharing a maximum code-ratio. In turn, this led to the merging of profiles.

Although profiles 1-3 have each a single most prevalent code-ratio (Table 5), it is not unique to each profile but the same for all three of them; the separation is not strong. Hence we combined these profiles into a single self-explaining profile (SE-Profile). Members of this merged profile (n=25) are characterised by responses composed mainly of bridging inference codes, BI, (>50% of response); consequently, we described this SE-profile as bridging inferential.

Profile 4 is non-homogenous: There is no single code type that characterises membership in this profile. Quite the contrary, it is the multiplicity in the nature of their behaviour that identifies members in this group; we described this SE-profile as mixed behaviour. Although not homogenous, this group is clearly separated from the others. Emergence of this profile is an example of the power of statistical tools such as LPA. An analysis based solely on the number of codes would have masked the behaviour of students in this SE-Profile who used all four explanatory codes in similar proportions.

Profiles 5, 6, and 7 have a single and unique predominant code-ratio and are homogenous and well separated from all other profiles. We assigned labels to these profiles in accordance with the code that predominates in each case. Therefore, Profile 5 became deductive inferential, Profile 6 became elaborative, and Profile 7 became summative (Table 5).

Table 5 Code-ratios and SE-profile descriptors for seven-profile model solution

Profiles	n	Mean code-ratio				SE-Profile descriptor
		BI	DI	E	P	
Profile 1	6	1.00*	0.00*	0.00*	0.00*	
Profile 2	4	0.70*	0.17*	0.05	0.08	Bridging Inferential
Profile 3	15	0.50*	0.22*	0.13*	0.15*	
Profile 4	12	0.29*	0.27*	0.19*	0.25*	Mixed-behaviour
Profile 5	20	0.00*	0.95*	0.03	0.02	Deductive Inferential
Profile 6	24	0.00*	0.23*	0.73*	0.05*	Elaborative
Profile 7	47	0.00*	0.35*	0.12*	0.53*	Summative

* $p < .05$.

SE-Profile and SE-Task association analysis. Once we established the student behaviours in terms of the SE-Profiles, we analysed the association between SE-Profiles and SE-Tasks. Table 6 shows the resulting cross-tabulation. The Chi-square test showed a significant association between the SE-Profile and SE-Task at a 95% confidence level, χ^2 (12, N = 128) = 22.75, p < .05. Inspection of Table 6 shows a trend in the percentage distribution of SE-Profiles across the SE-Tasks (dark and light grey shaded cells) that could explain this relationship.

Table 6 Percentage Distribution of SE-Profile across SE-Task

SE-Profile	n	SE-Task			
		%SEA	%SEO	%SEIA	%EADA
Bridging Inferential	25	36	32	20	12
Mixed-behaviour	12	50	25	25	-
Deductive Inferential	20	10	35	40	15
Elaborative	24	17	12	29	42
Summative	47	17	30	21	32

χ^2 (12, N = 128) = 22.75, p < .05.

In the trend in Table 6, the SEA task—self-explain-own-answer—has the highest percentage of students in the SE-Profiles associated with the more analytic self-explaining behaviours (i.e., bridging inferential and mixed-behaviour). Thus, more students in this SE-Task engaged in generating inferences and connecting ideas via more complex explanatory behaviours. Conversely, SEIA and EADA (explain-incorrect-answer and explain-agreement/disagreement with others, respectively) showed a higher percentage of students in the less analytical self-explaining behaviours (i.e., elaborative and summative).

We hypothesise that in the case of SEIA and EADA the constraint set for the students might have acted as an inhibitor of self-explaining. When presented with the solution, those participants in agreement may default to restating the solution while those in disagreement may simply rephrase it in opposite sense. We propose that by constraining students to agreeing/disagreeing we induced knowledge-telling episodes (i.e., unelaborated summaries and paraphrases) over knowledge-building episodes (integration of concepts and generation of inferences; Chi, 2009). Our original assumption was that considering solutions different from the student's own answer could engage students in a deeper reflection and a stronger commitment to self-explain. The nature of the task and the dichotomous nature of the answer (one thing or the other) might have obscured the original intended effect for this particular General Chemistry 2 sample.

In the case of SEA (self-explain-own-answer) and SEO (self-explain-for-others) tasks, we kept the task unconstrained for students. The fact that self-explaining directed to others was not more conducive to sophisticated behaviours is not entirely surprising. Roscoe and Chi (2008) compared self-explaining and *other*-directed explaining with students interacting with a tutor and found that the former was better, even when the tutor was virtual (i.e., computer generated). One possible explanation is that explaining to oneself focuses on repairing what one does not understand, without the distraction of focusing on others (Chi, 2009). However, it must be stressed that one cannot rely on a general description of an activity to judge its quality and outcomes as a learning experience (Chi, 2009). Otherwise neglected aspects may prove to be fundamental warrants for caution when generalising findings. For instance, Siegler (2002) observed that creating a social dimension by telling students to explain for others acted as an incentive to explain. This may seem sufficiently similar to our conditions as to try to extend their findings to ours, however, in that study, researchers utilised a think-aloud protocol in a laboratory setting where children interacted with an adult researcher. In our case, an ecologically natural learning environment at college level, the *other* was an anonymous peer.

Regardless of the SE-Task a number of students defaulted to paraphrasing which is evidenced by the distribution of students in the summative SE-Profile across all SE-Tasks (Table 6). We contend that this behaviour may be more attributable to long periods of conditioning supported by prior class norms (e.g. the classroom game; Lemke, 1990) than indicative of task effectiveness or lack thereof. Kohn (2004) has pointed out that students may become accustomed to and comfortable with learning environments in which they are not expected to actively engage with intellectual challenges. It may take some effort to persuade students to act differently. Far from discouraging, we deem understanding of this occurrence informative and constructive. Students are not blank slates; they bring their personal history, prior knowledge, attitudes, skills and strategies, etc. to the classroom, and naturalistic approaches to research intend to explore behaviour in the complex ecology of classroom learning. This finding highlights that students, in addition to not spontaneously engaging in explaining, may resist explaining when prompted. That is, student behaviour is not determined by the prompt provided—a stance that would evoke a behavioural approach to learning. It is not surprising that the same instructional activity may trigger varying self-explaining responses in different individuals. This individual-task interaction is consistent with the *"subtle interaction between a task and the individual struggling to find an appropriate answer or solution"* that Bodner and collaborators identified in problem solving research (2003). Acknowledging the effect of these individual differences and their interactions with the nature of the prompts is an important step in advancing instructional design.

There are no quick fixes in education, and the resistance or activation barrier associated with self-explaining will not be resolved with a single instantiation. Our emphasis is on the fact that a considerable proportion of students did engage in self-explaining upon prompting and that the sophistication of this engagement was, to some extent, tuneable by the design of the task.

Limitations

Some limitations of this study are worth mentioning and findings must be interpreted within these limitations. First and foremost, although we randomised condition assignment within our sample, this was a convenience sample. Students in this study were enrolled in the General Chemistry 2 section taught by one of the authors. Although basic demographic indicators are not significantly different from the rest of the General Chemistry cohort, we have no way to elucidate whether latent factors might have influenced students' choice of this particular section (e.g. instructor reputation, schedule convenience).

We removed prompts from the responses for the textual analysis; however, coders could infer the corresponding self-explaining condition from the structure of the responses. This limitation creates the potential for coder bias where coders may be prone to make code assignments based on the condition and not strictly on the analysis of the responses. Although not very practical, one may choose to assign the coding process to individuals who are not involved with the research study.

Despite the clear value of research in naturalistic environments, there are concomitant limitations. Understandably, unlike the case of tightly controlled experimental studies, in a natural setting control of exogenous variables is not possible and their effects unpredictable. Our randomisation within the sample contributed to minimize this limitation. Another possible concern related to the study design may be participants' impulse to behave as socially desirable or to adjust their behaviour some other way when they are under the impression of being observed. To minimize this effect, for this study we used procedures consistent with in-class assignment norms. Our assumption is that the sense of familiarity with the procedures prevented predisposition of any kind. From the students' perspective, this learning event was not different from other learning experiences in class, that is, there were no cues to interpret it as research. There were no unfamiliar individuals in the lecture hall while data gathering. Use of a convenience sample actually allowed us to frame the learning event in such a natural way.

Conclusions and implications

The ability to generate explanations of scientific phenomena is an essential learning outcome for all students (National Research Council, 2013). This study intended to gather evidence to establish whether tasks of different self-explaining demand induce observable, categorical differences in self-explaining behaviour. Students' self-explaining behaviours were categorised via analysis of textual data of their responses, followed by data transformation and modelling using Latent Profile Analysis. Data was reduced to five self-explanatory behaviours that proved to be associated with the tasks we created as study conditions.

Independently of the rationale one may generate to explain the behaviour of this particular cohort of students, results in this study reveal an association between the way the tasks are framed for students and their engagement in producing self-explanations of different sophistication levels. Caution is warranted in that we do not intend to be prescriptive and describe the type of prompts that should be used in chemistry classrooms to engage students in effective self-explaining. Such a goal would imply an over-simplistic, reductive view of the complexity of learning environments. Those involved in instructional design should understand this complexity and the effect of contextual and other situational factors. As cited by O'Donnell (2008), Berman and McLaughlin observed: *"The bridge between a promising idea and the impact on students is implementation, but innovations are seldom implemented as intended."*

We hypothesised that an association between self-explaining tasks and overt self-explaining behaviour would strongly suggest that instruction in the naturalistic classroom setting can effectively modify self-explaining practices. In other words the qualities of student responses could be modulated through the design of learning experiences in tune with the instructor's goals (Chi, 2009). Considering the different responses from students, a varied array of prompts may be more effective than searching for a single, one-size-fits-all type of prompt. Chemistry educators may use this and other supporting evidence to decide whether to integrate self-explaining activities to the repertoire of their instructional design. Identifying the evidence-based active ingredients that promote learning in natural learning environments may lower the activation barrier associated with undertaking innovations. This is especially true for novice instructors, who may find integrating such strategies into their instructional design less intimidating and invasive than relinquishing control to a pre-packaged pedagogical model.

Our work did not use strategy training or direct-instruction, that is, we did not teach students to self-explain to later test their adherence to a particular behaviour that may vanish once the stimulus is removed. Self-explaining behaviour was effectively elicited by the learning event. Thus, we put forth cultivating constructive learning strategies such as self-explaining as components of well-designed instruction. Deeper engagement in self-explaining may become habitual upon practice and hopefully develop into the new norm in students' relationship with chemistry knowledge. This stance is consistent with related research in the field that suggests *"meaningful learning may help students progress from a stage in which re-description and functional explanations are dominant, to a phase in which connections between parts are emphasised, to a point in which cause-effect relationships are frequently used as the basis for explanations"* (Talanquer, 2010).

Several future lines of work arise from findings in this study. Whereas in this study we focused on learning strategies as the learning outcome of interest, currently we are engaged in the assessment of how different levels of self-explaining influence learning of specific chemistry content. In this work we observed variability in the sophistication of student responses to the same self-explaining task. Investigating what individual characteristics may be associated with this differential behaviour is another potential line of work. Likewise, we are interested in the investigation of change in self-explaining behaviour by using latent variable models on longitudinal data collected across multiple learning activities.

Notes and references

[a] *Department of Chemistry CHE205, University of South Florida, Tampa, Florida 33620, USA. Email: ssandi@usf.edu*

† [1]Hereafter we reserve the use of the term problem to novel situations for which one does not have a set of rules or a procedure to produce an answer (Wheatly, 1984).

[2]A detailed description of Latent Profile Analysis is outside the scope of this report. For detailed descriptions and exemplar applications see the following: Magidson, and Vermunt, 2004; Pastor et al., 2007; Marsh et al., 2009; Collins and Lanza, 2010; and Muthén and Muthén, 2010.

Electronic Supplementary Information (ESI) available: Research instruments, interview protocol, and Supplemental data analysis and results. See DOI: 10.1039/b000000x/

Ainsworth, S., and Burcham, S. (2007). The impact of text coherence on learning by self-explanation. *Learning and Instruction, 17*(3), 286-303.

Ainsworth, S., and Loizou, A. T. (2003). The effects of self-explaining when learning with text or diagrams. *Cognitive Science, 27*(4), 669-681.

Atkinson, R. K., Renkl, A., and Merrill, M. M. (2003). Transitioning from studying examples to solving problems: Effects of self-explanation prompts and fading worked-out steps. *Journal of Educational Psychology, 95*(4), 774-83.

Best, R. M., Rowe, M., Ozuru, Y., and McNamara, D. S. (2005). Deep-level comprehension of science texts: The role of the reader and the text. *Topics in Language Disorders, 25*(1), 65-83.

Bielaczyc, K., Pirolli, P. L., and Brown, A. L. (1995). Training in self-explanation and self-regulation strategies: Investigating the effects of knowledge acquisition activities on problem solving. *Cognition and Instruction, 13*(2), 221-252.

Bodner, G., and Herron, J. (2003). In Gilbert J., Jong O., Justi R., Treagust D. and Driel J.(Eds.), *Problem-solving in chemistry.* Springer Netherlands.

Butcher, K. R. (2006). Learning from text with diagrams: Promoting mental model development and inference generation. *Journal of Educational Psychology, 98*(1), 182-197.

Brewer, M. B. (2000). Research design and issues of validity. *Handbook of research methods in social and personality psychology*, 3-16.

Chamizo, J. A. (2012). Heuristic diagrams as a tool to teach history of science. *Science* and *Education*, *21*(5), 745-762.

Chi, M. T. H. (2000). Self-explaining expository texts: The dual processes of generating inferences and repairing mental models. *Advances in Instructional Psychology, 5*, 161-238.

Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. Topics in Cognitive Science, *1*(1), 73-105.

Chi, M. (2011). Differentiating four levels of engagement with learning materials: The ICAP hypothesis. *Presentation at the 19th International Conference on Computers in Education*. Chiang Mai, Thailand.

Chi, M. T., Bassok, M., Lewis, M. W., Reimann, P., and Glaser, R. (1989). Self-explanations: How students study and use examples in learning to solve problems. *Cognitive science*, *13*(2), 145-182.

Chi, M. T. H., De Leeuw, N., Chiu, M. H., and LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science, 18*(3), 439-477.

Collins, L. M., and Lanza, S. T. (2010). *Latent class and latent transition analysis: With applications in the social, behavioral, and health sciences* (Vol. 718). Wiley.

Crippen, K. J., and Earl, B. L. (2004). Considering the efficacy of web-based worked examples in introductory chemistry. *The Journal of Computers in Mathematics and Science Teaching, 23*(2), 151-167.

Crippen, K. J., and Earl, B. L. (2007). The impact of web-based worked examples and self-explanation on performance, problem solving, and self-efficacy. *Computers & Education, 49*(3), 809-821.

Deutsch, D. (2011). *The Beginning of Infinity: explanations that transform the world*. Penguin UK.

Durst, R. K. (1987). Cognitive and linguistic demands of analytic writing. *Research in the Teaching of English*, 347-376.

Ford, M. J., and Wargo, B. M. (2012). Dialogic framing of scientific content for conceptual and epistemic understanding. *Science Education*, *96*(3), 369-391.

Fonseca, B. A. and Chi, M. T. H., Instruction based on self-explanation. In: Mayer, R. E. and Alexander, P. A. (editors), *The Handbook of Research on Learning and Instruction*, (296-321), New York: Routledge, 2010.

Franke, R. H., and Kaul, J. D. (1978). The Hawthorne experiments: First statistical interpretation. *American sociological review*, 623-643.

Gadgil, S., Nokes-Malach, T. J., and Chi, M. T. H. (2012). Effectiveness of holistic mental model confrontation in driving conceptual change. *Learning and Instruction, 22*(1), 47-61.

Henderson, C., Beach, A. and Finkelstein, N. (2011). Facilitating change in undergraduate STEM instructional practices: An analytic review of the literature, *Journal of Research in Science Teaching*, *48*(8), 952-984, 2011.

IBM Corp. Released 2012. IBM SPSS Statistics for Windows, Version 21.0. Armonk, NY: IBM Corp.

Jones, S. R. (1992). Was there a Hawthorne effect?. *American Journal of Sociology*, 451-468.

Kohn, A. (2004). Challenging students-and how to have more of them. *Phi Delta Kappan*, *86*, 184-194. Available online (12-17-13): http://www.alfiekohn.org/teaching/challenging.htm.

Lemke, J. L. (1990). *Talking science: Language, learning, and values*. Ablex Publishing Corporation, 355 Chestnut Street, Norwood, NJ 07648 (hardback: ISBN-0-89391-565-3; paperback: ISBN-0-89391-566-1).

Magidson, J., and Vermunt, J. K. (2004). Latent class models. *The Sage handbook of quantitative methodology for the social sciences*, 175-198.

Marsh, H. W., Lüdtke, O., Trautwein, U., and Morin, A. J. (2009). Classical latent profile analysis of academic self-concept dimensions: Synergy of person-and variable-centered approaches to theoretical models of self-concept. *Structural Equation Modeling*, *16*(2), 191-225.

McNamara, D. S. (2004). SERT: Self-explanation reading training. *Discourse Processes: A Multidisciplinary Journal, 38*(1), 1-30.

McNamara, D. S., Magliano, J. P. (2009). 5 Self-Explanation and Metacognition. In Hacker, D. J., Dunlosky, J., & Graesser, A. C. (Eds.), *Handbook of metacognition in education* (pp. 60-81). Hoboken: Routledge.

Microsoft. (2010). Microsoft Excel [computer software]. Redmond, Washington: Microsoft.

Muthén, L. K., and Muthén, B. O. (1998-2010). Mplus User's Guide. Sixth Edition. Los Angeles, CA: Muthén and Muthén.

National Research Council (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: The National Academies Press.

National Research Council (2013). *Next Generation Science Standards: For States, By States*. Washington, DC: The National Academies Press.

National Survey Student Engagement, NSSE. (2013). Accessed on February 17th, 2014: *http://nsse.iub.edu/html/summary_tables.cfm.*

O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. *Review of Educational Research*, *78*(1), 33-84.

Pastor, D. A., Barron, K. E., Miller, B. J., and Davis, S. L. (2007). A latent profile analysis of college students' achievement goal orientation. *Contemporary Educational Psychology*, *32*(1), 8-47.

Ravid, R. (2010). *Practical statistics for educators*. Rowman & Littlefield Publishers.

Roscoe, R. D., and Chi, M. T. H. (2008). Tutor learning: The role of explaining and responding to questions. *Instructional Science, 36*(4), 321-350.

Ross, K. N. (2005). Educational Research: Some Basic concepts and Terminology. *International Institute for Educational Planning, UNESCO. Available: http://www. unesco. org/iiep*.

Roy, M., and Chi, M. T. (2005). The self-explanation principle in multimedia learning. In Mayer, R. E. (Ed.), *The Cambridge handbook of multimedia learning* (pp. 271-286). Cambridge, England: Cambridge University Press.

Ruscio, J., and Ruscio, J. (2008). Advancing psychological science through the study of latent structure. *Current Directions in Psychological Science*. *17*(3), 203-207.

Schwab, J. J. (1962). The teaching of science as inquiry. In P. F. Branwein (Ed.), *The teaching of science* (3-103). Cambridge, MA: Harvard University Press.

Schworm, S., and Renkl, A. (2006). Computer-supported example-based learning: When instructional explanations reduce self-explanations. *Computers and Education, 46*(4), 426-445.

Schworm, S., and Renkl, A. (2007). Learning argumentation skills through the use of prompts for self-explaining examples. *Journal of Educational Psychology, 99*(2), 285-296.

Siegler, R. S. (2002). Microgenetic studies of self-explanation. In N. Granott and J. Parziale (Eds.), *Microdevelopment: Transition processes in development and learning* (pp. 31-58). Cambridge, England: Cambridge University Press.

Siegler, R., and Lin, X. (2009). Self-explanations promote children's learning. In H. S. Waters, and W. Schneider (Eds.), *Metacognition, strategy use, and instruction* (pp. 85-112). Guilford Press.

Shadish, W. R., Cook, T. D., and Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Houghton-Mifflin; Boston.

Taber, K. S. (2000). Chemistry lessons for universities?: A review of constructivist ideas. *University Chemistry Education*, *4*(2), 63-72.

Talanquer, V. (2010). Exploring dominant types of explanations built by general chemistry students. *International Journal of Science Education*, *32*(18), 2393-2412.

Talanquer, V., and Pollard, J. (2010). Let's teach how we think instead of what we know. *Chemistry Education Research and Practice*, *11*(2), 74-83.

van Teijlingen, E., and Hundley, V. (2001). The importance of pilot studies. *Social research update*, (35), 1-4.

Vanlehn, K. (2006). The behavior of tutoring systems. *International journal of artificial intelligence in education*, *16*(3), 227-265.

Villalta-Cerdas, A. (2014). *Development and Assessment of Self-explaining Skills in College Chemistry Instruction*. (Unpublished doctoral dissertation). University of South Florida, Florida, USA.

Villalta-Cerdas, A., and Sandi-Urena, S. (2013). Self-explaining and its Use in College Chemistry Instruction. *Educación química*, *24*(4), 431-438.

Wheatley, G. H. (1984). Problem solving in school mathematics. *MEPS Technical Report*, 84.01, School Mathematics and Science Center, Purdue University, West Lafayette, IN.

Woloshyn, V. and Gallagher, T. (2009). Self-Explanation. Available online (11-06-2013):http://www.education.com/reference/article/self-explanation.

Chemistry Education Research and Practice RSCPublishing

Paper

Appendix 1: Research Instruments

The learning event consisted of a textbook passage with a general description of the Second Law of Thermodynamics and common to all the participants. A self-explaining task, SE-Task, followed this passage. There were four different SE-Tasks, each defining one of the study conditions: (1) Self-explaining own answer, SEA; (2) Self-explaining agreement/disagreement, EADA; (3) Self-explaining for others, SEO; and (4) Self-explaining incorrect answer, SEIA. Following information provided to students during the learning event.

<u>Second Law of Thermodynamics (General description)</u>

We have seen that both the system and surroundings may undergo changes in entropy during a process. The sum of the entropy changes for the system and the surroundings is the entropy change for the universe:

$$\Delta S_{univ} = \Delta S_{sys} + \Delta S_{surr}$$

The Second Law of Thermodynamics says that for a process to be spontaneous as written (in the forward direction), ΔS_{univ} must be positive ($\Delta S_{univ} > 0$). Therefore, the system may undergo a decrease in entropy as long as the surroundings undergo a larger increase in entropy making the resulting ΔS_{univ} positive, and vice versa. A process for which ΔS_{univ} is negative is not spontaneous as written.

<u>Self-explaining tasks</u>

1. **Self-explaining own answer (SEA): Working on a problem and explaining one's own answer.**

When water freezes below 0ºC, its change in entropy is negative ($\Delta S_{sys} < 0$). However, this process is <u>spontaneous</u>. How do you explain this? Please be as thorough in your response as possible.

2. **Self-explaining agreement/disagreement (EADA): Considering others' answers to a problem and explaining one's agreement/disagreement.**

When water freezes below 0ºC, its change in entropy is negative ($\Delta S_{sys} < 0$). Despite this observation, your group members maintain that this process is <u>spontaneous</u>. Therefore, they say, no energy input from the outside is necessary to make this change happen. Do you agree with your classmates? Please explain and be as thorough in your response as possible.

3. **Self-explaining for others (SEO): Explaining answer to a problem for others to use in their studying.**

When water freezes below 0ºC, its change in entropy is negative ($\Delta S_{sys} < 0$). However, this process is spontaneous. Explain this in writing so that a classmate of yours can use your explanation as reference when answering a similar problem. Your answer will be used by your classmate. Please be as thorough in your response as possible.

4. **Self-explaining incorrect answer (SEIA): Explaining others' incorrect answers to a problem.**

When water freezes below 0ºC, its change in entropy is negative ($\Delta S_{sys} < 0$). Your group members maintain that this process will not occur spontaneously. Therefore, they say, there must be an energy input from the outside to make this change happen; otherwise, water will not freeze. This stance is incorrect. What do you think led your classmates to this incorrect conclusion? Please be as thorough in your response as possible.

Appendix 2: Interview protocol

Think-aloud Interview protocol

Students' assessment of research materials

1. Introductory aspects [read to interviewee]

We are testing an instrument that has questions that may be difficult to understand, hard to answer, or that make little sense. We would like you to answer the questions as carefully as possible. *We are primarily interested in the ways that they arrived at those answers, and the problems they encountered*. Therefore, any detailed help you can give us is of interest, even if it seems irrelevant or trivial.

We are not looking for correct answers; we just want to listen to your comments. I didn't write these questions, so don't worry about hurting my feelings if you criticize them -my job is to find out what's wrong with them.

The conversations will be audio taped just as a means for us to go back and review *what* was said and not *who* said *what*. This interview is confidential; you will not be identified by name and only the transcriber will listen to this tape. The transcriber is bound to confidentiality, as well. During the conversation, I may take notes, which most probably will be reminders to myself of something I want to inquire about later, or something especially interesting you said. I will not jot down things *about* you, you are not under observation.

Please feel free to spend as much time as you need or want on any given topic. You do not have to reply to a question if for any reason you do not feel comfortable. We may stop the conversation at any time you wish or need to. Do not feel like I am being too insistent if I ask some follow up questions to your comments. It is our interest to clearly understand what you mean; we are trying to get to a deeper level of understanding.

Once again, **this interview is absolutely confidential.** We very much appreciate your taking the time for this conversation. We will start with some general background information and then we will move on to aspects related to the instrument.

2. Background [use these to strengthen rapport with interviewee and set a comfortable environment]

a) What is your undergraduate major in?

b) What chemistry courses have you taken in the past?

c) Are you taking any chemistry classes this semester?

3. Think-aloud training exercise:

"Try to visualize the place where you live, and think about how many windows there are in that place. As you count up the windows, tell me what you are seeing and thinking about."

4. Instrument assessment. One self-explaining task condition. [Use prompts and follow-ups as necessary]

1. The following instrument is intended for students taking general chemistry 2. Please read the following information.
2. Give student information sheet "Entropy definition". Give time to read. Then remove information.
3. Give student information sheet "Second law of thermodynamics. Give time to read. Let student keep this information for the rest of the interview.
4. I will read a question to you and I would like you to think out loud when you answer the following questions.
5. Read prompt "Self-explaining own answer, SEA" to student to think-aloud while solving it.

Verbal Probes during resolution:

a. Please repeat the question I just asked in your own words?
b. How did you arrive at that answer?
c. I noticed that you hesitated - tell me what you were thinking.

6. Verbal Probes after resolution:
    a. How difficult was this question to answer?
    b. How sure are you of your answer?

5. **Instrument assessment. Comparison with two other intervention conditions. [Use prompts and follow-ups as necessary]**
    1. Now I am going to give you another question.
    2. Provide prompt "Self-explaining agreement/disagreement, EADA" to student. Give time to read. Let student keep this information for the rest of the interview. Verbal probe technique is used.

Verbal Probes:

a. What does the term "energy input from the outside" mean to you?
b. How hard is it to think of reasons for your classmates to get the incorrect conclusion?
c. What other reasons can you think of?
d. Overall, how difficult was this question to answer?

3. Provide another prompt to student. Give time to read. Let student keep this information for the rest of the interview.

Verbal Probes:

a. How difficult is this question to answer?
b. How is this question related to the previous two questions?
c. Please arrange the three questions in order of difficulty. (Give student time to arrange questions).
d. What do you understand as "difficult" when arranging these questions?

6. **Wrap up**

Thank you again for your valuable collaboration. Once more, **this interview is confidential,** you will not be identified by name and only the transcriber will listen to this tape. The transcriber is bound to confidentiality, as well.

Information sheets given to students during the interview:

Sheet 1: Entropy definition

Entropy (S) is a thermodynamic function that increases with the number of energetically equivalent ways to arrange the components of a system to achieve a particular state. It may be thought of as a measure of the dispersion of the energy in a system and it is associated with disorder or randomness at the molecular level.

Sheet 2: Second Law of Thermodynamics

We have seen that both the system and surroundings may undergo changes in entropy during a process. The sum of the entropy changes for the system and the surroundings is the entropy change for the universe:

$$\Delta S_{univ} = \Delta S_{sys} + \Delta S_{surr}$$

The Second Law of Thermodynamics says that for a process to be spontaneous as written (in the forward direction), ΔS_{univ} must be positive ($\Delta S_{univ} > 0$). Therefore, the system may undergo a decrease in entropy as long as the surroundings undergo a larger increase in entropy making the resulting ΔS_{univ} positive, and vice versa. A process for which ΔS_{univ} is negative is not spontaneous as written.

Prompts evaluated during the interview:

Self-explaining own answer, SEA

When water freezes below 0ºC, its change in entropy is negative ($\Delta S_{sys} < 0$). However, this process is spontaneous. How do you explain this? Please be as thorough in your response as possible.

Self-explaining agreement/disagreement, EADA

When water freezes below 0 ºC, its change in entropy is negative ($\Delta Ssys < 0$). Despite this observation, your group members maintain that this process is spontaneous. Therefore, they say, no energy input from the outside is necessary to make this change happen. Do you agree with your classmates? Please explain and be as thorough in your response as possible.

Self-explaining for others, SEO

When water freezes below 0ºC, its change in entropy is negative ($\Delta S_{sys} < 0$). However, this process is spontaneous. Explain this in writing so that a classmate of yours can use your explanation as reference when answering a similar problem. Your answer will be used by your classmate. Please be as thorough in your response as possible.

Self-explaining incorrect answer, SEIA

When water freezes below 0ºC, its change in entropy is negative ($\Delta S_{sys} < 0$). Your group members maintain that this process will not occur spontaneously. Therefore, they say, there must be an energy input from the outside to make this change happen; otherwise, water will not freeze. This stance is incorrect. What do you think led your classmates to this incorrect conclusion? Please be as thorough in your response as possible.

Appendix 3: Supplemental Data Analysis and Results

Data analysis

In addition to the textual analysis of the responses, we conducted a structural analysis. Linguistic studies have reported the structure of essays (i.e. use of types of conjunctions and length of text) vary with their summative-analytical nature (Durst, 1987). Although the responses in our case were relatively short to be considered short essays, we decided to investigate whether structural difference would be noticeable as a function of level of self-explaining.

Structural analysis of Learning Event data. For each *response*, we counted the total number of words and cohesive conjunctions (i.e. words used in text construction to connect sentences). For the total number of words, we considered symbols representing individual concepts such as *change in entropy of the system*, ΔS_{sys}, as a single unit. Other examples are: –ΔS_sys, +, →, H_2O. We tallied mathematical sentences using the same principle; therefore, the word count for an equation corresponded to the number of elements used in the mathematical sentence. The word count for the following two examples is three: $\Delta S_{sys} < 0$; ΔS_universe = 0. We tallied contractions as two words. In the case of cohesive conjunctions we used the categories shown in Table A1. Linguistic studies have shown the prevalence of causal and adversative cohesive conjunctions in analytical essays, and additive and temporal cohesive conjunctions in summative essays (Durst, 1987). We compared the mean word count by SE-Tasks using ANOVA. We calculated the ratio of each cohesive conjunction-type by dividing the frequency by the total word count. We used these cohesive conjunction-type ratios as observed variables for the subsequent Latent Profile analysis. We postulated that the overt explanatory behaviour of the students would be associated with the structural characteristics (total word count and cohesive conjunction-type ratios).

Table A1 Cohesive Conjunction Categories as Described by Durst (1987)

Cohesive Conjunction categories	**Description and examples**
Additive	Indicates coordination; two sentences are given equal weight. Examples include conjunctions such as "and," "also," "furthermore," "or", "plus", "that".
Temporal	Conjunctive relation showing chronological connection. Examples include "after," "then," "when," "once", "while".
Causal	Indicate cause and effect relation. Examples include "because," "so," "therefore," "thus", "since", "due", "as", "if".
Adversative	Indicate that what follows contrasts with what has just been said. Examples include "in fact," "but," "however," "instead", "although", "whereas", "though", "yet".

We performed LPA using the conjunction-type ratios and total word count as observed variables (i.e., five observed variables). The output of the LPA analysis of the conjunction-type ratios was the categorization of students into distinct profile classes based on their text construction, the Text Construction Profiles, TC-Profile. Table A2 shows examples of coded responses from the learning event data, and the corresponding observed variables we used in the latent profile analyses.

Table A2 Example Report of Textual Analysis and Structural Analysis of Learning Event Responses

Response Examples	Textual Analysis							Structural Analysis					
								Value	Total words	Cohesive Conjunctions			
	Value	BI	DI	E	P	U	NR			Additive	Temporal	Causal	Adversative
The change in entropy for the system is less than zero **therefore** the ΔS_{surr} must have a larger increase of entropy than the negative decrease of the entropy of the system for the process to be spontaneous, meaning ΔS_{univ} will still be greater than 0.(DI) **When** water freezes the system loses entropy, **but** the outside surroundings gain more entropy than what was lost by the system.(BI)	Count	1	1	0	0	0	0	Count	66	0	1	1	1
	Code--ratio	0.5	0.5	0	0	0	0	Count per 100 words	-	0	1.52	1.52	1.52
Water freezes at 0˚C.(P) After frozen, no matter how much more energy is lost (E = heat) it's still just **as** frozen – it can never be "more frozen" w/ more cold.(E) The process is spontaneous, **yet** negative **because** the system is more negative than the surroundings are positive.(DI)	Count	0	1	1	1	0	0	Count	48	0	0	2	1
	Code-ratio	0	0.33	0.33	0.33	0	0	Count per 100 words	-	0	0	4.17	2.08

Association analysis between self-explaining tasks and self-explaining profile. We used Chi-square tests to determine the association between Text Construction Profiles membership, TC-Profile, and the Self-Explaining Task, SE-Task. We used IBM SPSS Statistics (Version 21.0.0.0) for the Chi-square tests.

Results and discussion

Pilot Study

Following we describe results corresponding to the analysis of the pilot study dataset. These results informed the implementation of the main study and also supported findings from the main study.

Code type distribution. Students' responses were coded following the coding scheme described in the methodology section to produce a tally of the codes in each response. Initially, we analysed the association between these code types and the self-explaining task, SE-Task (Table A3). For this analysis the codes "unclassifiable, U" and "non-relevant, NR" were excluded because they did not contribute valuable insight about the self-explaining behaviour. From the total count of codes it is interesting to notice that for this dataset, the "deductive inference" code presents the highest count (Table A3). This suggests that students were actively engage in the generation of inferences. It is also noticeable that the "elaboration, E" and "paraphrasing, P" codes have high number of occurrences. These E and P codes are associated with lower explanatory sophistication as they describe responses that recount information. High counts in E and P codes thus suggested that students relied heavily on recounting information when prompted to write explanations.

The Chi-square test showed no significant association between the code types and the SE-Tasks, at a 95% confidence interval, χ^2 (9, N = 269) = 15.83, p = .07. Nonetheless, inspection of Table A3 showed evidence of a trend: higher percentages (dark grey shaded cells) of the "bridging inference, BI" code are found in the self-explaining-to-others task, SEO, and self-explaining agreement/disagreement task, EADA. This finding suggests that the SEO and EADA tasks prompted students to generate more bridging inferences to link chemistry concepts (i.e., entropy and the Second Law of Thermodynamics). In the case of the code "deductive inference, DI" the higher percentage was found in the EADA task, which suggests that students generated more deductive inferences while working on this SE-task. In the case of the DI code the SEA (self-explain own answer) and SEIA (self-explain incorrect answer) tasks had moderately high percentages too, thus engaging students in the generation of inferences as well. In the case of the "elaboration, E" code, the results suggest a similar percentage in the SEO, EADE and SEA tasks, but a higher percentage in the SEIA task. Finally, the "paraphrasing, P" code presents higher percentages in the EADA and SEIA tasks, suggesting that responses on these SE-tasks were heavily composed of recounted information. In summary, these results suggest that, although not statistically significant, the code types in the students' responses were associated with the self-explaining tasks.

Table A3 Pilot study learning event code type distribution by SE-Task

Code Type	Total	SE-Task			
		%SEO	%EADA	%SEA	%SEIA
BI	25	40	32	16	12
DI	95	17	33	26	24
E	77	25	22	21	32
P	72	14	33	18	35
U	21	14	24	38	24
NR	28	29	29	14	29

Without codes "U" and "NR": χ^2 (9, N = 269) = 15.83, p = .07.

The results in Table A3 show the total number of code types per SE-Task, but for 80% of the students the responses were not composed of only one code. Therefore the follow-up analysis considered the code types and the total number of codes in each response. This way, we accounted for the relationship among all the code types present in each response for the categorization of student's self-explaining behaviour.

Latent profile analysis. We used Latent Profile Analysis (LPA) to identify patterns in code-ratios (i.e., number of code type divided by total codes in response) in students' responses. These analyses required the selection of the best model for the data. In order to make that decision the following information was used: number of profiles selected; goodness of fit indexes (Loglikelihood (LLH), Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), sample-size adjusted Bayesian Information Criteria (SSA–BIC)); likelihood ratio tests (Voung–Lo–Mendell–Rubin and parametric bootstrap); and homogeneity (referred to as entropy value) and cases size in each profile. Next we describe the model selection process for the pilot study.

Consistent with common practice, we explored solutions with varying numbers of profiles and selected the one that made the most sense in terms of interpretability and model fit information. We evaluated one- to ten-profile solutions models in relation to indexes of fit commonly used for this purpose (Table A4). For the information criteria Loglikelihood, LLH, the value increased as the number of profiles increased thereby indicating progressive model fitness from the model with only one profile up to the model with ten profiles. Thus, the LLH value did not provide useful information for the model selection. For the three information indexes (AIC, BIC, and SSA-BIC), lower values indicate better model fit. Our results showed that all information indexes progressively became lower as the model solution incorporated more profiles (Figure A1). This indicated that model solutions with more profiles seemed to better fit the data. In the case of the Voung–Lo–Mendell–Rubin (VLMR) test and parametric Bootstrap Likelihood Ratio Test (BLRT) the p-value reflects how significant it is to have a model with n-profiles against a model with (n-1)-profiles ("n" being the number of profiles within each model). Therefore, if the p-values are lower than .05 this means that the model solution with n-profiles is favourable over the model solution with (n-1)-profiles. The p-values for the BLRT were all lower than .05 suggesting that any model solutions are significantly better than the corresponding previous model solution, thus these results did not help in selecting a model solution. The results in Table A4 showed that only the model solutions with two-profiles and five-profiles have a p-value lower than .05 for the VLMR test, suggesting that these model solutions were favourable over the other model solutions. Finally, the last criteria for the model solution selection are the homogeneity (i.e. entropy value) of the profiles in the model and the number of cases in each profile within the model solutions. In our case both the two-profile and five-profile solution models had high homogeneity. We selected the five-profile model solution (grey shaded cells in Table A4) because larger number of profiles increased our categorization power of students. This was a judgment call based on the model fitness, parsimony and interpretability of the five-model solution.

Table A4 Goodness of fit for LPA models based on code-ratios, pilot study dataset (N=103)

Number of Profiles	Number of Parameters	LLH	AIC	BIC	SSA–BIC	*p* VLMR	*p* BLRT	Entropy	Group Sizes	
									LT1 %	LT5 %
LP 1	8	-85	186	207	182	—	—	—	0	0
LP 2	13	-29	85	119	78	.00	.00	.99	0	0
LP 3	18	31	-26	22	-35	.49	.00	1.0	0	0
LP 4	23	89	-131	-71	-143	.20	.00	1.0	0	2
LP 5	28	138	-221	-147	-235	.03	.00	.97	0	2
LP 6	33	183	-300	-214	-318	.79	.00	.97	0	2
LP 7	38	218	-359	-259	-379	.46	.00	.95	0	2
LP 8	43	253	-420	-307	-443	.62	.00	.98	1	3
LP 9	48	276	-457	-330	-482	.85	.00	.99	0	2
LP 10	53	305	-503	-364	-531	.25	.00	.99	0	5

Note: LLH = Loglikelihood; AIC = Akaike's Information Criterion; BIC = Bayesian Information Criterion; SSA–BIC = sample-size adjusted Bayesian Information Criteria; p VLMR = p-value for the Voung–Lo–Mendell–Rubin likelihood ratio test for K versus K - 1 classes; p BLRT = p-value for the parametric bootstrap likelihood ratio test for K versus K - 1 classes. Group sizes refer to the number of groups with less than 1% and less than 5% of the cases, N = 103.

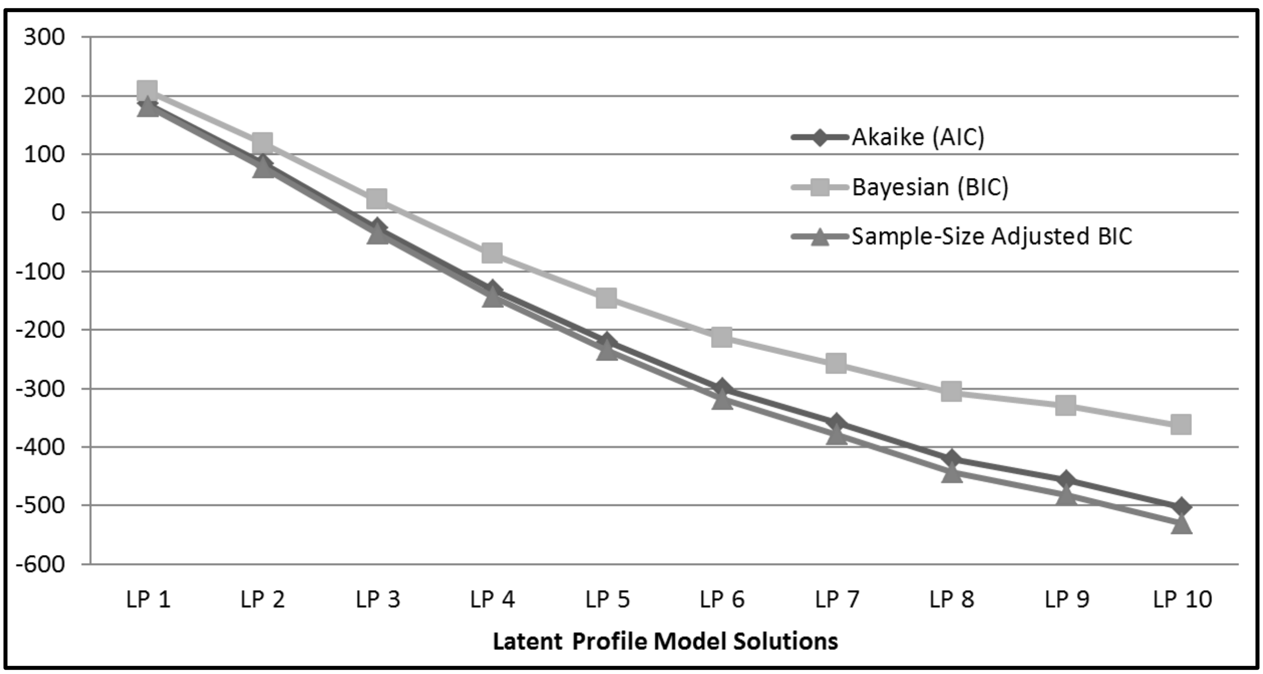

Fig. A1 AIC, BIC and Sample-size adjusted BIC values for explanation sophistication one- to ten-profile model solutions.

The profiles in the five-profile model are presented in Table A5. For each profile Table A5 shows the number of students in the group and the mean values for each of the four code-ratios. Figure A2 presents the mean values of code-ratios for each profile as a visual aid for discussion. In the case of Profile 1, this group was composed of five students whose responses only contained "bridging inference, BI" codes. Therefore, we described this profile (SE-profile) as bridging inferential. Profile 2, a three-member group, presented responses mainly using bridging and deductive inferences (89% of the response) so we described them as bridging/deductive inferential. Profile 3 is an interesting group of fourteen students whose responses used a mixture of all codes in evenly distributed ratios. We described this profile as "mixed-behaviour." Profile 4, 63 students, had a high code-ratio for DI but also had a significant code-ratio of "paraphrasing, P". In other words, deductive inferences predominated in these responses but students also relied significantly on recounting information. We described this profile as "deductive inferential." Finally, the 18 students in Profile 5 relied heavily on elaboration statements, having a mean value of 92% of the response coded as "elaborations, E". Thus, we described this profile as "elaborative."

Table A5 Code-ratios and SE-profile descriptors for five-profile model solution (N=103)

Profile Group	n	Mean code-ratio				SE-Profile descriptor
		BI	DI	E	P	
Profile 1	5	1.00*	0.00*	0.00*	0.00*	Bridging Inferential
Profile 2	3	0.56*	0.33*	0.11	0.00	Bridging/Deductive Inferential
Profile 3	14	0.29*	0.26*	0.30*	0.15*	Mixed-behaviour
Profile 4	63	0.00	0.49*	0.15*	0.36*	Deductive Inferential
Profile 5	18	0.00	0.04	0.92*	0.04	Elaborative

* $p < .05$.

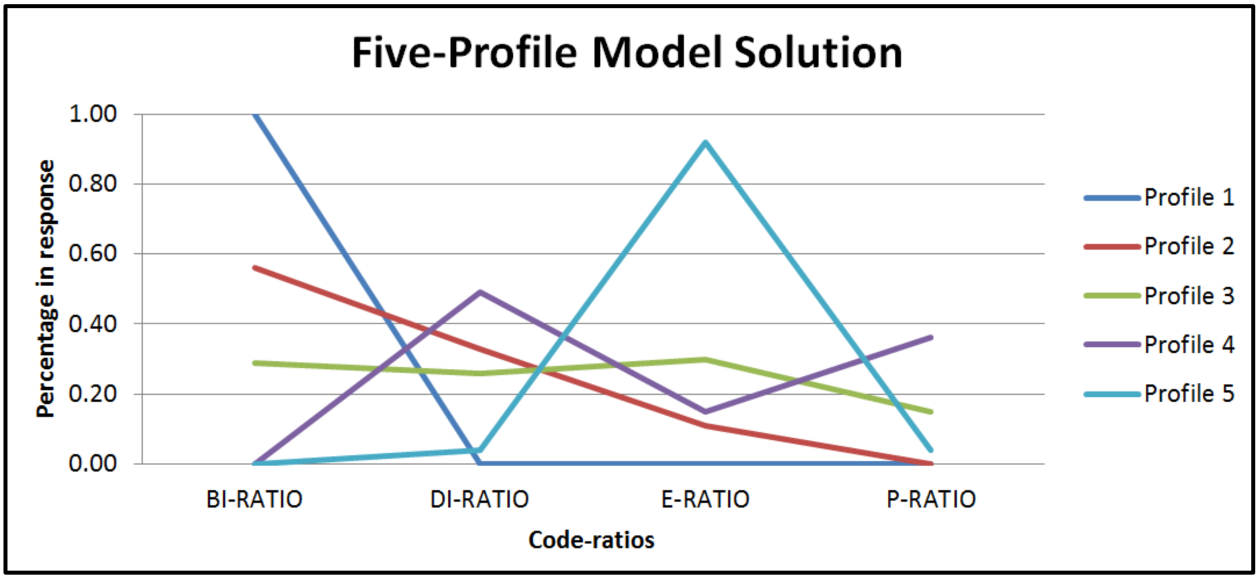

Fig. A2 Profile plot for the five-profile solution model, code-ratios.

The results in Table A5 support the interpretability of the LPA outcome (i.e., five-profile model solution) under the theoretical framework for the construct of self-explaining. This is because the LPA outcome allows the categorization of students' self-explaining behaviour in five clearly distinct groups. This finding directly addresses the first research question: Do tasks that require different self-explaining engagement induce observable categorical differences in self-explaining behaviour in the context of a General Chemistry classroom? Findings support the emergence of observable categorical differences in self-explaining behaviour when tasks prompted pupils to provide written explanations. However, to fully answer this research question we studied the association of these self-explaining behaviours (SE-Profiles) with the self-explaining task (SE-Task).

SE-Profile and SE-Task association analysis. Table A6 shows the cross tabulation of SE-Profile and SE-Task. The Chi-square test was not applicable in this case due to low sample size. This is because in the Chi-square calculation, 16 cells (80.0%) had an expected count value lower than five which is in violation of the Chi-square test requirements (less than 20% cells with expected count lower than five). Therefore the result from the Chi-square analysis was not conclusive, χ^2 (12, N = 103) = 11.69, p = .47. Nonetheless, inspection of Table A6 shows an apparent trend. SEA and SEO tasks have a higher proportion of students in the SE-Profile associated with a more analytic behaviour (i.e., bridging inferential and bridging/deductive inferential). That is, more students coming from these SE-Tasks engaged in drawing inferences and connecting ideas. Conversely, SEIA and EADA SE-tasks have higher proportions of students in the least analytic behaviours (i.e., elaborative, and deductive inferential). The apparent trend in Table A6 suggests that, although not statistically significant, the self-explaining tasks (SE-Tasks) were associated with the self-explaining behaviours (SE-Profiles).

Table A6 Percentage distribution of SE-Profile across SE-Task (N=103)

SE-Profile	n	SE-Task			
		%SEA	%SEO	%EADA	%SEIA
Bridging Inferential	5	40	40	20	-
Bridging/Deductive Inferential	3	33	33	33	-
Mixed-behaviour	14	7	36	36	21
Deductive Inferential	63	27	16	30	27
Elaborative	18	22	28	11	39

Main Study

Following we describe additional results corresponding to the analysis of the main study dataset. This information goes to a level of technical detail deeper than that discussed in the manuscript.

LPA seven-profile model solution selection. We used Latent Profile Analysis (LPA) to identify patterns in code-ratios (i.e., number of code type divided by total codes in response) in students' responses. These analyses required the selection of the best model for the data. In order to make that decision the following information was used: number of profiles selected; goodness of fit indexes (Loglikelihood (LLH), Akaike's Information Criterion (AIC), Bayesian Information Criterion (BIC), sample-size adjusted Bayesian Information Criteria (SSA–BIC)); likelihood ratio tests (Voung–Lo–Mendell–Rubin and parametric bootstrap); and homogeneity (referred to as entropy value) and cases size in each profile.

Consistent with common practice, we explored solutions with varying numbers of profiles and selected the one that made the most sense in terms of interpretability and model fit information. We evaluated one- to ten-profile solutions models in relation to indexes of fit commonly used for this purpose (Table A7). For the information criteria Loglikelihood, LLH, the value increased as the number of profiles increases indicating progressive model fitness from the model with only one profile up to the model with ten profiles. Thus, the LLH value showed that a model with more profiles is favoured. For the three information indexes (AIC, BIC, and SSA-BIC), lower values indicate better model fit. Our results showed that all information indexes progressively became lower as the model solution incorporated more profiles (Figure A3). As the figure shows, the values decreased as the number of latent profiles increased up to seven and then they started to level off. This indicated that model solutions higher than seven-profiles are favoured with no much improvement after seven profiles.

In the case of the Voung–Lo–Mendell–Rubin (VLMR) test and parametric Bootstrap Likelihood Ratio Test (BLRT) the p-value reflects how significant it is to have a model with n-profiles against a model with (n-1)-profiles ("n" being the number of profiles within each model). Therefore, if the p-values are lower than .05 this means that the model solution with n-profiles is favourable over the model solution with (n-1)-profiles at a 95% confidence interval. In the case of the VLMR test the model solutions for two, three, six showed values lower than .05, also the seven-profile model solution showed a low p-value indicating that this solution is considerably good. The p-values for the BLRT were all lower than .05 suggesting that any model solutions are significantly better than the corresponding previous model solution, thus these results did not help in selecting a model solution. The entropy value for all model solution was close to the highest possible value of one, which means the homogeneity of the profiles in each solution is high, which is favourable.

Inspection of the eight- to ten-profile model solutions showed three or more group sizes with less than 5% of the total cases. We considered that the eight- to ten-profile model solutions did not add valueable insight into the categorization of the students. Based on the results from Table A7 we selected the seven-profile model solution.

Table A7 Goodness of fit for LPA models based on code-ratios, main study data (N=128)

Number of Profiles	Number of Parameters	LLH	AIC	BIC	SSA–BIC	*p* VLMR	*p* BLRT	Entropy	Group Sizes	
									LT1 %	LT5 %
LP 1	8	-94	205	228	202	-	-	-	0	0
LP 2	13	-41	108	145	104	.04	.00	.95	0	0
LP 3	18	21	-6.2	45	-12	.03	.00	.99	0	0
LP 4	23	86	-127	-61	-134	.26	.00	1.0	0	0
LP 5	28	153	-250	-170	-258	.36	.00	1.0	0	2
LP 6	33	199	-332	-238	-342	.01	.00	.96	0	2
LP 7	38	259	-442	-334	-454	.07	.00	.98	0	2
LP 8	43	273	-460	-337	-473	.54	.00	.98	0	3
LP 9	48	329	-562	-425	-577	.35	.00	.98	0	4
LP 10	53	343	-580	-429	-597	.60	.00	.99	0	5

Note: LLH = Loglikelihood; AIC = Akaike's Information Criterion; BIC = Bayesian Information Criterion; SSA–BIC = sample-size adjusted Bayesian Information Criteria; p VLMR = p value for the Voung–Lo–Mendell–Rubin likelihood ratio test for K versus K - 1 classes; p BLRT = p value for the parametric bootstrap likelihood ratio test for K versus K - 1 classes. Group sizes refer to the number of groups with less than 1% and less than 5% of the cases, N = 128.

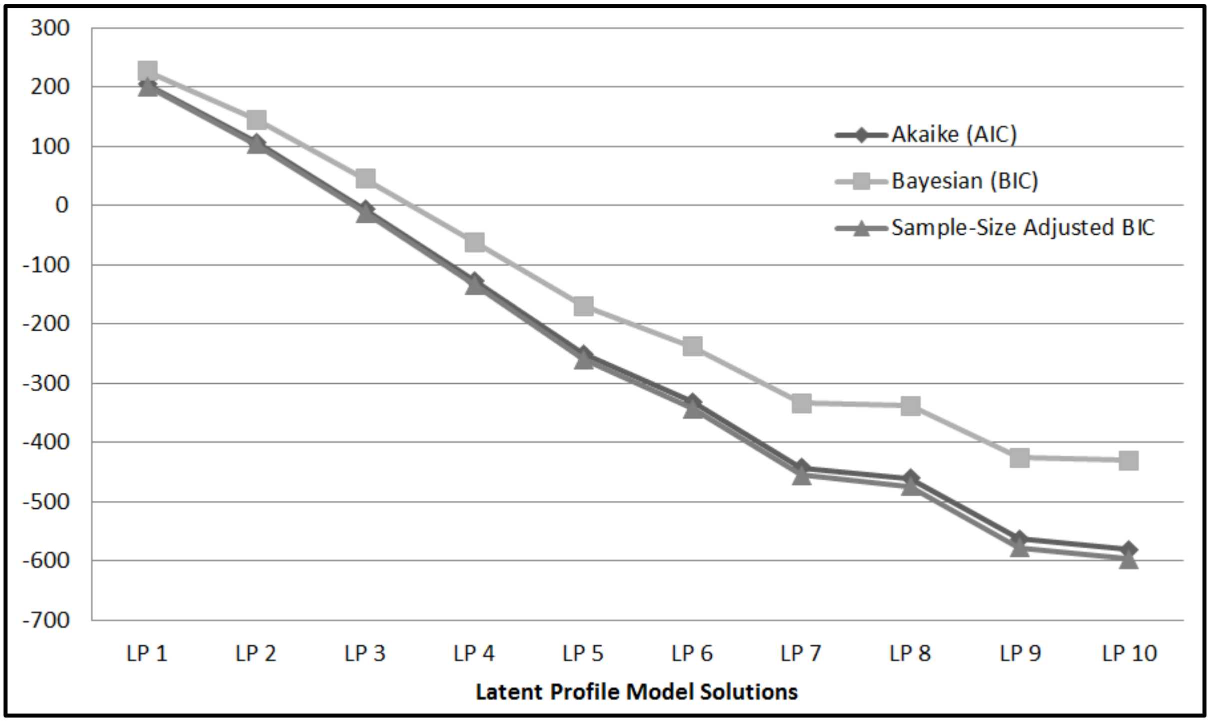

Fig. A3 AIC, BIC and Sample-size adjusted BIC values for explanation sophistication one- to ten-profile solutions.

Structural Analysis. We analysed the responses of the students in the main study phase for their structural composition in terms of (1) the total word count and (2) the cohesive conjunction type count (Table A8). Analysis showed no relevant differences among the self-explaining tasks, SE-Task, and self-explaining profiles, SE-Profiles, across these two counts. These results suggest that the text construction of the written responses is not different between participants doing different self-explaining tasks, or behaving differently when self-explaining. However, we acknowledge that the extension of the participants' responses is not as extensive as in the case of other research studies (e.g., mean total words = 500) (Durst, 1987), and this may impact our resolution.

Table A8 Descriptive statistics of word counts by SE-Task (N=128)[a]

SE-Task	n	Total words Mean (SD)	Cohesive conjunctions per 100 words Mean (SD)			
			Additive	Temporal	Causal	Adversative
SEA	29	61 (24)	2.4 (2.0)	1.5 (1.7)	4.3 (2.7)	0.72 (0.97)
EADA	31	63 (23)	2.3 (2.0)	0.9 (1.1)	3.7 (2.5)	0.42 (0.85)
SEO	35	55 (24)	2.1 (2.0)	1.3 (1.6)	4.0 (3.1)	0.49 (0.77)
SEIA	33	64 (19)	3.0 (2.6)	1.1 (1.1)	2.6 (2.1)	0.76 (1.12)
Total	128	61 (22)	2.2 (2.2)	1.2 (1.4)	3.6 (2.7)	0.60 (0.94)
ANOVA	*F*	1.17	2.03	.96	2.68	1.00
(3, 124)	*p*	.33	.11	.41	.05	.40

SE-Profile	n	Total words Mean (SD)	Cohesive conjunctions per 100 words Mean (SD)			
			Additive	Temporal	Causal	Adversative
Bridging Inferential	25	60 (21)	2.5 (2.3)	1.3 (1.6)	4.0 (3.3)	0.60 (0.78)
Mixed behavior	12	66 (22)	3.2 (2.8)	1.2 (1.4)	2.4 (1.8)	0.66 (1.14)
Deductive Inferential	20	46 (24)	4.2 (3.8)	1.5 (1.8)	4.3 (3.1)	0.42 (0.96)
Elaborative	24	62 (19)	3.5 (2.4)	1.4 (1.2)	3.3 (1.9)	0.67 (0.98)
Summative	47	63 (22)	3.4 (2.4)	1.0 (1.2)	3.7 (2.7)	0.64 (0.99)
Total	128	59.9 (22)	3.4 (2.7)	1.22 (1.4)	3.7 (2.7)	0.60 (0.95)
ANOVA	*F*	2.65	1.3	0.65	1.24	.24
(3, 124)	*p*	.04	.28	.63	.30	.92

[a] Six responses were unintelligible and therefore removed from the analysis.

To further analyse these data we used LPA to investigate categorical differences among students' text construction behaviours, TC-Profiles. The idea of this LPA study was to identify groups of students with similar text construction styles (in terms of the length of the explanation and the used of cohesive conjunction words) when writing explanations. In contrast with the previously shown analysis of variance, ANOVA, which used only one word count in each analysis, the LPA used all of the five different word count values in each response as observed variables to classified students into text construction profiles, TC-Profiles. This allowed an in-depth categorization of students' text construction behaviours. Next we studied potential differences of the text construction styles (TC-Profiles) across the experimental conditions (SE-Tasks) and self-explaining behaviours (SE-Profiles).

Latent profile analysis: Text construction profiles. The selection process for the best latent profile fit model solution followed the procedures described previously (see above). Results for the goodness of fit indexes are shown in Table A9 and Figure A4. The three-profile model solution showed the best fit for the data. Also the high value of homogenity (i.e., Entropy value = .998) in the three-profile model solution suggested that students' membership within each of the three profiles was well established. This meant that all students within each profile had low uncertainty of belonging to other profile within the model solution.

Table A9 Goodness of fit for LPA models based on word counts, main study data (N=128)

Number of Profiles	Number of Parameters	LLH	AIC	BIC	SSA–BIC	*p* VLMR	*p* BLRT	Entropy	Group Sizes	
									LT1 %	LT5 %
LP 1	10	-1566	3151	3180	3148	-	-	-	0	0
LP 2	16	-1537	3105	3151	3100	.55	.00	.94	0	0
LP 3	22	-1479	3002	3065	2996	.04	.00	1.0	0	1
LP 4	28	-1434	2923	3003	2915	.50	.00	.99	0	1
LP 5	34	-1412	2893	2990	2882	.60	.00	.98	0	1
LP 6	40	-1386	2851	2965	2839	.26	.00	.99	1	1
LP 7	46	-1365	2823	2954	2808	.13	.00	.96	1	1
LP 8	52	-1348	2801	2949	2785	.35	.00	.94	1	1
LP 9	58	-1331	2779	2944	2761	.81	.60	.94	1	2
LP 10	64	-1305	2738	2921	2718	.81	.02	.95	1	3

Note: LLH = Loglikelihood; AIC = Akaike's Information Criterion; BIC = Bayesian Information Criterion; SSA–BIC = sample-size adjusted Bayesian Information Criteria; p VLMR = p value for the Voung–Lo–Mendell–Rubin likelihood ratio test for K versus K - 1 classes; p BLRT = p value for the parametric bootstrap likelihood ratio test for K versus K - 1 classes. Group sizes refer to the number of groups with less than 1% and less than 5% of the cases, N = 128.

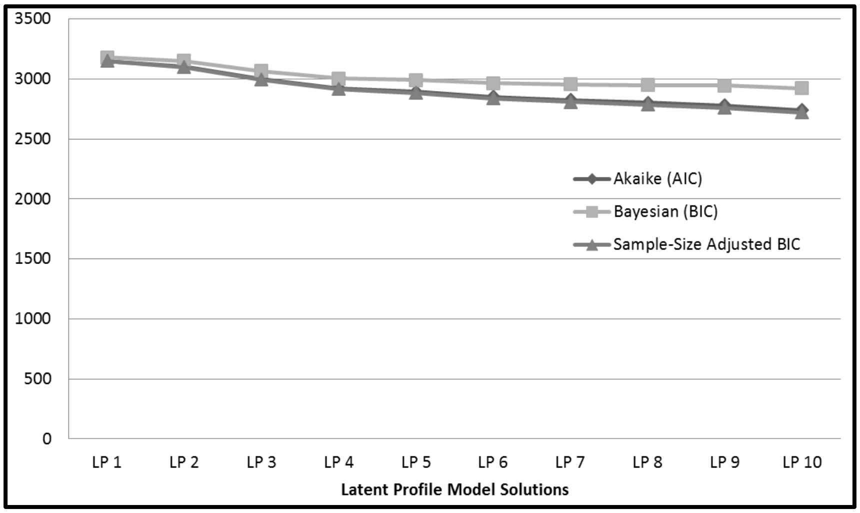

Fig. A4 AIC, BIC and Sample-size adjusted BIC values for text construction one- to ten-profile model solutions.

The results for the three-profile solution are shown in Table A10. Most profiles presented substantial differentiation among word counts (i.e., total word and cohesive conjunction counts). Figure A5 shows the cohesive conjunction counts for each profile as visual aid. Profile 1 showed the highest count of adversative cohesive conjunctions of the three profiles. We described this profile as "Adversative." This profile size is small in comparison with the other and for practical reasons subsequent analyses did not consider it. Profile 2 showed the highest count of causal cohesive conjunctions, therefore we described it as "Causal." Finally, Profile 3 showed the highest number of total word count. We described this profile as "longer-texts."

Table A10 Word counts for text construction three-profile model solution

Profile	n	Total words Mean (SD)	Cohesive conjunction per 100 words Mean (SD)				TC-Profile Descriptor
			Additive	Temporal	Causal	Adversative	
Profile 1	4	63 (21)*	1.3 (2.2)*	0.8 (1.4)*	2.2 (2.6)*	3.7 (0.3)*	Adversative
Profile 2	84	54 (21)*	2.2 (2.2)*	1.3 (1.4)*	4.0 (2.6)*	0.0 (0.3)	Causal
Profile 3	40	74 (21)*	2.5 (2.2)*	1.1 (1.4)*	2.9 (2.6)*	1.5 (0.3)*	Longer-texts

* $p < .05$.

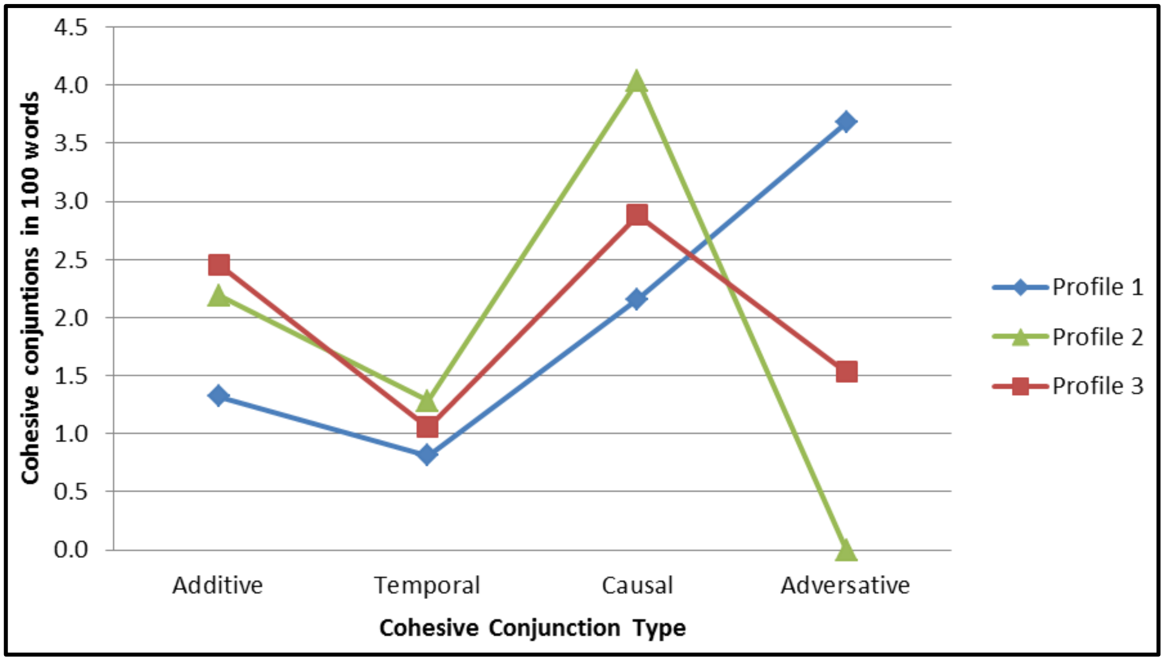

Fig. A5 Profile plot for the three-profile solution model, cohesive conjunctions in 100 words.

TC-Profile and SE-Task association analysis. For the association analysis of the TC-Profiles and the SE-Task, we did not consider the "adversative" profile due to its low number of cases. The Chi-square test showed no significant association between the remaining two TC-Profiles and SE-Tasks at a 95% confidence interval, $\chi^2(3, N = 124) = 3.30$, $p = .35$ (Table A11). This result suggests that the text construction behaviour of the students was not found to be different depending of the self-explaining task. This finding provides further support to the previously discussed ANOVA study (see above) as no significant differences were found among the writing styles of the students across the SE-Tasks.

Table A11 Percentage distribution of TC-Profile across SE-Task[a]

TC-Profile		SE-Task			
	n	%SEA	%EADA	%SEO	%SEIA
Adversative	4	25	25	-	50
Causal	40	28	15	28	30
Longer-texts	84	20	29	29	23

[a] Without "Adversative" profile: χ^2 $(3, N = 124) = 3.30$, $p = .35$.

SE-Profile and TC-Profile association analysis. As in the previous analysis, for the association analysis of the SE-Profiles and the TC-Profiles, we did not consider the "adversative" profile due to its low number of cases. The Chi-square test showed no significant association between the remaining two TC-Profiles and SE-Profile at a 95% confidence interval, $\chi^2(4, N = 124) = 3.37$, $p = .50$ (Table A2). This result suggests that the self-explaining behaviour of the student is not significantly associated to students' text construction behaviour in terms of use of cohesive conjunction types and text extension. We acknowledge that the low mean values for total words in the students' response presents a limitation for the resolution and power of this result.

Table A12 Percentage distribution of SE-Profile across TC-Profile[a]

SE-Profile	n	TC-Profile		
		%Adversative	%Causal	%Longer-text
Bridging Inferential	25	-	60	40
Mixed behaviour	12	8	67	25
Deductive Inferential	20	5	80	15
Elaborative	24	4	63	33
Summative	47	2	64	34

[a] Without "Adversative" profile: χ^2 (4, N = 124) = 3.37, p = .50.