International Journal of STEM Education
a SpringerOpen Journal

RESEARCH Open Access

Developing a computer-based assessment of complex problem solving in Chemistry

Ronny Scherer[1*], Jenny Meßinger-Koppelt[2] and Rüdiger Tiemann[3]

Abstract

Background: Complex problem-solving competence is regarded as a key construct in science education. But due to the necessity of using interactive and intransparent assessment procedures, appropriate measures of the construct are rare. This paper consequently presents the development and validation of a computer-based problem-solving environment, which can be used to assess students' performance on complex problems in Chemistry. The test consists of four scales, namely, understanding and characterizing the problem, representing the problem, solving the problem, and reflecting and communicating the solution. Based on this four-dimensional framework, the computer-based assessment has been evaluated with the data of $N = 395$ 10th grade high school students.

Results: Result showed that students' complex problem-solving competence could be modelled by four related but empirically distinct factors with moderate to high intercorrelations. The construct showed substantial relations with fluid intelligence and prior domain knowledge in Chemistry, indicating that construct validity and domain specificity were given. Processes of understanding and characterizing the problem were substantially related to subsequent processes in complex problem solving.

Conclusions: Due to the complexity of complex problem-solving processes in Chemistry, multidimensionality of the construct could be assumed. Consequently, science educators should take into account abilities of understanding, representing, solving the problem, and finally reflecting and communicating the solution when developing instructional approaches and valid computer-based assessments.

Keywords: Chemistry education; Complex problem solving; Computer-based assessment; Interactivity; Validity

Background

The assessment of competences has been shifted towards the use of computer-based procedures (Jurecka 2008). Even in the field of science education, there are different approaches to use computers in order to facilitate the work on complex problems (Jonassen 2004). Van Merriënboer (2013), one of the leading educational researchers in problem solving, stressed the importance of using computers as assessment and instructional tools, as they simulate real-world problems, which are ill-structured and complex in nature.

The major advantage of computer-based tests lies in the assessment of new content areas and constructs (Drasgow and Chuah 2006; Wirth 2008). Furthermore, different kinds of skills such as scientific processing and the ability to design and execute scientific investigations can be fostered with the help of computer-based assessments (CBAs) (Honey and Hilton 2011). Moreover, due to the availability of different representation modes and item formats, students' conceptual understanding of scientific phenomena can be assessed. Furthermore, computer-based assessments enable teachers and researchers to collect different types of data. In addition to traditional scores on multiple-choice or constructed-response items, data on the time needed to perform a number of interactions and the sequence of operations are accessible (Wirth 2008). Hence, researchers have the possibility to design meaningful and motivating real-life scenarios, in which students can solve complex and interactive problems (Funke 2010; Greiff et al. 2013). Additionally, the use of CBAs is advantageous due to test economics, improvements in objectivity, and test reliability (Wirth 2008). But the efficiency of CBA procedures strongly depends on the application of common design characteristics, which determine reliability and validity measures

* Correspondence: ronny.scherer@cemo.uio.no
[1]Centre for Educational Measurement at University of Oslo (CEMO), Faculty of Educational Sciences, Postbox 1161 Blindern, 0318 Oslo, Norway
Full list of author information is available at the end of the article

(Rigas et al. 2002). Also, CBAs need to be based on educational frameworks, taking into account different types of knowledge and cognitive processes (Van Merriënboer 2013). In science education, there is still a need for conceptual models of inquiry- and problem-oriented abilities which could be used to design meaningful and valid assessments (Abd-El-Khalick et al. 2004; Scherer 2012). However, there are difficulties within the evaluation process. Due to a high complexity of computer-based assessments, it might be difficult for test takers to find an optimal or correct solution (Sager et al. 2011). Furthermore, the amount of collected data could become problematic within the evaluation process.

Understanding the cognitive processes, which are involved in problem solving, and analyzing the structure of the problem solving construct have been a challenge in psychology and science education (Greiff et al. 2013; Kind 2013). Consequently, the present study investigates students' complex problem-solving competence (CPS) by taking into account different abilities, which determine their problem solving success.

The present study, first, proposes a theoretical model of CPS in Chemistry, which formed the basis for developing a computer-based assessment. Second, the design and characteristics of the assessment tool are described. By means of item response theory, the CBA is empirically evaluated and tested for construct validity in a third step (Quellmalz et al. 2012). In this regard, we check whether or not a theoretical model, which distinguishes between four components of CPS, is supported by the data. This empirical evaluation is mainly concerned with the dimensionality of the construct. Although there have been approaches to describe *cross-curricular* CPS by empirical means (e.g., Bühner et al. 2008; Kröner et al. 2005), this study incorporates *domain-specific* operationalizations of the construct. We also address the importance of taking into account the many components of designing computer-based assessments in science (Kuo and Wu 2013; Quellmalz et al. 2012).

The paper, thus, contributes to the development of a theory-driven assessment of students' complex problem-solving competences which reflects different components of the construct and provides meaningful insights into the determining factors and the opportunities to model CPS in Chemistry. Such an assessment could be used to provide a detailed feedback for teachers and learners on the different abilities involved in problem solving. Consequently, this approach systematically extends the domain-general problem-solving framework of problem solving within the Programme for International Student Assessment (PISA) studies towards the domain of Chemistry. In light of the upcoming importance of theoretically sound and empirically valid assessment in technology-rich environments (OECD 2013; Quellmalz et al. 2012), we present a framework and an assessment which systematically combine cognitive psychology, science education, and modern assessments.

Problem solving from a psychological perspective

In this study, we focus on the evaluation of complex problem-solving competence in the domain of Chemistry. According to the PISA problem-solving framework (OECD 2004, 2013), this project refers to 'problem solving competence' as

> […] an individual's capacity to engage in cognitive processing to understand and resolve problem situations where a method of solution is not immediately obvious. It includes the willingness to engage with such situations in order to achieve one's potential as a constructive and reflective citizen. (OECD 2013, p. 122).

This definition underlines the term 'competency' by taking into account its domain-specific, contextual, and situational characteristics. In contrast to analytical problem solving, complex problem solving (CPS) is defined by the following characteristics: complexity and connectivity of system variables, temporal dynamics and system changes, interactivity, disclosed structure of the system or the problem situation (intransparency), and polytely in the presence of competing goals (Funke 2010; OECD 2013). Obviously, traditional paper-and-pencil tests are not able to assess CPS due to its dynamic and interactive character (Wirth and Klieme 2004).

While performing a problem-solving process, different kinds of cognitive operations and influences of covariates such as prior knowledge, experience, and motivation come together. For instance, due to the complex character and intransparency, problem solvers must interact with given systems in order to obtain information about variables and their connectivity, and, finally, use these information to solve the problem successfully (Funke 2010). Consequently, feedback and supportive information are needed to reach a given goal state by using problem-solving strategies (Taasoobshirazi and Glynn 2009). These strategies mostly involve controlling for variables in order to obtain information on their effects on the outcome. In doing so, students build up a mental model which represents the structure of variables and their relations (Künsting et al. 2011). This knowledge can subsequently be used to achieve a goal state, representing a problem solution. Finally, after finding an appropriate solution, students must evaluate and communicate their solutions (Kapa 2007; OECD 2004). These (meta) cognitive skills are essential in order to monitor the problem-solving process, and, subsequently, publish the results (Scherer and Tiemann 2012). These processes

and cognitive requirements are essential in CPS and were, at least to some degree, studied in domain-general settings (Greiff et al. 2013; Sonnleitner et al. 2013). Taken together, there are at least four problem-solving components that can be distinguished: exploring and understanding, representing and formulating, planning and executing, and monitoring and reflecting (OECD 2013).

There have been further attempts to describe the structure of cross-curricular problem solving (Funke 2010). For instance, Kröner et al. (2005) operationalized the construct as a measure of intelligence and distinguished between three components: rule identification, rule knowledge, and rule application. In the first step, students have to identify the connections between variables in order to acquire system knowledge. Subsequently, they apply their knowledge to solve a complex problem. In this regard, it has also been investigated whether interactive problem solving could be regarded as a component of intelligence (Danner et al. 2011; Funke and Frensch 2007; Kröner et al. 2005; Leutner 2002). But so far, the results on the relationship between the two constructs have been quite contradictory. The correlations differed according to the types and factors of intelligence (Leutner 2002). Danner et al. (2011) argued that dynamic decision-making, which is often regarded as complex problem solving, significantly correlated with general intelligence but required further abilities. Therefore, they concluded that problem-solving competence was distinct from intelligence, although it determines processes of knowledge acquisition. Accordingly, psychological research on problem solving identified, first, domain-general cognitive processes involved in problem solving, and, second, analyzed the empirical distinction between intelligence and complex problem-solving competence.

Other approaches such as the MicroDYN framework distinguish between three types of competences which are necessary to solve complex problems: model building, forecasting, and information retrieval (Wüstenberg et al. 2012). This approach was implemented for cross-curricular problem solving rather than domain-specific dimensions of the construct. Although the MicroDYN framework captures essential cognitive abilities, it does not reflect educational demands of problem solving. For instance, processes of monitoring, reflecting, and communicating a solution in science are essential parts of the scientific problem-solving process (e.g., Bernholt et al. 2012; Klahr 2000). These components have not been taken into account explicitly in domain-general models such as MicroDYN. Furthermore, the approaches described above have rarely been transferred to complex problem situations in the domain of science, in which prior knowledge and 'strong' solution strategies gain importance (Jonassen 2004).

Problem solving in science

Contextualized and domain-specific assessment procedures have gained importance in many subjects (Funke and Frensch 2007; Jonassen 2004). They are regarded as powerful tools to evaluate problem-solving skills and curriculum-related competences (Koeppen et al. 2008). As some researchers discussed (Gabel and Bunce 1994; Jonassen 2004; Kind 2013), the concept of domain specificity does not only manifest in the effects of domain knowledge on performance but also in specific problem-solving strategies. Scherer and Tiemann (2012) further argued that even knowledge *about* strategies would be domain-specific. It is, thus, indicated that domain-general and domain-specific CPS are related but distinct constructs (Molnár et al. 2013). This argument stressed the need for contextualized assessments (Koeppen et al. 2008) and led, for instance, to the development of simulations in which students could explore structure–property relationships and basic concepts in Chemistry (Cartrette and Bodner 2010). These concepts focused on scientific inquiry and the conceptual understanding in science (Abd-El-Khalick et al. 2004; Flick and Lederman 2006). Moreover, Taasoobshirazi and Glynn (2009) showed that affective and motivational constructs play an important role in problem solving in science. It appears reasonable that students' attitudes towards science affect their success in domain-specific problem solving. These relationships are often mediated by students' goal orientations (Künsting et al. 2011).

Besides domain specificity, different cognitive variables affect the results of problem-solving processes (Kröner et al. 2005). As mentioned previously, the understanding and characterization of the problem requires reasoning abilities and relates to the analytical properties of problem-solving competences. The interpretation of system outputs and information represented by tables, texts, and diagrams allow students to understand complex tasks. Furthermore, the ability to extract and apply information is regarded as one of the key components within the PISA framework of scientific literacy (Nentwig et al. 2009) and is an integral part of science education (Jones 2009). Kind (2013) stressed the importance of these competences as components of scientific reasoning. In his review, he identified different aspects and curricular demands in reasoning and problem solving situations: hypothesizing, experimenting, and evaluating evidence. He also argued that these processes could be regarded as *domain-general*, whereas different types of knowledge involved in scientific reasoning are highly *domain-specific*. In this context, content knowledge and epistemological knowledge about science interfere with solution strategies (see also Abd-El-Khalick et al. 2004). In contrast, domain-general CPS was often found to predict grades and school achievement in specific domains (e.g. Wüstenberg et al. 2012). But these findings might be

due to the strong relationships among reasoning, domain-specific, and domain-general CPS (e.g. Molnár et al. 2013).

Gilbert and Treagust (2009) further argued that the role of adequate problem representations at a macroscopic, microscopic, and sub-microscopic level was significant in chemistry education in order to foster conceptual understanding. Furthermore, Lee (2010) and Lee et al. (2011) supported this argumentation by stressing that developing a mental model that represents the problem structure is crucial for subsequent solution steps. Consequently, one can argue that if students' problem-solving strategies are built upon an adequate representation of their conceptual knowledge in a specific domain, they are more likely to develop expertise (Taasoobshirazi and Glynn 2009).

As mentioned above, the transfer of problem-solving competences into domain-specific areas gains importance, especially in science education because the underlying processes and skills are closely related to scientific inquiry (Friege and Lind 2006; Klahr 2000; Künsting et al. 2011; Schmidt-Weigand et al. 2009). Especially the features of interactivity of experimental systems are powerful tools for the assessment of scientific process skills (Jurecka 2008; Rutten et al. 2011; Kim and Hannafin 2011; Wu and Pedersen 2011). Together with Jonassen (2004), we argue that problem-solving skills are domain-specific and embedded in specific contexts. Van Merriënboer (2013) supported this argumentation and distinguished between 'weak' and 'strong' problem-solving strategies. The latter refer to strategies and sequences of actions which only occur in specific domains or contexts (Gabel and Bunce 1994). However, until now, the issue of domain specificity has not yet been clarified for solving complex and ill-defined problems (Scherer 2012).

To sum up, there are different processes involved in scientific problem solving. These refer to the abilities such as information retrieval, problem representation, model building, strategic behavior, knowledge application, and the phases of scientific inquiry such as planning an experiment and evaluating the results (e.g., Bernholt et al. 2012; Klahr 2000). Based on these and the outcomes of domain-general studies, which provided evidence on fundamental cognitive processes independent from the domain, we propose a model of complex problem solving consisting of four factors (based on Koppelt 2011, OECD 2013 and Scherer 2012): First, students have to understand, characterize, and simplify the problem situation (PUC) in order to build an adequate mental model which is represented in the second step (PR). Based on this model, strategies and methods for solving the problem are developed (PS). Finally, students evaluate and communicate the problem solution (SRC). We note that this model is cyclic and that students could repeat the different steps. A detailed summary of these factors is given in Table 1.

The present study

The purpose of this study is twofold: First, we develop a computer-based problem-solving environment with interactive features according to the proposed model of complex problem solving in Chemistry. Second, we evaluate this tool by checking the fit between the empirical data and the theoretical framework. As a first attempt, we analyze whether or not the differentiation of four problem-solving steps leads to adequate measurement models. We analyze whether the theoretically implied structure of the construct could be represented by the data and, thus, address construct validity

Table 1 Cognitive dimensions of problem-solving processes

Cognitive dimension	Description
Understanding and characterizing the problem (PUC)	-Understanding the situation in which the problem occurs
	-Identifying relevant information
	-Extracting information from scientific texts, tables, and/or figures
	-Developing scientific hypotheses
Representing the problem (PR)	-Creating adequate representations of the problem situation (e.g., mind maps, structural representations of chemical substances)
	-Shifting between different types of problem representations (graphical, verbal, symbolic, and tabular information)
Solving the problem (PS)	-Performing systematic and strategic methodologies in order to achieve a goal state (e.g., by controlling variables in a scientific experiment)
	-Planning and executing scientific investigations and experiments; testing hypotheses
Reflecting and communicating the solution (SRC)	-Evaluating and reflecting problem solutions and scientific evidence
	-Finding alternative solutions
	-Communicating solutions and addressing different audiences (e.g., the scientific community, the public)
	-Distinguishing between scientific and everyday language

(Messick 1995). Finally, relationships among CPS and related constructs such as intelligence and domain knowledge are assessed in order to obtain evidence on external validity. The newly developed computer-based assessment is used to evaluate the theoretical framework of CPS in Chemistry. Our modelling approach provides a combination between the assessment of cognitive abilities of problem solving and educational demands in Chemistry lessons. It therefore contributes to the areas of educational assessment in science and the development of frameworks of higher-order thinking skills.

Methods

Sample and procedure

Participants were 10th grade students attending the German Gymnasium in the federal state of Berlin ($N = 420$; 52.4% female). These students worked on a computer-based assessment of complex problem-solving competence. Additionally, paper-and-pencil tests were administered, which assessed covariates such as prior domain knowledge in Chemistry, fluid intelligence (Gf), and general interest. The pupils' mean age was 15.8 years (SD = 0.7 years) ranging from 14 to 18 years. Some of the students reported that their mother language was not German (23.8%). However, for the present data, Koppelt (2011) has shown that this did not affect the results of the assessment. The assessment procedure was divided into two sessions of 90 min each, which were conducted on two adjacent days, leading to 395 complete data sets.

Measures

Dependent variable: complex problem-solving competence

The following section provides a detailed description of the CBA's design characteristics and shows examples of items and indicators which were used to assess the four dimensions of CPS.

Development of a computer-based assessment of CPS The computer-based assessment was implemented with the easy-to-use developmental environment *ChemLab-Builder* (Meßinger 2010). In this environment, one can define various laboratory tools such as chemical substances, machines for syntheses or analyses, and different forms of information materials. The tool requires the operationalization of inputs, outputs, and the data which is shown in the resulting log files. Furthermore, the degree of interactivity can be adapted according to the number of variables and their relationships.

In order to assess complex problem-solving competence, different tasks were implemented which could be assigned to one of the four problem-solving steps presented by Koppelt (2011) (see 'Problem solving in science' section). In these tasks, students were able to solve single items independently from their performance on previous ones. After completing two evaluation-free exploration phases of 10 min each, students, first, had to identify unknown chemicals by using an analysis and a synthesis machine (for a discussion on the importance of exploration phases in CPS assessments, see Leutner et al. 2005). Second, students had to identify and synthesize a flavoring substance (in this case, methyl butyrate) fulfilling different criteria. In order to design an attractive and motivating problem-solving environment, the task was embedded into a contextual framework.

Within the computer-based assessment, there were different machines representing complex systems. Students had to interact with these systems during the problem-solving process in order to obtain information about their functionalities. This design feature requires the acquisition of system knowledge, which is crucial for solving complex problems (Goode and Beckmann 2010; Sonnleitner et al. 2013). As the relationships between dependent and independent variables were disclosed at the beginning (Funke 2010), the systems allowed different adjustments to identify the number of correlated variables and their complexity.

In order to utilize system interactivity, the identification of unknown chemical substances had to be accomplished with the help of analytical spectra. This kind of supportive and indirect feedback was necessary in order to foster students' interactions with the systems. Nevertheless, system interactivity played an important role in assessing CPS and in simulating domain-specific problems by computational means (Jonassen 2004; Scherer and Tiemann 2012). Due to the administration of different subtasks, which referred to the problem-solving steps, students had to focus on the main task in the presence of competitive goals. This program feature referred to goal orientation and polytely (Blech and Funke 2010).

Moreover, students had to overcome some difficulties within the environment: The name of the chemical substance was unknown, and students had to suggest a plausible synthesis. Furthermore, the reaction yield had to be optimized. The problem-solving task was, thus, complex and constructed in a way that students with low prior knowledge would also be able to solve the task successfully (Scherer and Tiemann 2012).

To sum up, the main characteristics of complex problem-solving environments were taken into account within the test development procedure (see, for example, Funke 2010). However, due to our focus on a fixed chemical system with defined variables, temporal dynamics have not been implemented. Accordingly, our environment was based on curricular demands of teaching the concepts of chemical equilibria and the chemistry of esters in grade 10, which did not refer to time-varying settings.

Measurement of 'understanding and characterizing the problem' The evaluation of students' performance on 'understanding and characterizing the problem' included

the analysis of unknown chemical substances. Students' responses were evaluated against a direct solution (variables *PUC01* to *PUC07* in Table 2).

To identify the substances, students had to consider the information given in the lab books ('spectra' and 'molar mass'), extract relevant properties, and, finally, relate them to each other in order to suggest the substance's name. For example, a chemical substance showed a molar mass of 46 g/mol. This substance could either be ethanol or formic acid. Only by the determination of the substance class with the help of the spectral output and information depicted in a table, students were able to suggest the correct name. Figure 1 shows a screenshot of this subtask.

The items of PUC varied in difficulty according to the amount of information needed for the identification process (see Table 2). The evaluation of the students' answers was applied by using computer-based means. Each of the seven items (i.e., unknown substances) was dichotomously coded, which resulted in a maximum score of seven.

Measurement of 'representing the problem' The computer environment enabled us to develop interactive and static tasks. In order to provide measures of representing the problem (PR), we implemented static items with multiple-choice options, in which students had to complete concept maps or tables with different formats of structural representations of chemical substances. The resulting items were polytomously coded (for further information, refer to Koppelt 2011 and Scherer 2012).

Measurement of 'solving the problem' To analyze the performance on 'solving the problem (PS)', the achievement of a goal state (*PS01a*, *PS01b*) and students' general problem-solving behavior (*PS02a* to *PS04*) were examined. In this operationalization, the goal state referred to an optimal solution, which could be achieved by different types of system settings. Subsequently, the students' solution was directly compared with an optimal goal state. Table 3 contains examples of variables measuring PS, which were evaluated by analyzing the students' log files.

Within this record of problem-solving behavior, a separate table was given that contained a list of analyses and syntheses. With the help of these entries, it was possible to determine the number of replicated analyses and syntheses. As students' activities in the laboratory as well as system inputs and outputs were recorded, they were given the opportunity to monitor or recall settings and outputs during the tasks. We consequently coded the *PS02* items to zero if an analysis or synthesis was applied more than twice. This feature was in line with the operationalization of systematic problem-solving behavior proposed by Künsting et al. (2011). Moreover, if students performed syntheses with substances, which were not identified previously, the entries of 'syntheses (student's view)' appeared with a question mark, and the variable *PS04* was coded to zero (see Figure 2).

Taken together, the PS items required different cognitive processes such as tactical and goal-oriented actions, the achievement of required conditions, and the systematic variation of factors (Scherer and Tiemann 2012). Further aspects of measuring PS are reported in Koppelt (2011).

Measurement of 'reflecting and communicating the solution' In order to operationalize this step, we developed static tasks, which assessed the various abilities of reflecting and communicating the solution (SRC). First, the students had to answer multiple-choice items, in which they had to recall their solution strategies. Second, they had to choose among given representations of the problem solution (e.g., a journal article) and decide whether they were appropriate for different audiences (for further information on this approach, see Bernholt et al. 2012). Again, the items were dichotomously and polytomously coded.

Log file data For each student, a single log file has been obtained, which contained the following information: responses on multiple-choice, multiple-select, and constructed response items (provided as numbers or nominal entries), time needed to solve the task, number of actions, a list of chemical substances used in the analyses and syntheses, the sequence of action within the computer-based assessment, and the sequence of analyses and syntheses. To filter these raw data, we set up variables that were assigned to measure the abovementioned factors of CPS. In doing so, the entries on constructed response items and sequences of actions have been coded according to their correctness and efficiency. For further details, refer

Table 2 Examples of variables measuring the competence of 'understanding and characterizing the problem (PUC)'

Variable	Analysis of	Description	Coding
PUC01-02	Methanol and acetic acid	Analyzing an unknown substance by taking into account system outputs (spectra, molar mass) and additional information given by tables	0, 1 each
PUC03-06	Ethanol, pentanol, formic acid, and butyric acid	Analyzing an unknown substance by taking into account system outputs (spectra, molar mass) and further information given by tables	0, 1 each
		Different types of information have to be combined (tabular and graphical sources)	
PUC07	Sulfuric acid	Analyzing an unknown substance with information that is explicitly given (molar mass)	0, 1

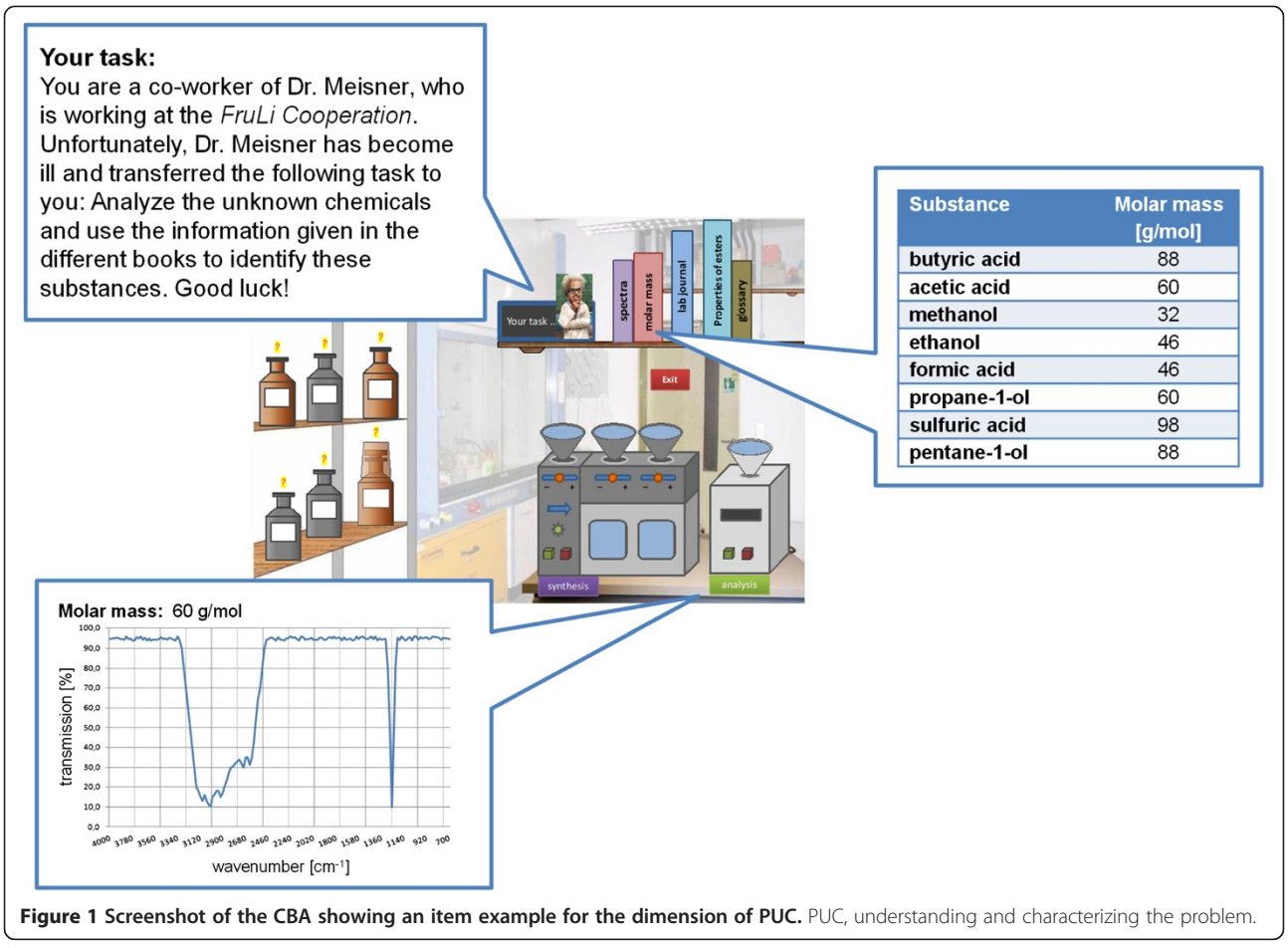


Figure 1 Screenshot of the CBA showing an item example for the dimension of PUC. PUC, understanding and characterizing the problem.

to Koppelt (2011), Scherer (2012), and Scherer and Tiemann (2012).

Quality of the coding scheme In order to check the coding scheme for explicitness and reliability, 20% of the log files were coded by two independent raters. By these means, objectivity of the coding procedure has been ensured. We determined Cohen's κ as a measure for interrater reliability and attained statistically significant values ($p < 0.05$) ranging from 0.85 to 1.00 across the four steps. Values of 1.00 occurred due to the fact that selected tasks have been analyzed by computer-based means only. Consequently, there was no need for further interpretation of students' responses in these tasks.

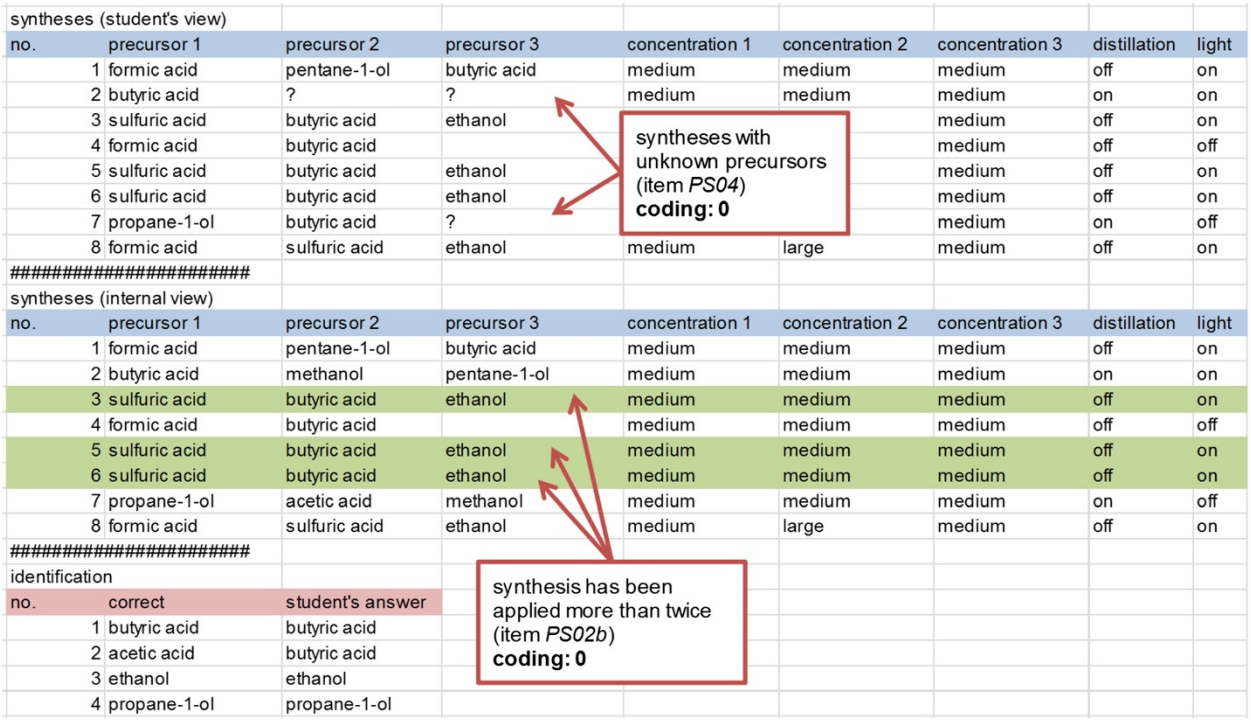
Table 3 Examples of variables measuring the competence of 'solving the problem (PS)'

Variable	Description	Coding
PS01a	Choosing three correct substances in order to synthesize the required substance	0, 1, 2, 3
PS01b	Choosing the correct system settings of concentration and distillation in order to maximize the reaction yield	0, 1, 2
PS02a	Replicating analyses only once	0, 1
PS02b	Replicating syntheses only once	0, 1
PS03	Applying optimization steps of the reaction yield for meaningful reactions only	0, 1
PS04	Synthesizing substances with previously identified and analyzed precursors only	0, 1

Covariates of complex problem-solving competence

In order to investigate the external validity of the CPS assessment, we administered tests on related constructs of CPS. This approach is common in evaluating discriminant validity and provides information on the uniqueness of constructs and the quality of the assessment tools. In this context, validity is referred to as a test characteristic, supporting the interpretation of the relationships between test scores and empirical evidence (Messick 1995). Reliabilities and descriptive statistics of the tests on covariates are reported in the Table 4.

First, a *domain-specific prior knowledge test* was administered, which comprised three scales in different content areas: the chemistry of esters, structure–property relationships, and the nature of chemical equilibria. The test consisted of 22 multiple-choice items, which



syntheses (student's view)

no.	precursor 1	precursor 2	precursor 3	concentration 1	concentration 2	concentration 3	distillation	light
1	formic acid	pentane-1-ol	butyric acid	medium	medium	medium	off	on
2	butyric acid	?	?	medium	medium	medium	on	on
3	sulfuric acid	butyric acid	ethanol			medium	off	on
4	formic acid	butyric acid				medium	off	off
5	sulfuric acid	butyric acid	ethanol			medium	off	on
6	sulfuric acid	butyric acid	ethanol			medium	off	on
7	propane-1-ol	butyric acid	?			medium	on	off
8	formic acid	sulfuric acid	ethanol	medium	large	medium	off	on

syntheses with unknown precursors (item *PS04*) **coding: 0**

######################

syntheses (internal view)

no.	precursor 1	precursor 2	precursor 3	concentration 1	concentration 2	concentration 3	distillation	light
1	formic acid	pentane-1-ol	butyric acid	medium	medium	medium	off	on
2	butyric acid	methanol	pentane-1-ol	medium	medium	medium	on	on
3	sulfuric acid	butyric acid	ethanol	medium	medium	medium	off	on
4	formic acid	butyric acid		medium	medium	medium	off	off
5	sulfuric acid	butyric acid	ethanol	medium	medium	medium	off	on
6	sulfuric acid	butyric acid	ethanol	medium	medium	medium	off	on
7	propane-1-ol	acetic acid	methanol	medium	medium	medium	on	off
8	formic acid	sulfuric acid	ethanol	medium	large	medium	off	on

######################

identification

no.	correct	student's answer
1	butyric acid	butyric acid
2	acetic acid	butyric acid
3	ethanol	ethanol
4	propane-1-ol	propane-1-ol

synthesis has been applied more than twice (item *PS02b*) **coding: 0**

Figure 2 Lab journal taken from a student's log file.

were dichotomously scored. The resulting sum score was regarded as a measure of students' prior domain knowledge.

Second, the students had to work on a test of *fluid intelligence*, which contained 36 items. These referred to a figural, a numerical, and a verbal scale of 12 items each. The students had to work on figural and verbal analogies (e.g., 'apples:juice = potatoes:?') as well as mathematical reasoning problems (for details, see Schroeders et al. 2010). All items were dichotomously scored and comprised to a general factor (Gf).

Finally, we checked the students' general interest by administering two scales of the *AIST* test (German: *Allgemeiner Interessens-Struktur-Test*; Bergmann and Eder 2005). We chose the realistic (AIST-r) and the investigative (AIST-i) scales for further analyses, as they reflect facets of scientific interest with acceptable psychometric properties. The test on general interest contained 20 statements, which were ranked on a four-point Likert scale and summed up to a final score for each scale (0 = *I totally disagree* to 3 = *I totally agree*).

Data analyses

Application of probabilistic measurement models

To estimate students' problem-solving abilities, raw scores were scaled with the help of item response theory (IRT) models, which are implemented in the software package *ACER ConQuest 2.0* (Wu et al. 2007). These models are advantageous in modelling multidimensional competences because they allow direct comparisons between competing models and show the relationship between item difficulties and person ability parameters. They have become prominent in test development and science education (Neumann et al. 2011). We chose the IRT analysis as a state-of-the-art approach in modelling complex and multidimensional constructs (Bond and Fox 2007). The IRT models are appropriate in modelling categorical data without the assumption of normally

Table 4 Descriptive statistics and correlations of tests on covariates

Variables	N_{Items}	*M*	SD	Min	Max	*α*	Correlations		
							1.	2.	3.
1. Domain-specific prior knowledge	22	6.53	3.18	0	18	0.73	1.00		
2. Fluid intelligence (Gf)	36	12.03	5.53	0	30	0.84	0.30**	1.00	
3. General interest (AIST)	20	43.60	16.37	0	74	0.94	0.02 ns	−0.15***	1.00

Correlations were based on the sum scores. N_{Items}, number of items; Min, Max, achieved minimum/maximum; ns, statistically insignificant ($p > 0.05$); *α*, Cronbach's alpha; Gf, general factor; AIST, Allgemeine Interessens-Struktur-Test. **$p < 0.001$; ***$p < 0.01$.

distributed variables and allow for non-linear relationships between latent variables and items (Wirth and Edwards 2007). Following the modelling approaches of problem solving and inquiry-based abilities proposed by many science educators and educational psychologist (e.g., Kuo and Wu 2013; Neumann et al. 2011; Scherer 2012; Quellmalz et al. 2012; Wüstenberg et al. 2012), these analyses appeared appropriate in applying a confirmatory analysis of dimensionality, especially because of the categorical nature of students' scores (Bond and Fox 2007). Since items were dichotomously and polytomously scored, the partial credit model has been used in this study. This model accounts for thresholds between categories and can be generalized to multidimensional analogues, interpreting person abilities and item difficulties as multidimensional vectors. Additionally, the structure of the variance-covariance matrix was taken into account. We further note that item dependencies within the students' responses were neglectable (for further information, refer to Scherer 2012).

The IRT scaling was applied in two steps: The first step included an analysis based on all administered items with the aim of identifying items which did not fit the IRT model. Uni- and four-dimensional models were used to investigate the latent structure of the construct and reliabilities of each factor (see Figure 3).

Subsequently, only items with a moderate discrimination, a weighted mean square value (wMNSQ) between 0.75 and 1.33 and an absolute t value lower than 1.96, were included in the second step (Adams and Khoo 1996). Again, the uni- and four-dimensional models were applied to the data. The resulting person and item parameters were used to obtain descriptive statistics. In this procedure, the Markov chain Monte Carlo algorithm for multidimensional IRT models has been applied (Wu et al. 2007).

To evaluate model fit, common information criteria such as the Akaike's Information Criterion (AIC) and the Bayesian Information Criterion (BIC) were taken into account. These indexes are given as follows:

$$\mathrm{AIC} = \mathrm{dev} + 2n_\mathrm{p}$$
$$\mathrm{BIC} = \mathrm{dev} + \log(N) \cdot n_\mathrm{p},$$

where 'dev' represents the final deviance of the model, N the sample size, and n_p the number of parameters for model estimation. Models with smaller values of AIC and BIC are empirically preferred. In order to compare competing models, the information criteria and a χ^2 likelihood ratio test of final deviances were used.

Referring to our theoretical assumption of four problem-solving factors, we only tested whether or not the four-dimensional model outperformed the model with one factor. It would have been possible to test other models of CPS with fewer dimensions, but we focused on the validation of the proposed model with four factors because there was no conceptual and empirical evidence for combining or differentiating these steps (Koppelt 2011; OECD 2013; Scherer 2012).

Structural equation modelling

In order to analyze the relationships between CPS and related constructs, we established structural equation models, which were analyzed in *Mplus 6.0* (Muthén and Muthén 2010). In these analyses, the person parameters of the IRT scaling procedure have been used as indicators of CPS. Within the estimation process, missing

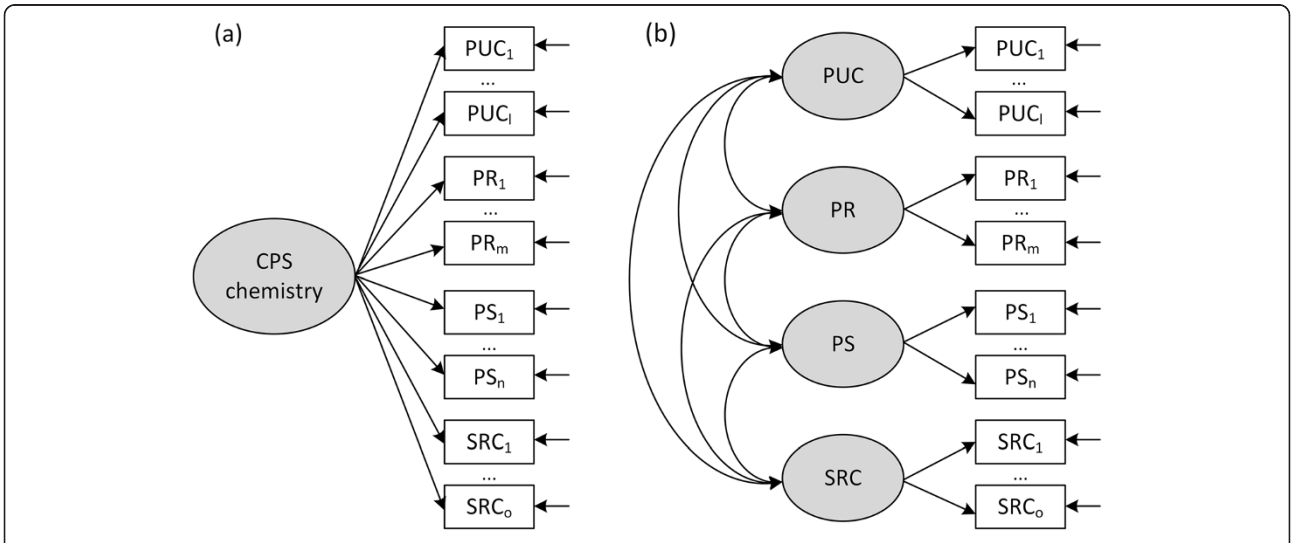


Figure 3 Two competing model structures. (a) Single-factor model and **(b)** faceted model with four correlated factors representing the problem-solving steps. Letters *l*, *m*, *n*, and *o* denote the numbers of items in each of the four steps.

values were handled by applying the full information maximum likelihood procedure (FIML) (Enders 2010). According to Little's MCAR test, the data were likely to follow the missing completely at random mechanism (χ^2 (309) = 335.16, $p = 0.15$), which legitimized the use of FIML. In this modelling approach, CPS served as the outcome and covariates as predictors. We have finally chosen structural equation modelling (SEM) to obtain a quite parsimonious model that reflects the relations among CPS and covariates. In light of the relatively small sample size, alternative approaches such as IRT modelling with latent regression could have yielded more biased estimates.

Results

In this section, we address the result in relation to our research goals. First, the results on the IRT scaling are presented. Second, the relationships between CPS and related constructs are described, obtaining information on the external validity of the CPS test.

IRT scaling outcomes

Item selection and estimation of reliability

One of the prerequisites of IRT analyses is that the categorical items are locally independent from each other (Bond and Fox 2007). Therefore, we conceptually checked whether or not the solutions of two adjacent tasks affected each other. After this evaluation process, 32 out of 39 independent items resulted, which were used for further analyses.

Further results on unidimensional IRT scaling of the 32 items suggested that only 1 item of the PUC step did not meet the item fit criteria (wMNSQ value above 1.33; Bond and Fox 2007) and was therefore excluded. The exclusion of this item led to a significant improvement in final deviance and favored the IRT model with 31 items (model with 32 items: final deviance dev = 18,085.76, n_p = 56; model with 31 items: dev = 17,344.36, $n_p = 53$; model comparison: Δdev = 741.1, $\Delta n_p = 3$, $p < 0.001$). Since the remaining items showed a sufficient fit (wMNSQ values between 0.79 and 1.26, absolute t values below 1.96), we accepted the model with 31 items. In item response theory modelling, one assumption refers to the local independence of items (Bond and Fox 2007). Major violations of this assumption could lead to biased estimates of item difficulties, thresholds, and further model parameters (Wainer et al. 2007). In the present study, we addressed this issue in two ways: First, we designed items that did not necessarily require correct responses from previous items (Koppelt 2011). Second, Scherer (2012, 2014) quantified item dependencies and showed that they were neglectable for the present assessment.

The application of the unidimensional partial credit model revealed a sufficient expected *a posteriori* over persons variance (EAP/PV) reliability of 0.83. Furthermore, the internal consistency, which is based on assumptions of classical test theory, was acceptable for this model ($\alpha = 0.79$). However, these data did not provide information on the measurement accuracy for each problem-solving step. Therefore, we conducted an IRT analysis by establishing a four-dimensional partial credit model and checked for EAP/PV reliability of these four scales. Table 5 contains the scaling outcomes of this analysis. We note that we constrained the means of item parameters of each scale to zero in order to identify the scale of person abilities (Bond and Fox 2007). In this context, students showed lowest performance in PUC and PS, whereas they performed better in representing the problem and communicating the solution. By and large, the ability distributions were broad.

The PUC and PS scales showed acceptable reliabilities above 0.70, whereas the PR and SRC scales provided reasonable values of 0.65. However, we argue that these values were substantial in order to assess a quite complex construct which is composed of further factors (Brunner and Süß 2005; Yang and Green 2011).

Structure of CPS

By establishing uni- and four-dimensional models of CPS, we evaluated the differences in model fit criteria (Table 6). A χ^2 likelihood ratio test of final deviances (dev) was applied to test for significant differences between the models. Again, the faceted model with four correlated dimensions outperformed the single-factor approach, as the difference in the final deviances was statistically significant (Δdev = 595.17, $\Delta df = 9$, $p < 0.001$).

Information criteria supported this finding: The AIC value of the unidimensional model was greater than the AIC of the four-dimensional model ($AIC_{1dim} > AIC_{4dim}$). Exactly the same relation was found for the BIC, which took into account the sample size: $BIC_{1dim} > BIC_{4dim}$. Both indices favored the CPS model with four separable steps. Finally, this model has been accepted.

Furthermore, the latent correlations between the four problem-solving factors were statistically significant and ranged between low and moderate values (Table 7). The strongest relationship has been found between PUC and SRC, followed by PS and SRC. The lowest value occurred for the dimensions of PR and PS.

Relationships among CPS and covariates

In order to analyze the relationships between CPS and related constructs, we established measurement models within a structural equation framework with CPS as a latent variable, measured by the four scales PUC, PR, PS, and SRC. We used person parameters as indicators of students' performance on these scales. By introducing fluid intelligence, domain-specific prior knowledge, and general

Table 5 Descriptive statistics of item and person parameters

Descriptive statistics	Item parameters					Person parameters				
	PUC	PR	PS	SRC	SUM	PUC	PR	PS	SRC	SUM
M	0.00	0.00	0.00	0.00	0.00	−0.45	0.12	−0.47	0.60	−0.04
SD	0.77	1.15	0.89	1.14	0.87	1.11	1.51	1.70	0.95	0.72
Min	−0.83	−1.49	−1.37	−1.91	−2.56	−3.17	−3.81	−3.70	−2.14	−2.01
Max	1.14	2.32	1.51	2.77	1.56	3.30	4.01	2.52	3.96	1.89
EAP/PV reliability	0.71	0.65	0.76	0.65	0.83					
N_{Items}	7	7	11	6	31					

SUM, overall problem-solving performance; EAP/PV, expected *a posteriori* over persons; N_{Items}, number of items. In order to identify the scale, the means of the item parameters were constrained to zero.

interest, we checked for their effects on CPS. This approach has the major advantage of correcting from measurement error (Rigas et al. 2002). In this test, fluid intelligence was measured by three scales which are based on the manifest raw scores. Finally, we specified a regression model in which CPS ability was the criterion, fluid intelligence, prior knowledge, and general interest (AIST) were the predictors (Figure 4).

In order to check whether or not this model represented the data, we used different fit indexes such as the χ^2 value, the comparative fit index (CFI), the root mean square error of approximation (RMSEA), and the standardized root mean square residual (SRMR). The models show a reasonable fit in the case of a CFI above 0.90, an RMSEA below 0.08, and a SRMR below 0.09 (Hu and Bentler 1999). The resulting model showed a reasonable goodness of fit ($\chi^2(49, N = 395) = 108.91$, $p < 0.001$, $\chi^2/df = 2.22$; CFI = 0.93, RMSEA = 0.06, SRMR = 0.05) and was thus accepted. In this model, only prior domain knowledge and Gf showed substantial regression weights, whereas interest (AIST) did not significantly predict CPS. In sum, 18% of variance in CPS could be explained, whereby prior knowledge in Chemistry showed larger effects than Gf. Interestingly, fluid intelligence and interest were negatively associated, meaning that students with higher values of fluid intelligence showed lower interest in realistic and investigative actions.

Table 6 Final deviances and information criteria of the uni- and the four-dimensional partial credit model

Model	Unidimensional model	Four-dimensional model
Final deviance (dev)	17,344.36	16,749.19
Number of parameters (n_p)	53	62
AIC	17,450.36	16,873.19
BIC	17,481.98	16,910.18

AIC, Akaike's Information Criterion; BIC, Bayesian Information Criterion.

Discussion

Test development and the structure of complex problem solving

The present study focused on the development of a computer-based assessment of CPS, which was based on a theoretical framework. In this regard, a coding scheme has been proposed which referred to the four problem-solving components and enabled us to score the data of students' log files. Moreover, the evaluation of these log files was substantially objective and resulted in a reliable IRT scale. However, as students had to cope with different virtual devices and tasks, the assessment procedure was quite complex. Therefore, the psychometric properties could be affected by interactions between different assessment modes, item formats, or biases deriving from missing data (Bond and Fox 2007; Enders 2010). In future research, these problems need further attention and might be modelled explicitly (Scherer 2012). Also, due to the complex operationalization of the problem-solving steps, there is a need for in-depth analyses of their structure. It should be investigated whether different competences are subsumed by each step or the steps are strictly unidimensional. A further differentiation in the model structure might be beneficial (Kröner et al. 2005; Scherer 2012; Sonnleitner et al. 2013).

The computer-based assessment was implemented for one content area of the German National Curriculum for Chemistry, namely, esters and organic compounds, and was evaluated for a sample of $N = 395$ students. Therefore, it is of importance to validate our CPS model with further content areas in order to check for content specificity by empirical means. Further research should also focus on the transfer of the assessment procedure to different age groups with a much greater sample size.

In our study, we also focused on the evaluation of construct validity. In this regard, we first analyzed the dimensionality of CPS by referring to two different assumptions: (a) CPS as a single factor (unidimensionality) and (b) CPS as a construct which comprises four separable factors (multidimensionality). In order to address

Table 7 Latent correlations among the four problem-solving components

	1.	2.	3.	4.
1. PUC	1.00	0.55***	0.57***	0.78***
2. PR		1.00	0.20***	0.43***
3. PS			1.00	0.60***
4. SRC				1.00

***$p < 0.001$.

this aspect, model fit criteria of partial credit models were evaluated and compared. It was found that the four-dimensional model outperformed the model with a single factor. Therefore, the conclusion is eminently reasonable that CPS can be regarded as a competence consisting of multiple factors, which represent the four problem-solving steps (Koppelt 2011; Scherer 2012). Although the multidimensional framework has been supported empirically, we argue that other frameworks of CPS could also work better than the unidimensional approach. For instance, Scherer (2014) was able to show that two- and three-dimensional models of CPS, which were based on the 'problem solving as dual search' assumption (Klahr 2000), significantly outperformed the unidimensional model. However, the four-dimensional model was superior to less dimensional approaches. Furthermore, Koppelt (2011) clarified that from a theoretical and conceptual perspective, a further differentiation of the problem-solving factors would lead to unreliable fits of items to these factors. To this extent, we regard our model as appropriate for describing the structure of CPS in domain-specific settings.

Moreover, the relationships between the four latent factors indicated their empirical distinction. Interestingly, the steps of PUC and SRC showed the highest correlation, meaning that processes of understanding the structure of a problem are strongly associated with processes of reflecting and evaluating a solution. This relationship appears reasonable, as students need to recall criteria of an appropriate solution and activate their knowledge about the problem (Jones 2009). Also, solving the problem and evaluating a solution showed a moderate correlation. Again, this finding appears reasonable, as evaluating a solution against given criteria requires that a solution has been generated previously. In terms of metacognition, the processes of monitoring and elaboration are related to PS and SRC (Kapa 2007; Scherer and Tiemann 2012). Unexpectedly, the dimensions of PR and PS showed the lowest correlation. This finding might be due to the design of items in the computer-based assessment. Items referring to PR required students to, first, shift between different levels of representation (e.g., transforming a structural formula into a sum formula) and, second, build representations of the problem structure (e.g., by using concept maps). As these items do not necessarily involve aspects of systematicity or the strategy of controlling variables, their direct relation to solution processes was comparably weak. However, further attention on this relationship is needed, as current research on domain-general CPS identified the two processes as strongly associated (e.g., Wüstenberg et al. 2012).

Furthermore, it can be concluded that the process of understanding and characterizing the problem is strongly associated with subsequent problem-solving steps. Jones (2009), Goode and Beckmann (2010), and Greiff et al. (2013) argued that building an adequate mental model of the complex system by understanding the relationships among variables is crucial for solving the complex problem. Hence, the empirical data obtained from the present study supported this argumentation. Although low to moderate correlations were found, the structure of CPS could confound with a general factor, which underlies the

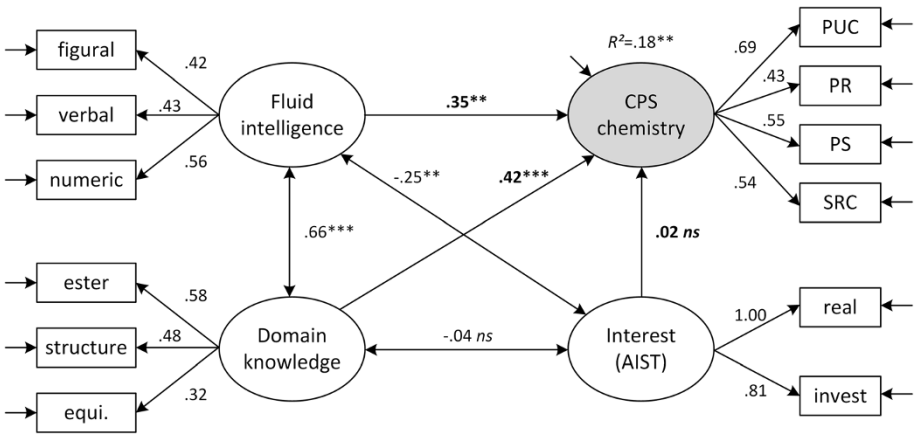


Figure 4 Regression model with correlated predictors of CPS. equi., equilibrium; real, realistic AIST scale; invest, investigative AIST scale; numeric, numerical scale of fluid intelligence. ns, statistically insignificant ($p > 0.05$). **$p < 0.01$, ***$p < 0.001$.

four latent dimensions. Therefore, further analyses are necessary in order to validate the higher-order structure of our model (Scherer 2012; Sonnleitner et al. 2013).

External validation of the CPS assessment

In our study, we were able to replicate the findings on the relationship between intelligence and complex problem solving by obtaining a latent correlation of $\rho = 0.57$ which was comparable to the findings obtained by Bühner (2008) and Kröner et al. (2005). Although Rigas et al. (2002) found partial correlations between fluid intelligence and CPS performance between 0.38 and 0.51, they argued that there was a unique competence which accounted for the specificity of computer-based assessments. Furthermore, they concluded that general CPS measures would become more similar to intelligence tests if researchers changed the characteristics of CBAs (Funke 2010). We, finally, argue that intelligence has a moderate effect on CPS performance, but both constructs could be separated. This result is in line with research conducted by Danner et al. (2011) and Scherer (2012). However, different forms of intelligence could be taken into account in future research.

Moreover, we argue that students' complex problem-solving ability was related to their prior domain knowledge. This result underpins the significance of knowledge acquisition by interacting with a given system (Goode and Beckmann 2010; Sonnleitner et al. 2013). Although the assessments were designed in such a way that students with low prior knowledge could successfully solve the problem, prior knowledge on the chemical concepts, implemented in the CBA, was beneficial (supporting Friege and Lind 2006; Hambrick 2005; Schmidt-Weigand et al. 2009). To some degree, this finding indicates the domain dependence of CPS. To sum up, our findings suggested that intelligence and prior knowledge were determining factors of CPS. The effect of prior domain knowledge is often underestimated in problem solving (Jonassen 2004) but revealed moderate relationships in our study.

As considered, the analysis of the relationships between CPS and related constructs was conducted by using person parameters which resulted from IRT scaling. By establishing latent variables, the resulting correlations were corrected from measurement error. These analyses were considered as adequate for the investigation of discriminant validity (Kuo and Wu 2013). Furthermore, model fit indexes were obtained, which provided additional information on how well the data represented the proposed theoretical framework. But due to high and statistically significant correlations among CPS, intelligence, and prior domain knowledge, further models were established in order to control for the relationships among the covariates of CPS. The resulting regression model revealed that domain knowledge was the strongest predictor of CPS in Chemistry, underlining the importance of knowledge about the system and Chemistry as a domain (e.g., Jonassen 2004; Koppelt 2011). In this model, the effect of fluid intelligence was strong, as expected and proposed by previous research (e.g., Funke 2010). Again, general interest, as measured by an investigative and a realistic scale, did not show significant regression coefficients on CPS, indicating that motivational variables do not necessarily play an important role in solving complex problems in Chemistry. This is in contrast to the argumentation of science educators such as Taasoobshirazi and Glynn (2009) who proposed strong effects of students' attitudes towards science in problem solving. However, we argue that previous results on this relationship were mainly focused on *analytical* and *static* problems, whereas our study used *complex* and *interactive* problems with different task characteristics (Wirth and Klieme 2004). In computer-based scenarios, it appears more likely that students already have a certain level of interest which subsequently leads to a weaker relationship with performance (Jonassen 2004). Again, further research on the different effects of interest and motivational variables on CPS is necessary.

Another issue that needs to be addressed in future research is that students' personal background (e.g., the mother language) could affect the results of the present assessment. As discussed by Scherer (2012), measurement invariance across different subgroups of students might be compromised by differential item functioning. It would consequently be desirable to investigate these effects for a larger and more representative sample of German students.

Conclusions

As a conclusion, our model of complex problem solving with four factors represents a theoretical framework which describes the complex structure of CPS. We exemplarily showed how this framework could be transferred to specific tasks and item responses which subsequently led to appropriate measurement models (Kuo and Wu 2013). The underlying construct map served as a guideline for developing the computer-based assessments (Kuo and Wu 2013; Pellegrino 2012). CPS could, thus, be assessed by taking into account all four steps in order to investigate students' strengths and weaknesses within the problem-solving process. Consequently, we argue that differentiating into the factors of CPS yields more diagnostic information than a unidimensional approach. Finally, computer-based assessments are powerful measurement tools, but researchers must keep an eye on psychometric properties within the process of test development in order to establish valid and meaningful assessments (Quellmalz et al. 2012; Wüstenberg et al. 2012).

The results on the relationships among covariates and CPS in Chemistry provided evidence on construct validity (Messick 1995). Based on our framework, which systematically combined approaches of scientific inquiry and psychological theories of problem solving, the theoretical assumptions on the structure of CPS and the relationships with covariates such as intelligence and domain knowledge have been confirmed. Hence, fostering problem-solving abilities requires knowledge acquisition and the development of reasoning abilities (Kuhn 2009). Furthermore, students' abilities to understand the complex problem (PUC) and find an appropriate solution by applying systematic strategies (PS) are crucial in structuring problem-solving processes. It might therefore be beneficial to specifically enhance these two factors (Kim and Hannafin 2011; Klahr 2000).

The present study provided a new computer-based assessment of complex problem-solving competence in Chemistry, which was developed by taking into account domain-general and domain-specific processes of problem solving. This assessment could be used in science classrooms in order to evaluate students' competences and to give specific feedback to learners (Kim and Hannafin 2011; Quellmalz et al. 2012). As the domain-general PISA 2012 assessment of problem solving was based on processes similar to those described in our study (OECD 2013), the CBA could add more diagnostic information for students' competences in science. Using such an assessment could also foster the incorporation of computers in science. It also shows how evidence-centered assessments and model-based learning approaches could be combined for the construct of complex problem solving (Kuo and Wu 2013; Quellmalz et al. 2012). Also, the proposed model of problem solving could serve as a teaching tool which forms a guideline for science lessons and developing instructional material (Van Merriënboer 2013).

Abbreviations

AIC: Akaike's information criterion; AIST: Allgemeiner-Interessens-Struktur-Test (*German* general test on the structure of interest); BIC: Bayesian information criterion; CBA: computer-based assessment; CFA: confirmatory factor analysis; CFI: comparative fit index; CPS: complex problem solving; dev: final deviance; EAP/PV: expected *a posteriori* over persons variance; FIML: full information maximum likelihood; Gf: general factor (of fluid intelligence); IRT: item response theory; MCAR: missing completely at random; PR: representing the problem; PS: solving the problem; PUC: understanding and characterizing the problem; RMSEA: root mean square error of approximation; SRC: reflecting and communicating the solution; SRMR: standardized root mean square residual.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JM-K and RS participated in the design of the study, performed the statistical analyses, and drafted the manuscript. RT conceived the study, and participated in its design and coordination. All authors read and approved the final manuscript.

Author details

[1]Centre for Educational Measurement at University of Oslo (CEMO), Faculty of Educational Sciences, Postbox 1161 Blindern, 0318 Oslo, Norway. [2]Joachim Herz Stiftung, Langenhorner Chaussee 384, 22419 Hamburg, Germany. [3]Department of Chemistry, Humboldt-Universität zu Berlin, Brook-Taylor-Str. 2, 12489 Berlin, Germany.

Received: 16 October 2013 Accepted: 29 January 2014
Published: 27 August 2014

References

Abd-El-Khalick, F, Boujaoude, S, Duschl, R, Lederman, NG, Mamlok-Naaman, R, Hofstein, A, Niaz, M, Treagust, D, & Tuan, H-L. (2004). Inquiry in science education: International perspectives. *Science Education, 88*, 397–419.

Adams, RJ, & Khoo, ST. (1996). *ACER Quest [computer software]*. Melbourne: ACER.

Bergmann, C, & Eder, F. (2005). *Allgemeiner Interessen-Struktur-Test mit Umwelt-Struktur-Test (UST-R) – Revision (AIST-R)*. Göttingen: Beltz.

Bernholt, S, Eggert, S, & Kulgemeyer, C. (2012). Capturing the diversity of students' competences in science classrooms: Differences and commonalities of three complementary approaches. In S Bernholt, K Neumann, & P Nentwig (Eds.), *Making it tangible – learning outcomes in science education* (pp. 173–200). Münster: Waxmann.

Blech, C, & Funke, J. (2010). You cannot have your cake and eat it, too: How induced goal conflicts affect interactive problem solving. *The Open Psychology Journal, 3*, 42–53.

Bond, TG, & Fox, CM. (2007). *Applying the Rasch Model: Fundamental measurement in Human Sciences* (2nd ed.). Mahwah: Lawrence Erlbaum.

Brunner, M, & Süß, H-M. (2005). Analyzing the reliability of multidimensional measures: An example from intelligence research. *Educational and Psychological Measurement, 65*(2), 227–240.

Bühner, M, Kröner, S, & Ziegler, M. (2008). Working memory, visual-spatial-intelligence and their relationship to problem-solving. *Intelligence, 36*, 672–680.

Cartrette, DP, & Bodner, GM. (2010). Non-mathematical problem solving in organic chemistry. *Journal of Research in Science Teaching, 47*(6), 643–660.

Danner, D, Hagemann, D, Holt, DV, Hager, M, Schankin, A, Wüstenberg, S, & Funke, J. (2011). Measuring performance in dynamic decision making. Reliability and validity of the tailorshop simulation. *Journal of Individual Differences, 32*(4), 225–233.

Drasgow, F, & Chuah, SC. (2006). Computer-based testing. In M Eid & E Diener (Eds.), *Handbook of multimethod measurement in Psychology* (pp. 87–100). Washington, DC: American Psychological Association.

Enders, CK. (2010). *Applied missing data analysis*. New York: The Guilford Press.

Flick, LB, & Lederman, NG (Eds.). (2006). *Scientific inquiry and nature of science*. Dordrecht: Springer.

Friege, G, & Lind, G. (2006). Types and qualities of knowledge and their relations to problem solving in physics. *International Journal of Science and Mathematics Education, 4*, 437–465.

Funke, J. (2010). Complex problem solving: A case for complex cognition? *Cognitive Processing, 11*, 133–142.

Funke, J, & Frensch, PA. (2007). Complex problem solving: The European perspective – 10 years after. In DH Jonassen (Ed.), *Learning to solve complex scientific problems* (pp. 25–47). New York/London: Lawrence Erlbaum.

Gabel, DL, & Bunce, DM. (1994). Research on problem solving: Chemistry. In DL Gabel (Ed.), *Handbook of research on science teaching and learning* (pp. 301–326). New York: Macmillan.

Gilbert, JK, & Treagust, D (Eds.). (2009). *Multiple representations in chemistry education*. New York: Springer.

Goode, N, & Beckmann, JF. (2010). You need to know: There is a causal relationship between structural knowledge and control performance in complex problem solving tasks. *Intelligence, 38*, 345–352.

Greiff, S, Holt, DV, Wüstenberg, S, Goldhammer, F, & Funke, J. (2013). Computer-based assessment of complex problem solving: Concept, implementation, and application. *Educational Technology Research & Development, 61*, 407–421.

Hambrick, DZ. (2005). The role of domain knowledge in higher-level cognition. In O Wilhelm & RW Engle (Eds.), *Handbook of understanding and measuring intelligence* (pp. 361–372). Thousand Oaks: Sage Publications.

Honey, MA, & Hilton, ML (Eds.). (2011). *Learning science through computer games and simulations*. Washington, DC: The National Academic Press.

Hu, L, & Bentler, PM. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1–55.

Jonassen, DH. (2004). *Learning to solve problems: An instructional design guide*. San Francisco: John Wiley & Sons.
Jones, GJF. (2009). An inquiry-based learning approach to teaching information retrieval. *Information Retrieval, 12*, 148–161.
Jurecka, A. (2008). Introduction to the computer-based assessment of competencies. In J Hartig, E Klieme, & D Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 193–214). Cambridge/Göttingen: Hogrefe & Huber Publishers.
Kapa, E. (2007). Transfer from structured to open-ended problem solving in a computerized metacognitive environment. *Learning and Instruction, 17*, 688–707.
Kim, MC, & Hannafin, MJ. (2011). Scaffolding problem solving in technology-enhanced learning environments (TELEs): Bridging research and theory with practice. *Computers & Education, 56*, 403–417.
Kind, PM. (2013). Establishing assessment scales using a novel disciplinary rationale for scientific reasoning. *Journal of Research in Science Teaching, 50*(5), 530–560.
Klahr, D. (2000). *Exploring science*. Cambridge: MIT Press.
Koeppen, K, Hartig, J, Klieme, E, & Leutner, D. (2008). Current issues in competence modeling and assessment. *Journal of Psychology, 216*(2), 61–73.
Koppelt, J. (2011). *Modellierung dynamischer problemlösekompetenz im Chemieunterricht [Modeling complex problem-solving competence in Chemistry]*. Berlin: Mensch & Buch.
Kröner, S, Plass, JL, & Leutner, D. (2005). Intelligence assessment with computer simulations. *Intelligence, 33*, 347–368.
Kuhn, D. (2009). Do students need to be taught how to reason? *Educational Research Review, 4*(1), 1–6.
Künsting, J, Wirth, J, & Paas, F. (2011). The goal specificity effect on strategy use and instructional efficiency during computer-based scientific discovery learning. *Computers & Education, 56*, 668–679.
Kuo, C-Y, & Wu, K-H. (2013). Toward an integrated model for designing assessment systems: An analysis of the current status of computer-based assessments in science. *Computers & Education, 68*, 388–403.
Lee, CB. (2010). The interactions between problem solving and conceptual change: System dynamic modelling as a platform for learning. *Computers & Education, 55*, 1145–1158.
Lee, CB, Jonassen, DH, & Teo, T. (2011). The role of model building in problem solving and conceptual change. *Interactive Learning Environments, 19*(3), 247–265.
Leutner, D. (2002). The fuzzy relationship of intelligence and problem solving in computer simulations. *Computers in Human Behavior, 18*, 685–697.
Leutner, D, Klieme, E, Meyer, K, & Wirth, J. (2005). Die Problemlösekompetenz in den Ländern der Bundesrepublik Deutschland. In D PISA-Konsortium (Ed.), *PISA 2003 – Der zweite Vergleich der Länder in Deutschland* (pp. 125–146). Münster: Waxmann.
Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practice, 14*(4), 5–8.
Meßinger, J. (2010). *ChemLabBuilder [computer software]*. Chemnitz: mera.bit Meßinger & Rantzuch GbR.
Molnár, G, Greiff, S, & Csapó, B. (2013). Inductive reasoning, domain specific and complex problem solving: Relations and development. *Thinking Skills and Creativity, 9*, 35–45.
Muthén, B, & Muthén, L. (2010). *Mplus 6 [computer software]*. Los Angeles: Muthén & Muthén.
Nentwig, P, Roennebeck, S, Schoeps, K, Rumann, S, & Carstensen, C. (2009). Performance and levels of contextualization in a selection of OECD countries in PISA 2006. *Journal of Research in Science Teaching, 46*(8), 897–908.
Neumann, I, Neumann, K, & Nehm, R. (2011). Evaluating instrument quality in science education: Rasch-based analyses of a nature of science test. *International Journal of Science Education, 33*, 1373–1405.
OECD. (2004). *Problem solving for tomorrow's world – First measures of cross curricular competences from PISA 2003*. Paris: OECD.
OECD. (2013). *PISA 2012 assessment and analytical framework*. Paris: OECD.
Pellegrino, JW. (2012). Assessment of science learning: Living in interesting times. *Journal of Research in Science Teaching, 49*(6), 831–841.
Quellmalz, ES, Timms, MJ, Silberglitt, MD, & Buckley, BC. (2012). Science assessments for all: integrating science simulations into balanced state science assessment systems. *Journal of Research in Science Teaching, 49*(3), 363–393.
Rigas, G, Carling, E, & Brehmer, B. (2002). Reliability and validity of performance measures in microworlds. *Intelligence, 30*, 463–480.
Rutten, N, Van Joolingen, WR, & Van der Veen, JT. (2011). The learning effects of computer simulations in science education. *Computers & Education, 58*, 136–153.
Sager, S, Barth, CM, Diedam, H, Engelhart, M, & Funke, J. (2011). Optimization as an analysis tool for human complex problem solving. *Journal of Optimization, 21*(3), 936–959.
Scherer, R. (2012). *Analyse der struktur, messinvarianz und ausprägung komplexer problemlösekompetenz im fach Chemie [Analyzing the structure, invariance, and performance of students' complex problem-solving competencies in Chemistry]*. Berlin: Logos.
Scherer, R. (2014). Psychometric challenges in modeling scientific problem-solving competency: An item response theory approach. In H Bock (Ed.), *Studies in classification, data analysis, and knowledge organization*. New York: Springer. in press.
Scherer, R, & Tiemann, R. (2012). Factors of problem-solving competency in a virtual chemistry environment: The role of metacognitive knowledge about strategies. *Computers & Education, 59*(4), 1199–1214.
Schmidt-Weigand, F, Hänze, M, & Wodzinski, R. (2009). Complex problem solving and worked examples. *Zeitschrift für Pädagogische Psychologie, 23*, 129–138.
Schroeders, U, Wilhelm, O, & Bucholtz, N. (2010). Reading, listening, and viewing comprehension in English as a foreign language: One or more constructs? *Intelligence, 38*, 562–573.
Sonnleitner, P, Keller, U, Martin, R, & Brunner, M. (2013). Students' complex problem-solving abilities: Their structure and relations to reasoning ability and educational success. *Intelligence, 41*(5), 289–305.
Taasoobshirazi, G, & Glynn, SM. (2009). College students solving chemistry problems: A theoretical model of expertise. *Journal of Research in Science Teaching, 46*(10), 1070–1089.
Van Merriënboer, JJG. (2013). Perspectives on problem solving and instruction. *Computers & Education, 64*, 153–160.
Wainer, H, Bradlow, E, & Wang, X. (2007). *Testlet response theory and its applications*. Cambridge: Cambridge University Press.
Wirth, J. (2008). Computer-based tests: Alternatives for test and item design. In J Hartig, E Klieme, & D Leutner (Eds.), *Assessment of competencies in educational contexts* (pp. 235–252). Cambridge/Göttingen: Hogrefe & Huber Publishers.
Wirth, RJ, & Edwards, MC. (2007). Item factor analysis: Current approaches and future directions. *Psychological Methods, 12*, 58–79.
Wirth, J, & Klieme, E. (2004). Computer-based assessment of problem solving competence. *Assessment in Education: Principles, Policy and Practice, 10*(3), 329–345.
Wu, H-L, & Pedersen, S. (2011). Integrating computer- and teacher-based scaffolds in science inquiry. *Computers & Education, 57*, 2352–2363.
Wu, ML, Adams, RJ, Wilson, M, & Haldane, S. (2007). *ACER Conquest 2.0: Generalized item response modeling software [computer software]*. Hawthorn: ACER.
Wüstenberg, S, Greiff, S, & Funke, J. (2012). Complex problem solving – more than reasoning? *Intelligence, 40*(1), 1–14.
Yang, Y, & Green, SB. (2011). Coefficient alpha: A reliability coefficient for the 21st century? *Journal of Psychoeducational Assessment, 29*(4), 377–392.

doi:10.1186/2196-7822-1-2
Cite this article as: Scherer *et al.*: **Developing a computer-based assessment of complex problem solving in Chemistry.** *International Journal of STEM Education* 2014 **1**:2.