

MEASURING EXPERIMENTAL SKILLS IN LARGE-SCALE ASSESSMENTS: DEVELOPING A SIMULATION-BASED TEST INSTRUMENT

Martin Dickmann¹, Bodo Eickhorst², Heike Theyßen¹, Knut Neumann³, Horst Schecker² and Nico Schreiber¹

¹ University of Duisburg-Essen

² University of Bremen

³ Leibniz Institute for Science and Mathematics Education

Abstract: Fostering students' experimental skills is widely regarded as a key objective of science education (e.g. KMK, 2005; NRC, 2012). Hence, assessment tools for measuring experimental skills are required. In most large-scale assessments experimental skills are measured by paper and pencil tests. This is due to the efficiency of administering and scoring such tests. However, several studies show only low correlations between students' achievements in paper and pencil tests and their achievements in hands-on experiments (e.g. Shavelson, Ruiz-Primo, & Wiley, 1999; Stecher et al., 2000; Hammann et al., 2008; Emden, 2011; Schreiber, 2012). This calls the validity of the paper and pencil approach into question. On the other hand, testing 'hands-on' with real experiments is costly and time-consuming concerning logistics, data acquisition and data analysis, especially in large-scale-assessments. More manageable tools for measuring experimental skills in large-scale assessments seem to be computer-based experiments with interactive simulations (cf. Schreiber, 2012). Our goal is a simulation-based test instrument that measures experimental skills validly and reliably. It should cover all the three phases of experimenting: preparation, performance and evaluation. Schreiber (2012) found that comprehensive experimental tasks posed in an open format cause a high dropout rate during the test. We thus develop consecutive tasks that divide the complex experimental demands into a sequence of smaller items. Each item operationalizes a specific experimental skill. To avoid follow-up errors, each item contains a sample solution of the preceding item. In order to secure test quality, extensive validation studies are carried out.

Keywords: large-scale assessment, computer-based testing, assessment of competence, performance assessment, scientific experimentation

THEORETICAL FRAMEWORK

Science education standards emphasize the importance of experimental skills for scientific literacy (e.g., KMK, 2005; NRC, 2012). Students' abilities to plan and carry out experimental investigations are included in evaluations of national standards as well as in international student assessments (OECD, 2007). Theories of the experimental process typically distinguish between three phases of experimenting: preparation (e.g. planning experimental procedures), performance (e.g. setting up the apparatus) and evaluation (e.g. interpreting results) (cf. Emden, 2011). A test instrument measuring experimental skills should cover all the three phases.

Testing experimental skills has to address several problems, especially in large-scale assessments. A process-based assessment, analyzing students' actions during hands-on experiments, is resource-consuming. Supplying standardized apparatus for hands-on tests poses problems of logistics. Paper and pencil tests can hardly cover experimental skills of the performance phase. Thus, paper-pencil tests are often narrowed to the preparation of experiments and the evaluation of data (e.g., Glug, 2009).

Previous studies on the exchangeability of test formats for experimental skills show only low correlations between students' achievements in paper and pencil tests and their achievements in hands-on experiments (e.g., Shavelson, Ruiz-Primo, & Wiley, 1999; Stecher et al., 2000; Hammann et al., 2008; Emden, 2011; Schreiber, 2012). On the other hand, studies indicate that computer-simulations might be valid substitutes for hands-on experiments in tests (cf. Shavelson et al., 1999; Schreiber, 2012).

Schreiber (2012) found no significant difference between the distributions of achievement scores gained from computer-based testing with mouse-on experiments and hands-on testing, whereas the distributions differed significantly between a paper and pencil test and a hands-on test. Schreiber (2012) also found that broad experimental tasks posed in an open format cause a high dropout rate during the test. This leads to ground-effects and missing data.

RATIONALE AND METHODS

Aims of the study

Our overall aim is to develop a test instrument that can reliably and validly measure experimental skills and that is suitable for large-scale assessments. We assume that the problems and restrictions discussed above can be reduced by computer-based testing. This approach allows us to comprise the performance phase of experimental investigations. In contrast to hands-on experiments, the logistics of computer-based testing is easier for large-scale studies. Students' actions can be recorded automatically and evaluated on the basis of log files. In order to secure test quality, extensive validation studies are carried out.

Test instrument

The test instrument refers to typical experimental tasks in secondary school physics instruction. The target group are students at the end of lower secondary education (aged 14 to 16). The test instrument consists of several units. Each unit deals with a specific experimental task. The students have to perform a complete experimental investigation, i.e. plan the experiment, prepare the setup, perform the measurements, analyze experimental data and draw conclusions. To minimize drop-out caused by comprehensive experimental tasks (Schreiber, 2012), each unit is split up into a sequence of items each referring to one experimental skill (e. g. plan the experiment or perform the measurements). Furthermore, each item starts with a sample solution of the preceding item. Thus, students' experimental skills can be assessed across the full range of the phases of an experimental investigation. For instance, students who do not succeed in assembling an appropriate experimental set-up can still proceed with the measurement item, because it provides them with a functional set-up. The

preparation	performance	evaluation
describe basic idea	assemble and test an experimental setup	plan the evaluation of data
specify procedure <div style="border: 1px dashed black; padding: 5px; display: flex; justify-content: space-around;"> <div style="border: 1px dashed black; padding: 2px; text-align: center;">select from a given set of apparatus</div> <div style="border: 1px dashed black; padding: 2px; text-align: center;">sketch the set-up</div> <div style="border: 1px dashed black; padding: 2px; text-align: center;">describe the course of action</div> </div>	perform and document measurements	process data
prepare measurement report		draw conclusions

Figure 1: Model of experimental tasks (grey: phases of experimentation; light: components)

intermediate solutions are presented by two fictitious students (“Alina and Bodo”) who are said to have worked on the same experimental task.

Altogether, twelve units are being developed and tested. They cover content areas in electric circuits, mechanics and geometrical optics.

Task development is based on the model shown in Figure 1. The model integrates previously developed models for experimental skills (cf. Schreiber et al., 2012; Nawrath, Maiseyenka & Schecker, 2011). The model describes eight experimental skills (light boxes in Fig. 1), grouped into the three phases of experimentation. As the test is intended to focus on the actual performance of an experimental investigation, we do not consider more general components of scientific knowledge generation like ‘develop questions’ and ‘phrase hypotheses’.

A unit consists of six items (out of eight), with two items for each phase of experimentation. The two components of the performance phase are included in each unit.

Figure 2 shows a sample item of the unit ‘Elongation of a rubber band’ from the content area mechanics. In this unit Alina and Bodo want to test the hypothesis: “The expansion of a rubber band is proportional to the attached weight.” The students have to choose the right material, describe the measuring procedure, assemble the setup etc. In the particular item shown in Figure 2 the students get a functional setup to perform their measurements and a properly prepared table to document the data.

For the choice of suitable material (preparation phase), the assembling and testing of the experimental setup and the performance of the measurement (performance phase) simulations are provided that enable the students to interact with the material, to observe the effects and to measure data. In the simulation shown in Figure 2 students can for example attach pieces of mass to the rubber band, adjust the scale, and observe and measure the elongation of the rubber band.

Table 1

Validation aspects, corresponding research questions and studies to answer the research questions

Validation Aspects	Research questions	Studies
Content	Do the tasks represent experiments that students are likely to have seen or worked on? Are the tasks consistent with typical demands posed in classroom practices of experimenting?	Analyses of syllabi and schoolbooks; expert ratings
Individual strategies (cognitive processes)	Do experimental considerations dominate in students' thinking while working on the tasks? Do the tasks offer adequate support to compensate for deficits in physics content knowledge?	Think aloud (intro- and retrospective)
Relationship with external variables	Is the cognitive load of mouse-on experimenting comparable to a hands-on test format?	Comparative studies in the science education lab
	Is the mouse-on test performance a good predictor for performance in hands-on tests?	mouse-on vs. hands-on
Internal test structure	Do experimental skills differ sufficiently from physics content knowledge and cognitive abilities?	Large-scale (400 students per unit)
	Do the items form a reliable scale? Are the three phases of experimentation empirically separable in students' performances?	

CONTENT ANALYSIS

Methods

Syllabi and schoolbooks were initially analyzed to identify physics content areas and experimental challenges that are in accordance with aims and practices of physics instruction (content validity). The analysis was done in two steps. In the first step, key terms were identified in an inductive process by going through curricula and schoolbooks. To ensure comparability across the syllabi of the 16 German federal states, similar terms were clustered in 35 'term groups'. The term group 'electrical resistance' for example includes the terms *electrical resistance*, *specific resistance*, *I-U characteristics*, *Ohm's law*, and *electrical conductivity*. The quality of this method was verified for the content areas mechanics, optics, electricity, and thermodynamics. The inter-rater reliability (Cohen's kappa) of assigning terms to term groups is at least satisfactory ($.78 < \kappa < .95$). In the second step, a criteria-based investigation of the 16 syllabi and of selected schoolbooks was carried out. The curricula and schoolbooks were searched for the term groups, differentiating between general occurrence and occurrence in (explicit) conjunction with an experimental action (preparation, performance, evaluation). In addition, the syllabi were analyzed with regard to the grades in which a term group occurs and whether it is obligatory or optional content. In the schoolbooks all the experimental tasks referring to a term group were

identified.

Based on these analyses 22 suggestions for typical experimental tasks were generated. 53 experts (experienced teachers) rated to which extent these tasks comply with typical demands posed in classroom practices of experimenting (four-level Likert-scales). This was done by an online-survey. We e.g. asked the experts how likely students would have had appropriate learning opportunities, enabling them to solve the task. We also asked the experts how likely students could plan, perform or evaluate just this or a very similar experiment at the end of lower secondary education.

Evaluating the online-survey, we ranked the experimental tasks for each content area separately. Our main criterion was that, according to the experts' estimations, it is likely or very likely that the students can perform the experimental task.

Results

Our syllabus analysis confirmed the central role of experiments in physics teaching (cf. Tesch, 2005). The analysis yielded a high consistency of the obligatory content (measured by the occurrences of the term groups) to be dealt with during lower secondary physics instruction across the 16 federal states of Germany. Minor differences were found with regard to the grade in which the content is taught. Comparing the experiments presented in the schoolbooks we were able to identify a consistent set of widely used topics for student experiments. For 12 of the 22 tasks our main criterion ($M \geq 3.0$ on a scale from 1 to 4) was fulfilled. Especially in the domains electric circuits and optics our tasks comply with typical demands posed in classroom practices of experimenting. In mechanics the ratings of two out of four tasks were not satisfactory. We thus developed two more tasks for this domain.

As the result of our content analysis we can build on a set of twelve experimental tasks with high content validity for the physics domains electric circuits, geometrical optics and mechanics. The tasks provide a solid basis for the investigation of further validation aspects, in particular cognitive validity. We have designed twelve complete units around these tasks (together with the interactive simulations).

COGNITIVE VALIDATION

Methods

For the aspect of cognitive validation we focus on the students' cognitive processes while working on the units. The key issue is whether experimental considerations dominate in students' thinking while they try to solve the items: Are their actions driven by reflections on the experiment to be conducted or by other aspects, like operating the simulation software? To answer this research question, four out of twelve units from the three content areas are analyzed with think-aloud techniques (intro- and retrospective). About 40 students worked on each unit.

The verbalizations of the students are rated in a deductive mode of qualitative content analysis. We use indicators to distinguish between students' considerations that are related to the process of experimentation (e.g. safety issues,

measurement accuracy etc.) and considerations that are based on non-experimental arguments (e.g. plausibility considerations). Table 2 shows examples for the item *assemble and test the experimental setup* of the unit ‘elongation of a rubber band’.

Table 2

Examples of experimental and non-experimental considerations in the unit ‘Elongation of a rubber band’

Experimental considerations	Non-experimental considerations
“In order to measure accurately, the ruler has to be very close to the rubber band.”	“For assembling the setup I am looking at which parts match together”
“Ah, the elongation changes more than the weight I attach. This can’t be proportional.”	“The last device doesn’t fit in anywhere. The setup should be working now.”

The inter-rater reliability (Cohen’s kappa) of differentiating between experimental and non-experimental considerations is at least satisfactory for the item *assemble and test the experimental setup* (assembling experimental setup: 88 % agreement ($\kappa = .714$); test experimental setup: 100 % agreement). The analysis of further units and items is in progress.

Results

The analysis of the cognitive processes for the item *assemble and test the experimental setup* of the rubber band unit shows that most students dominantly express experimental considerations (see figure 3) while working on this item. Further analyses will show whether this result can be confirmed for other items and units.

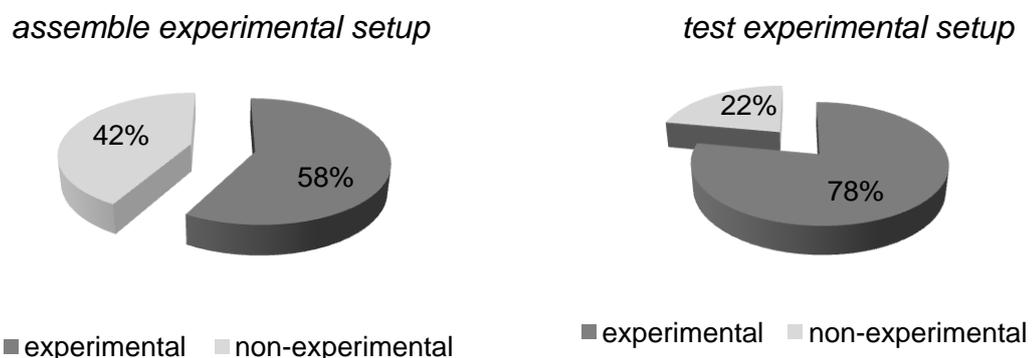


Figure 3: Percentage of students with experimental and non-experimental considerations for the item *assemble and test experimental setup*

SUMMARY AND OUTLOOK

In our project the development of the test instrument and validation studies are closely intertwined. Content analysis and expert panels have led to a set of experimental tasks that are adequate challenges for secondary students. Around these tasks, twelve test units with items for specific experimental skills have been developed. The units are realized in an online test environment with embedded simulations (mouse-on test). First empirical studies indicate that the test is cognitively valid.

Besides further analyses of cognitive validity – with consequences for test improvement – we will, as a next step, put a focus on studies of convergent validity. Paper and pencil tests with hands-on experiments serve as benchmarks. Structural validity will be researched on the basis of a large-scale data sampling in 2014.

REFERENCES

- Emden, M. (2011). *Prozessorientierte Leistungsmessung des naturwissenschaftlich-experimentellen Arbeitens*. Berlin: Logos.
- Glug, I. (2009). *Entwicklung und Validierung eines Multiple-Choice-Tests zur Erfassung prozessbezogener naturwissenschaftlicher Grundbildung*. Christian-Albrechts- Universität zu Kiel.
- Hammann, M., Phan, T. T. H., Ehmer, M. & Grimm, T. (2008). Assessing pupils' skills in experimentation. *Journal of Biological Education* 42 (2), 66-72.
- KMK. Sekretariat der Ständigen Konferenz der Kultusminister der Länder in der Bundesrepublik Deutschland. (2005). *Bildungsstandards im Fach Physik für den Mittleren Schulabschluss*. München: Luchterhand.
- Nawrath, D., Maiseyenko, V., & Schecker, H. (2011). Experimentelle Kompetenz - Ein Modell für die Unterrichtspraxis. *Praxis der Naturwissenschaften – Physik in der Schule*, 60(6), 42–49.
- NRC. National Research Council. (2012). *A Framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas*. Washington, DC: The National Academies Press.
- OECD (ed.) (2007). *PISA 2006 - Schulleistungen im internationalen Vergleich: Naturwissenschaftliche Kompetenzen für die Welt von morgen*. Bielefeld: Bertelsmann.
- Schreiber, N. (2012). *Diagnostik experimenteller Kompetenz. Validierung technologie- gestützter Testverfahren im Rahmen eines Kompetenzstrukturmodells*. Berlin: Logos.
- Schreiber, N., Theyßen, H., & Schecker, H. (2012). Experimental Competencies in science: a comparison of assessment tools. In C. Bruguière, A. Tiberghien, & P. Clément (Eds.), *E- Book Proceedings of the ESERA 2011 Conference: Science*

learning and Citizenship. Part 10 (co-ed. R. Millar), 66–72. Lyon: European Science Education Research Association.

Shavelson, R. J., Ruiz-Primo, M. A., & Wiley, E. W. (1999). Note on Sources of Sampling Variability in Science Performance Assessments. *Journal of Educational Measurement*, 36(1), 61–71

Stecher, B. M., Klein, S. P., Solano-Flores, G., McCaffrey, D., Robyn, A., Shavelson, R. J. (2000). The effects of Content, Format and Inquiry Level on Science Performance Assessment Scores. *Applied Measurement in Education*, 13(2), 139-160.

Tesch, M. (2005). *Das Experiment im Physikunterricht – Didaktische Konzepte und Ergebnisse einer Videostudie*. Berlin: Logos.

Wilhelm, O. & Kunina, O. (2009). Pädagogisch-psychologische Diagnostik. In Wild, E., & Möller, J. (Eds.), *Pädagogische Psychologie*. Heidelberg: Springer.