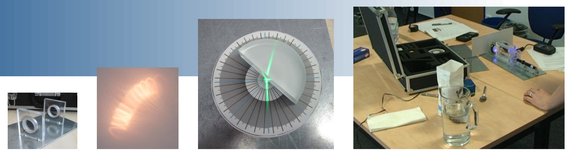
Investigating the cognitive validity of an optics performance assessment using think-alouds

Jan Heidrich, Knut Neumann, Stefan Petersen

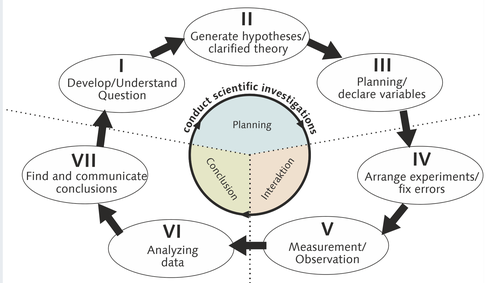


Subject/Problem

-One important aim of science education is for students to learn how to conduct scientific investigations (NRC, 2012).
-It seems that performance assessments have validity-problems to measure the ability of conduct scientific investigations (Shavelson, Gao & Baxter, 1993; Shavelson, Ruiz-Primo & Wiley, 1999).
-There are six sources of validity-problems that have to be ensured (Messick, 1989, 1995).

Conducting scientific investigations requires:

-**Correct application** of inquiry skills [e.g. analyzing, all skills shown in the figure]
-**Logical order** of application of inquiry skills [e.g. in order of numbers in the figure]
-**Efficient strategy** in application of skills to solve the problem [e.g. trial-and-error or planning-based]



Inquiry Skills of conduct scientific investigations

(Klahr, 2000; Hodsen, 1996; Kipnis & Hofstein, 2008; Hanauer, Hatfull & Jacobs-Sera, 2009; NRC, 2012)

Purpose

Develop a performance assessment that is valid for measuring the ability of university-level physics students to conduct scientific investigations in optics.

Presented aspect: Validity of analyses

(Substantive- and structural Validity)

Analysis	Scoring based on			Economic
	Outcomes (documentation or answer)	**Actions** (live rating or analysis of videos)	**Thinking** (interviews or think aloud)	
Product-oriented	X			+
Process-oriented	X	X		-
Reference	X	X	X	--

Question: *To what extent can a product-analysis, a process-analysis and a reference-analysis generate valid data?*

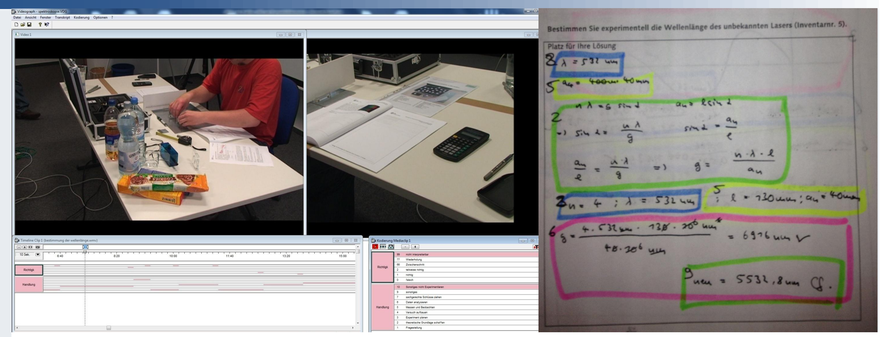
Study – Think-Alouds

Design

Participants (*N=11*) completed an performance assessment with 6 tasks, filled out an lab notebook, get filmed and have to think aloud (Thelk et al. 2006).

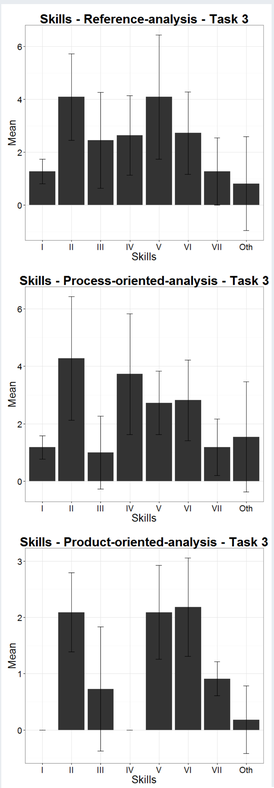
Analyses

(1) Identify inquiry skills and identify order of inquiry skills
(2) Rate the correct apply of inquiry skills
(3) Calculate the score for correct apply, logical order and efficient strategy
-Rater agreement of 2 Raters are adaptable to good (κ=.66-.90).



Identifying and sorting of inquiry skills

Results



Skills identified for the example (task 3, similar for other tasks)

Reference-Analysis
-Participants used all inquiry skills.
-The influence of construct-irrelevant skills is minor.

Process-oriented-analysis
-Process-analysis can identify ~95% of all skills (compared with ref-analysis).
-Process-analysis have problems to differ the inquiry skills III and IV adequately.
-If participants do not know what to do, analysis without *thinking* can not identify skills adequately. Especially this situations show the quality of logical order and strategy (Outcomes of qualitative video analysis).

Product-oriented-analysis
-Product-analysis can only identify the parts of conduct scientific investigations where participants create outcomes (~60% of all skills, compared with ref-analysis).
-There are inquiry skills that can not be identified by product-analysis (e.g. IV).

Reliability as indicator for problems of validity: Cronbachs α and item discrimination r_{it}

	Correct apply	Logical order	Efficient Strategy
product-oriented	α = .65 r_{it} = .40	α = .10 r_{it} = .08	α = .31 r_{it} = .11
process-oriented	α = .69 r_{it} = .45	α = .47 r_{it} = .34	α = .31 r_{it} = .22
reference	α = .78 r_{it} = .58	α = .57 r_{it} = .34	α = .78 r_{it} = .61

-Internal consistency is acceptable for correct apply.
-logical order and efficient strategy are only consistent for ref-analysis.

Conclusion

-The ability to conduct scientific investigations can be measured validly, if all inquiry skills can be identified adequately.
-To identify all inquiry skills adequately it requires a scoring based on *outcomes*, *actions* and *thinking*.

Next step

Develop a valid and economic analysis:
-Scoring of *outcomes* with lab notebook
-Scoring of *actions* with logging by participants themselves
-Scoring of *thinking* with: (1) logging by participants themselves, (2) request the participants to document their conducting so that it can be replicated by other students.

New analysis will be investigated in April-July '14.

Contact: heidrich@ipn.uni-kiel.de

IPN Leibniz Institute for Science and Mathematics Education

Titel
Investigating the cognitive validity of a performance assessment using think alouds

Abstract
Scientific inquiry is one central aspect of science education. The ability to engage in scientific inquiry includes performing scientific investigations. Successfully performing scientific investigations means (I) to correctly apply a series of (inquiry) skills, (II) in a logical order and (III) using the right strategy. Current assessment instruments for evaluating inquiry performance often show problems with respect to cognitive validity. The aim of the presented project is to develop an assessment that can generate cognitive valid data about how well university level physics students can perform scientific investigations. Investigation of cognitive validity requires an analysis of reliability and a matching of performance with a reference. Analysis based on think alouds was used as reference. In the study *N=11* participates carried out a developed performance assessment. Their perfomance was assessed using either only the lab-notebook, a video recording of their actions without audio or the full video including a think-aloud. The reliability is calculated for all three analyses. Correlations between analysis and reference are calculated. The results show, that the performance assessment can generate cognitive valid data for two of the three above mentioned aspects of performing scientific investigations with reference analysis and for one of three aspects with other analyses.

Subject/Problem

One important aim of science education is to provide students with the ability to engage in scientific inquiry (NRC, 2012). In Labwork-courses this includes the ability to plan, perform and analyze scientific investigations (Hodson, 1996). The process of performing scientific investigations can be divided into three phases: planning, performing and analyzing (Klahr, 2000). Past research has identified a variety of skills that play a part in these phases (Hodsen, 1996; Klahr, 2000; Kipnis & Hofstein, 2008; Hanauer, Hatfull & Jacobs-Sera, 2009; NRC, 2012; cf. Figure 1). But there is more required than correctly applying the right skills to perform scientific investigations. Klahr (2000) for example suggests that there are different strategies to conduct scientific investigations such as trial & error or planning. Also in K-12 framework the skills are only a part of practice (NRC, 2012).

Therefore based on past research a model of performance in scientific investigations is developed. The model includes skills (see fig. 1) which have to be performed correctly and in a logical order for a high level of performance. Hodsen (1996) states that there is not "the one" right order for every situation and investigation. The choice of a skill depends on a specific situation in the investigation. In a complete sequence of skills (see fig. 1) the step from one skill to the next is logically right.

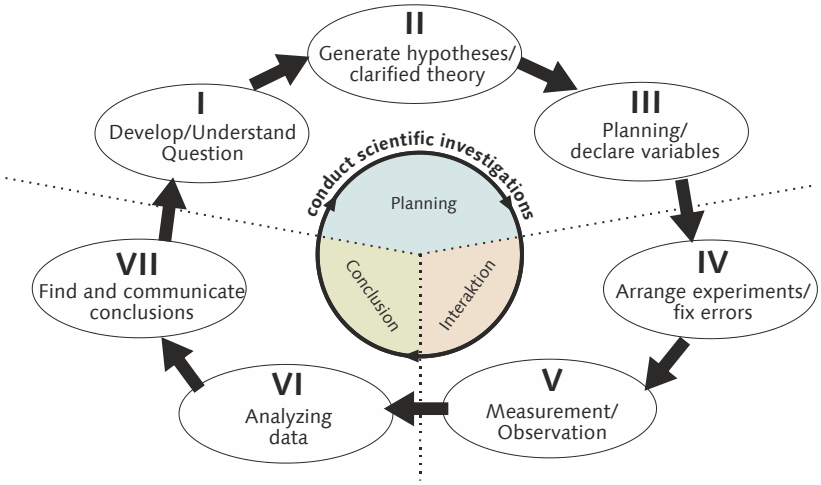


Figure 1: Skills and potential iterative process of scientific inquiry.

But in specific situations it can be logical to skip a skill, e.g. if the skill is already performed correctly in previews investigations. Also backward-steps can be useful to find errors or to control performance. To know the next skill in a specific situation schematic knowledge of scientific investigations and methods is required (Hodson, 1996; Shavelson & Ruiz-Primo, 1998). The correct performance of skills as well as their right order is the basis for a research cycle. Research cycle means that one investigation is finished and data is generated. E.g. if a first hypothesis must be rejected by data, in a new cycle a second hypotheses will be investigated. Complex investigations can include a row of single research cycles. Strategies of scientific investigation are used to control series of research cycles and the interaction between single iterations. Strategies range from "trial and error" to a "planned -" or "analogy based" research (Klahr & Dunbar, 2000). Experts plan their research and need a lower count of research cycles

than novices which are more likely to use trial and error (Klahr, 2000). Based on this theoretical background we evaluate the following aspects of scientific investigations (I) performing skills correctly; (II) using skills in a logical order and (III) research cycle strategy.
With the interest in scientific inquiry growing, the assesment of students' performance in scientific investigations became of increasing interest. Test instruments were developed to measure the performance of scientific investigations (Lawson, 1978; Shavelson, Solano-Flores, & Ruiz-Primo, 1998; Olson, Martin, Mullis, & Arora, op. 2008). However, Shavelson, Gao and Baxter (1993) find that only a small percentage in the variance of student performance can be explained through person ability. This leads to the question how valid the conclusions drawn from perfomance assessments are. Therefore any newly developed performance assessment must aim to minimize non-person-related (NPR) variance. Both, test design and analysis, may be sources for NPR variance. The higher-level design of instruments can be paper & pencil, simulations or performance. Tests with different formats measure different constructs (Shavelson, Ruiz-Primo, & Wiley, 1999). Performance is nearest to practice. Therefore performance can be assumed to be the test design most likely to provide valid conclusions on the performance in scientific investigations. In performance assessment a lot of aspects can generate NPR variance in different tasks, e.g. problems with different material in tasks. A framework for test development can help to keep these aspects constant (Stecher et al., 2000). Also the analysis of test varies, too. In some tests the outcomes of skills are subjects of analysis (product orientated). Other tests have focused on the process of performing (process orientated). To minimize NPR variance it is important to know how appropriate the different analyses can capture the cognitive processes at performance. Valid conclusions are required to compare teaching techniques and evaluate Labwork-courses. But furthermore test economy is important espacially for large-scale assessment. Process-orientated analysis can probably generate the most cognitive valid data. But this analysis is time-consuming. For product-orientated analysis the reverse holds.

The aim of this research project is to develop a performance assessment for university-level that is able to generate cognitive valid data. Approaches to minimize the NPR variances generated by test design are presented in the theoretical background. But NPR variance generated by analysis is still problematic. This NPR variance can be generated in two ways. (1) Tasks which measure different constructs. In this case the reliability of analysis is poor. (2) Analysis which does not cover the complete construct or depends on additional constructs. In this case the correlation between product- or process-orientated analysis and a reference analysis are poor. As reference an analysis is required that can describe the cognitive processes in the best way. Think alouds may be assumed to achieve this requirement (Thelk & Hoole, 2006). To control the two sources of NPR variance we investigate the following research questions:
F1: To which extent can performance of scientific investigations be measured reliably with product- resp. process-orientated analysis and analysis based on think alouds?
F2: To which extent does measuring performance of scientific investigations match between the product- resp. process-orientated analysis and the analysis based on think alouds?

Design

In order to work towards a valid perfomance assessment, a respective test instrument was developed and evaluated with respect to cognitive validity. This performance assessment embraced six performance tasks. To reduce variance from the characteristics of the individual task, a framework was used for task development (Stecher et al., 2000). This framework also helped to ensure that the tasks were designed such that all the skills listed in Figure 1 had to be applied at least once in order to solve the task. In order to prevent variance from different content, all tasks where from the domain of optics. Additionally in order to prevent variance from different experimental material in each task, the same material was provided for all tasks.

The evaluation of cognitive validity was done in three steps. In the first step the performance assessment was carried out by a group of nine physics students and two physics research assistants. Each participant was filmed and had to think aloud. Participants also documented their work in a lab notebook. In the second step the Lab-notebooks (products) and students actions documented by videos (process) were analyzed. For product-orientated analysis the products in Lab-notebooks were mapped to skills of scientific investigations. Subsequently the correctness of the outcomes was rated. For process-orientated analysis instead of outcomes the actions in videos were mapped. The correctness is rated with outcomes like the product-oriented analysis. Analysis of think aloud is similar to process-orientated with additional information from the audio of videos. The mapping of products resp. actions to skills and the rating of correctness are performed by two raters. The agreement of the raters is shown in table 1. The agreement of the raters for Lab-notebooks is good. The agreement of high inference coding with processes and think alouds are acceptable. Scores for the three aspects of performing scientific investigations are calculated based upon these data.

Table 1: Interrater agreement for assignment products/actions and rating of correctness

	Product-orientated (lab-notebook)	Process-oriented (videos)	Reference (think alouds)
Assignment of products/actions	$\kappa = .90$	$\kappa = .66$	$\kappa = .71$
Rating of correctness	$\kappa = .87$	$\kappa = .68$	$\kappa = .66$

In the third step the reliability of analysis and correlation between product- resp. process-oriented analysis and analysis based on think aloud is calculated. In the analysis of reliability one task has been excluded. Item discrimination brings the conclusion that it is measuring a different construct due to a low discrimination.

Analysis and Findings

The aims were to investigate the cognitive validity of an optics performance assessment aimed at university level physics students. Two research questions were formulated. One question focuses on reliability of analyses and one focuses on matchting between analysis and reference. For the first question cronbachs α and item discrimination i_{rt} were calculated. For the different analysis and the three aspects of performing scientific investigations the findings are shown in table 2. Aspect of correctness can be measured reliably using all three analyses. The aspect of logical

order cannot be measured by any analysis. Aspect of strategy can be measured reliably with think alouds only.

Table 2: Cronbachs α and item discrimination i_{rt} of aspects from performing scientific investigations for different analyses

	Product-orientated (lab-notebook)	Process-oriented (videos)	Reference (think alouds)
Correctness	α = .622 i_{rt} = .375	α = .685 i_{rt} = .453	α = .781 i_{rt} = .579
Logical order	α = .098 i_{rt} = .077	α = .474 i_{rt} = .341	α = .569 i_{rt} = .344
Strategy	α = .309 i_{rt} = .113	α = .311 i_{rt} = .224	α = .783 i_{rt} = .608

For the second research question the correlation between analyses for aspects of performing scientific investigations were calculated. The matching between product- resp. process-orientated analysis and reference are shown in table 3. The matching at the other aspects of performing scientific investigations cannot generate scientific information while missing quality of reliability generates a high amount of not identified variance.

Table 3: Correlation between analyses for aspect of correctness

	Product-orientated (lab-notebook)	Process-oriented (videos)	Reference (think alouds)
Product-orientated (lab-notebook)		.84**	$.56^{+}$
Process-oriented (videos)			.65*
Reference (think alouds)			

The matching between product- resp. process-orientated analysis and analysis based on think aloud shows medium to high correlation for both. But the significance is rather low. This might be explained by small population. Effects from .73 can be expected to be significant with population of N=11. The correlation between product- and process-orientated analyses is high and significant. Based on these outcomes cognitive validity in aspect of correctness can be accepted.

Contribution

The aim of the project is to develop a performance assessment that generates cognitive valid data about the ability to perform scientific investigations in the field of optics. There is the potential risk of generating NPR variance. The agreement of raters is acceptable to good (tab. 1). The generated NPR variance is small, especially for analysis with Lab-notebooks. As far as correctness is concerned the reliability is acceptable to good (tab. 2). The generated NPR variance is small, especially for analysis with reference. The correlation between analyses and reference is acceptable. That implies that the generated data for aspect of correctness can be accepted as cognitive valid. Therefore it is possible to evaluate the correctness of performing skills in scientific investigations with developed performance assessment. Problems with reliability are identified for aspects of logical order and strategy. The scale for logical order must be redesigned. Aspect of strategy is reliable for reference analysis. A possible explanation for

that difference might be that students do not know what to do in a situation. They have a cognitive problem to solve and do not use scientific skills for this time. Therefore detailed problems of generating cognitive valid data with performing scientific investigations are identified. Future research can evaluate approaches to solve these detailed problems. The analysis must be designed in such a way that they can capture these cognitive processes.

General Interest

The K-12 framework puts a strong emphasis on scientific inquiry. One aspect of scientific inquiry is performing scientific investigations. But there are problems with assessment of performance in scientific investigations, especially with cognitive validity. In the research project a performance assessment with different analyses was developed and evaluated for cognitive validity. With the test cognitive valid data can be generated for performing skills correctly with product- and process-oriented analyses. Generated data for strategies in performing scientific investigations were cognitive valid for think alouds only. Product-oriented analysis must be redesigned. In the presentation the problems with cognitive validity are addressed and approaches to overcome them are presented.

References

National Research Council [NRC]. (2012). A framework for K-12 Science Education: Practices, Crosscutting Concepts, and Core Ideas. Committee on a Conceptual Framework for New K-12 Science Education Standards. Board on Science Education, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Hanauer, D. I., Hatfull, G. F., & Jacobs-Sera, D. (2009). *Active assessment: Assessing scientific inquiry*. *Mentoring in academia and industry: Vol. 2*. New York, London: Springer.

Hodson, D. (1996). Laboratory work as scientific method: three decades of confusion and distortion. *Journal of Curriculum Studies*, *28*(2), 115–135. doi:10.1080/0022027980280201

Kipnis, M., & Hofstein, A. (2008). The Inquiry Laboratory as a Source for Development of Metacognitive Skills. *International Journal of Science and Mathematics Education*, *6*(3), 601–627. doi:10.1007/s10763-007-9066-y

Klahr, D., & Dunbar, K. (2000). *Exploring science: The cognition and development of discovery processes*. Cambridge, Mass.: MIT Press. Retrieved from http://www.gbv.de/dms/bs/toc/269299742.pdf

Lawson, A. E. (1978). The development and validation of a classroom test of formal reasoning. *Journal of Research in Science Teaching*, *15*(1), 11–24. doi:10.1002/tea.3660150103

Olson, J. F., Martin, M. O., Mullis, I. V., & Arora, A. (op. 2008). *TIMSS 2007 technical report*. Boston, MA: IEA TIMSS & PIRLS.

Shavelson R. J., Gao X., & Baxter G. P. (1993). *Sampling Variability of Performance Assessments: CSE Technical Report 361*. Retrieved from http://www.cse.ucla.edu/products/Reports/TECH361.pdf

Shavelson, R. J. & Ruiz-Primo, M. A. (1998). *On the assessment of science achievement: CSE Technical Report 491*. Retrieved from http://research.cse.ucla.edu/Reports/TECH491.pdf

Shavelson, R. J., Solano-Flores, G., & Ruiz-Primo, M. A. (1998). Toward a science performance assessment technology. *Evaluation and programm planning : an international journal*, (21), 171–184.

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. doi:10.1111/j.1745-3984.1999.tb00546.x

Stecher, B. M., Klein, S. P., Solano-Flores, G., McCaffrey, D., Robyn, A., Shavelson, R. J., & Haertel, E. The Effects of Content, Format, and Inquiry Level on Science Performance Assessment Scores.

Thelk, A. D., & Hoole, E. R. (2006). What Are You Thinking? Postsecondary Student Think-Alouds of Scientific and Quantitative Reasoning Items. *The Journal of General Education*, *55*(1), 17–39. doi:10.1353/jge.2006.0019