Chapter 20
Modelling and Assessing Experimental Competencies in Physics

Heike Theyßen, Horst Schecker, Christoph Gut, Martin Hopf, Jochen Kuhn, Peter Labudde, Andreas Müller, Nico Schreiber, and Patrik Vogt

1 Introduction

Experimental work is regarded as an essential part of physics education. Science curricula from primary school to university request that students learn to carry out experiments (e.g., England (QCA 2004 (revised)), Canada (NRC 1996) and Germany (KMK 2004)). In contrast to its importance, experimental work of students is often criticised as ineffective. Most findings from science education research do not support the hypothesis that students working with experimental materials grasp science more easily (Singer et al. 2006). In a review of research, Hofstein and Lunetta (2004) demand more support for teachers to overcome the limitations of labwork in teaching science. Section 3 of this chapter presents a problem-based guided enquiry approach and describes its effects.

H. Theyßen (✉) • N. Schreiber
University of Duisburg-Essen, Essen, Germany
e-mail: heike.theyssen@uni-due.de; nico.schreiber@tu-dortmund.de

H. Schecker
University of Bremen, Bremen, Germany
e-mail: schecker@uni-bremen.de

C. Gut
Zurich University of Teacher Education, Zurich, Switzerland
e-mail: christoph.gut@fhnw.ch

M. Hopf
University of Vienna, Vienna, Austria
e-mail: martin.hopf@univie.ac.at

J. Kuhn
University of Kaiserslautern, Kaiserslautern, Germany
e-mail: kuhn@uni-landau.de

C. Bruguière et al. (eds.), *Topics and Trends in Current Science Education: 9th ESERA Conference Selected Contributions*, Contributions from Science Education Research 1,
DOI 10.1007/978-94-007-7281-6_20, © Springer Science+Business Media Dordrecht 2014

Experimenting as an important domain of scientific literacy needs to be structured by explicit models. This chapter introduces two approaches for modelling the phases of experimenting and the development of experimental competencies (Sect. 4). The focus of this chapter lies on assessment procedures for students' performances in experimenting. Section 5 shows product-based as well as process-oriented analyses. Data from empirical studies serve to evaluate chances and limitations of the assessment tools. The chapter starts with a brief overview of the theoretical background.

2 Theoretical Background

2.1 Modelling Experimental Competencies

Although the significance of experimental competence is non-controversial, the construct needs to be clarified as a basis for teaching and assessment. As a starting point, normative models have to be developed, which are then to be validated empirically. Most approaches to experimental competencies, like those presented by Hammann (2004) and Walpuski (2006), distinguish three phases of the experimental process: The first phase comprises the "preparation", e.g. to generate questions for an experimental investigation and to plan an adequate experiment. The second phase contains the performance of the experiment and the third phase the conclusions, e.g. to answer the initial question with regard to the experimental results.

Most models of experimental competencies do not get more detailed concerning the second phase, the actual performance of the experiment. This may account for the fact that most of the tests used to validate these models neither measure a student's ability to actually perform a real experiment nor do they differentiate between components of this ability. Klahr and Dunbar (1988), e.g., used very simple computer-based experiments in which the test persons only had to choose the right constellation of experimental parameters and the system displayed the results that the corresponding experiment would have yielded. Others (e.g. Möller et al. 2007; Hammann 2004) used written tests only.

Section 4 of this chapter presents and compares two models for structuring the components of experimental competence in more detail.

P. Labudde
University of Applied Sciences North-West Switzerland, Basel, Switzerland
e-mail: peter.labudde@fhnw.ch

A. Müller
University of Geneva, Geneva, Switzerland
e-mail: Andreas.Mueller@unige.ch

P. Vogt
University of Education Schwäbisch Gmünd, Schwäbisch Gmünd, Germany
e-mail: patrik.vogt@ph-gmuend.de

2.2 Assessing Experimental Competencies

Usually experimental competencies are assessed by written tests (e.g. Henke and Sumfleth 2006; Grube et al. 2007) or hands-on tests (e.g. Harmon et al. 1997; Shavelson et al. 1999; Chen and Klahr 1999). In some studies, computer-based simulations are used as assessment tools (e.g. Shavelson et al. 1999; Klahr and Dunbar 1988). The choice of tools depends on the underlying model of experimental competencies and the focus of the investigation. Studies focusing on *process* aspects typically use tasks with hands-on experiments and a process-oriented analysis of students' actions (e.g. Walpuski 2006). Since the method is very resource consuming, a product-based analysis is often preferred to evaluate hands-on tests (e.g. Shavelson et al. 1999; Harmon et al. 1997). There is a need to further develop and evaluate assessment procedures. Section 5 focuses on this topic.

3 A Problem-Based Guided Enquiry Approach

There has been a long discussion about effects of labwork in science education with a lot of research and very heterogeneous results. One reason for this heterogeneity is that labwork is a very broad term for all sorts of activities. For the purpose of measuring experimental competencies, this is a real problem, which will be discussed in the following.

Hopf (2011) studied the effects of labwork on conceptual understanding, interests and motivation. One of the most cited criticisms of traditional "cookbook"-style labwork is that students are not fully engaged in thinking about their labwork while doing it. In the literature one can find many ideas how to improve labwork learning environments to change this unwanted student behaviour. Hopf evaluated one of these approaches in depth (for more details see Hopf 2011). Adapting an idea from van Heuvelen (1995) several typical labwork activities from introductory optics and electricity were reformulated into an "ill-defined" problem. For example, instead of finding the right additional resistance for a voltmeter, the problem stated was to measure an unknown voltage with an antique instrument without damaging it. The problems were constructed in a way that trial-and-error-strategies could not successfully be applied for the solution. Interesting topics were chosen for contextualisation.

To assess the effects of this problem-based learning environment, a comparative empirical study in 9th- and 10th-grade classrooms was planned and carried out. 410 students from 17 classes participated in this study. In the experimental group, students worked for ten 45-min lessons with the new materials. In the control group cookbook-style labs were used. In all the groups, knowledge tests and a questionnaire including adapted and translated subscales of SLEI (Fraser et al. 1995) as well as a subscale of CLES on personal relevance (Taylor et al. 1997) to measure students' perception of the learning environments were used as pre- and posttests. To compare students' activities during labwork, a small sample of students from the control and from the experimental group were videotaped. These videos were analysed in 10-s loops with an adapted version of the CBAV-category system (Niedderer et al. 1998).

The results are very heterogeneous: On the one hand the measurement of students' activities with video analysis showed advantages of the problem-based labwork. Students spent significantly more time on task and talked much more about physics and of the relationship between their experimental data and the physics theory than students from the control group. On the other hand, practically no effects could be found on students' conceptual understanding or on students' attitudes. The scales were of limited use for measuring effects in the learning environments studied: Only one scale of the SLEI ("student cohesiveness") could be replicated. Thus, in terms of learning *processes*, the problem-based labwork activities were successful, but not in terms of *outcomes*.

A possible interpretation of these results is that measuring effects of labwork is very context dependent. While standard instruments, e.g., for labroom perception (like the SLEI) work very well in standard laboratory classrooms, they do not work in non-standard learning environments as the one discussed above. It can be assumed that this could also be a problem for the measurement of experimental competencies. Assessing the effects of students' activities in the lab is difficult – the more so as a lot of different activities are subsumed under the term "labwork".

4 Modelling Experimental Competencies

The following Sects. 4 and 5 focus on experimental competencies as important instrumental and epistemological components of scientific literacy. The overall research question joining these parts is: How can experimental competencies be described and assessed based on normative models? Section 4 presents and compares two approaches for modelling experimental competencies. These models form the theoretical background of the assessment studies described in Sects. 5.2 and 5.3.

4.1 Modelling the Performance of Experiments

Competencies related to the actual performance of experiments are difficult to operationalise and to measure. Nevertheless, the performance of experiments should be modelled in detail in order to support adequate teaching and assessment. If performance aspects are neglected, students may learn theoretically about scientific reasoning, but they will not be enabled to carry out and learn from own experimental investigations.

Within the German project eXkomp (diagnostics of experimental competencies), a model of experimental competence (Fig. 20.1) was developed that accentuates performance components (Schreiber et al. 2009, 2012). Like most proposals to

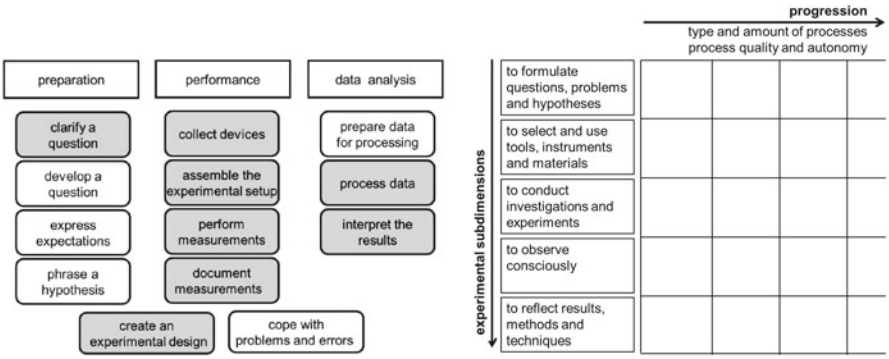

preparation
performance
data analysis
clarify a question
develop a question
express expectations
phrase a hypothesis
collect devices
assemble the experimental setup
perform measurements
document measurements
prepare data for processing
process data
interpret the results
create an experimental design
cope with problems and errors
progression
type and amount of processes
process quality and autonomy
experimental subdimensions
to formulate questions, problems and hypotheses
to select and use tools, instruments and materials
to conduct investigations and experiments
to observe consciously
to reflect results, methods and techniques

Fig. 20.1 Models of experimental competencies (*left*: according to Schreiber, Theyßen and Schecker (2012), for the *grey* components see Sect. 5.2; *right*: according to Gut and Labudde (2010))

structure experimental competencies (see Sect. 2.1), this description distinguishes between three phases: preparation, performance and analysis. Each phase consists of several components. The model focuses on the middle phase: assembling devices and setting up the experiment, measuring and documenting data. The model does not suggest a linear order of steps to be followed in the experimental process. The components can be passed in an iterative and a not necessarily complete sequence. The normative model was validated with an expert rating (teachers and teacher educators; cf. Schreiber 2012). The results confirm the comprehensiveness of the model and the high relevance of all its components for the description of labwork in science teaching.

4.2 Modelling the Progression of Experimental Competence

While the eXkomp model was designed as a starting point for empirical research, the aim of the Swiss project HarmoS (harmonisation of compulsory school) was to formulate national science education standards. For that purpose a competence model had to be developed. The HarmoS model includes three dimensions: skills, domains of contents and achievement levels (Labudde et al. 2012; Ramseier et al. 2011). The model distinguishes eight skills addressing cognitive and social competencies as well as skills for practical laboratory work. For each skill, a priori achievement levels were formulated, ranging consecutively from the 2nd to the 9th grade (Labudde et al. 2009). In correspondence to the achievement levels, based on the evaluation of students' performances in large-scale tests, standards were defined by the political authorities (Labudde 2007; EDK 2011).

The HarmoS model distinguishes five sub-dimensions that refer directly to experimental competence as given on the right part of Fig. 20.1. For each sub-dimension standards were formulated. The progression of the standards depends on the type and the amount of processes that are addressed by a task as well as on the quality and the autonomy with which a task has to be accomplished.

The five sub-dimensions of the HarmoS model describe quite similar abilities as the components of the eXkomp model (Fig. 20.1). They can also be assigned to the three main phases "preparation", "performance" and "data analysis". The eXkomp model is more differentiated in its components, especially those concerning the performance phase, whereas the HarmoS model describes the *progression* within each sub-dimension. The latter point is important for setting standards.

5 Assessment Tools for Experimental Competencies

With regard to the research question given in the introduction of Sect. 4 ("How can experimental competencies be described and assessed based on normative models?"), this part is dedicated to model-based assessment. In the following we present a process-based and two product-based approaches to the assessment of experimental competencies. Issues of validity are discussed.

5.1 Rubrics as Tools for Measuring Experimental Competence

Rubrics are standardised tools for performance measurement of multidimensional, composite competencies. They explicitly state the relevant criteria (or assessment components) and the levels of attainment for each of them. The term "rubric" goes back to the red colour usually used for correction (cf. Latin *rubrica*, red ochre or chalk). Rubrics take the form of two-dimensional matrices, with lines (usually) given by criteria, and columns by level. Both for their explicitness of the assessment components and attainment levels and for their clear 2×2 arrangement, rubrics are widely used and discussed as practical and reliable way of assessment (Jonsson and Svingby 2007).

The aim of this study was to develop a viable and reliable rubric for experimental competencies.

5.1.1 Design

As for experimental competencies, several rubrics have been proposed in the literature, which partially overlap (e.g. Nadji et al. 2003; RPAEG 2008 for a practitioner and a research point of view, respectively). In order to arrive at a comprehensive, well-validated instrument, Vogt, Müller and Kuhn proceeded in two steps:

1. Conceptual analysis and synthesis of existing rubrics.
2. Psychometric analysis and characterisation of synthesis rubrics of step 1.

For Step 1 the existing rubrics were considered as a background of expert statements about important aspects of experimental competence. In order to subsume this expert knowledge, a qualitative analysis was undertaken, where all the criteria were entered into a common text-base and treated as follows:

- *Splitting*: combined criteria (such as "response summarizes data, draws conclusions from these results, and evaluates them relative to the problem", Nadji et al. 2003, under heading "conclusions") were first split into separate statements.
- *Matching and Formulation*: On the basis of a one-by-one comparison, it was then decided, whether a given criterion statement matched another one from a different source; for a group of matching statements, the clearest formulation was then sought (sometimes by paraphrasing, if none of the existing formulations seemed adequate).
- *Subsuming*: For clarity and usability of the instrument, appropriate groupings of the assessment components are necessary. We decided to group them according to the different stages of the experimentation process (see below). Other groupings may be possible, but this is one which is both appropriate for the practical scoring process and conceptually sound, as it conveys the basic conceptual and chronological structure of experimentation.

Step 1 led to a first version of a synthesis rubric of experimental competence (ShREC) with 50 assessment criteria in five subgroups ("preparation", 5; "design", 3; "procedure", 13; "analysis", 14; "presentation", 15). The procedure described above still involves a subjective element on the side of the researcher (as length and formulation of an instrument always do). In order to address questions like whether formulations are understandable or whether criteria are missing or superfluous, a subsequent quantitative study within a larger expert community was undertaken. This is Step 2 and described below.

5.1.2 Methods: Analysis of Rubric Validity and Reliability

Within a first round of expert rating, ten raters (physics teachers and university lecturers) assessed whether the items of the instrument represent a criterion, which is either "essential", "useful but not essential", or "not useful" for measuring experimental competence (content validity). To determine furthermore such items which could be dropped without impairing content validity, the content validity ratio (CVR; Lawshe 1975) for each item was calculated. Moreover, experts were asked to add items they considered as pertinent for measuring experimental competencies.

After adapting the instrument to the rating feedback (by deleting and adding items as described above) a second round of the expert-rating process had to be conducted to determine the content validity of the modified instrument. Items with $\mathrm{CVR} < 0.62$ were deleted again. Then, the mean of the remaining items was calculated, obtaining the content validity index (CVI), which represents a measure of the content validity of the instrument in total (Lawshe 1975).

Moreover, convergent validity was tested as cross-validity with traditional lab report assessment, based on the following sample: From the labwork courses taught by two of the authors, there is a continuously growing sample of lab reports by physics teacher students for which traditional assessment reports exist, i.e. plain text comments together with a final grading mark. In the 2010 summer term, 122 such reports and their evaluations were collected and analysed; no kind of selection on this sample was carried out. For the traditional and ShREC assessment, the measure of cross-validity according to standard regression-correlation techniques (Diehl and Staufenbiel 2002) was determined, yielding a measure of concurrent validity of rubrics and traditional labwork assessment.

The last step was a reliability analysis of the synthesis rubric for experimental competence under two perspectives. First, reliability in the sense of internal consistency was calculated (Cronbach's α). Second, inter-rater reliability was assessed: ten raters (physics teachers and university lecturers; the same rater sample as mentioned above) evaluated the same three lab reports (a bad one, an average one and an excellent one).

5.1.3 Results and Discussion

In a pilot study on validation of ShREC according to above methods and based on the sample above, the following results were obtained:

The reliability of the instrument as a whole was satisfactory (Cronbach's $\alpha > 0.70$), and the reliabilities of the item subgroups were smaller but still acceptable (Cronbach's $\alpha > 0.64$). The cross-validation between the instrument and traditional lab report evaluation reveals reasonable convergent validity (correlation = 0.75). The subgroups based on the practical and chronological structure of labwork did not express themselves in a corresponding factor analytic structure. Still, they are justified for both practical and conceptual reasons (see above).

Based on the work so far, the experimental competence rubric investigated here shows satisfying reliability and validity (the latter, in particular, being based on a systematic synthesis of expert knowledge and corroborated by psychometric analysis). Henceforth, it could be considered as a step towards meeting the triple requirement of conceptual analysis and adequacy, of practical feasibility and of psychometric validation.

5.2 Process-Oriented Assessment by Hands-On and Mouse-On Tests

While rubrics are useful for rating lab protocols as the final *products* of experimenting, a *process*-oriented analysis of labwork activities has to take students' actions in the lab into account. Process-studies usually refer to video recordings of students' hands-on activities. Video analysis consumes much more resources than a product-based approach. Schreiber, Theyßen and Schecker present a method that may allow for a process-oriented assessment with reduced effort (for details see Schreiber et al. 2012). The idea is to replace assessment procedures looking at hands-on experimenting by mouse-on experiments, i.e. workbenches that simulate real experimental environments. Log files are used for a time-efficient analysis of the experimental process. Thus, two research questions arise:

1. Are mouse-on tests based on computer-simulation workbenches a suitable substitute for hands-on tests?
2. Are ratings of process-qualities based on log-file analysis as reliable as ratings based on the qualitative analysis of screen recordings?

5.2.1 Design, Assessment Tools and Data Collection

The research questions were answered in a comparative study within a larger project on the diagnostics of experimental competence (cf. Schreiber et al. 2012). Two similar physics tasks in the domain of electric circuits were implemented both as hands-on experiments and as mouse-on experiments. Based on a pretest, two groups of students (grades 10-12, upper secondary schools) with about 35 students each were put together according to selected cognitive parameters like content knowledge and general and spatial intelligence. After a training session, in which the students were introduced into the experimental devices and the simulation workbench, they performed the actual tests.

One of the tasks was “Here are three bulbs. Find the one with the highest power at 6 V”. The second task was to find out which of three given wire materials has the lowest specific resistance. Each student worked on one task in the hands-on setting and the other in a simulation environment (“mouse-on”). The order of the assessment settings was varied systematically. The simulation provided the same experimental apparatus as the hands-on experiment and allowed for all the relevant manipulations of the devices (selecting and assembling devices, measuring, even the destruction of equipment in case of incorrect use). A pre-structured worksheet prompted the students to clarify the question (task), create and conduct a suitable experiment, document their measurements and interpret the result. The students had 30 min to plan, perform and analyse each experiment. Within this period, the timing and the course of action were up to the students; no intermediate results had to be presented.

Figure 20.2 shows a screenshot of the simulation workbench (right) and a photograph of the corresponding hands-on experiment (left). In the hands-on test, students’ actions were videotaped. Their paper and pencil worksheets were collected. In the “mouse-on” experiment, students’ actions were documented by screen recordings. The students filled in an online worksheet. Detailed log files captured each manipulation of the virtual experiment and each entry to the online worksheet.

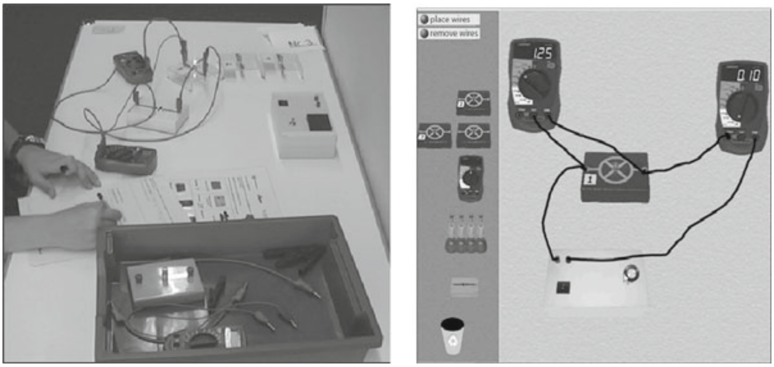


Fig. 20.2 The hands-on environment (*left*) and a screenshot of the mouse-on environment (*right*)

5.2.2 Methods: Process-Oriented Data Analysis

For a process-oriented data analysis, a set of categories was developed for coding students' experimental actions and their sequence. The categories represent the eight model components covered in the assessments (highlighted in Fig. 20.1). They can be coded with a high inter-coder reliability (Cohen's kappa = 0.84). For instance, when a student connects electrical devices on the table or on the screen, the category "assemble the experimental setup" is attributed. Selecting devices from the pool and placing them on the table (or on the screen) is coded as "collect devices".

A subsequent step of analysis takes into account the *quality* of actions, e.g. whether the experimental setup is correct or incorrect. The *substructure* of actions within a model component is also evaluated, e.g. whether the setup improves during several trials. This detailed *sequence analysis* yields a score on a five-stage ordinal scale for students' actions within each model component.

Sequence analysis was applied to the videos of the hands-on experiments as well as to the screen recordings of the mouse-on experiments and even to the log data of the mouse-on experiments. The inter-coder reliability is high (e.g. Cohen's kappa = 0.91 for the setup phase). The validity was confirmed by comparison with a high inferent expert rating of students' experimental competencies (Dickmann et al. 2012).

5.2.3 Results

Research question 2 refers to the reliability of log-file analyses for coding students' experimental actions in a simulation environment. Our data show that algorithms can provide dependable action diagrams. Human coders then only have to make the qualitative coding decisions according to the correctness (e.g. of the setup). This semiautomatic approach significantly simplifies the analysis and leads to reliable final scores (Cohen's kappa > 0.77).

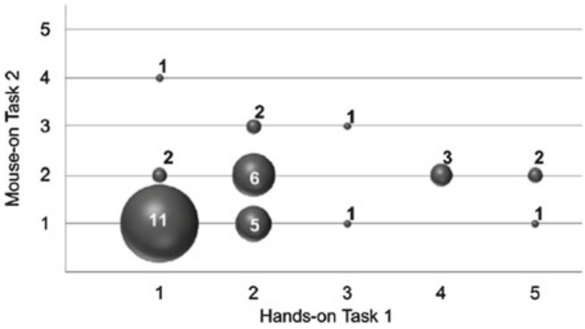


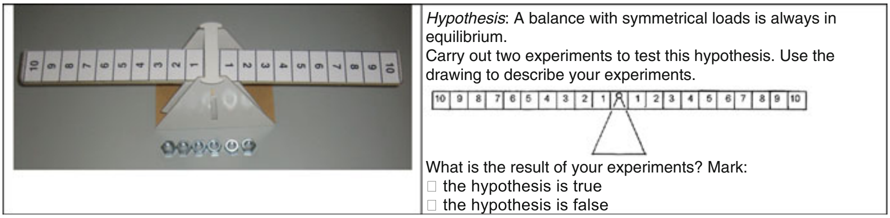
Fig. 20.3 Bubble diagram showing the scores students achieved in the hands-on task and in the mouse-on task (Kendall-tau-b = 0.337*)

While research question 2 can be answered positively, this is not the case for research question 1.

The sequence analysis yields two scores for a student's experimental competence: one score from the hands-on test and one from the mouse-on test. Although from the physics point of view the tasks are very similar as far as experimental actions are concerned, the correlations between the achievement scores are much lower than expected (Kendall-Tau-b < 0.4). There are dramatic changes of individual students' scores between tasks (cf. Fig. 20.3).

5.2.4 Discussion

At first glance, it seems that mouse-on tests are no valid substitutes for hands-on tests with regard to the diagnostics of experimental competence. Figure 20.3 shows a ground effect (both tasks seem to be too difficult/there are too many low achievers) as well as the instability of students' achievements. Several students reach high scores for the one task and only low scores for the other task. On the other hand, a qualitative comparison of the hands-on and mouse-on tests across all test persons shows very similar patterns of successful experimenting and very similar mistakes in both environments, e.g. students use a battery in addition to the power supply or choose wrong settings for the multimeters in both settings. Obviously, the challenges and difficulties that lead to success or failure rather result from the physics content than from the test format. Although the two experimental tasks require very similar physics content knowledge as well as similar experimental actions, students could not apply their knowledge consistently in both experimental situations. Subsequent analyses (cf. Schreiber 2012) show that the results of hands-on and mouse-on tests do correlate, when one compares the distributions of the achievements in the hands-on and the mouse-on assessment instead of comparing individual students' scores. Thus, the "mouse-on" technique appears to be promising for large-scale assessments.



Hypothesis: A balance with symmetrical loads is always in equilibrium.
Carry out two experiments to test this hypothesis. Use the drawing to describe your experiments.

What is the result of your experiments? Mark:
□ the hypothesis is true
□ the hypothesis is false

Fig. 20.4 The balance experiment (*left*) and an excerpt from the paper and pencil worksheet (*right*; reduced reproduction)

5.3 Item Difficulties of Large-Scale Hands-On Tests

Students' performances in an experimental competence test are influenced by more aspects than only the demands of the experimental task or the students' content-related subject matter knowledge. Item characteristics like the format of the problem presentation and the answer format can be expected to affect item difficulties. The aim of Gut's study (2012) was to find out to what extent the HarmoS experimental test is sensitive to item features that are directly relevant to the experimental competencies compared to other features like format aspects.

5.3.1 Design, Assessment Tools and Data Collection

As part of the project HarmoS (Labudde et al. 2012), a hands-on test was carried out with 738 students of the 6th and the 9th grade in the German-speaking part of Switzerland in spring 2008. Fifteen experimental units, similar to the units of the TIMSS performance assessment in Harmon et al. (1997), covering different content domains within physical or living systems were used. The units contain 95 items that relate to one of the five sub-dimensions of laboratory work mentioned in Sect. 4.2.

The aim of the test was to provide evidence about the experimental competencies of Swiss students and to formulate basic standards for labwork on the base of these test results. Therefore, the tasks had to ensure an authentic assessment of a wide range of meaningful classroom activities (see Messick 1994). Furthermore, the results of the test were used to validate the HarmoS competence model (see above; Gut and Labudde 2010; Labudde et al. 2009). In order to make the test sensitive for low achievers, the tasks were developed according to Solano-Flores and Shavelson (1997) by repeated pilot tests with small groups of students. The task sheets were simplified after each repetition. The simplifications addressed item features closely related to experimental competencies as well as general item features (e.g. language, answer format). In the main test, each student worked on two experimental tasks during a session of 60 min.

Figure 20.4 shows an example of a hands-on experiment with an excerpt from the corresponding paper and pencil worksheet. In this task, the students had to

Table 20.1 Item features used for the analysis of item difficulty (for the grey features see below)

Features irrelevant to exp. competencies		Features relevant to exp. competencies	
"To grasp the task"	"To give the answer"	"To solve the problem"	"To code the solution"
Language	Gap formats	Problem	Correctness
Text length	Empty space	Processes	Heuristics
Sentence structure	Empty lines	Task type	Theory
Text coherence	Figures	Subject/context	Evidence
	Multiple choice		Logic
	…		Technique/practical
Content input	Fill formats	Solution	Quality
Textual inputs	Describing	Structuredness	Precision of measurement
Figural inputs	Terming	Openness	Precision of observation
	Marking		
	Drawing		
Description	Description		Completeness
Problem description	Task description		Gaps
			Specifications

"verify" or "falsify" a given hypothesis about the equilibrium of a lever by performing two experiments with a mathematical balance and six weights.

5.3.2 Methods: Analysis of Item Difficulty

Explorative analyses of the labwork tasks were done post hoc (Gut 2012). For clarifying the content validity of the test, the correspondence of item features and item difficulties was studied.

For the analysis, a conceptual framework was developed, distinguishing four dimensions of item difficulties with corresponding sets of item features. These dimensions relate to the four stages of performing and evaluating a task: "to grasp the task", "to solve the problem", "to give the answer" and "to code the solution", comprising about three dozens of item features (cf. Table 20.1). The two stages "to grasp the task" and "to give the answer" correspond to demands such as reading and writing difficulties that are not relevant to the experimental competencies. The stage "to solve the problem" relates to the complexity of the experimental task, whereas in the stage "to code the solution" the quality of the experimental processes is rated. Item features corresponding to both stages are relevant to the experimental competencies. The rather large number of analysed features was necessary to cope with the heterogeneity of the tasks with respect to the experimental problems, the task description and the coding systems.

The analysis was performed in three steps. First, the difficulties of the 95 items were calculated within a one-dimensional Rasch analysis by the program ConQuest 2.0. Second, the variables corresponding to the item features were dichotomised, mainly by a median split. At last, a multiple regression was calculated with the item difficulty as the dependent variable and the item features as the predictor variables. According to the method described by Prenzel et al. (2002) the item variables were reduced by iterative regression calculations.

5.3.3 Results

As result of the data analyses, 12 item features, highlighted in Table 20.1, can be identified that explain significantly 44 % of the variance of the item difficulties (Gut 2012). The analysis shows that the HarmoS hands-on test is sensitive to demands relevant to experimental competencies: For instance, the coding of the adequacy of practical-technical handling and the precision of measurement make items more difficult. But, as shown in Table 20.1 with the significant item features highlighted, the test is not sensitive to the crucial working step “to solve the problem”. Therefore, the analysis of item difficulty post hoc does not help to explain competence progression. On the other hand, the test is also sensitive to demands not relevant to practical experimenting. The majority of significant item features belongs to one of the two working steps “to grasp the task” and “to give the answer”. We assume that the test scores are influenced by the reading and writing competencies of the students to a rather large extent.

5.3.4 Discussion

From a psychometric point of view, the test analysis is tenuous to a certain extent. In relation to the 12 significant predictor variables, the number of 95 explained items is rather small. Furthermore, some results cannot be explained plausibly, e.g. that the use of the gap format “empty space” makes items easier. Both results may be a consequence of the difficulty to model heterogeneous experimental tasks with respect to the types of problem and the scoring systems. On the other hand, the heterogeneity of the test is necessary when authentic and meaningful classroom activities are to be assessed.

6 Conclusion

The studies reported in this chapter emphasise that the assessment of experimental competencies as well as of the effects of labwork still pose challenges to science education research. The study presented in Sect. 3 applied elaborate empirical research methods for comparing a problem-based guided enquiry approach with a conventional “cookbook”-style labwork setting. The conclusion is that the effects of non-standard laboratory learning environments cannot be measured with standard instruments. The validation of rubrics (Sect. 5.1) yields that these instruments are useful tools for the assessment of experimental competencies in science education studies with respect to content validity and reliability. However, their potentials and limitations have to be considered adequately. Rubrics can be used to analyse the *documentations* and *products* of labwork (summative data). The analysis of *processes* in the lab, i.e. the quality of students' *actions*, needs formative data and further methods. In the eXkomp study (Sects. 4.1 and 5.2) students worked on non-trivial tasks compatible with the syllabus of physics instruction. Using hands-on and mouse-on tests, the researchers found that individual students often worked on two consecutive similar experimental tasks in rather different ways. The very particular physics content of the experimental task seemed to play an important role for the individual student. This leads to the thesis, that the more advanced the physics topic of an experimental task is, the more challenging becomes the reliable measurement of experimental competencies distinct from content knowledge. A similar conclusion is drawn from the evaluation of the HarmoS test (Sects. 4.2 and 5.3): The more authentic an experimental test is, in the sense of real and meaningful classroom activities, the more challenging becomes the valid measurement of experimental competence. The analysis of the relationships between item features and item difficulties in a large-scale hands-on test yielded that students' performances were very sensitive to the formulation of the task and the prescribed answer format. The findings of the projects eXkomp and HarmoS correspond with other research reports showing a volatile picture of students' achievements in studies on experimental competence (e.g. Shavelson et al. 1999; Rosenquist et al. 2000). The question is to which extend this instability is due to the assessment methods or to inherent performance variations on the side of the students. Thus further investigations are necessary with a broader variety of experimental tasks within a specific content area and across different content areas.

In summary, the results of the studies presented in this chapter emphasise that the various objectives of experimenting in physics education and the complexity of the construct "experimental competencies" itself demand the development of highly adapted assessment tools. Summative and formative data should be included in the analyses of students' performances. The tools have to be validated carefully. Researchers should cooperate closely to exchange instruments and to share experiences.

References

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development, 70*(5), 1098–1120.

Dickmann, M., Schreiber, N., & Theyßen, H. (2012). Vergleich prozessorientierter Auswertungsverfahren für Experimentaltests. In S. Bernhold (Ed.), *Konzepte fachdidaktischer Strukturierung für den Unterricht* (pp. 449–451). Berlin: Lit-Verlag.

Diehl, J. M., & Staufenbiel, T. (2002). *Statistik mit SPSS – Version 10 + 11*. Eschborn: Verlag Dietmar Klotz.

EDK, Schweizerische Konferenz der kantonalen Erziehungsdirektoren. (2011). *Grundkompetenzen für die Naturwissenschaften. Nationale Bildungsstandards*. Bern: EDK. http://edudoc.ch/record/96787/files/grundkomp_nawi_d.pdf. Accessed 30 Apr 2012.

Fraser, B. J., Giddings, G. J., & McRobbie, C. J. (1995). Evolution and validation of a personal form of an instrument for assessing science laboratory classroom environments. *Journal of Research in Science Teaching, 32*, 399–422.

Grube, C., Möller, A., & Mayer, J. (2007). Dimensionen eines Kompetenzstrukturmodells zum Experimentieren. In H. Bayrhuber et al. (Eds.), *Ausbildung und Professionalisierung von Lehrkräften, internationale Tagung der Fachgruppe Biologiedidaktik im Verband Biologie, Biowissenschaften, Biomedizin* (pp. 31–34). Kassel: Universität Kassel.

Gut, C. (2012). Modellierung und Messung experimenteller Kompetenz: Analyse eines large-scale Experimentiertests. Ph.D. thesis. Berlin: Logos.

Gut, C., & Labudde, P. (2010). Assessment of students' practical performance in science: The Swiss HarmoS project. In G. Çakmaki & M. Taşar (Eds.), Contemporary science education research: Learning and assessment. ESERA proceedings 2009 (pp. 295–298). Istanbul: Pegem Akademi.

Hammann, M. (2004). Kompetenzentwicklungsmodelle: Merkmale und ihre Bedeutung – dargestellt anhand von Kompetenzen beim Experimentieren. *MNU, 57*, 196–203.

Harmon, M., Smith, T. A., Martin, M. O., Kelly, D. L., Beaton, A. E., Mullis, I. V. S., Gonzalez, E. J., & Orpwood, G. (1997). *Performance assessment in IEA's third international mathematics and science study*. Chestnut Hill: TIMSS International Study Center, Boston College.

Henke, C., & Sumfleth, E. (2006). Leistungsmessung in der Oberstufe mit chemischen Experimentalaufgaben. In A. Pitton (Ed.), *Gesellschaft für Didaktik der Chemie und Physik: Lehren und lernen mit neuen Medien* (pp. 340–342). Berlin: Lit-Verlag.

Hofstein, A., & Lunetta, V. N. (2004). The laboratory in science education: Foundations for the twenty-first century. *Science Education, 88*(1), 28–54.

Hopf, M. (2011). Measuring effects of non conventional labwork. In C. Bruguière, A. Tiberghien, & P. Clément (Eds.), *E-book proceedings of the ESERA 2011 conference*, Lyon. http://lsg.ucy.ac.cy/esera/e_book/base/ebook/strand10/ebook-esera2011_HOPF-10.pdf. Accessed 1 May 2012.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review, 2*, 130–144.

Klahr, D., & Dunbar, K. (1988). Dual space search during scientific reasoning. *Cognitive Science, 12*, 1–48.

KMK. (2004). *Bildungsstandards im Fach Physik für den Mittleren Schulabschluss*. München: Luchterhand.

Labudde, P. (2007). How to develop, implement and assess standards in science education, 12 challenges from a Swiss perspective. In D. Waddington, P. Nentwig, & S. Schanze (Eds.), *Making it comparable: Standards in science education* (pp. 277–301). Münster: Waxmann.

Labudde, P., Nidegger, C., Adamina, M., & Gingins, F. (2012). The development, validation, and implementation of standards in science education: Chances and difficulties in the Swiss project HarmoS. In S. Bernholt, K. Neumann, & P. Nentwig (Eds.), *Making it tangible: Learning outcomes in science education* (pp. 235–259). Münster/New York/München/Berlin: Waxmann.

Labudde, P., Metzger, S., & Gut, C. (2009). Bildungsstandards: Validierung des Kompetenzmodells. In D. Höttecke (Ed.), *Chemie- und Physikdidaktik für die Lehramtsausbildung, GDCP Jahrestagung 2008* (pp. 307–317). Berlin: LIT Verlag.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology, 28*, 563–575.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessment. *Educational Researcher, 23*(2), 13–23.

Möller, A., Grube, C., & Mayer, J. (2007). Kompetenzniveaus der Erkenntnisgewinnung bei Schülerinnen & Schülern der Sekundarstufe I. In H. Bayrhuber et al. (Eds.), *Ausbildung und Professionalisierung von Lehrkräften, internationale Tagung der Fachgruppe Biologiedidaktik im Verband Biologie, Biowissenschaften & Biomedizin* (pp. 55–58). Kassel: Universität Kassel.

National Research Council (NRC) (1996). *National science education standards*. Washington, DC: National Academy Press.

Nadji, T., Lach, M., & Blanton, P. (2003). Assessment strategies for laboratory reports. *The Science Teacher, 4*(1), 56. ftp://ftp.aip.org/epaps/phys_teach/E-PHTEAH-41-022301/Rubric.doc 2-12-2010. Accessed 7 May 2012.

Niedderer, H., Tiberghien, A., Buty, C., Haller, K., Hucke, L., Sander, F., Fischer, H. E., Schecker, H., Aufschnaiter, S., & Welzel, M. (1998). Category based analysis of videotapes from labwork (CBAV) – Method and results from four case-studies. http://www.idn.uni-bremen.de/pubs/Niedderer/1998-WP9.pdf. Accessed 8 August 2013.

Prenzel, M., Häussler, P., Rost, J., & Senkbeil, M. (2002). Der PISA-Naturwissenschaftstest: Lassen sich die Aufgabenschwierigkeiten vorhersagen? *Unterrichtswissenschaft, 30*(1), 120–135.

QCA (Qualifications and Curriculum Authority) (Eds.) (2004). *Science – The national curriculum for England* (Revised 2004). London: Department for Education and Skills.

Ramseier, E., Labudde, P., & Adamina, M. (2011). Validierung des Kompetenzmodells HarmoS Naturwissenschaften: Fazit und Defizite. *Zeitschrift für Didaktik der Naturwissenschaften, 17*, 17–33.

Rosenquist, A., Shavelson, R. J., & Ruiz-Primo, M. A. (2000). *On the "exchangeability" of hands-on and computer-simulated science performance assessment.* Technical report, Center for the Study of Evaluation; National Center for Research on evaluation, Standards, and Student Testing. Graduate School of Education and Information Studies; University of California, Los Angeles.

RPAEG. (2008). *Scientific ability rubrics (Rutgers physics and astronomy education group).* http://paer.rutgers.edu/ScientificAbilities/Rubrics/default.aspx. Accessed 7 May 2012.

Schreiber, N. (2012). Diagnostik experimenteller Kompetenz. Validierung technologiegestützter Testverfahren im Rahmen eines Kompetenzstrukturmodells. Berlin: Logos.

Schreiber, N., Theyßen, H., & Schecker, H. (2009). Experimentelle Kompetenz messen?! *Physik und Didaktik in Schule und Hochschule, 8*(3), 92–101.

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, Lyon. http://lsg.ucy.ac.cy/esera/e_book/base/ebook/strand10/ebook-esera2011_SCHREIBER-10.pdf. Accessed 1 May 2012.

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.

Singer, S. R., Hilton, M. L., & Schweingruber, H. A. (2006). *America's lab report: Investigations in high school science*. Washington: National Academies Press.

Solano-Flores, G., & Shavelson, R. J. (1997). Development of performance assessments in science: Conceptual, practical and logistical issues. *Educational Measurement: Issues and Practice, 16*(3), 16–25.

Taylor, P. C., Fraser, B. J., & Fisher, D. L. (1997). Monitoring constructivist classroom learning environments / advances in research on educational learning environments. *International Journal of Educational Research, 27*(4), 293–302.

Van Heuvelen, A. (1995). Experiment problems for mechanics. *The Physics Teacher, 33*(3), 176–180.

Walpuski, M. (2006). *Optimierung von Kleingruppenarbeit durch Strukturierungshilfen und Feedback*. Berlin: Logos.