MODELLING AND ASSESSING EXPERIMENTAL COMPETENCE: AN INTERDISCIPLINARY PROGRESS MODEL FOR HANDS-ON ASSESSMENTS

Susanne Metzger, Christoph Gut, Pitt Hild, and Josiane Tardent
Zurich University of Teacher Education

Abstract: On the lower secondary level in Swiss schools, biology, chemistry, and physics usually are taught as one subject. Accordingly, the national education standards do not differentiate between the three sciences. In order to assess standards of experimental competence, we developed a model that spans all three sciences. In this model, experimental competence is structured by sub-dimensions referring to experimental problem types such as «categories conducted observation», «measurement with a given scale», «scientific investigation», «experimental comparison», or «constructive problem solving». The progression of competence is modelled for each problem type separately, differentiating three to five levels in terms of quality standards. In a first attempt to validate the progress model six tasks for «categories conducted observation» and «measurement with a given scale» have been developed. A pilot test was administered to 250 students (grade 7, 8, and 9) of different levels. The results of the pilot test affirm that the progression model can be applied reasonably to all three sciences. The tasks can be standardised well with respect to the test sheet structure, question formats, and textual demands. Both, the structure of our model and the hierarchy of quality standards seem to be applicable.

Keywords: science, experimental competence, hands-on assessment, lower secondary school, Switzerland

INTRODUCTION
In Swiss schools, science usually is taught as one subject on the lower secondary level. Accordingly, the national education standards do not differentiate between biology, chemistry, and physics. Within the Swiss project HarmoS (Labudde et al., 2012) an interdisciplinary structure model of scientific competence was developed. The validation of the model by a large-scale hands-on test showed that the progression of experimental competence cannot be explained post hoc (Gut, 2012). Based on these results, we developed a new interdisciplinary normative progression model for experimental competence for practical assessments. In this paper, the model and first results of the validation by pilot assessments are presented.
RATIONALE
According to Schecker & Parchmann (2006) there are different purposes of normative competence models: Such models should help to define competence by determining structure and progression. They should be practical in formulating adequate standards for experimenting in school science and useful for teacher education. They should also provide an appropriate basis for the development of valid and reliable assessments of students’ performance (e.g. Kauertz et al., 2012; Lunetta et al., 2007). In order to attain these goals, four kinds of a priori decisions have to be made when experimental competence is modelled. First, one has to define which scientific problems are to be standardised and assessed. Second, one has to decide how competence may be decomposed in sub-dimensions. In the often-used process approach, problem solving is conceived as a linear chain of processes (Murphy & Gott, 1984), such as formulating a hypothesis, planning and carrying out experiments, and analysing data (Emden & Sumfleth, 2011). Alternatively, one can differentiate between types of problems. The solution of each type demands specific knowledge and skills (Millar et al., 1996). Therefore, each solution is scored by typical scoring schemes and criteria (Ruiz-Primo & Shavelson, 1996). Thirdly, it has to be decided whether the progression of competence is modelled in terms of task complexity (e.g. Wellnitz et al., 2012), or in terms of quality criteria that standardised problems are solved with (Millar et al., 1996), or in terms of both simultaneously. The fourth decision concerns the kind of assessment (hands-on, simulation, or paper and pencil test) by which the competence is to be measured. Of course, these four decisions are not independent of each other. The process approach leads to a restricted view of experimental activities, excluding engineering tasks for instance. Also, considering the variety of problem types, the complexity cannot be empirically explained based upon one single progression model (Gut, 2012).

MODEL OF EXPERIMENTAL COMPETENCE
In order to attain the goals above, we developed a competence model on the basis of the following principals: In our conception, experimental competence refers only to problems with an authentic hands-on interaction, involving scientific questions as well as engineering tasks. Experimental competence is structured by sub-dimensions referring to various problem types such as «categories conducted observation», «measurement with a given scale», «scientific investigation», «experimental comparison», or «constructive problem solving». As we already mentioned before, the progression of competence is modelled for each problem type separately differentiating three to five levels in terms of quality standards for the solution of a standardised problem. These quality standards correspond to different subtasks. For example, the basic problem of «measurement with a given scale» is to measure a property of an object with given instruments as precisely as possible. The students have the options to repeat measurements and to select an instrument. The options correspond to different dimensions of the openness of a problem (Priemer, 2011). Using these options appropriately, students can reach higher progression levels, as shown in figure 1.
METHOD

For the first internal validation six tasks have been developed: three tasks for «categories conducted observation» and three tasks for «measurement with a given scale» within selected topics in biology, chemistry, and physics. An example for the problem type «measurement with a given scale» in physics is the task «thread», where students should find out at what force a thread breaks. They get a thread, scissors, two different spring scales (A and B) and a calculator. Students are asked to think about how they could answer the question, with which instrument they can do it best and how many measurements would be necessary. After finding the result they have to draw or write down, how they did the measurement. In addition they are explicitly asked to tell with which instrument they did the measurement. Afterwards they are asked to estimate how exact the measurement is and how they could improve it. At the end, there are some control questions, such as which spring scale they used or if they calculated a mean and how they did it.

The pilot assessment has been administered to 250 students of different grades (7, 8, and 9) and levels of the lower secondary school, especially to low achievers. In the following, «low level» indicates the lowest achievement level, whereas «high level» indicates the highest level of the so called «Sekundarschule» (in Switzerland, «Gymnasium» is even higher). The students had to solve three tasks, each in 20 minutes. They worked on their own with printed test sheets on which they were asked to write down their answers and a brief report. Each task was coded by minimum two persons and as one single item, i.e. all answers have been evaluated as a whole: Several dichotomous quality criteria could be

Figure 1: Progression model for the problem type «measurement with a given scale»; the hierarchy of quality standards is set a priori.
achieved or not. These quality criteria were clustered to the quality standards (see figure 1) which could be achieved or not. To achieve a quality standard 50% or more of the quality criteria had to be achieved.

The item score could be measured in two ways: On the one hand one can sum all achieved quality standards (unconditional level score: uLev). On the other hand one can use the hierarchy of quality standards and set the item score to that level, where all lower quality standards are achieved (conditional level score: cLev). E.g. if a student achieves quality standards 1, 2, 3 and 5 in a task, his unconditional level score would be uLev = 4 and his conditional level score would be cLev = 3. The advantage of conditional levelling is to get back information, which is lost by summing up scores.

RESULTS
The results of the first pilot assessment affirm that the progression model for the two problem types «categories conducted observation» and «measurement with a given scale» can be applied reasonably to all three science subjects. 1-dimensional Rasch analyses (with the program Winsteps) for each problem type show good item fits. For conditional level score we found for the three tasks of «categories conducted observation».96 < infit < 1.13 and .82 < outfit < 1.24 and for the three tasks of «measurement with a given scale».81 < infit < 1.22 and .76 < outfit < 1.32. At least for the «measurement with a given scale» sufficiently high reliability (> .6) is achieved.

Structure of the model
The low correlations between the two latent variables for «categories conducted observation» and for «measurement with a given scale» (unconditional levels: .404**; conditional levels: .368**) indicate that the structure of the model (i.e. differentiation of problem types) seems to be reasonable. For the problem type «categories conducted observation» three quality standards could be distinguished:
1. single observation: correctness and completeness,
2. identification of differences,
3. identification of similarities.
For «measurement with a given scale» five levels as showed in figure 1 could be found, but the a priori hierarchy of quality standards had to be changed: identifying sources of measurement errors seems to pose higher difficulties to students than documentation of results, measurement repetitions, and instrument selection. Therefore the new hierarchy of the progression model is:
1. single measurement,
2. documentation of results,
3. measurement repetition,
4. instrument selection,
5. sources of measurement errors.
To validate these progressions we compared the frequencies of achieved quality standards for each task. As an example, figure 2 shows the frequencies of achieved quality standards for the task «thread».

![Figure 2: Frequencies of achieved quality standards for the task «thread», differentiated in four groups (low and high levels of grade 7 and 9 of the «Sekundarschule»).](image)

Figure 2 shows some misfits of the model: First, frequencies of every group should be smaller the higher (i.e. the more right on the x-axis) a quality standard is. Secondly, frequencies of high level groups of the same grade and also higher grade groups of the same level should be higher for every quality standard. Not each task shows the same misfits as shown in figure 2. But generally the differentiation and the hierarchy of level 3 (multiple measurement) and 4 (instrument selection) is critical. More tests have to show, if these two levels have to be changed or put together.

**Validation of tasks**

The tasks can be standardised well with respect to the structure of the test sheet, the question formats, and textual demands. All tasks were coded by minimum two persons. On the level of quality criteria as well as on quality standards a high interrater correlation (> .8) can be achieved. Nevertheless, for «categories conducted observation» tasks more rater training is needed than for «measurement with a given scale» tasks. Separate 1-dimensional Rasch analyses with unconditional and with conditional levels show high correlations between the two scoring alternatives («categories conducted observation»: .940**; «measurement with a given scale»: .847**). Therefore, in order to gain more information it seems to be reasonable to work with the unconditional levels instead of the conditional ones.
**Students’ performance**

To make the results comparable, we transformed the item parameters to skill points. Based on PISA and other large scale assessments, we set the mean of all students of grade 9 to 500 with a standard deviation of 100. Table 1 presents the results for the tasks of the problem types «categories conducted observation» and «measurement with a given scale». For «categories conducted observation», increases from grade 7 to 9 (high level) and from low to high level in grade 9 are significant with a value of almost a standard deviation. All increases from grade 7 to 9 (both levels) and from low to high level (in grade 7 and 9) are significant with a value of almost a standard deviation.

Table 1

*Students’ performance for tasks of the problem types «categories conducted observation» (left) and «measurement with a given scale» (right). In each case is given the mean and the standard deviation in parenthesis. U stands for Mann-Whitney-U-test, t for t-test and n.s. for not significant.*

<table>
<thead>
<tr>
<th>«categories conducted observation»</th>
<th>low level</th>
<th>Δ</th>
<th>high level</th>
</tr>
</thead>
<tbody>
<tr>
<td>grade 7</td>
<td>437 (79)</td>
<td>n.s.</td>
<td>452 (93)</td>
</tr>
<tr>
<td>Δ</td>
<td>n.s.</td>
<td>U: .006</td>
<td></td>
</tr>
<tr>
<td>grade 9</td>
<td>459 (86)</td>
<td>t: .011</td>
<td>537 (99)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>«measurement with a given scale»</th>
<th>low level</th>
<th>Δ</th>
<th>high level</th>
</tr>
</thead>
<tbody>
<tr>
<td>grade 7</td>
<td>365 (101)</td>
<td>U: .000</td>
<td>450 (107)</td>
</tr>
<tr>
<td>Δ</td>
<td>U: .000</td>
<td>U: .008</td>
<td></td>
</tr>
<tr>
<td>grade 9</td>
<td>458 (83)</td>
<td>t: .009</td>
<td>538 (101)</td>
</tr>
</tbody>
</table>

With these results we can assign the students to the levels shown in table 2.

Table 2

*Allocation of students’ performance to quality standard levels for tasks of the problem types «categories conducted observation» (left) and «measurement with a given scale» (right).*

<table>
<thead>
<tr>
<th>«categories conducted observation»</th>
<th>low level</th>
<th>high level</th>
</tr>
</thead>
<tbody>
<tr>
<td>grade 7</td>
<td>I</td>
<td>I</td>
</tr>
<tr>
<td>grade 9</td>
<td>I</td>
<td>II</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>«measurement with a given scale»</th>
<th>low level</th>
<th>high level</th>
</tr>
</thead>
<tbody>
<tr>
<td>grade 7</td>
<td>I</td>
<td>II</td>
</tr>
<tr>
<td>grade 9</td>
<td>II</td>
<td>IV</td>
</tr>
</tbody>
</table>

For «categories conducted observation», students in grade 7 and in the low level of grade 9 achieve level I, i.e. they observe a phenomenon correct and complete. Only high level students of grade 9 achieve level II, i.e. they also identify differences between two
observations. For «measurement with a given scale», low level students in grade 7 achieve level I, i.e. they measure correct. High level students of grade 7 and low level students of grade 9 achieve level II, i.e. they also prepare a well documentation of results. Finally, high level students of grade 9 attain level IV, i.e. in addition they repeat measurements and select the right instruments.

**CONCLUSION AND OUTLOOK**

The results of the first pilot assessment suggest that the model works: On the one hand, «categories conducted observation» and «measurement with a given scale» can be discerned as two different problem types, so the structure of our model (i.e. differentiation of problem types) seems to be applicable. On the other hand, we can show a hierarchy of quality standards for «measurement with a given scale» as well as for «categories conducted observation», so the progression of out model seems also to be applicable. Nevertheless, it has to be mentioned that these results are based on a sample of 250 Swiss students. Currently further internal validation are undertaken for other problem types such as «scientific investigation» and «experimental comparison». These results are expected in spring 2014.

Because of the low number of test items a quantitative analysis of the pilot test is problematic. For a valid assessment, it is planned to develop more tasks and to assess these with larger student samples.

**REFERENCES**


