Development of an Assessment Tool to Probe Students' Understanding of Measurement Uncertainty

Johannes Schulz and Burkhard Priemer
Humboldt-Universität zu Berlin

Abstract

On the basis of the development of a model that structures and describes the relevance of measurement uncertainties for secondary school levels (Hellwig, 2012; Priemer & Hellwig, 2016) we look at how the concepts and dimensions – proposed by the model – can be operationalized and measured. Selected dimensions and concepts given by the model are formulated by learning progressions and are seen as latent constructs. For these constructs, scales are developed based on Item Response Theory. The main aim is to construct a high-quality assessment tool to probe students' understanding of measurement uncertainties. Such a tool can help to evaluate learning progressions and teaching instructions. In this paper we will give a short overview about parts of the results of our first pilot study concerning 5 of the 10 concepts proposed by the above mentioned model.

Keywords

Measurement Uncertainties, Secondary education; scientific literacy

1 Introduction

Practical work with measurements is important in science and engineering where capturing and discussing measurement uncertainties is essential to assess the quality of an investigation. This is true for education as well when inferences are drawn from observations and measurements. We are able to assess the quality of our empirical results in science and engineering lessons only by looking at uncertainties. However, measurement uncertainties are an often-neglected topic in school curricula (Hellwig 2012). Hence, research is needed that develops and analyzes learning progressions and teaching instructions in order to bring this topic to teachers' attention and to facilitate effective teaching of these concepts.

When looking at research on measurement uncertainties in science education contexts we can find in general research of classifications and curriculum development (for example Deardorff 2001, Munier et al. 2012) and studies about students' views and problems when learning about measurement uncertainties (for example Buffler et al. 2001, Masnick & Morris 2008). Most of these studies mainly address upper secondary education and they consist of selections of different aspects of measurement uncertainties. Also, these assessments did not develop a comprehensive scientific model. In their work Priemer & Hellwig (2016) suggest a content structure model for the field of measurement uncertainty for secondary education which is structured in four main dimensions and ten concepts (fig. 1). The model contains additional subconcepts which are not shown here.

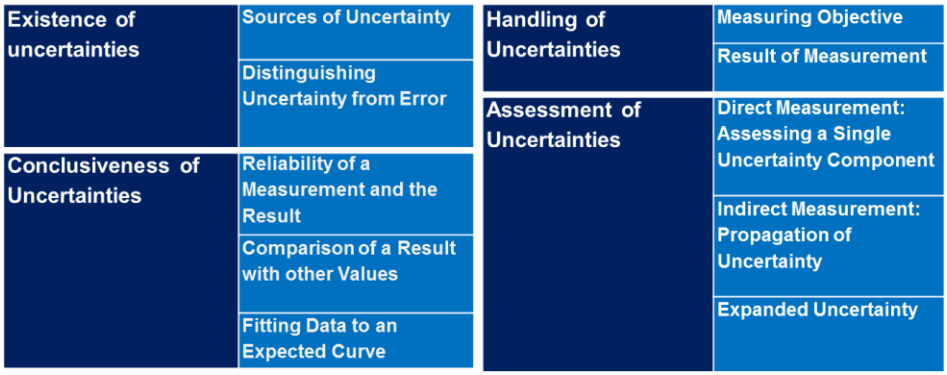

Fig. 1: A content structure model for the field of measurement uncertainty (Hellwig, 2012; Priemer & Hellwig, 2016)

If we assume that this model gives a valid structure of the content for secondary education, a next step is to identify the concepts as latent constructs with corresponding learning progressions. On this basis learning and teaching instructions can be developed and a test instrument can probe competencies achieved by students. In this paper we focus on the test instrument and the operationalization which is needed to develop the instrument fully.

Based on the model outlined above we pose the following two research questions:

1. How can the concepts of the model be operationalized and measured?

2. How well do the empirically-obtained results regarding the structure of the content fit with the theoretically developed model?

In this paper we give an overview of our work to answer question 1 and we present parts of the results of a pilot study in which we operationalized and measured 5 of the 10 concepts. The work contains the validation of items by expert-rating and the results of the Rasch-analysis.

2 Method

To tackle question 1 we had to assure that there are no intersections regarding the content of the different categories when formulating test items. This led us to a table of content knowledge for each of the concepts (and their subconcepts) which can be used for measuring the addressed competencies (an example is given in fig. 2).

Handing of Uncertainties		
Measuring Objective	Striving for an adequate measurement uncertainty	Setting a maximum value for the measurement uncertainty
		Adjusting the measurement

Measuring Objective	The students are able to…
	- **choose capable measurement instruments** for a certain process of measurement - **adjust a measurement process** to minimize measurement uncertainties, e.g. by changing the setting/process/method of measurement, the environment or measurement instruments - **set a maximum value** for the measurement uncertainty in a certain measurement process

Fig. 2: Competencies for the concept "Measuring Objective"

Then we constructed items for each of the given concepts according to the formulated competencies. An item as an example is shown in fig. 3:

Example Item:

Laura wants to measure the mass of a crown cap. Unfortunately her scale shows a value of 0g.

How can Laura adjust her measurement?

(More than one answer might be correct)

- ○ Laura can use a more sensitive scale to measure the weigh of one crown cap.
- ○ She can repeat her measurement (one cap on the scale) until the scale shows a value.
- ○ Laura can wait until the scale shows a value that is different from zero.
- ○ She can measure the mass of ten caps and divide this mass by ten.

Fig. 3: Example Item: Scale with crown cap

For every concept on the test we added also an introduction page which gives an overview of the content to assure that students can understand the scientific terms used. All of the items are given in multiple choice format. The set of items contains questions where students can only choose one answer and questions where students can choose more than one answer. For the multiple-answer questions, there was for every item at least one wrong answer and at least one correct answer.

To explore the quality of the items we used an expert rating for validation and gave the test to school and university students for a Rasch-analysis.

2.1 Expert Rating

In order to validate the items we presented a subset of 52 items to three experts in the field of metrology together with the table of all competencies for the concepts. The items that were given to the experts were chosen at random with the exception that there was at least one item in the subset for every concept. We then asked the experts to assign the given items to the concepts. We also added an eleventh category to the concepts for items that did not fit into any of the provided categories. Finally, the experts had room to give comments.

2.2 A Pilot Study with Students

In order to get an overview how the test instrument works we chose two different groups of students:

First, we presented two concepts to 143 pupils from the 8^{th} grade to the 12^{th} grade of six different classes of three different German schools. These were the concepts "Reliability of a Measurement and the Results" (17 items) and "Comparison of a result with other values" (17 items). All students were asked to answer all of the items. For half of the participants the order of the two concepts was changed. The order of the items was the same for every concept.

Second, we also presented three concepts ("Measuring Objective", "Sources of Uncertainty" and "Direct Measurement: Assessing a single uncertainty component") to 364 university students in their first year at a German university. In this study we use nine different subsets of the test items. Each student was asked to answer one set which contains six items for two of the concepts. The nine different subsets of items were designed that there were intersections between all of the items in all of the concepts (multimatrix design). The test was used in a biology lecture (116 students) and in two other lectures with no background in natural sciences (248 students).

In both cases the participants had as much time as they needed to answer. No school student took longer than 90 minutes and no university student took longer than 30 minutes. The answers were coded either 0 (false) or 1 (correct). This also was applied for the multi-answer items as well. They were coded as 1 (correct) if and only if all selections for the item were correct.

3 Results

First, we present the results of the expert rating followed by the results of the pilot study.

3.1 Results of the expert rating

The three experts gave the following rating: 31 items were sorted to the same concept, 14 items were sorted to the same concept by two of the three experts, and 7 items were sorted to three different concepts by the three experts. The inter rater agreement of the experts was $\kappa=0.67$ (Fleiss kappa).

The experts recommended some minor changes in the formulation of some of the items. It would take too long to list all the changes here, but most of them were just about changing word order, terms, or formulations in the language.

3.1 Results of the pilot study with students

The Rasch-analysis and additional tests were calculated with Winsteps and R (also R Studio). In the study with the university students the items with very low agreement among the experts (sorted to three different concepts by all experts) were deleted for the study and replaced by other items of the same concept. In the study with the school students this was unfortunately impossible due to time constraints.

Study with the school students. Figure 4 shows the Wright-Maps for the concepts "Reliability of a Measurement and the Results" and "Comparison of a result with other values". The EAP reliability of the Rasch-analyses for the concept "Reliability of a Measurement and the Results" is r = 0.54 and for the concept "Comparison of a result with other values" it is r = 0.8.

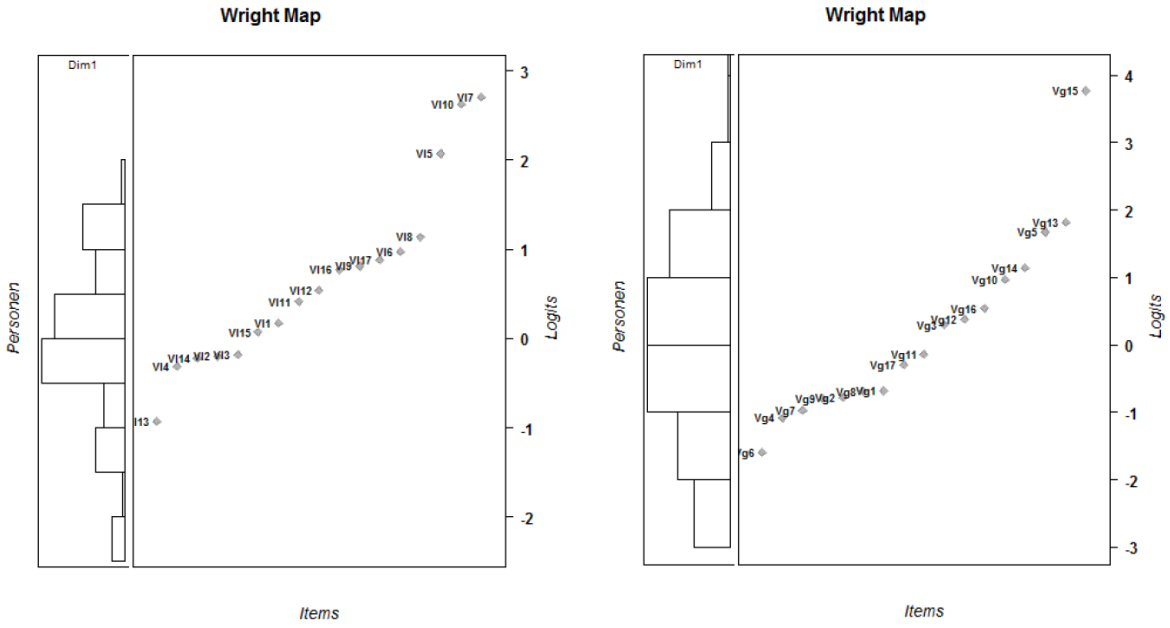

Fig. 4: Wright-Map for the concepts "Reliability of a Measurement and the Results" (left) and "Comparison of a result with other values" (right)

We also calculated the MNSQ Outfit values and the difficulty of the items (see fig. 5 and 6).

Item	estimate	MNSQ Outfit	Item	estimate	MNSQ Outfit
Vl1	0,17	0,99	Vl10	2,62	1,18
Vl2	-0,20	0,99	Vl11	0,41	1,06
Vl3	-0,18	0,98	Vl12	0,54	1,01
Vl4	-0,31	1,03	Vl13	-0,93	0,93
Vl5	2,07	1,11	Vl14	-0,22	1,03
Vl6	0,97	0,92	Vl15	0,07	0,93
Vl7	2,70	0,93	Vl16	0,77	1,01
Vl8	1,13	0,96	Vl17	0,88	0,98
Vl9	0,81	1,04			

Fig. 5: Items for concept "Reliability of a Measurement and the Results"

Item	estimate	MNSQ Outfit	Item	estimate	MNSQ Outfit
Vg1	-0,68	1,37	Vg10	0,95	0,95
Vg2	-0,78	1,02	Vg11	,095	0,95
Vg3	0,30	0,76	Vg12	0,94	0,94
Vg4	-1,08	0,78	Vg13	0,76	0,76
Vg5	0,85	0,85	Vg14	0,84	0,84
Vg6	0,79	0,79	Vg15	1,09	1,09
Vg7	0,99	0,99	Vg16	0,98	0,98
Vg8	0,81	0,81	Vg17	1,18	1,18
Vg9	1,26	1,26			

Fig. 6: Items for concept "Comparison of a result with other values"

Study with the university students. We calculated the difficulty of the items and the MNSQ values (see fig. 7 at the end of the paragraph). The Wright-Maps for the different concepts are shown in fig. 8. The EAP reliabilities are r = 0.22 for the concept "Sources of uncertainty", r = 0.18 for the concept "measurement objective" and r = 0.28 for the concept "Direct Measurement: Assessing a single uncertainty component".

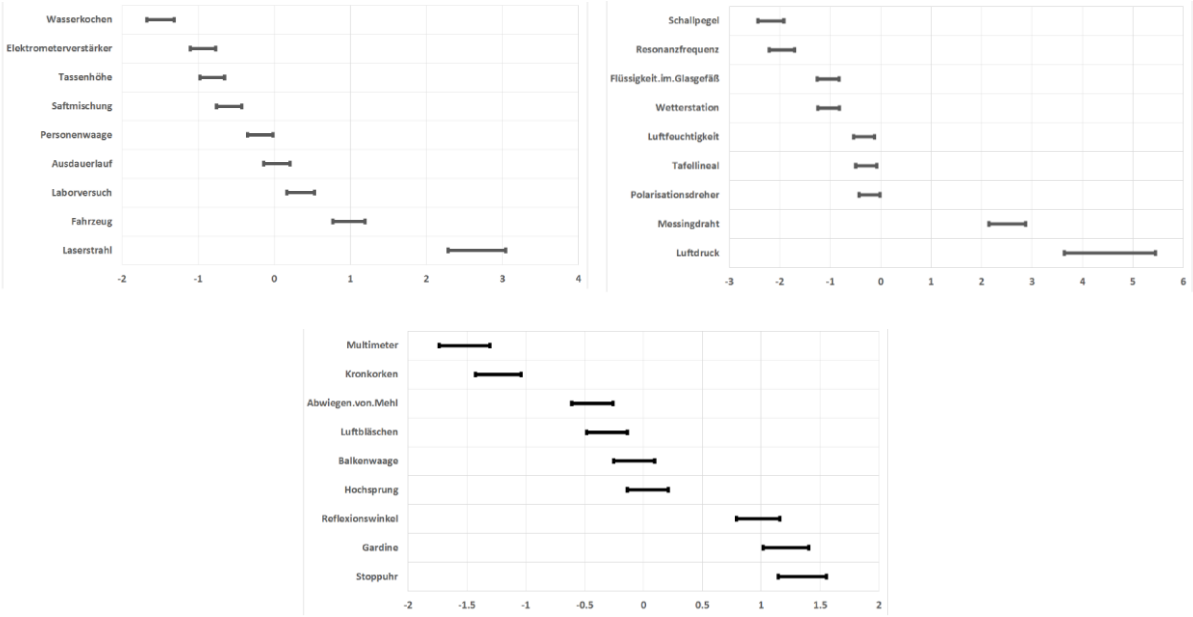

Fig. 8: Wright-Map of the Concepts "Sources of uncertainty" (top left), "Direct Measurement: Assessing a single uncertainty component" (top right) and "measurement objective" (bottom)

Further, we calculated the correlations between the person scores of the three concepts (see fig. 9).

Concepts	Correlation
Sources of Uncertainty and Direct Measurement: Assessing a single uncertainty component	0.29

Sources of Uncertainty and Measuring Objective	0.39
Measuring Objective and Direct Measurement: Assessing a single uncertainty component	0.28

Fig.9: Correlation of Concepts

Item	Parameterschätzung	Standardfehler	Infit	Outfit
Wasserkochen	-1.496	0.179	0,768	0,843
Elektrometerverstärker	-0.937	0.168	1,062	1,064
Tassenhöhe	-0.817	0.165	0,91	0,944
Saftmischung	-0.593	0.166	1,129	1,028
Personenwaage	-0.187	0.169	0,998	1,003
Ausdauerlauf	0.034	0.174	0,903	0,946
Laborversuch	0.347	0.181	0,929	0,962
Fahrzeug	0.983	0.212	0,713	0,83
Laserstrahl	2.666	0.38	0,524	0,67
Schallpegel	-2.178	0.26	0,699	0,747
Resonanzfrequenz	-1.959	0.249	0,727	0,985
Flüssigkeit im Glasgefäß	-1.045	0.214	0,776	0,891
Wetterstation	-1.035	0.215	0,873	0,987
Luftfeuchtigkeit	-0.33	0.208	1,184	1,165
Tafellineal	-0.286	0.209	0,877	0,901
Polarisationsdreher	-0.222	0.212	0,735	0,834
Messingdraht	2.51	0.363	0,913	0,756
Luftdruck	4.545	0.903	0,039	0,262
Multimeter	-1.52	0.216	0,711	0,869
Kronkorken	-1.234	0.196	0,824	0,903
Abwiegen von Mehl	-0.434	0.176	0,92	0,951
Luftbläschen	-0.308	0.173	0,867	0,898
Balkenwaage	-0.079	0.174	0,885	0,919
Hochsprung	0.037	0.175	1,011	1,031
Reflexionswinkel	0.975	0.186	0,932	0,878
Gardine	1.211	0.194	1,288	1,066
Stoppuhr	1.351	0.205	0,751	0,881

Fig. 7: Item difficulties and MNSQ for testing students (Concepts from top to bottom: “Sources of Uncertainty”, “Direct Measurement” and “Measurement objective”)

4 Discussion

4.1 Expert rating

The expert rating lead to a Fleiss kappa of κ=0.67 which is a “substantial agreement” according to Landis & Koch (1977). However, only a subset of the given items was rated. But since we removed the items with no agreement we can assume that overall the items are quite valid for the given concepts. Further, we have to keep in mind that assigning the items to the concepts is just one way of a validation. It doesn’t ensure for example that the content of the concept is covered completely. So even if the value of Fleiss kappa suggests a decent validity we have to look at other forms of validity.

The experts also suggested some minor changes in the formulation of some items, which will be considered in future tests of the material. Due to their low agreement, six items in the concept

"Reliability of a Measurement and the Results" and "Comparison of a result with other values" will be reconsidered and reformulated.

4.2 A pilot study with students

The Wright-Maps shows that for most of the concepts the distribution of the difficulty of the items generally fits the competencies of the school and university students quite well. There is neither a huge ceiling nor a floor effect. However, we see a lot of things that can be improved. For the concept "Reliability of a Measurement and the Results" we should add more items with low difficulty. And for the concepts "Sources of uncertainty" and "Direct Measurement: Assessing a single uncertainty component" more difficult items should be added to cover the competencies of the students better.

Some of the MNSQ values are not satisfactory, so some items need a review. For the concept "Reliability of a Measurement and the Results" the MNSQ-Outfit values are all inside the interval of 0.7 - 1.3 as recommended by Linacre and Wright (1994). Also for the concept "Comparison of a result with other" all items fall into the interval (except item Vg1). Since item Vg1 received high agreement by all three of the experts, we will keep it in the test but we will analyze it regarding misunderstandings by the participants. The items "Luftdruck" of the concept "Direct Measurement: Assessing a single uncertainty component" and "Laserstrahl" of the concept "Sources of Uncertainty" have also MNSQ-Outfits outside the suggested interval by Linacre and Wright (1994). In case of the item "Laserstrahl" this is likely caused by an error in the formulation of the item. This error of formulation would also explain the high difficulty of this item. In this case a change of the formulation could help to fix this problem. The item "Luftdruck" is valid for the given concept due to the expert rating. It seems that it is too difficult and causes people to answer at random. Therefore it should be replaced by an easier item or another multi-answer item with a lower chance of answering correctly by chance.

In the test with the university students there is also a problem with the uncertainty of the person score given by the Rasch-model which is greater than 0.6 for almost every person. On the other hand, the error for the difficulty of the items seems good. This suggests that there is a problem computing the person score. One reason can be that there is an insufficient number of anchor items in the multimatrix design since most of the items seem to be Rasch-conforming (the corresponding analysis are not shown here). This problem concerning the multimatrix design can also be the reason for the insufficient reliability of the test in all three concepts. However, the low reliability could be also explained by an insufficient number of items. Future work has to improve the reliability.

The correlation between the person scores of the given concepts suggests that the competencies among students are not caused by the same latent construct since the values are moderate. On the other hand the concepts are not completely independent. Unfortunately, due to the high uncertainty in the person scores the correlation between the concepts can just be seen as an indicator. To support our assumption, a more detailed factor analysis is needed.

Finally, there is also to mention that the results of the pupils and student tests highly depend on the sample set. We tried to find a sample set which we expected to cover most of the spectrum of competencies. But of course the given items might give other results in other sample sets. For

our main investigation we will try to tackle this problem by enlarging the sample set so that more students, classes and courses can be taken into account.

5 Conclusion

The results show that our test instrument to probe students' understanding of measurement uncertainty works well for some of the content covered. The scales and items have desired levels of difficulty and they cover the students' competencies (with a few exceptions) quite well. But the reliability of our measurement is insufficient. Therefore we have to improve the "problematic" items and the corresponding scales that reflect the concepts. However, the results from our first pilot study suggest that the concepts of the model are measurable by using multiple choice items. And the results also suggest that the concepts indeed trigger different competencies of the students. But further research is needed to confirm this hypothesis and to improve the quality of our test instrument.

Acknowledgments

We would like to thank Sarah Heydemann, Laura Kemnitzer and David Winderlich for their support in the pilot study. Also we would like to thank Amy Masnick for her helpful comments on an earlier draft of the paper.

References

1 Hellwig, J. (2012). Messunsicherheiten verstehen – Entwicklung eines normativen Sachstrukturmodells am Beispiel des Unterrichtsfaches Physik. Dissertation: http://www-brs.ub.ruhr-uni-bochum.de/netahtml/HSS/Diss/HellwigJulia

2 Priemer, B. & Hellwig, J. (2016). Learning About Measurement Uncertainties in Secondary Education: A Model of the Subject Matter. *International Journal of Science and Mathematics Education.* doi:10.1007/s10763-016-9768-0.

3 Linacre, JM & Wright, B. (1994). Reasonable mean-square fit values http://www.rasch.org/rmt/rmt83b.htm

4 Landis, J.R. & Koch, G.G. (1977). The measurement of observer agreement for categorical data: In: Biometrics, 33.

5 Masnick, A. M., & Morris, B. J. (2008). Investigating the development of data evaluation: The role of data characteristics. *Child Development, 79*, 1032-1048. doi:10.1111/j.1467-8624.2008.01174.x

6 Munier, V., Merle, H. & Brehelin, D. (2011). Teaching Scientific Measurement and Uncertainty in Elementary School. *International Journal of Science Education, iFirst Article*, 1-32. doi: 10.1080/09500693.2011.640360

7 Deardorff, D. (2001). Introductory physics students' treatment of measurement uncertainty (Diss., North State University, Raleigh, NC). https://www.ncsu.edu/PER/Articles/DeardorffDissertation.pdf

8 Buffler, A., Allie, S., Lubben, F. & Campbell, B. (2001). The development of first year physics students' ideas about measurement in terms of point and set paradigms. *International Journal of Science Education*, 23 (11), 1137-1156.

Name of the paper's First Author

Johannes Schulz, M.Ed. Dipl.-Math., Humboldt-Universität zu Berlin, Department of Physics

Name of the paper's Second Author

Burkhard Priemer, Prof. Dr., Humboldt-Universität zu Berlin, Department of Physics