Chapter 8
Assessing Sub-competencies of Mathematical Modelling—Development of a New Test Instrument

Corinna Hankeln, Catharina Adamek and Gilbert Greefrath

Abstract The distinction between different phases of a modelling process and thus of different sub-competencies for carrying out these processes is widespread in the modelling literature. In this chapter, we present our research on the assessment of these modelling sub-competencies. Based on a conceptual clarification of sub-competencies, we consider various ways of operationalising them into test items and present examples. With the help of psychometric models, we show that the sub-competencies of modelling, simplifying, mathematising, interpreting and validating, can be treated as separate dimensions, rather than being subsumed in a two-dimensional model, in which *simplifying* and *mathematising,* as well as *interpreting* and *validating,* have been combined.

Keywords Interpreting · Mathematising · Simplifying · Sub-competencies · Test instrument · Validating

8.1 Theoretical Background

8.1.1 Mathematical Modelling Competency

In 2003, the German ministers of education in the various federated states stipulated mathematical modelling as a mandatory part of each school's mathematics curriculum (see KMK 2003). The German national standards require students to possess

Corinna Hankeln: née Hertleif.

C. Hankeln (✉) · C. Adamek · G. Greefrath
Institut für Didaktik der Mathematik und der Informatik, University of Münster, Fliednerstr. 21, 48149 Münster, Germany
e-mail: c.hankeln@wwu.de

C. Adamek
e-mail: c.adamek@uni-muenster.de

G. Greefrath
e-mail: greefrath@uni-muenster.de

G. A. Stillman and J .P. Brown (eds.), *Lines of Inquiry in Mathematical Modelling Research in Education*, ICME-13 Monographs,
https://doi.org/10.1007/978-3-030-14931-4_8

abilities to translate real situations into mathematical problems and vice versa. As is widely accepted in the modelling debate, these processes can be represented in an idealized manner as a modelling cycle (e.g. Blum and Leiß 2006; Maaß 2006). Being competent in mathematical modelling is being able to autonomously and insightfully carry out all aspects of a mathematical modelling process in a certain context (Blomhøj and Kjeldsen 2006; Niss 2004). Accordingly, the German standards require students to be able to translate a situation in mathematical terms, structures and relations, to work within the respective mathematical model as well as to interpret and check results in relation to the corresponding situation (KMK 2003).

In the research literature, there is a broad debate on how mathematical modelling competency can be defined (see Kaiser and Brand 2015). First of all, two different perspectives (holistic and analytical) can be identified, which are also evident through the use of certain terms. Firstly, from a holistic perspective, the term *modelling competence* is used and interpreted in relation to experiencing an entire modelling of a situation. Some authors who adopt this perspective, propose competence models that incorporate different levels. Greer and Verschaffel (2007), for example, distinguish between three levels of mathematical modelling: implicit (in which the student is essentially modelling without being aware of it), explicit (in which attention is drawn to the modelling process), and critical modelling (whereby the roles of modelling within mathematics and science, and within society, are critically examined). Blomhøj and Jensen (2003) also take up the distinction of different competency dimensions formulated by Niss and Højgaard (2011). They distinguish the *degree of coverage*, which relates to the part of the modelling process with which the students' work and the level of their reflection, the *technical level*, which refers to the kind of mathematics students use, and the *radius of action*, which describes the domain of situations in which students are able to perform modelling activities (see Kaiser and Brand 2015).

Secondly, in addition to these holistic approaches to mathematical modelling competence, other authors adopt an analytical perspective and refer to a *modelling competency* that can be subdivided into different elements or sub-competencies. This analytic view on competencies thus focuses on identifying different elementary competencies that are part of a more general modelling competency. Therefore, researchers who follow this perspective formulate models that focus more on the competency structure and not so much on its levels. Within this perspective, "competencies should be defined by the range of situations and tasks which have to be mastered" (Klieme et al. 2008, p. 9). The distinction between different subcompetencies, according to the different phases of the modelling cycle, is an example of this view within the modelling debate. Several authors (e.g. Kaiser 2007; Maaß 2006) formulate definitions of sub-competencies that are necessary for performing a single step in the modelling cycle.

Maaß (2006), for example, distinguishes between the following five subcompetencies: The competencies needed to understand the real problem and to build a model based on reality are referred to as *Simplifying*. This sub-competency includes the competency to make assumptions, identify relevant quantities and key variables, construct relationships between these variables and to find available information.

Mathematising refers to "competencies to set up a mathematical model from the real model" (Maaß 2006, p. 116). This includes competencies to translate relevant quantities and their relationships into mathematical language by choosing appropriate mathematical notations or by representing situations graphically. *Working mathematically* describes competencies for solving mathematical questions within the mathematical model by using mathematical knowledge or heuristic strategies. Again following Maaß (2006, p. 116), *Interpreting* can further be seen as the "competencies to interpret mathematical results in a real situation". This includes being able to relate results back to the specified extra-mathematical situation. Finally, competencies for verifying the solution and for critically reflecting on the solution, the assumptions made or the model used, are subsumed under the term *Validating*.

Even though these sub-competencies form the indispensable basis for a more general modelling competency, their mere existence is not sufficient. As research has shown, several additional factors such as metacognition or social competencies might be necessary for solving a complete modelling problem and carrying through a whole modelling cycle (see e.g. Blomhøj and Jensen 2003; Maaß 2006). Research has additionally shown that modelling competency is different from a technical mathematical competence and can also be empirically distinct (Harks et al. 2014).

8.1.2 Assessment of Modelling Competencies

The assessment of competencies generally depends on the underlying concept of a competency. Since we base our research on the functional concept of competencies as used, for example, in the Program for International Student Assessment (PISA), we assume that "modelling competencies include, in contrast to modelling abilities, not only the ability but also the willingness to work out problems, with mathematical aspects taken from reality, through mathematical modelling" (Kaiser 2007, p. 110). Therefore, "assessment might be done by confronting the student with a sample of … (eventually simulated) situations" (Klieme et al. 2008, p. 9). This confrontation can either be done with a written test, or with the help of observations or interviews (Dunne and Galbraith 2003; Maaß 2007). Written forms of assessment however, have the advantage that they can easily be applied to a huge number of students at the same time, that they are often more objective than interviews or observations (Smith et al. 2005) and that they can be confidential and anonymous.

Written tests do not necessarily have to be limited to solving tasks on paper, as Vos (2007) shows. In her hands-on tests, students even experimented with tangible material such as rubber bands, and afterwards responded to open-ended tasks. However, such tests require specific testing situations in which such activities are possible. Furthermore, coding of students' responses might pose difficulties as well (Smith et al. 2005). A more common way is to employ test items that can be solved on paper.

One of the most important distinctions for test items or tasks is the difference between holistic and atomistic tasks (Blomhøj and Jensen 2003). While in holis-

tic tasks, students have to proceed through a complete modelling cycle to solve a problem, atomistic tasks pre-structure a modelling problem and focus on one or two sub-processes. Both forms of tasks can be used in written assessment, either of which has its own benefits and disadvantages.

If the aim is to assess students' ability to complete a modelling process (which is often called the *general modelling competency*), it is preferable to use holistic tasks. In atomistic tasks students only have to deal with problems that require a limited range of modelling competencies, so these tasks cannot be used to obtain information about whether a person is generally capable of completing a modelling process. Holistic items have been used by several researchers to measure students' modelling competency (e.g. Kreckler 2015, 2017; Rellensmann et al. 2017; Schukajlow et al. 2015).

The disadvantage in using holistic items lies in the interdependence of the modelling steps. If, for example, a person is weak in simplifying a problem, he or she might not reach the point of interpreting a mathematical result. Thus, this person would not be regarded as having a high modelling competency, despite being capable of conducting the modelling process once the problem has been simplified. To avoid this problem, some authors have employed atomistic tasks to assess different sub-competencies of mathematical modelling and interpreted the sum of the measured sub-competencies as a general modelling competency (Haines et al. 2001; Kaiser 2007; Maaß 2004), even though the sub-competencies are not sufficient for general modelling competence (as stated above). Therefore, some researchers have tried to combine both forms of task and evaluated their data with the aid of Item Response Theory (Brand 2014; Zöttl 2010; Zöttl et al. 2011).

However, if the aim is not to assess general modelling competency, but rather several modelling sub-competencies, it is preferable to use atomistic tasks in a test. Since the different steps of the modelling cycle are intertwined and based on one another, it is almost impossible to rate the different sub-competencies separately in holistic tasks. If, for example, a person fails to simplify a situation adequately, he or she might not even reach the point of validating a solution, since none was found. Therefore, it is then impossible to judge that person's competencies in validating a result.

Even though some researchers have focused on assessing sub-competencies of mathematical modelling (Brand 2014; Haines et al. 2001; Zöttl 2010), there is no sound empirical evidence that the theoretically assumed division into different sub-competencies adequately describes the structure of mathematical modelling competency. The two authors who assessed different sub-competencies of mathematical modelling, namely Brand (2014) and Zöttl (2010), summarized different sub-competencies. They subsumed the sub-competencies of simplifying and mathematising, as one dimension of modelling competency, and interpreting and validating as another. Additionally, they examined working mathematically and general modelling competency. Even though both authors use the same structure of combining sub-competencies they do not give any reason other than time economy.

We therefore wanted to determine whether it is possible to measure the sub-competencies of simplifying, mathematising, interpreting and validating as sepa-

rate dimensions of modelling competency. If this proves not to be the case and the demands made in the different phases of the modelling cycle are very similar to each other, is the aggregated view of Brand (2014) and Zöttl (2010) the more suitable to depict the structure of mathematical modelling competency?

Based on the theoretical work concerning the sub-competencies of mathematical modelling, we expected it to be possible to assess these sub-competencies separately and hoped to create a test instrument that could be used, for example, to evaluate experimental interventions at the level of sub-competencies.

8.2 Methods

8.2.1 Item Construction

Based on the theoretical considerations, as well as existing test items, we began to construct atomistic test items that were intended to assess each sub-competency of mathematical modelling separately. As the basis for operationalisation, we used the familiar definitions of the sub-competencies as explained above (Kaiser et al. 2015, an English translation can, for example, be found in Maaß 2006).

As we had in mind using the new test instrument in further studies, we aligned our work with the requirements of these studies. For example, we focused on geometric modelling problems and chose grade 9 students (15–16 years old) to be our target group. There were no content-related reasons for these choices concerning the research questions formulated above, and we expect the results of our study to be transferable, to a certain extent, to other mathematical domains. However, as Blum (2011) states, learning is always dependent on the specific context, and hence, a simple transfer from one situation to another cannot be assumed. He emphasises that this applies to the learning of mathematical modelling in particular, so that modelling has to be learnt specifically. Thus, if a student is a good modeller in the field of geometry, he or she is not necessarily a good modeller in the field of functions. Of course, the restriction to geometric modelling problems limits the generalizability of our results, but allows us at the same time to gain more reliable and meaningful findings regarding the chosen topic.

Next, we present an example of a test-item for each of the four sub-competencies we measured, and explain, to what extent this item actually measures the sub-competency. To provide some evidence for the quality of the items, the solution frequency and the item-total-correlation as an indicator for its selectivity are given, as found in an implementation of the test in a large sample (3300 completed tests).

During their summer vacation, Marcus and Irina are standing on top of a lighthouse and enjoying the view. "How far is it to the horizon?" Irina asks.

Mark all of the following information that you consider to be important to calculating the distance to the horizon.

https://upload.wikimedia.org/wikipedia/commons/b/bf/Louisbourg_Lighthouse.jpg

☐	Between the lighthouse and the ocean, there are 25 m of sandy beach.	☐	The two are standing on the Atlantic coast in France.
☐	There are no clouds in the sky.	☐	The radius of the earth measures 6370 km.
☐	The lighthouse is 83 m high.	☐	The lighthouse's light shines as far as 10 km.

Fig. 8.1 The *Lighthouse Task*: multiple-choice item that measures competencies in simplifying a problem (translated task)

8.2.1.1 Simplifying: *Lighthouse* Item

An example of a test item that was used to measure the sub-competency of *simplifying* is the *Lighthouse Task* (see Fig. 8.1, translation). It is a modification of the well-known lighthouse question (Kaiser et al. 2015), which requires the use of a geometrical model and is suitable for grade nine students. The given situation is depicted by a picture of a lighthouse. The students' task is to select all the information that is relevant to calculate the distance to the horizon. Thus, the item measures competencies for identifying relevant quantities and key variables, which are part of the definition of the sub-competency simplifying.

The fact that more than one answer has to be selected, namely the radius of the earth and the height of the lighthouse, reduces the probability of selecting the correct answer by guessing. The alternative answers represent misconceptions, for example the answer "There are no clouds in the sky" reflects confusing the distance to the horizon with the visibility. The first two alternatives show different misconceptions of the dependence on location of the lighthouse, and the last alternative represents a misunderstanding of the question, or rather the misconception that the distance to the horizon depends on the range of the light.

The distractors were developed with the help of experts in the field of modelling. We collected various items of information we thought students might select as relevant, even though they are not. In our pilot studies, as well as in the implementation of the test with a large sample, we checked these distractors and found that all of them were chosen by at least some students. The two most common mistakes were to select the distractor: *Between the lighthouse and the ocean, there are* 25 m *of sandy beach* (25.2% of wrong answers) and not to select the second correct option: *The radius of the earth measures* 6370 km (13.5% of wrong answers).

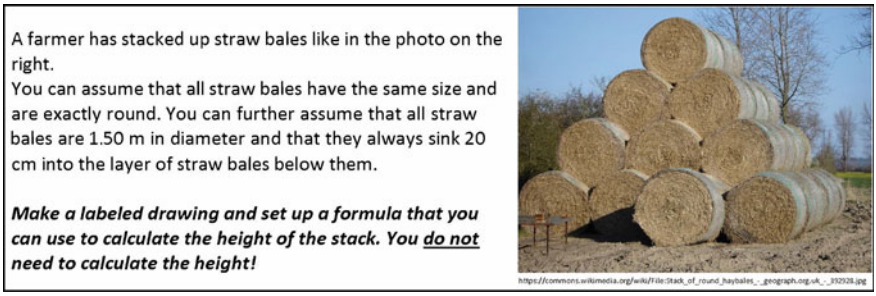

Fig. 8.2 The *Straw Bale Task*: short answer-item that measures competencies in setting up a mathematical model (translated task)

The item was used in a study with a large sample which led to 1473 responses to this item. A total of 45.35% of the students was able to answer this question correctly and received 2 points. Students who selected one additional distractor or forgot to select the second correct answer without selecting one of the distractors still received 1 point. This was the case with 27.70% of the students. Even though it is thus a relatively easy item, its Item-Total-Correlation of $r = 0.43$ yields a satisfactory selectivity of this item in this sample.

8.2.1.2 Mathematising: *Straw Bale* Item

The item in Fig. 8.2 was used to assess the competencies for setting up a mathematical model from a simplified real situation (i.e. *mathematising*). This item is inspired by the *Straw Bale Task* in Borromeo Ferri (2011, pp. 84–85), which confronts students with a real-life situation of a stack of straw bales in a field. The idealizing assumptions that all straw bales are the same size and that they are evenly and exactly round are given in the text. So are the diameter of 1.50 m and the depth that the straw bales sink into the layer below them. The student's task is to convert this situation into a mathematical representation, both graphically in a labelled drawing and symbolically as a formula with the aim of calculating the stack's height. The item thus measures the competencies required for choosing appropriate mathematical notations or by representing situations graphically.

A correct answer must include the stack's diameter, the depth of sinking in and, as the unknown quantity, the height of the stack. Answers using the specific sizes and those using abstract variables to denote these quantities were acceptable. Students could achieve a maximum of two points for this item, one for the correct drawing and one for a correct formula.

Use in a large sample produced 1143 responses to this item, which was correctly solved by 24.58% of the students, 36.05% scored one point and 39.37% gave a completely incorrect answer. With an item-total-correlation of $r = 0.50$, its selectivity is also within a satisfactory range. Since this item is in a short-answer format, approx-

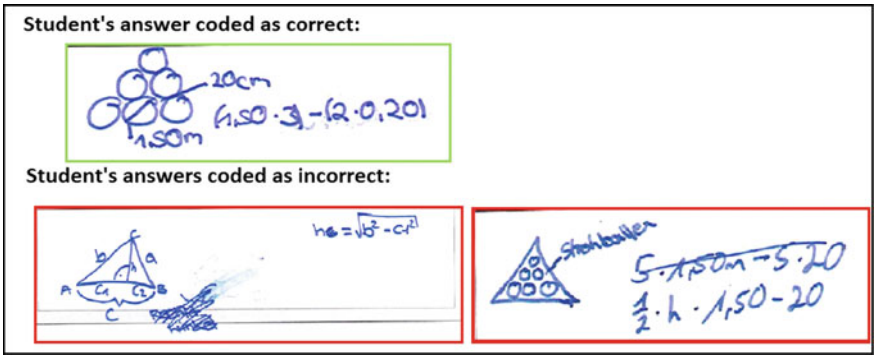

Fig. 8.3 Student responses to the *Straw Bale Task*

imately 40% of students' answers were rated by two independent raters according to a coding manual. The interrater-reliability Cohen's Kappa was $\kappa = 0.86$ and thus very good.

Figure 8.3 gives an example of how this item was coded. The first solution shows a correct solution given by a student. He or she was able to use the given relevant information to build a graphical and a symbolic mathematical model. The answers below show incorrect responses. The answer on the left shows that the student tried to apply Pythagoras' theorem and was not able to transform the given data into a mathematical model, with which it would have been possible to solve the problem. The response on the right shows a graphical representation of the situation where the straw bales are still shown (which is written next to the drawing). The formula used is an attempt to incorporate the given data, but on one hand does not pay attention to the units, and on the other hand, employs the formula for the area of a triangle. This mathematical model thus cannot be used for solving the task and was therefore coded zero.

8.2.1.3 Interpreting: *Dresden* Item

The sub-competency of *interpreting* a mathematical result and relating it back to the extra-mathematical context was measured with items such as Fig. 8.4. In this item, students are confronted with an extra-mathematical situation, which has been simplified and converted into a mathematical model. In the Dresden item in Fig. 8.4, a boy takes a look at a photograph, where he identifies his father standing in front of a giant arch at a Christmas fair. He mathematises the situation by measuring the height of his father and of the arch in the photo, and by setting up a mathematical term that combines all given numbers and yields the numerical result 3.8. In other words, the modelling cycle has already been carried out up to the point where the mathematical result has to be related back to the context. The student's task is to explain what the result 3.8 means in relation to the specified situation. Since the mathematical term

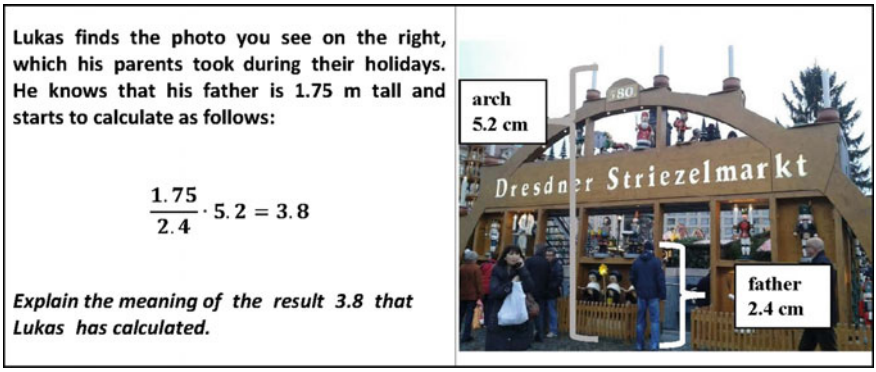

Fig. 8.4 The *Dresden Task*: short answer item that measures competencies relating a mathematical result back to reality (translated task)

represents the father's height in reality, divided by his size on the photo, multiplied by the size of the arch, the correct answer, which was rewarded one point, is that the arch is in reality 3.8 m high.

In our study, 56.05% of the students gave a correct answer. The most common incorrect response was that 3.8 represents the difference between the size of the father and the arch. This is probably due to the fact that the numbers given in the picture have a difference of 2.8. Students who do not pay attention to the 'borrowing' in the subtraction thus confuse the given result with the difference. These students clearly display a deficit in their competencies for interpreting a mathematical result, and subsequently did not receive a point for their answer. The selectivity for this item was satisfactory with a value of $r = 0.48$. The interrater-reliability (Cohen's Kappa) was very good with a value of $\kappa = 0.95$.

8.2.1.4 Validating: *Rock* Item

The sub-competency *validating* was perhaps the most difficult to assess. As the definition of this sub-competency shows, it consists of different facets, namely to critically check solutions, reflect on the choice of assumptions or of the mathematical model and also to search for alternative ways to solve the problem. To measure this sub-competency, we therefore employed a broader variety of items, which means that the items measuring the sub-competency validating were not as similar to each other as the items in the other sub-competencies. Figure 8.5 gives an example of an item that assessed the competencies for critically reflecting a result. In this item, students are confronted with a photo of a girl standing beside a rock. Without presenting a mathematical model, students are given the result of a calculation, namely the assertion that the rock is 8 m tall. They are asked to explain whether or not this result is plausible. To solve this task, students must use the photo and compare the size of the girl with that of the rock. As the rock is approximately three times as high as

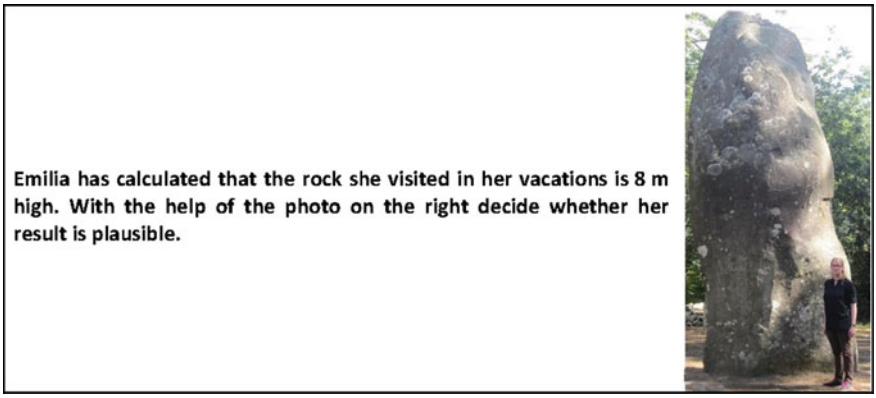

Fig. 8.5 The *Rock Task*: short answer item that measures competencies for critically checking a solution (translated task)

the girl, she would have to be more than two metres tall if the result was correct. A student's response, which clearly stated that the assertion is wrong and justified this answer by comparing the size of the girl and the rock, and additionally identified a maximum size for the girl was coded with two points. Answers like "No, since the rock is just approximately three times as big as the girl" which did not give a maximum size for the girl were still awarded one point. Answers that were coded as wrong mostly either were not justified at all or the result was found to be plausible.

Approximately half of the students (51.05%) acquired one point in this item, 27.56% were given two points. The selectivity was $r = 0.40$ and the interrater-reliability was again very good with $\kappa = 0.88$. Other items that assessed this sub-competency did not focus so strongly on checking a result, but confronted students with the choice of a mathematical model and asked them to decide whether the mathematical model would fit the given extra-mathematical situation. Additionally, there were items that assessed student abilities to find objects that help in determining the plausibility of a result. For example, students were given a photo of a dog and the claim that this dog is 28 cm high. They were asked to name one object that is approximately 28 cm high with which they could mentally compare the dog. In contrast to the Rock item in Fig. 8.5, students in this item were not asked to actually check the given result. This item assessed whether students were able to fall back on supporting knowledge (in German "Stützpunktwissen") as a basis for checking their results. We therefore had a broad variety of difficulty levels of items that assessed the different facets of the sub-competency of validating.

8.2.2 Testing of Items

Before constructing test booklets, we had a phase of intensive item testing. We first presented the items to experts in the field of modelling and asked them to comment on the tasks and to indicate what they thought the items would assess. All experts classified the items as we expected, but there were some that tended to assess more than just the one sub-competency. We reworked those items and related them more closely to the definitions of the respective sub-competency. Special attention was paid to the multiple-choice items and the choice of distractors. We asked the experts to comment on all answers that were part of the items and to add an answer to the item if they thought an answer or a typical mistake would be missing.

Subsequently, we gave the items to 36 students in a class, observed their working processes and asked them afterwards in groups what problems they had solving the tasks. Most of their answers referred to the poor quality of a photo which was then changed. In this phase, we identified formulations that were too complicated and made items too difficult to understand. With the help of students' comments, we simplified the language and made clear references for students who would subsequently be expected to use a picture as in the Rock item in Fig. 8.5. Students found some of the items "easy and interesting to solve, since they are different from conventional maths exercises", but "had to think intensively" about some of the items. These comments, as well as the analysis of their answers to the exercises, revealed a wide range of item difficulties, with a large number of items having a medium solution frequency, but also with a substantial number of items with a high as well as with a low solution frequency. No item remained unsolved, but no item was solved by all participants either. The qualitative analysis of students' answers made it possible to identify possible difficulties in coding the answers, which led to small changes in formulation. It was also the basis of a first draft of a coding manual for the test instrument.

Afterwards, we conducted a second pilot study with the aim of acquiring quantitative data to check the test's quality and to generate solution frequencies of the various items. In this study, no item was solved by all, or by none, of the 189 students. The answers the students gave additionally helped us to improve the coding manual.

8.2.3 Combining Items into a Test

One of the most difficult challenges in constructing a test that can be used in an experimental design is to ensure the comparability of pre- and post-tests. This challenge of creating parallel tests becomes redundant if one uses psychometric models and interprets responses to items as manifest indicators of one or several latent variables. The central idea is that the more distinct a person's latent variable is, the greater his or her probability of solving an item. Thus, in the simplest model, only the difficulty of the items and the person's ability are taken into account. The great advantage of

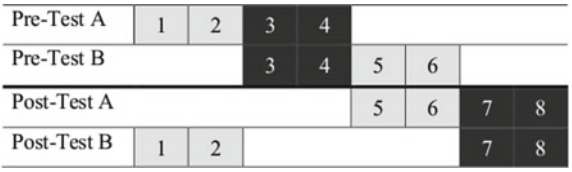

Fig. 8.6 Multi-matrix design of pre- and post-test: light grey boxes show the linkage between the pre- and post-test, dark grey boxes show the linkage between booklets at one point of measurement

this model is that the person's ability can even be determined if not all items are presented, which makes it possible to use a multi-matrix-design.

Figure 8.6 illustrates the test structure. Firstly, we constructed eight item blocks consisting of one item per sub-competency, a total of four items per block. No items were in more than one block. Secondly, we combined the item blocks into four test booklets, two for each point of measurement, so that each test booklet consisted of 16 items. We thereby paid attention to a similar average difficulty of the test booklets so as to avoid motivational problems for some groups of students. The fourteen multiple choice items were also equally distributed over the different booklets, so that all test booklets contained both item formats.

The two booklets we used at the first point of measurement were linked to each other via two blocks (blocks 3 and 4 in Fig. 8.6). Additionally, booklet A contained items that were not part of booklet B and vice versa. The same linking method was used for the post-test, where new items (blocks 7 and 8 in Fig. 8.6) were used to link the booklets. A person who answered test-booklet A in the pre-test also received post-test A, and the same for booklet B. By so doing, no student answered the same items twice. Nevertheless, since the item blocks 1, 2, 5 and 6 were used at both points of measurement, it was possible to link the two points of measurement. We determined the item difficulties using the data of all points of measurement, and then calculated the person's abilities for each point of measurement separately.

8.2.4 Methods of Data Collection

We implemented the test in 44 classes of grade 9 students who completed the test instrument three times each. This led to a total of 3300 completed tests which was the basis for the evaluation of the test instrument presented in this chapter.

Each testing lesson had a duration of 45 min, and since each student had to answer a set of just 16 items, no time pressure was observed. The testing was performed by the teachers strictly following a written test-manual, in which all details for conducting the testing process, as well as instructions to be read out, were recorded. In this way, it was possible to have a standardized execution in each of the participating classes. The correct implementation was controlled at random.

The completed test sheets were coded according to the coding manual. Some items were coded dichotomously and some had a Partial Credit scoring, receiving two points for a completely correct solution and one point for a partially correct solution. A sample (40%) of the completed test sheets were rated by two independent coders. The interrater-reliability for the open tasks was within a range of $0.81 \leq \kappa \leq 0.96$ (Cohen's Kappa) which reflects very good agreement.

The data were scaled using a one-parameter Rasch model with the help of the software ConQuest (Wu et al. 2007). For the estimation of item- and person-parameters, weighted likelihood estimations were used. To determine item parameters and to evaluate the test instrument, all three points of measurement in the main study were treated as if they were independent observations of different people, even though the same person could appear in up to three different rows in the data matrix. This approach is called 'using virtual persons' (Rost 2004) and is used in PISA (OECD 2012) and TIMSS (Martin et al. 2016), since it is unproblematic for the estimation of item parameters. These item parameters are the basis for evaluating the test instrument reported in this chapter.

8.2.5 Statistical Analyses to Answer the Research Questions

To be able to use the outcome of a probabilistic model for empirical data, it is necessary to check whether the a priori chosen model fits the data. Since the model that fits the data best is regarded as the best reproduction of the structure of the latent variable, this check of model fit can be used to gain more information about the competence structure itself. We therefore calculated various different models and compared the respective model fits. As we were interested to know whether it is possible to measure the different sub-competencies separately, we compared three models shown in Fig. 8.7. The first Model is a four-dimensional one in which each sub-competency is measured as a separate dimension. The second Model reflects the aggregation of sub-competencies as Brand (2014) and Zöttl (2010) chose for their research. The third Model is one-dimensional. If this was found to be the best fitting model for the empirical data this would mean the abilities students need to solve the different types of items, as presented in Sect. 8.2.1, were so similar that it would not be appropriate to model them as different dimensions.

When scaling empirical data with the help of item response theory (IRT), there are different ways to check how well a model fits the data. In the case of estimating item- and person-parameters, the algorithm used, iterates until the likelihood of observed responses reaches its maximum under the constraints of the given model. Therefore, the fit of two models can be compared by analysing their likelihood (L). After estimating the parameters, the programme ConQuest displays the final deviance (D) of the estimation, which derives from the likelihood by $D = -2ln(L)$. The smaller the final deviance, the greater the likelihood and the better the model fits the data. This measure does not take into account the sample size and the number of

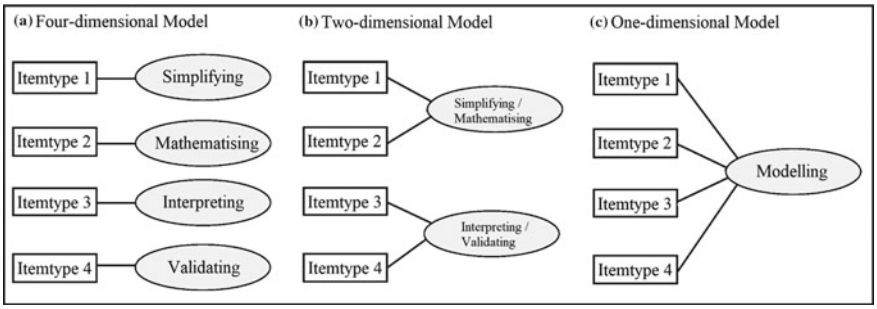

Fig. 8.7 Models used and compared to gain information about the competence structure

Table 8.1 Information criteria to compare models

	Model		
	Four-dimensional	Two-dimensional	One-dimensional
Sample size	3300	3300	3300
No. of estimated parameters	61	54	52
Final deviance	85,471.21	85,821.33	85,838.86
AIC	85,593.21	85,929.33	85,942.86
BIC	85,965.41	86,258.82	86,260.15

Note Lower values indicate a better fit for the model

estimated parameters. Therefore, the AIC and BIC are also reported.[1] AIC tends to prefer models that are too large whereas BIC prefers smaller models. If both criteria prefer the same model, this is likely to be the best of the candidate models (Kuha 2004, p. 223).

8.3 Results

Table 8.1 shows the final deviance, AIC and BIC for the three different models in Fig. 8.7. The four-dimensional model, for which each of the different types of items, which were designed to measure the sub-competencies separately, loads on one dimension each, fits the data best. All three measures are lower than in the two-dimensional model, in which *Simplifying* and *Mathematising,* as well as *Interpreting* and *Validating*, have been combined, and lower than in the one-dimensional model, where all types of items load on just one factor.

[1] AIC (Akaike Information Criterion) defined by $AIC = -2\ln(L) + 2n_p$ and BIC (Bayes Information Criterion) by $BIC = -2\ln(L) + \ln(N) \cdot n_p$ where n_p is the number of estimated parameters and N the sample size. A smaller value indicates a better fit.

The results indicate that scaling the test items used (which aimed to measure different sub-competencies) with a one-dimensional model, is the poorest of the tested options. The two-dimensional model fits the data slightly better than the one-dimensional model, as all measures have a lower value. That the four-dimensional model fits the data best, with a considerable margin concerning the information criteria, provides quantitative empirical evidence supporting the theoretically assumed, and qualitatively observed, sub-competencies of mathematical modelling.

Another possibility for checking the fit of the model is to look not only at the overall model fit, but to check the fit of the different items separately. There are also different measures that indicate the fit of items from which the weighted mean square fit (WMNSQ) is the most frequently reported value. This fit index reflects how much the empirically determined responses to an item differ from the solution probabilities that the model predicted (Wilson 2008). Since those whose abilities lay near the item parameters provide more information than persons with a more extreme ability value, it is sensible to weight their residua more strongly (Bond and Fox 2007). The WMNSQ can be z-standardised and tested statistically for significance. Following Bond and Fox (2007, p. 243) and PISA (OECD 2012), WMNSQ should be within a range of 0.8–1.2 for high stakes tests, and 0.7–1.3 for "run of the mill" tests. For the four-dimensional model, the WMNSQ was within a range of 0.93–1.11 and thus quite near to the generating value of 1, which reflects a perfect model fit.

As mentioned above, there are two parameters calculated in the models we used: those that reflect the *item difficulty* and the parameters that indicate the *degree of competency*. The item parameters within this approach were determined with an (item-separation)-reliability of 0.996, which is an excellent result, even though such high values are not unusual for large samples (Wu et al. 2007). The EAP/PV-reliabilities for the ability parameters lie within a range of 0.66 and 0.80 for the different sub-competencies at different points of measurement. Since the EAP/PV-reliabilities can be compared to Cronbach's Alpha, these values are within a satisfactory to good range, and certainly sufficient to compare groups, as planned for further studies.

8.4 Summary and Discussion

This chapter focused on the assessment of modelling competencies, taking up the notion of different sub-processes in a modelling process that requires different competencies. Previous research has already shown that modelling requires different competencies than purely technical mathematical competencies (Harks et al. 2014). Concerning the sub-competencies of simplifying, mathematising, interpreting and validating, it was still unknown whether it is possible to assess them separately. Those test instruments that assessed sub-competencies of mathematical modelling subsumed different sub-competencies, instead of treating them as independent dimensions of a more general modelling competency. We therefore constructed a new test instrument with specific test items for each of the four chosen sub-competencies. This chapter presented an exemplary item for each type. We explained to what extent these

items are able to measure the sub-competencies of mathematical modelling. It was clear from the beginning that the new test instrument was not designed to measure a more global modelling competency, but to enable making statements concerning one or several sub-competencies. This is especially fruitful for empirical research, for example if the effects of certain interventions have to be evaluated. A test that objectively, reliably, and validly measures the sub-competencies can be used to compare student achievements before and after having experienced a certain treatment. If the treatment being examined is expected to have effects that differ from one sub-competency to another, it is preferable to evaluate the intervention at the level of sub-competencies, instead of building average scores.

We have shown that the test we presented can be used to do so. The psychometric model with four separate dimensions fits the data best. Thus, we took up the idea of Haines et al. (2001) to assess sub-competencies separately, continued the work done by Brand (2014) and Zöttl (2010), and succeeded for the first time in measuring the sub-competencies empirically as independent latent variables.

These results underline the different demands that a modelling process imposes on students and further confirms the empirically assumed division of a modelling process into different steps. However, our study of course has its limitations. Firstly, we limited our research to the field of geometric modelling. This was due to other studies for which the test was constructed and not grounded in reasons with regard to content. Further studies should expand the content areas, as well as test further age groups. Secondly, even though there are already some tests that assess sub-competencies of modelling, we have not yet had the opportunity to check the correlation of those tests with our new test, so as to control the validity statistically. This is also valid for discriminant validity.

Thirdly, we did not aim to assess general modelling competency and hence cannot make assumptions on the interplay between a general competency and the sub-competencies. The question arises as to how much we know about a person's general modelling competency, if we know his or her strengths and weaknesses in the sub-competencies. By answering this question, it could be possible to substantiate the assumptions, for example concerning meta-knowledge, that have been derived from qualitative studies with quantitative data.

References

Blomhøj, M., & Jensen, T. (2003). Developing mathematical modelling competence: Conceptual clarification and educational planning. *Teaching Mathematics and its Applications, 22*(3), 123–139.

Blomhøj, M., & Kjeldsen, T. H. (2006). Teaching mathematical modelling through project work. *ZDM Mathematics Education, 38*(2), 385–395.

Blum, W. (2011). Can modelling be taught and learnt? Some answers from empirical research. In G. Kaiser, W. Blum, R. Borromeo Ferri, & G. Stillman (Eds.), *Trends in teaching and learning of mathematical modelling* (pp. 15–30). Dordrecht: Springer.

Blum, W., & Leiß, D. (2006). "Filling up"—The problem of independence-preserving teacher interventions in lessons with demanding modelling tasks. In M. Bosch (Ed.), *Proceedings of the Fourth Congress of the European Society for Research in Mathematics Education* (pp. 1623–1633). Barcelona: Universitat Ramon Llull.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model*: *Fundamental measurement in the human sciences* (2nd ed.). Mahwah, NJ: Routledge.

Borromeo Ferri, R. (2011). *Wege zur Innenwelt des mathematischen modellierens. Kognitive analysen zu modellierungsprozessen im mathematikunterricht*. Wiesbaden: Vieweg+Teubner.

Brand, S. (2014). Erwerb von Modellierungskompetenzen. Empirischer Vergleich eines holistischen und eines atomistischen Ansatzes zur Förderung von Modellierungskompetenzen. In G. Kaiser, R. Borromeo Ferri, & W. Blum (Eds.), *Perspektiven der Mathematikdidaktik*. Wiesbaden: Springer Spektrum.

Dunne, T., & Galbraith, P. (2003). Mathematical modelling as pedagogy—Impact of an immersion program. In Q. Ye, W. Blum, K. S. Houston, & Q. Jiang (Eds.), *Mathematical modelling in education and culture* (pp. 16–30). Chichester: Horwood.

Greer, B., & Verschaffel, L. (2007). Modelling competencies—Overview. In W. Blum, P. L. Galbraith, H.-W. Henn, & M. Niss (Eds.), *Modelling and applications in mathematics education: The 14th ICMI study* (pp. 219–224). New York: Springer.

Haines, C., Crouch, R., & Davis, J. (2001). Understanding students' modeling skills. In J. F. Matos, S. K. Houston, & W. Blum (Eds.), *Modelling and mathematics education: ICTMA 9—Applications in science and technology* (pp. 366–380). Chichester: Horwood.

Harks, B., Klieme, E., Hartig, J., & Leiss, D. (2014). Separating cognitive and content domains in mathematical competence. *Educational Assessment, 19,* 243–266.

Kaiser, G. (2007). Modelling and modelling competencies in school. In C. Haines, P. Galbraith, W. Blum, & S. Khan (Eds.), *Mathematical modelling: Education, engineering and economics* (pp. 110–119). Chichester: Horwood.

Kaiser, G., Blum, W., Borromeo Ferri, R., & Greefrath, G. (2015). Anwendungen und modellieren. In R. Bruder, L. Hefendehl-Hebeker, & H.-G. Weigand (Eds.), *Handbuch der mathematikdidaktik* (pp. 357–383). Berlin: Springer.

Kaiser, G., & Brand, S. (2015). Modelling competencies: Past development and further perspectives. In G. A. Stillman, W. Blum, & M. S. Biembengut (Eds.), *Mathematical modelling in education research and practice* (pp. 129–149). Cham: Springer.

Klieme, E., Hartig, J., & Rauch, D. (2008). The concept of competence in educational contexts. In J. Hartig, E. Kliene, & D. Leutne (Eds.), *Assessment of competencies in educational contexts* (pp. 3–22). Göttingen: Hogrefe & Huber.

KMK. (2003). *Bildungsstandards im Fach Mathematik für die Allgemeine Hochschulreife*. Online available http://www.kmk.org/fileadmin/veroeffentlichungen_beschluesse/2003/2003_12_04-Bildungsstandards-Mathe-Mittleren-SA.pdf (Last access October 07, 2015).

Kreckler, J. (2015). *Standortplanung und Geometrie. Mathematische Modellierung im Regelunterricht*. Wiesbaden: Springer.

Kreckler, J. (2017). Implementing modelling into the classroom: Results of an empirical research study. In G. A. Stillman, W. Blum, & G. Kaiser (Eds.), *Mathematical modelling and applications: Crossing and researching boundaries in mathematics education* (pp. 277–287). Cham: Springer.

Kuha, J. (2004). AIC and BIC. Comparisons of assumptions and performance. *Sociological Methods and Research*, *33*(2), 188–229.

Maaß, K. (2004). *Mathematisches modellieren im Unterricht*. Ergebnisse einer empirischen Studie: Franzbecker Verlag.

Maaß, K. (2006). What are modelling competencies? *ZDM Mathematics Education, 38*(2), 113–142.

Maaß, K. (2007). Modelling in class: What do we want the students to learn? In C. Haines, P. Galbraith, W. Blum, & S. Khan (Eds.), *Mathematical modelling: Education, engineering and economics* (pp. 63–78). Chichester: Horwood.

Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.). (2016). *Methods and procedures in TIMSS 2015*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website http://timssandpirls.bc.edu/publications/timss/2015-methods.html.

Niss, M. (2004). Mathematical competencies and the learning of mathematics: The Danish KOM project. In A. Gagtsis & S. Papastavridis (Eds.), *3rd Mediterranean conference on mathematical education* (pp. 115–124). Athens: The Hellenic Mathematical Society.

Niss, M. & Højgaard, T. (Eds.). (2011). *Competencies and mathematical learning. Ideas and inspiration for the development of mathematics teaching and learning in Denmark* (English ed.). Roskilde: Roskilde University, IMFUFA.

OECD. (2012). *PISA 2009 technical report*. Paris: OECD.

Rellensmann, J., Schukajlow, S., & Leopold, C. (2017). Make a drawing. Effects of strategic knowledge, drawing accuracy, and type of drawing on students' mathematical modelling performance. *Educational Studies in Mathematics*, *95*(1), 53–78.

Rost, J. (2004). *Testtheorie - Testkonstruktion*. Bern: Hans Huber Verlag.

Schukajlow, S., Krug, A., & Rakoczy, K. (2015). Effects of prompting multiple solutions for modelling problems on students' performance. *Educational Studies in Mathematics, 89*(3), 393–417.

Smith, B., Brown, S., & Race, P. (2005). *500 tips on assessment*. Abingdon: Routledge Falmer.

Vos, P. (2007). Assessments of applied mathematics and modelling: Using a laboratory like environment. In W. Blum, P. L. Galbraith, H. Henn, & M. Niss (Eds.), *Modelling and applications in mathematics education. The 14th ICMI study* (pp. 441–448). New York: Springer.

Wilson, M. (2008). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Taylor & Francis.

Wu, M. L., Adams, R., Wilson, M. & Haldan, S. (2007). *ACER ConQuest version 2.0: Generalised item response modelling software*. Melbourne: ACER.

Zöttl, L. (2010). *Modellierungskompetenzen fördern mit heuristischen Lösungsansätzensungsbeispielen*. Ph.D. thesis, Ludwig-Maximilians-Universität München.

Zöttl, L., Ufer, S., & Reiss, K. (2011). Assessing modelling competencies using a multidimensional IRT approach. In G. Kaiser, W. Blum, R. Borromeo Ferri, & G. Stillman (Eds.), *Trends in the teaching and learning of mathematical modelling* (pp. 427–437). Dordrecht: Springer.