16

Taxonomies and Competency Models

Peter Hubwieser and Sue Sentance

Chapter outline

Chapter synopsis

In this chapter, we compare two different methods used to describe the learning outcomes of students, learning objectives, including taxonomies and competencies. Traditionally, learning outcomes were described and monitored by the acquisition of certain knowledge elements or by the achievement of predefined learning objectives. During the last decade, mainly stimulated by the surprising results of the Programme for International Student Assessment (PISA) studies, the focus of the outcomes of school education has shifted – in some countries – towards target competencies. Competencies describe which 'real-world' problems or tasks students should be able to solve. Item response theory can be used in preference to classical testing approaches to assess competencies. Finally, the development of educational standards for regional or national assessment is described. To assure quality, such standards have to be based on properly defined competency models.

16.1 Introduction

In Chapter 11 we read about some techniques that an individual teacher can use to formatively assess progress in computer science. The question remains, how do we measure the learning summatively in a rigorous way? This relates not just to individual students and their progress, but to whole cohorts or even countries of students. Where we may use informal measures such as the average score on a test to measure the progress of our students, more rigorous methods can give us more accurate information. They also provide an alignment between educational goals and the instruction provided. This chapter addresses this issue by considering the current trend from learning outcomes to competencies, along with some statistical methods that are used to ensure accuracy and objectivity of this measurement.

Learning outcomes are assessed for many different purposes. For example, an individual teacher may want to assess his/her students to get feedback about the effectiveness of his/her lessons or to confer a certain qualification. Such an assessment could be performed either by a written or online test or by oral examinations. In these cases, one single teacher will examine a certain number of students that could range from one in the oral case up to hundreds in university examinations. A national school administration, several collaborating governments or even a community of many countries like the OECD[1] may want to compare the learning outcomes of their educational systems. In these cases, sometimes more than 100 people will assess a tremendous number of students, which sometimes (e.g. in the OECD PISA[2] surveys) reaches nearly 1 million. These assessment cases will differ in many respects. On the one hand, a single teacher assessing his/her students might be well informed about the learning content and teaching methods of the assessed lectures, which is unknown in large assessment projects like PISA. On the other hand, despite the very different scales of these cases, there are many common requirements. In particular, all tests and examinations should meet the basic three requirements *objectivity, reliability* and *validity*, as described by Adams nearly a century ago:

> When a test measures a function, simple or complex, as completely as possible, it is a valid measure of that function regardless of whether it measures with high or low accuracy … Reliability … is associated fundamentally with absence of systematic errors. … When test and retest measure the same function twice, then the test is reliable … Objectivity exists only when all errors of measurement are random. With the advent of correlated errors, subjectivity appears.
>
> Adams, 1936: 348–49

In this chapter, we present some general aspects and methods for assessment in computer science education (CSE) which affect particularly the *construct validity* of the assessment outcome. Construct validation identifies the constructs that account for the way students' performance in test varies (Cronbach and Meehl, 1955). In education, the constructs to be assessed in most cases are learning outcomes: certain changes in the knowledge and behaviour of students. Until the 1960s, learning outcomes were regarded predominantly as an increase of knowledge. Triggered by the modernization of education systems in the 1970s (see Robinsohn, 1967) and acknowledging

[1] Organisation for Economic Co-operation and Development.
[2] Programme for International Student Assessment.

that no learning progress can be observed without changes in behaviour, learning achievements were defined and measured mostly in terms of learning objectives.

During the last decade, mainly stimulated by the PISA of the OECD, the intended outcomes of learning are increasingly defined by target competencies as defined by Weinert (1999). These two paradigms rely on very different educational and psychological approaches and in consequence, show many differences. The most important might be that competencies must be very strictly based on empirical research (Klieme et al., 2004), while learning objectives tend to be set by educators according to personal beliefs or assumptions. In terms of the cognitive structure and the aspiration level of learning outcomes, learning objectives are usually organized by general, subject-independent taxonomies, while domain specific competency models are applied in the second case (Leutner, Hartig and Jude, 2008). The staistical approaches used are very different. In the case of learning objectives, the Classical Test Theory is usually applied, while competencies are measured using Item Response Theory (Hartig, Klieme and Leutner, 2008).

In CSE, many publications have focused on the suitability of learning object taxonomies (e.g. those of Anderson-Krathwohl (Anderson and Krathwohl, 2001) or Biggs (Biggs and Collis, 1982)). There have been some attempts to design specific taxonomies for CSE (Fuller et al., 2007, Meerbaum-Salant, Armoni and Ben-Ari, 2010). However, the paradigm shift from learning objectives to competencies has only just started. Only a few research projects have investigated the cognitive structure of competencies (e.g. the large German MoKoM project (Neugebauer, Magenheim, Ohrndorf, Schaper and Schubert, 2015)). Some recent progress has been made in the definition of competency models for programming (Kramer, Hubwieser and Brinda, 2016).

16.2 Learning objectives and taxonomies

The application of learning objectives in education has a long and complicated history, starting with Mager (1961). We can define a learning objective as 'a statement that tells what learners should be able to do when they have completed a segment of instruction' (Smith and Ragan, 2005: 969).

Key concept: Learning objectives

Learning objectives describe the goals that educators aim to achieve in terms of the learning progress of their students. Learning objectives may be located on very different abstraction levels. On the highest level, global objectives give the overall goals of education (e.g. the ability to act in a responsible way in the digital society). On an intermediate level, educational objectives describe the goals of some weeks or months of teaching, like 'being able to implement class diagrams in an imperative programming language'. On the most concrete level, instructional objectives detail the intended learning progress during some few lessons, e.g. 'to be able to combine different control structures to simple'. Usually, instructional objectives are formulated as combinations of a certain knowledge element and a description of observable behaviour (e.g. 'being able to implement variables'). To provide structure, learning objectives are classified by Learning Taxonomies (e.g. the Blooms Revised Taxonomy).

Bloom's taxonomy and its revision (by Anderson and Krathwohl)

The structure and hierarchy of learning objectives is described by category systems that are usually called 'taxonomies' in this context. The most famous taxonomy of learning objectives was presented by Bloom in 1956 (Bloom, 1956). Firstly, he separated three domains of objectives:

- Cognitive: mental skills (knowledge)
- Affective: growth in feelings or emotional areas (attitude or self)
- Psychomotor: manual or physical skills (skills).

Secondly, he presented a hierarchy of six levels for the cognitive domains:

1. *Knowledge*: Student recalls or recognizes information, ideas and principles in the approximate form in which they were learned.
2. *Comprehension*: Student translates, comprehends or interprets information based on prior learning.
3. *Application*: Student selects, transfers and uses data and principles to complete a problem or task with a minimum of direction.
4. *Analysis*: Student distinguishes, classifies and relates the assumptions, hypotheses, evidence or structure of a statement or question.
5. *Synthesis*: Student originates, integrates and combines ideas into a product, plan or proposal that is new to him or her.
6. *Evaluation*: Student appraises, assesses or critiques on a basis of specific standards and criteria.

Bloom's taxonomy has been widely accepted and applied in schools. Based on their experience with this taxonomy, Anderson and Krathwohl adopted it to a more outcome-focused modern education approach (Anderson and Krathwohl, 2001). They split the originally one-dimensional hierarchy into two dimensions, regarding a learning objective as a paired combination of (1) a certain type of *knowledge* and (2) an observable *behaviour* (called cognitive process). For the first dimension, the knowledge was portioned into four categories. The levels on the behaviour dimension were derived from Bloom's original taxonomy by switching from nouns to active verbs, reversing and renaming several levels. The result was the well-known 'Blooms Revised' taxonomy:

Table 16.1 The revised Bloom's taxonomy (Anderson and Krathwohl, 2001)

	Cognitive process					
Knowledge	**Remember**	**Understand**	**Apply**	**Analyse**	**Evaluate**	**Create**
Factual						
Conceptual						
Procedural						
Metacognitive						

Table 16.2 International comparison of key CS learning objectives (Hubwieser et al., 2015)

Global objective	Addressed by
Digital literacy (including use and handling of tools)	FI, USA, BY, KO, RUS, UK, SW, IN, IT, NRW, NZ
Computational thinking (including algorithmic and logical thinking)	FR, FI, USA, IS, RUS, UK, KO, SW, IN
Problem solving	NRW, USA, IS, KO, RUS, UK, SW, IN
Understanding of basic concepts of CS and IT	NZ, BY, IS, KO, SW, IN, FR, IT
Career preparation and choice	NRW, SW, BY, IN, FR, IT, KO
Support awareness of social, ethical, legal and privacy issues and impact of CS	NRW, KO, FR, RUS, UK, SW, NZ
General education to participate in society responsibly	NRW, BY, KO, SW, IN, RUS,
Prepare for university	NRW, KO, SW, IN

To describe the specificity of learning objectives, Anderson and Krathwohl proposed three levels (Anderson and Krathwohl, 2001):

- *global* objectives: 'complex, multifaceted learning outcomes that require substantial time and instruction to accomplish';
- *educational* objectives: 'derived from global objectives by breaking them down into more focused, delimited form';
- *instructional* objectives, 'focus teaching and testing on narrow, day-today slices of learning in fairly specific content areas'.

We will refer to this taxonomy from here on as AK. As examples of global objectives, we could take the most frequently addressed goals of CSE in K-12 according to the findings of a recent working group (Hubwieser et al., 2015). The group identified eight global objectives that were found in more than three of the analysed country reports on CSE in K-12 (see Table 16.2).

Obviously, the terms that describe educational objectives are often quite close to descriptions of potential competencies.

Example: Learning objectives and outcomes

Learning objectives may be phrased in terms of 'explain', 'program', 'design' and other 'doing' words. For example, over a period of time, you may want students to learn one or more of the following, which may be made more specific depending on the age group you are working with:

- to be able to program a type of search or sort
- to evaluate alternative models in order to choose one of them
- to write an algorithmic solution for a problem
- to explain and execute algorithms
- to exemplify how 2-D data structures can be implemented.

Taking the last example, working with 2-D data structures, a task can be designed to measure this understanding (e.g. to write a program to allow users to enter data into a

one-week timetable for a school). Completing the task successfully would demonstrate that the learning objective had been achieved: breaking the task down into sub-tasks (designing the data structure, writing an algorithm to populate it, initializing the data structure, and implementing the program) would give concrete learning outcomes from the task that could indicate the level of student performance. We will see later in the chapter that specifying learning objectives in this way can be problematic.

Pre-requisites

For students to write a simple object-orientated program that simulates a traffic light they would need to meet the following learning objectives:

- Understand a class definition
- Apply a for-loop
- Be able to implement a certain mathematical formula.

However, in many cases it is impossible to achieve a set of instructional objectives in any arbitrary order, because some of them have to be learned before certain others can be reached. For example, one has to *understand* the concept of *object* (O1) before one is able to *understand* the concept of *class* (O2). This connection can be described by a *prerequisite relation* on the set of learning objectives, in this case between O1 and O2: 'O1 *is prerequisite of* O2', meaning that 'O1 has to be achieved before O2' (see Hubwieser, 2007; Hubwieser, 2008). Closer considerations show that there are (at least) two different types of prerequisite relations:

1. 'Hard' pre-requisites forced by a substantial or logical dependency: in other words concept2 contained in objective O2 *is based on* concept1 contained in objective O1. This means that it is not possible to understand concept2 without having understood concept1.
2. 'Soft' pre-requisites suggested by didactical deliberations: it is necessary to reach objective O1 in order to apply teaching or working methods that support didactical principles. Therefore it is not *necessary* to reach O1 before objective O2, but it is *advisable* in order to ease or to improve the learning process towards O2.

Nevertheless, in many cases it is not easy to describe lessons by instructional objectives, primarily for two reasons:

- there are huge numbers of objectives; and/or
- there are many relations between these objectives.

The SOLO taxonomy

Following a totally different approach, Biggs proposed his SOLO taxonomy (Biggs and Collis, 1982). Based on his theory of meaningful learning, he put more emphasis on the learner and the actual learning outcome, instead of the learning material. In Table 16.3, *capacity* 'refers to the

Table 16.3 The SOLO taxonomy (Biggs and Collins, 1982: 24–25)

SOLO Level	Capacity	Relating operation
Prestructural	Minimal: Cue and response confused	Denial, tautology, transduction. Bound to specifics
Unistructural	Low: Cue and one relevant datum	Can 'generalize' only in terms of one aspect
Multistructural	Medium: Cue and isolated relevant data	Can 'generalize' only in terms of a few limited and independent aspects
Relational	High: Cue and relevant data and interrelations	Induction: Can generalize within given or experienced context using related aspects
Extended Abstract	Maximal: Cue and relevant data and interrelations and hypotheses	Deduction and induction. Can generalize to situations not experienced

amount of working memory, or attention span, that the different levels of SOLO require' (Biggs and Collis, 1982: 26). The relating operation refers to 'the way in which the cue and response interrelate'. Additionally, there is an attribute of 'Consistency and closure', referring to the felt need of the learner to come to a conclusion that is consistent with the data and other possible conclusions, which increases with the levels of the taxonomy (pp. 27–28).

Taxonomies for computer science

So far, we have looked at taxonomies in general: What about for computer science?

The SOLO taxonomy was applied to programming education by Hawkins and Hedberg (1986), who proposed different programming patterns of novices that correspond to the original categories of Biggs and Collis (1982), using as an example the task of drawing simple shapes as circles or rectangles. He associated:

- Pre-structural response: Immediate mode, commands are applied by trial and error, until the result is acceptable.
- Unistructural response: Immediate mode, the commands are entered in a planned and deliberated sequence.
- Multistructural response: Programming mode, structured sequences.
- Relational response: Functions are defined and control structures are used. Code is reused.
- Extended Abstract response: Parametrized functions.

More recently, a group of researchers investigated the fit of different taxonomies (Bloom, AK and SOLO) to the specific needs of computer science (Fuller et al., 2007). The group found that some concepts and structures of these taxonomies were difficult to transfer to CS, in particular, that *understand* and *apply* have specific meaning and an unclear hierarchical position in this domain. In summary, the group recommended the use of the AK taxonomy, but proposed a change of structure. The group suggested that the cognitive process dimension be split into two sub-dimensions: *Interpreting* and *Producing*. The latter represents the more active part of the learning process (e.g. all programming activities) and contains the levels *none, apply* and *create*. The remaining activities of the cognitive process dimension are arranged on the *interpreting* sub-dimension.

Table 16.4 Taxonomy for computer science education (Fuller et al., 2007)

Interpreting Producing	Remember	Understand	Analyse	Evaluate
None				
Apply				
Create				

Enabled by the division in sub-dimensions, it would be possible to express different levels of applying or creating by this way. For example, a programming concept like a repetition loop could be applied without any understanding on the lowest level or by considering aspects of efficiency on the highest level *Evaluate.*

Another attempt to propose a taxonomy for computer science was published in 2010 (Meerbaum-Salant et al., 2010). Here, a combination of the SOLO and the AK taxonomies was proposed based on their evaluation of a Scratch programming course. Looking for categories that were suitable for their specific context, they merged the AK categories Remember and Understand and subdivided levels 3 and 4 of Biggs according to AK levels:

1 Multistructural – Understanding.
2 Multistructural – Applying.
3 Relational – Applying.
4 Relational – Creating.

Limitations of learning objectives

Irrespective of their usefulness for specific purposes, the use of learning objectives has fallen into disrepute during the last decades. One of the reasons might be found in the suspicion that by elaborating a sequence of fine granular objectives for their lessons, teachers might be tempted to restrict the learning process of their students to a very tightly defined sequence (see Duffy and Jonassen, 1992). It might be suggested that teachers should restrict the use of learning objectives to purposes where these are really helpful, for example:

- To identify (one or more) possible learning paths through a specific subject area that is very complicated, very broad or very difficult.
- To arrange a set of concepts sequentially forced by certain circumstances, (e.g. to write a textbook).
- To design an assessment or examination which has to take into consideration which learning progress the students have made up to its point of time.

However, despite all reservations against them, there still is a strong need for learning objectives under certain circumstances. Without these didactical tools, we would struggle to measure progress. As a compromise for the practising teacher we suggest providing only the key learning objectives for each lesson to describe, communicate and evaluate the learning processes.

Example activity: Formulating and testing instructional objectives

Describe what your students are intended to learn during the next lesson. For this purpose, pick the 2–3 concepts of computer science that are most important for this lesson and combine them with descriptions of behaviour that you expect your students to be able to perform after the lesson. Examples may be 'explain the role of the IP address for the transfer of e-mails' or 'program quick sort in Java'. Having the learning goals formulated, design a test task for each of these goals. Write an exemplary solution of these tasks and mark where the intended learning objectives are applied in this solution.

16.3 Competencies

Driven by the upsetting results of the first large-scale studies of learning outcomes such as TIMSS (Trends in International Mathematics and Science Study, see (Mullis, Martin and Loveless, 2016)) and PISA during the first years of this century, the focus of education has shifted broadly from knowledge and learning outcomes towards competencies.

Unfortunately, the terms 'competence' and 'competency' are used in a manifold of senses, ranging from the popular understanding 'something that a person is able to do' to sophisticated definitions from the field of educational psychology. Additionally, there is no consistent differentiation between the terms *competence* and *competency* (Rychen, 2003). Dörge (2010) compared the different backgrounds and use of the terms *competency, skills* and *qualification* in the German and the English language area and found considerable differences.

Here we draw on the well-known definition of Weinert (2001), who defined competencies as 'the cognitive abilities and skills possessed by or able to be learned by individuals that enable them to solve particular problems, as well as the motivational, volitional and social readiness and capacity to use the solutions successfully and responsibly in variable situations' (pp. 27–28). Furthermore, Weinert stressed that competencies may be composed of several facets: ability, knowledge, understanding, skills, action, experience and motivation. It is clear that the combination of these different elements – cognitive ability and skill, motivation and readiness, and capacity to use – make competencies much more wide-ranging and complex than learning objectives, but give us the potential to describe and assess our subject in a more comprehensive way. Thus, the development of competencies is very relevant to teachers and a competency model can ensure more effective assessment.

Competency models

The main purpose of competency research is to define intended learning outcomes of educational processes, as required by the 'customers' of these processes. Obviously, there is a strong need to measure these outcomes to evaluate the educational processes. To align learning and teaching

processes and measure their success, these 'target' competencies must be defined and structured properly by suitable empirically validated competency models. For this purpose, different kinds of models are used (Klieme et al., 2004), which may focus on the structure, the different hierarchical levels or the development of the relevant competencies (Hartig, Klieme and Leutner, 2008). As regards the definition and measurement of competency models, much groundbreaking work was done in the context of the PISA studies (e.g. Seidel and Prenzel, 2008, OECD, 2013).

Klieme et al. (2004) describe three types of competency models:

- Competency *Structure* models, usually structured by dimensions (e.g. competency areas or competency characteristics) describing the cognitive dispositions that learning individuals need to solve tasks and problems in a specific content or requirement area.
- Competency *Level* models, giving information about the levels or profiles of the described competencies.
- Competency *Development* models aiming to describe, how competencies will develop over time.

Level or development models usually have to be based on structure models. As a suitable framework for the development of subject domain-specific competency models, the OECD has presented 'The Definition and Selection of Key Competencies (DeSeCo)' (Rychen, 2003).

Key concept: Competencies

Compared to learning objectives, competencies describe learning outcomes from the viewpoint of the 'customers' of educational institutions (e.g. the IT industry or universities). A competency depicts a quite complex disposal of behaviour that can be applied to solve a certain task or problem that is relevant in 'real' life (e.g. the 'ability to program a robot to move through a labyrinth').

Competencies are arranged in competency model, which come in three types as structure, level and development models.

In CSE, the development process of competency models is just beginning. As far as we know, the only serious attempt until now that could cope with the standards of PISA was the MoKoM project (see Magenheim et al., 2010; Schubert and Stechert, 2010; Neugebauer et al., 2014). The scope of MoKoM was very broad, covering the four dimensions:

- System application
- System comprehension
- System development and
- Dealing with system complexity.

The project aimed to develop an empirically-based competency model in the context of informatics in school. The work had started with a theory-driven model that was enriched through empirical data. In addition, the MoKoM-project aims to develop 'test instruments that are appropriate for

competence measurement and design, and the evaluation of learning environments that have been proven to be of high quality through competence measurement' (Schubert and Stechert, 2010).

Example: A competency model for object-orientated programming

Object-orientated programming (OOP) is usually introduced in upper secondary school or senior high school. A competency model for OOP has been proposed as follows:

1 OOP knowledge and skills
1.1 Data structure (graph, tree, array)
1.2 Class and object structure (object, attribute, association)
1.3 Algorithmic structure (loops, conditional statement)
1.4 Notional machine (data, working memory, processor, statement, program, automaton)
2 Mastering representation (language defined by syntax and semantics)
3 Cognitive Process
3.1 Problem solving stage (understanding the problem, determine how to solve the problem, translating the problem into a computer language program, testing and debugging the program)
3.2 Cognitive Process Type (Interpreting, Producing).

This proposal is based on an extensive literature study on competency models of different subject areas. So far, it has been validated through several surveys among researchers, teachers, and students (Kramer, Hubwieser and Brinda, 2016).

Measuring competencies

Due to their complex structure, it is apparent that the definition and the measurement of competencies are not an easy matter. According to Klieme et al. (2004), competence can only be assessed and measured in terms of performance and can be seen as an ability to deal with a task or particular situation. This means that concrete situations need to be presented to illustrate or assess a competence. In addition, Klieme et al. stress that one performance only does not indicate a competency. They refer to a 'spectrum of performance' and require that assessment should be broad and involve a range of tests to measure competence. This also means that the assessment is not just reflecting shallow and factual knowledge.

Obviously, we need to be convinced that the range of tests or tasks do actually focus on the competency that is being measured. In classical test theory we can do this by using a measure called internal consistency; this is calculated using a statistical test called Cronbach's Alpha Coefficient (Cronbach, 1951). The common rule of thumb for internal consistency is 'excellent' for $\alpha \geq 0.9$, 'good' for ≥ 0.8 and acceptable for $\alpha \geq 0.7$. This can be used to ensure that a test is reliable when testing learning outcomes, but a different type of statistical approach is needed for competencies. This is discussed in the next section.

Item demands and abilities

To assess competencies, we need to design suitable test instruments. A test in this context is a set of tasks, which themselves comprise one or more items. For example, a typical multiple-choice question will represent one task with a textual description and several possible answers to check. Each of these answers represents a dichotomous item in this case. In other cases, for example if the result of a certain formula has to be calculated and responded as an open answer, this task represents one item only. Whether the task is open or closed, each item requires certain skills or abilities, and these are known as the 'item demands'.

The item demands correspond roughly with instructional objectives in the sense of Anderson and Krathwohl (2001). As an example, imagine a test consisting of six tasks with open response format (e.g. submitting program code) that was designed to measure the (potential) competency 'being able to manage a sequence of data by implementing linked lists and their basic operations' (see Kramer et al., 2016). One of the tasks (representing one item) could be 'define a Java class that implements a linked list'. Then, among others, this task would have the following item demands:

(a) being able to write a class definition in Java
(b) being able to define methods in a Java class definition and
(c) being able to write a constructor for Java classes.

This competency definition meets the definition of educational objectives (see Anderson and Krathwohl, 2001): '. . . derived from global objectives by breaking them down into more focused, delimited form', p. 15). The intended global objective could be 'being able to manage write computer programs that store and process structured information'. Obviously, the item demands could be *instructional* objectives ('focus teaching and testing on narrow, day-today slices of learning in fairly specific content areas' (Anderson and Krathwohl, 2001: 16)).

Consider an example shown in Figure 16.1. If several items (e.g. Items 1, 2 versus Item 3 in Figure 16.1) differ in their demands, two types of differences have to be decided (see Hartig, 2008):

(1) The items differ in difficulty (e.g. in the empirical solution frequencies (case 1)).
(2) The items differ in relations between responses (i.e. correlations between scores for different items (case 2)).

Figure 16.1 demonstrates how the difficulty of an item can be identified. Case 1 shows equal correlations with different difficulty level, whereas case 2 shows different correlations with equal difficulties. In case 1, all three items have equally high correlations between each other. Items 1 and 2 are equally difficult, but item 3 is more difficult. This finding could mean that the ability to master the task demand (c) highly correlates with the ability to master demands (a) and (b) and can be regarded as the 'same ability' for measurement purposes, which has to be developed to a higher degree to master task demand (c).

In case 2, items 1 and 2 have a high positive correlation. However, the correlations of item 3 with items 1 and 2 are substantially lower than the correlation between items 1 and 2. In this case, an additional ability dimension would be needed to explain the specific variation caused by the

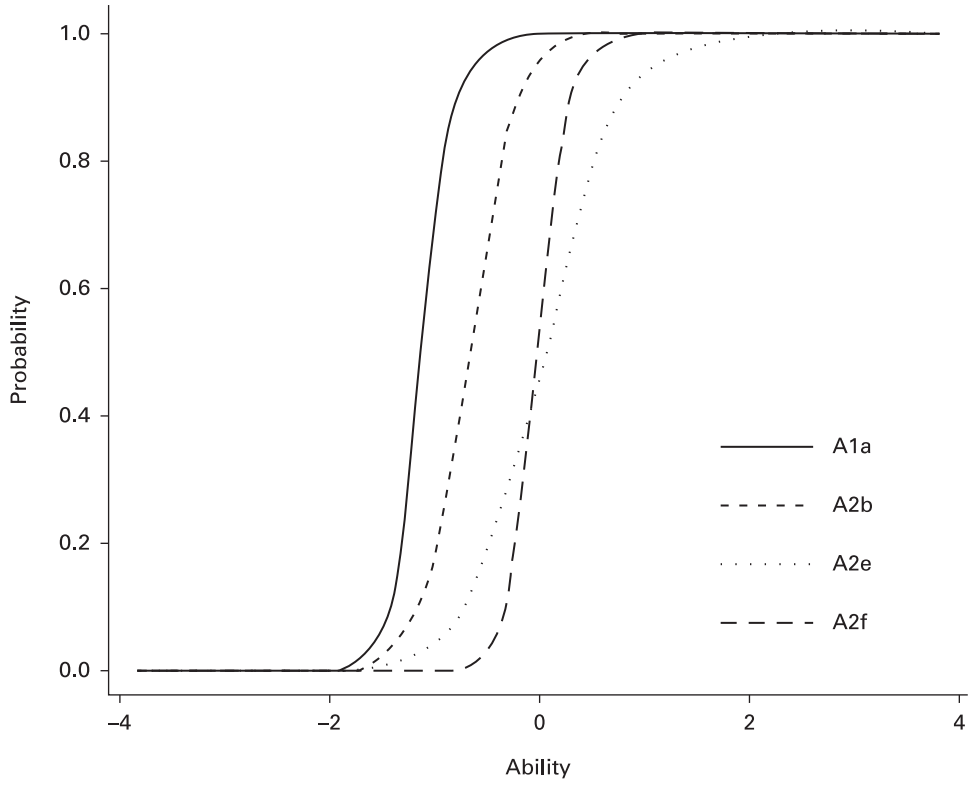

Figure 16.1 Differences in item demands

additional demand (c) of item 3. This could be regarded as a 'different ability' which is required to master item 3.

A third possibility would be that the task demand has no observable effect at all – items 1, 2 and 3 could turn out to be equally difficult and to have equal correlations among each other. In this case, the task demand would appear to be irrelevant for observable test performance.

In our linked-list example above, items 1 and 2 could require the demand (a) and (b), while only item 3 demands the implementation of a constructor (c). Then, case 1 would indicate that the implementation of a constructor is more difficult compared to (a) and (b), but nevertheless belongs to the same competency 'implementing a linked list'. In contrast, case 2 would require a separate competency dimension for implementing constructors.

Item response theory

Item Response Theory (IRT) is a way to analyse responses to tests or questionnaires with the goal of improving reliability and validity. It is a technique to ensure that the tests measure what they are supposed to measure (see Rasch, 1960). In terms of competencies, IRT is the current 'state of the art'. While in Classical Test Theory, the psychometric construct of interest (in our case a certain

competency) is considered to be measured directly by item scores, IRT considers this construct as latent and not directly measurable.

Instead, the probability of correct answers on a certain item depends on the competency in a certain way:

$$P(X_{ik} = 1 \mid \theta_i, \beta_k) = f(\theta_i, \beta_k) \tag{1}$$

Here θ_i is the *ability parameter* of person *i*, representing his/her level of competency β_k the *difficulty parameter* of Item *k,* and $f(\theta_i,\beta_k)$ a certain function that is determined by the *psychometric model* (e.g. the *Rasch Model*, see below) that is assumed to fit the observations. In most cases, these parameters have to be estimated by effortful numerical calculations. Depending on the structure of the psychometric constructs that are to be measured, several different models may be applied (e.g. *unidimensional* models that cover only one single competency or, alternatively, *multidimensional* models). One of the simplest and most widely used models is the basic unidimensional *Rasch Model* (*RM)* with one psychometrical factor and one parameter (1F1P):

$$P(X_{ik} = 1 \mid \theta_i, \beta_k) = \frac{\exp(\theta_i - \beta_k)}{1 + \exp(\theta_i - \beta_k)} \tag{2}$$

Due to its restriction on one factor and one parameter, the application of the RM requires three preconditions that have to be met:

1. *Homogeneity* of items: This means that all items must measure the same psychometric construct. In this case, we can call this set of items *homogenous*.
2. *Local stochastic independence*: the underlying psychometric construct is the only coupling factor between items.
3. *Specific objectivity*: for all samples from the population, the item parameters are independent of the specific person sample; the same holds for all samples of items and person parameters.

Provided that this model is applicable, some very convenient simplifications can be made. For example, the sum of the scores of all individual items is a sufficient statistic, which means that the (estimated) person parameter depends only on the *total number* of correct answers given by this person. It does not matter, *which* items the person has responded to correctly.

Key concept: Item response theory

While the Classical Test Theory assumes that the constructs of interest (e.g. motivation or intelligence) can be measured more or less directly with some errors, Item Response Theory (IRT) aims to give mathematical dependencies between the personal level of the constructs of interests (e.g. the level of competency) and the probability to solve a certain item. Thus it takes both the difficulty of the item and the proficiency of the student into account.

The graph of this function looks as displayed in Figure 16.2 for four different values of β ('Ability'). These graphs are called *Item Characteristic Curves* (ICCs).

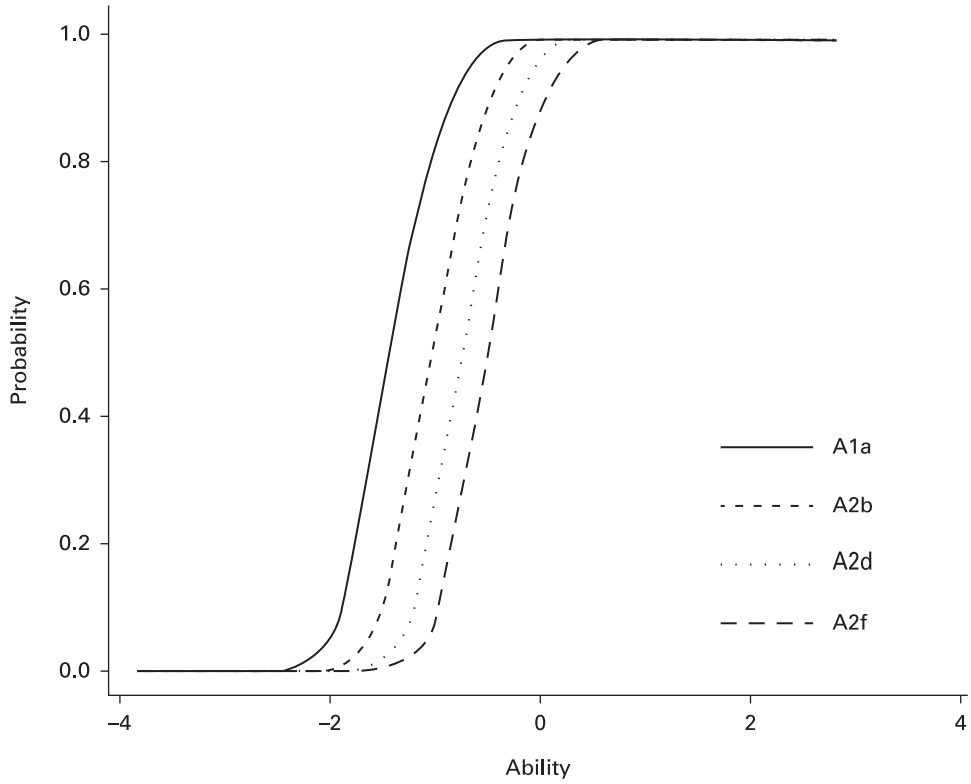

Figure 16.2 Example of an item-characteristic curve of the Rasch Model

If the difficulty (β) varies, the ICC shifts horizontally. This demonstrates a rather convenient advantage of IRT: that person and item parameters are located on the same scale. Thus, they can be directly compared to allow statements like 'the probability for a person x to solve item y is greater than 0.5 if $\theta_x > \beta_y$'.

Obviously, there might be cases where ICCs have different slopes. If so, we cannot model this situation using RM. Another model which can be used in this case is the Birnbaum Model (BM), which has an additional parameter called *Discrimination* δ_k. (Birnbaum, 1968). In the Birnbaum Model there is:

$$P(X_{ik} = 1 \mid \theta_i, \beta_k, \delta_k) = \frac{\exp(\delta_k(\theta_i - \beta_k))}{1 + \exp[\![(\delta]\!]_k(\theta_i - \beta_k))} \tag{3}$$

Figure 16.3 displays the ICCs of another (real existing) item set that varies in difficulty (horizontal position) β as well as in discrimination (slope) β. As the ICCs of item A2e and A2f in Figure 16.3 demonstrate, the variation of slope can cause intersections of ICCs. This would mean that the difficulty order of the regarded items depends on the person parameter, which would violate the requirement of specific objectivity (see above). The reason for this effect might be that the answers on these items might be influenced by other factors than the construct to be measured. On the other hand, low variation of slopes and missing intersections of a certain set of items in the BM can be regarded as a good indicator that the RM is applicable on this set.

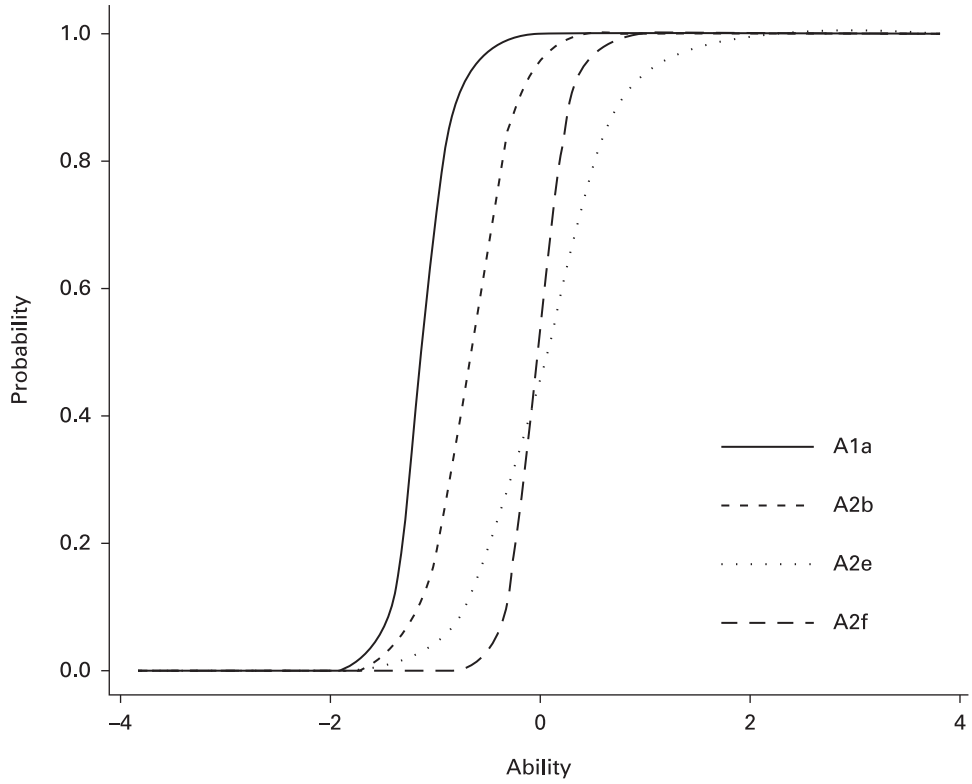

Figure 16.3 Example of an item-characteristic curve using the two-parameter Birnbaum Model

In the case that there is more than one psychometric construct to be measured, multidimensional models have to be applied. To read more about these we recommend Rost and Carstensen's original paper (Rost and Carstensen, 2002).

Factor analysis of dichotomous data

To investigate the homogeneity of a set of items, classical explorative factor analysis is applied traditionally. Yet, as the score format in competency measurement is often dichotomous (because things get much more complicated otherwise) this is not applicable. Latent trait analysis (LTA) offers a better alternative (Bartholomew, Steel, Moustaki and Galbrath, 2008).

With LTA, it is assumed that the responses of the students to a given set of items can be described by a certain psychometric model (e.g. by the *Rasch Model*). Under this assumption, one can estimate all person and item parameters based on the scoring matrix of the responses. Using the estimated values of the parameters, by calculating the probability P in equation (1) of section II.E, the expected number of occurrences $E(r)$ of all possible response patterns r (e.g. 01101 in the case of five items) can be calculated. For p dichotomous items, we have 2^p response patterns (i.e. combinations of 0s and 1s with the length p). For each pattern r, its expected frequency $E(r)$ is compared to the actually

observed pattern frequency *O(r)*. For the differences, the log-likelihood test statistic G^2 and the common X^2 statistic are calculated that both describe the differences of the expected and the measured values. As both statistics are approximately X^2 distributed, we could estimate the goodness of fit of the applied Rasch Model. The precondition for this calculation is a sufficient number of datasets, which assures that the frequency of each pattern has an expectation value of at least 5 (Bartholomew et al., 2008).

Rasch Model tests

In addition to LTA, a set of standard tests for the fit of the chosen psychometric model (e.g. the Rasch Model) are often applied. These standard tests for specific objectivity (remember that objectivity is one of the goals in improving summative assessment) are used to check that the model would produce the same results for different groups of participants. They use the idea of a splitting criteria to split participants into groups and test that the results fit for each group. Examples of three tests that can be used are as follows:

- *Likelihood-Ratio-Test* (Andersen, 1973) with the splitting criteria *median (*respectively *mean),* values of *combination score* and *gender* on the level of the total item set.
- *Martin-Löf-Test* (Martin-Löf, 1974) with the splitting criterion *median (*respectively *mean)* on the level of the total item set.
- *Wald-Test* (Wald, 1943) with the splitting criteria *median (*respectively *mean)* and gender on the level of single items.

These tests can be carried out using a statistical package such as R; we do not have space to explore these any further here.

16.4 Educational standards

Educational standards are sets of competencies depicted in detail that were decided by educational authorities to be the minimal or average learning outcomes of educational institutions (e.g. some algebraic competencies that should be achieved by all Year 9s of regional grammar schools). In an influential paper on the development of national educational standards, Klieme et al. state:

> Educational standards, as conceived of in this report, draw on general educational goals. They specify the competencies that schools must impart to their students in order to achieve certain key educational goals, and the competencies that children or teenagers are expected to have acquired by a particular grade. These competencies are described in such specific terms that they can be translated into particular tasks and, in principle, assessed by tests (Klieme et al., 2004: 15).

In computer science, a lot of work has to be done until our domain is ready for the definition of standard in this sense; CS runs far behind traditional subjects like mathematics. The *Principles and Standards of the National Council of Teachers of Mathematics* (National Council of Teachers of

Mathematics (NCTM), 2000), are the best-known and most influential example internationally. They describe framework conditions for instruction on all grade levels, from the beginning of primary education to the end of secondary schooling and provide guidelines for improving mathematics teaching by moving towards comprehension- and problem-based instruction. In particular the NCTM presents a definition of *problem solving* that might be transferred to CSE as well: 'Problem solving means engaging in a task for which the solution method is not known in advance. In order to find a solution, students must draw on their knowledge, and through this process, they will often develop new mathematical understandings. Solving problems is not only a goal of learning mathematics, but also a major means of doing so' (NCTM, 2000: 52).

Some proposals for educational standards in informatics have been published in Austria (Dorninger, 2005) and from the German *Gesellschaft für Informatik* (GI) (Gesellschaft für Informatik e V, 2008).

Recently the CSTA Standards Task Force presented its K-12 Computer Science Standards (Revised 2011) in a draft version (Seehorn et al., 31 March 2011). These standards may be comprised by the subcategory *standards* (of the category *intentions* in the DM). It defines three levels for the learning outcomes, where the highest is divided into three discrete 'courses':

- level 1 (recommended for grades K–6): Computer science and me
- level 2 (recommended for grades 6–9): Computer science and community
- level 3 (recommended for grades 9–12): Applying concepts and creating real-world solutions
- level 3A (recommended for grades 9 or 10): Computer science in the modern world
- level 3B: (recommended for grades 10 or 11): Computer science principles
- level 3C: (recommended for grades 11 or 12): Topics in computer science.

To avoid the perception that CSE should focus exclusively on programming, five complementary and essential strands throughout all three levels are distinguished:

- computational thinking
- collaboration
- computing practice
- computers and communication devices and
- community, global and ethical impacts.

These strands are further illustrated by lists of competencies that represent the proposed standards. Additionally the draft paper also offers a variety of activities, assigned to the levels and strands, respectively that show in detail what classroom teaching might look like.

16.5 Summary

In this chapter we have considered some quite challenging questions relating to how we can be sure we test and measure students' learning accurately and objectively. This is a complex field which needs to be addressed within CSE, although the development of standards and competencies is still

in the early stages of development. The change to a competency approach has been driven partly by PISA, an international assessment measure by which the achievements of young people in different countries can be compared. However, even at a local level, it is important to have an understanding of the potential inaccuracies of traditional testing methods and new methods for ameliorating the situation.

Key points

- In the area of summative assessment, there is a move from defining and measuring learning outcomes to being able to specify more broad-ranging competencies.
- Previously, learning outcomes have been categorized through the use of taxonomies such as Bloom's, the revised Bloom's taxonomy by Anderson and Krathwohl and the SOLO taxonomy.
- Competencies describe learning outcomes from the viewpoint of the 'customers' of educational institutions.
- More rigorous methods for measuring test results can give us accurate information. Item Response Theory (IRT) is starting to replace classical test theory as a method by which tests can be evaluated for reliability and validity.
- Educational standards are sets of competencies that explain in detail the minimal or average learning outcomes of educational institutions, regions or countries.

Further reflection

- Consider how students in your country participate in summative assessments, at the national, regional or school level. To what extent are the measurements of student performance in computer science reliable, valid and objective?
- Consider the educational standards in your country for CSE in schools. How are these made explicit to teachers and students?

References

Adams, HF (1936). 'Validity, Reliability and Objectivity' *47*(2) *Psychological Monographs* 329–350.

Andersen, EB (1973). 'A Goodness of Fit Test for the Rasch Model' *38*(1) *Psychometrika* 123–140.

Anderson, LW and DR Krathwohl (2001) *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives* Abridged edn (New York, Longman). Available at: wwwgbvde/dms/bowker/toc/9780321084057pdf

Bartholomew, DJ, F Steel, I Moustaki and JI Galbrath (2008). *Analysis of Multivariate Social Science Data* (2nd edn) Chapman and Hall/*CRC statistics in the Social and Behavioral Sciences Series* (Boca Raton Fl, CRC Press/Taylor & Francis).

Biggs, JB and KF Collins (1982). *Evaluating the Quality of Learning: The SOLO Taxonomy; Structure of the Observed Learning Outcome Educational Psychology Series* (New York, Academic Press).

Birnbaum, A (1968). 'Some Latent Trait Models and Their Use in Inferring an Examinee's Ability' in FM Lord, MR Novick and A Birnbaum (eds), *Statistical Theories of Mental Test Scores* (Reading, MA, Addision-Wesley) 395–479.

Bloom, BS (1956). *Taxonomy of Educational Objectives: The Classification of Educational Goals* (New York, David Mackay Company Inc). Available at: wwwworldcatorg/oclc/277182491.

Cronbach, LJ (1951). 'Coefficient Alpha and the Internal Structure of Tests' *16*(3) *Psychometrika* 297–334.

Cronbach, LJ and PE Meehl (1955). 'Construct Validity in Psychological Tests' 52 *Psychological Bulletin* 281–302.

Dörge, C (2010). 'Competencies and Skills: Filling Old Skins with New Wine' in *Key Competencies in the Knowledge Society* (Berlin, Springer) 78–89.

Dorninger, C (2005). 'Educational Standards in School Informatics in Austria' in RT Mittermeir (ed), *Lecture Notes in Computer Science: From Computer Literacy to Informatics Fundamentals* (Berlin, Springer) 65–69.

Duffy, TM and DH Jonassen (1992). *Constructivism and the Technology of Instruction: A Conversation* (Hillsdale, NJ: Lawrence Erlbaum Associates) Available at: wwwlocgov/catdir/enhancements/fy0745/92022781-dhtml

Fuller, U, CG Johnson, T Ahoniemi, D Cukierman, I Hernán-Losada, J Jackova, E Lahtinen, TL Lewis, D McGee, CR Thompson, E Thompson (2007). 'Developing a Computer Science-specific Learning Taxonomy' 39(4) *SIGCSE Bulletin* 152–170.

Gesellschaft für Informatik e V (ed), (2008). Grundsätze und Standards für die Informatik in der Schule Bildungsstandards Informatik für die Sekundarstufe I: Empfehlungen der Gesellschaft für Informatik e V, erarbeitet vom Arbeitskreis (Bonn, Bildungsstandards).

Hartig, J (2008). 'Psychometric Models for the Assessment of Competencies' in J Hartig, E Klieme and D Leutner (eds), *Assessment of Competencies in Educational Contexts* (Toronto: Hogrefe & Huber Publishers) 69–90.

Hartig, J, E Klieme and D Leutner (eds), (2008). *Assessment of Competencies in Educational Contexts* (Toronto, Hogrefe & Huber Publishers).

Hawkins, W and JG Hedberg (1986). 'Evaluating LOGO: Use of the SOLO Taxonomy' 2(2) *Australian Journal of Educational Technology* 103–109.

Hubwieser, P (2007). 'A Smooth Way Towards Object-oriented Programming in Secondary Schools' in D Benzie and M Iding (eds), *Informatics, Mathematics and ICT: A Golden Triangle*, Proceedings of the Working Joint IFIP Conference: WG31 Secondary Education, WG35 Primary Education; College of Computer and Information Science, Northeastern University Boston, Massachusetts, USA. 27–29 June 2007.

Hubwieser, P (2008). 'Analysis of Learning Objectives in Object-oriented Programming' in R T Mittermeir and M M Syslo (eds), *Lecture Notes in Computer Science, Informatics Education – Supporting Computational Thinking,* Third International Conference on Informatics in Secondary Schools – Evolution and Perspectives, ISSEP 2008, Torun, Poland, 1–4 July 2008, 142–150.

Hubwieser, P, MN Giannakos, M Berges, T Brinda, I Diethelm, J Magenheim, P Yogendra, J Jackova and E Jasute (2015). 'A Global Snapshot of Computer Science Education in K–12 Schools' in ITICSE-WGR '15, *Proceedings of the 2015 ITiCSE on Working Group Reports* (New York, NY, ACM) (65–83).

Klieme, E (ed), (2004). *The Development of National Educational Standards: An Expertise* (Berlin, Bundesministerium für BildungundForschung).

Kramer, M, P Hubwieser and T Brinda (2016). 'A Competency Structure Model of Object-oriented Programming' in *International Conference on Learning and Teaching in Computing and Engineering (LaTICE)* IEEE Xplore Digital Library, 1–8.

Leutner, D, J Hartig and N Jude (2008). 'Measuring Competencies: lntroduction to Concepts and Questions of Assessment in Education' in J Hartig, E Klieme and D Leutner (eds), *Assessment of Competencies in Educational Contexts* (Toronto, Hogrefe & Huber Publishers) 177–192.

Magenheim, J, W Nelles, T Rhode, N Schaper, SE Schubert and P Stechert (2010). 'Competencies for Informatics Systems and Modeling: Results of Qualitative Content Analysis of Expert Interviews' in *Education Engineering (EDUCON), 2010 IEEE*, 513–521.

Mager, RF (1961). *Preparing Objectives for Programmed Instruction* (San Francisco, Fearon Publishers).

Martin-Löf, P (1974). 'Exact Tests, Confidence Regions and Estimates in Memoirs' Vol 1 in *Proceedings of Conference on Foundational Questions in Statistical Inference* (Aarhus, 1973) 121–138.

Meerbaum-Salant, O, M Armonia and M Ben-Ari (2010). 'Learning Computer Science Concepts with Scratch in ACM' (ed), *ICER '10: Proceedings of the Sixth International Workshop on Computing Education Research* (New York, NY ACM) 69–76.

Mullis, IV, MO Martin and T Loveless (2016). '20 Years of TIMSS: International Trends in Mathematics and Science' Lynch School of Education, Boston College and International Association for the Evaluation of Educational Achievement (IEA): Achievement, Curriculum, and Instruction Boston, MA.

National Council of Teachers of Mathematics (2000). *Principles and Standards for School Mathematics* (Reston, VA).

Neugebauer, J, J Magenheim, L Ohrndorf, N Schaper and S Schubert (2015). 'Defining Proficiency Levels of High School Students in Computer Science by an Empirical Task Analysis'. Results of the MoKoM Project: Informatics in Schools Curricula, Competences, and Competitions: Proceedings in *8th International Conference on Informatics in Schools: Situation, Evolution, and Perspectives*, ISSEP 2015, Ljubljana, Slovenia, 28 September–1 October 2015, (New York, NY, Springer International Publishing) 45–56.

Neugebauer, J, P Hubwieser, J Magenheim, L Ohrndorf, N Schaper and S Schubert (2014). 'Measuring Student Competences in German Upper Secondary Computer Science Education' in Y Gülbahar nd E Karatas (eds), *Informatics in Schools Teaching and Learning Perspectives* (Heidelberg, New York, Springer) 100–111.

Organisation for Economic, Co-operation and Development (2013) Pisa 2012 'Results in Focus: What 15-year-olds know and what they can do with what they know' (Paris, OECD). Available at: wwwoecdorg/pisa/keyfindings/pisa–2012-results-overviewpdf

Rasch, G (1960). *Probabilistic Models for Some Intelligence and Attainment Tests Studies in Mathematical Psychology: Vol 1* (Copenhagen: Danmarks pædagogiske Institut).

Robinsohn, SB (1967). *Bildungsreform als Revision des Curriculum Aktuelle* (Pädagogik Neuwied aRh, Luchterhand).

Rost, J and CH Carstensen (2002). 'Multidimensional Rasch Measurement via Item Component Models and Faceted Designs' 26(1) *Applied Psychological Measurement* 42–56.

Rychen, DS (2003). 'Key Competencies for a Successful Life and a Well-functioning Society' (Toronto, Hogrefe & Huber Publishers).

Schubert, SE and P Stechert (2010). 'Competence Model Research on Informatics System Application' in IFIP (ed), *New Developments in ICT and Education Workshop of WG 31,* 28–30 June 2010, Amiens.

Seehorn, D, S Carey, B Fuschetto, I Lee, D Moix, D O'Grady-Cuniff and A Verno (2011). *CSTA K–12 Computer Science Standards: Revised 2011* (New York NY, Standards Task Force).

Seidel, T and M Prenzel (2008). 'Assessment in Large-Scale Studies' in J Hartig, E Klieme and D Leutner (eds), *Assessment of Competencies in Educational Contexts* (Toronto, Hogrefe and Huber Publishers) 279–304.

Smith, PL and TJ Ragan (2005). *Instructional Design* (3rd edn) (Hoboken, NJ, Wiley). Available at: wwwgbvde/dms/bowker/toc/9780471393535pdf

Wald, A (1943). 'Tests of Statistical Hypotheses Concerning Several Parameters When the Number of Observations Is Large' *Transactions of the American Mathematical Society*, 426–482.

Weinert, FE (1999). *Concepts of Competence: Definition and Selection of Competencies* (Elektronische Ressource (Sl), OFS).

Weinert, FE (2001). 'Concept of Competence: A Conceptual Clarification' in DS Rychen and L Salganik (eds), *Defining and Selecting Key Competencies* (Toronto, Hogrefe & Huber Publishers) 45–65.