Empirical Research in
Vocational Education and Training

RESEARCH

Open Access

Cognitive diagnosis models of students' skill profiles as a basis for adaptive teaching: an example from introductory accounting classes

Christoph Helm[1*], Julia Warwas[2] and Henry Schirmer[3]

*Correspondence:
christoph.helm@jku.at

[1] Johannes Kepler University of Linz, Altenbergerstraße 69, 4040 Linz, Austria
[2] University of Hohenheim, Schloss Hohenheim 1, 70599 Stuttgart, Germany
[3] Technical University of Munich, Arcisstr. 21, 80333 Munich, Germany

Abstract

A critical limiting factor for adaptive teaching is the availability of diagnostic tools that allow reliable and valid assessments of students' domain-specific skills in a way that produces detailed information for planning subsequent instructional strategies. The present study demonstrates how Cognitive Diagnosis Models (CDM) can deliver fine-grained diagnostic information on students' skills in dealing with domain-specific tasks, using introductory accounting as an exemplary field of application. Based on data from a sample of 773 students from secondary business schools in Austria, statistical analyses that incorporated several criteria for evaluating model fit corroborate theoretical assumptions on distinct skills as *multiple* dimensions of accounting competence. Moreover, they illustrate that CDMs allow not only to quantify the shares of students who have mastered or still lack each accounting skill but also to identify *individual skill profiles*, which can serve as reliable classification criteria to distinguish homogeneous or heterogeneous ability groups among the learners. We conclude by discussing the practical implications of different diagnostic information obtained from CDM outputs for generic strategies of adaptive teaching, again with an illustrative focus on introductory accounting instruction.

Keywords: Cognitive diagnosis model, Skill profiles, Introductory accounting education, Adaptive teaching, Within-class ability-grouping

Introduction

At the core of instructional designs such as adaptive teaching (e.g., Vogt and Rogalla 2009), individualized instruction (e.g., Waxman et al. 2012), and differentiated instruction (e.g., Valiandes 2015) lies the tailoring of instructional methods and materials to meet different learner needs and prerequisites. Although the latter span a wide range of psychological constructs, students' current cognitive skills and deficits in the relevant curricular domain count among the most important determinants of subsequent learning steps, and thus, represent a salient reference point for adjusting instructional designs (Blayney et al. 2015; Tomlinson and Jarvis 2009). Meta-analytic findings on predictors of academic achievement (Schneider and Preckel 2017; Steenbergen-Hu et al. 2016) further substantiate that students' current conceptual understanding and operational proficiency in handling domain-specific tasks should be considered carefully in providing adaptive teaching. Generic strategies of adaptiveness considering both within-class and between-class variation of domain-specific skills and deficits may include:

1. Delivering *individualized, elaborative feedback* that supports each student in reaching desired levels of understanding and proficiency (e.g., Hattie and Gan 2011). This strategy entails statements on (i) the student's proximal learning goals, (ii) his/her learning progress, as reflected in larger or smaller discrepancies between demonstrated task-processing skills and a defined standard or prior performance, and (iii) effective ways of handling the inherent demands of tasks that have not been mastered yet.
2. Drawing up intelligent plans of *within-class ability-grouping and differentiated task assignment*. This requires reliable identification of class members with specific profiles of skills and deficits to choose or construct appropriate learning tasks and materials for them (e.g., Park and Datnow 2017). While students with homogeneous skill profiles might be grouped to receive targeted interventions to meet specific task demands (such as additional explanations or worked examples; e.g., Paas and van Gog 2006), students with heterogeneous skill profiles might engage in reciprocal tutoring, thereby helping each other to enhance knowledge structures or fill knowledge gaps (e.g., Dioso-Henson 2012; Roscoe and Chi 2007).
3. Setting *priorities for deliberate practice* in different classes: Dependent on class-average proficiency levels, a teacher may choose to devote instructional time to extensive, reflective exercises for particular types of tasks that currently pose a major challenge for many students in one class but not in another (Bloom 1968; Fuchs et al. 2010; Lehtinen et al. 2017).

However, a critical constraining factor for adaptive teaching is the *availability of diagnostic tools* that allow *reliable and valid assessments* of students' domain-specific skills and deficits in a way that produces *rich and detailed information* for planning subsequent instructional strategies (Bennett 2011; Shepard 2005). More precisely, adaptive teaching necessitates both *evidence-based and fine-grained* diagnostic information on each learner's skills and deficits and the extent to which they converge or diverge with the skills and deficits of classmates.

This kind of information cannot be derived from unidimensional test scores or grades, which deliver summative assessments of the competence level a student has reached at the end of a longer instructional unit or even a school year (La Torre and Minchen 2014). Instead, *formative assessment* techniques are needed that accompany instructional processes to collect diagnostic information on student learning continuously (Black and Wiliam 2009). Although *using evidence* to guide instructional decisions and *involving students as the main source of evidence* (sometimes even as diligent evaluators of lessons) constitute key elements of all these assessment techniques (Lyon et al. 2019), they vary markedly in terms of formalization. *On-the-fly-assessments* represent the most informal category, as the teacher forms his impressions of student understanding and interest rather intuitively and spontaneously (e.g., Warwas et al. 2015). *Planned-for interactions* occupy a middle position, as the teacher uses prepared and elaborated interaction strategies, such as those of classroom dialogue, to probe into students' knowledge structures (e.g., Ruiz-Primo 2011). Standardized tests that are closely aligned with the curriculum or even an integral part of the educational program *(curriculum-based* or *curriculum-embedded assessments*; see Hopster-den Otter et al. 2019) follow strict rules and standards for their development and implementation.

Cognitive Diagnosis Models (CDMs) of processing domain-specific tasks (Dibello et al. 2007; Rupp et al. 2010) as set out in the present paper belong to this formal category. Based on a theoretically founded domain competence model, which delineates skills as multiple dimensions of competence, CDMs give insights into.

- the pattern of domain-specific skills each student possesses or lacks at any given point of assessment during an instructional unit *(individual skill profile)*;
- sub-groups of students with the same skill profiles *(skill classes)*;
- the share of learners within any given student body (such as students in a class or cohort) who possess a particular skill *(mastered skills).*

Thus far, assessments deploying statistical models from the CDM family mainly pertain to narrowly defined competencies in primary school curricula, such as reading comprehension ability or fraction subtraction ability (see for overviews Bley 2017; La Torre and Minchen 2014). Only recently they have been applied to complex professional competencies that exist in curricula of vocational education, namely intrapreneurship competency (George et al. 2019). However, to the best of our knowledge, no CDM-based assessments exist for the domain of introductory accounting instruction. Considering that conceptual understanding and operational proficiency in accounting are indispensable elements of any business education, the need for developing such an assessment is high.

The scientific literature on introductory accounting instruction offers a few analytical approaches that break up students' task processing into essential (ideal-type) cognitive operations. Among them is the contribution by Dillard et al. (1982), who propose a problem-solving model in bookkeeping based on a hierarchical representation of accounting knowledge, and the approach by Sithole and Abeysekera (2017), which adopts a cognitive load perspective on students' cognitive processes when dealing with accounting task demands. The skill-centered competency model elaborated in the present paper is fully compatible with such approaches. However, we take extant research one step further by demonstrating how the proposed model can guide the construction of diagnostic assessments and how the use of CDMs allows us to distinguish and sort individual, multidimensional skill profiles among learners *empirically*. Drawing on exemplary assessment results from introductory accounting classes in upper vocational schools, we thus illustrate how statistical CDM output can deliver reliable and highly detailed diagnostic data on the assessed learners' particular strengths and weaknesses.

Against this background, the present paper aims to describe and evaluate the design principles for and the diagnostic information from Cognitive Diagnosis Models (CDMs), with introductory accounting serving as an exemplary domain of application. To this

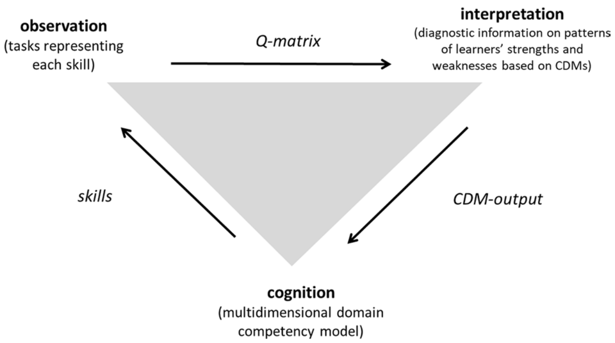
observation
(tasks representing each skill)
Q-matrix
interpretation
(diagnostic information on patterns of learners' strengths and weaknesses based on CDMs)
skills
CDM-output
cognition
(multidimensional domain competency model)

Fig. 1 Assessment triangle for diagnostic assessments applying CDMs (adapted from' George et al. 2019, p. 90)

end, we first outline a generic assessment framework for using CDMs and the essential psychometric properties of different models within the CDM family (Conceptual assessment framework). We then demonstrate how the assessment framework guides the construction and validation of CDMs in the curricular domain of interest (Constructing and validating CDMs for introductory accounting classes). This is done in several interdependent steps, including the specification of the assessment goal (3.1), the theoretical justification of a competence model for processing basic accounting tasks (3.2), the compilation of assessment items (3.3), the examination of their adequacy to trigger all of the skills postulated in the competence model (3.4), and finally, the examination of various criteria to select the most suitable model from the CDM family for the chosen domain of application (3.5). After presenting CDM outputs on skill profiles, skill classes, and skill mastery for 773 students in 31 introductory accounting classes (CDM output evaluation against the assessment context), we conclude by discussing the main findings and their practical implications for the three generic strategies of adaptive teaching set out at the beginning of this paper Discussion. We provide all data material (Additional files 1 and 2) and the R script (Additional file 3) for the analyses reported here in the supplementary material.

Conceptual assessment framework

To obtain detailed and substantive diagnostic data on each student's particular skills and deficits when dealing with domain-specific learning content, the procedure of constructing and validating diagnostic assessments must comply with a rigid design concept. The present paper draws on Pellegrino's *assessment triangle* (Pellegrino et al. 2001; Pellegrino 2010). This framework is consistent with central elements of the Evidence-Centered Design Approach (ECD; Mislevy and Haertel 2006) and often guides competence measurement in vocational education and training (Achtenhagen and Winther 2014; Klotz and Winther 2017). In general, the assessment triangle requires a stringent connection of three elements (i.e., the corners of the triangle, see Fig. 1) to provide evidence-based information for a defined diagnostic utilization:

1. the *cognition* corner, which entails justified assumptions on latent competencies for dealing successfully with typical cognitive demands of the focal domain following curricular and expert standards *(explication of the domain competency model)*;

2. the *observation* corner, which defines a set of tasks that are suitable for triggering the use of these competencies in the solution process *(specification of assessment tasks)*;
3. the *interpretation* corner, which includes reasoned decisions on the most appropriate statistical model, allowing inferences from a learner's observable task processing to his/her domain-specific competence *(methods for analyzing and interpreting observational data).*

When conducting Cognitive Diagnostic Assessments of students' task processing and employing *Cognitive Diagnosis Models* as the analytical method of choice, the connections between the corners of the assessment triangle can be established through skill definitions, a Q-matrix, and the CDM-output, as depicted in Fig. 1 (George et al. 2019).

In line with the graphical illustration, the development and validation of an assessment necessitate the definition of *skills* that give an exhaustive account of the content and structure of a multidimensional competence construct. Since distinct skills indicate essential dimensions of the domain competence model and their application should reliably be prompted by the assessment tasks, they establish the link between the cognition corner and the observation corner. Furthermore, a detailed *task-to-skill assignment* is needed to couple the observation corner and the interpretation corner. When CDMs serve as analytical tools, the *Q-matrix* (Tatsuoka 1983) accomplishes this connection, thereby assuming that each task necessitates the use of at least one skill to solve it correctly. Finally, the statistical *CDM output* relates the interpretation corner to the cognition corner by deriving statements about which skills of the domain competence model a student possesses or not, based on his/her demonstrated performance on the assessment tasks.

It has to be noted, however, that such inferences are only admissible under the condition that researchers have made an adequate choice from the family of CDMs. That is, valid interpretations presuppose the selection of a model that best aligns with theoretical assumptions of the domain competence model, extant knowledge about the assessed learner group and learning setting, and also empirical quality criteria (Dibello et al. 2007). We report CDM-model selection for diagnosing students' skills and deficits when dealing with basic accounting tasks in Selection of a psychometric model from the family of Cognitive Diagnosis Models. In the following, we introduce general assumptions of CDM-based assessments and their psychometric properties.

Model assumptions of CDMs

CDMs are statistical models that explain item responses by multiple underlying skills (multidimensional models). They represent a factor analytical approach and resemble traditional confirmatory factor analysis (CFA) in that they align items to underlying factors that were defined in advance. However, they differ in several aspects, of which the most important are:

(1) CDMs assume and specify unobservable groups of respondents (latent classes) with different skill patterns, that is, groups of students with a different set of mastered and non-mastered skills. For each latent class, a separate regression is estimated according to the skill profile of the latent class. Contrary to Latent Class Analysis

(LCA), which is mainly used in an explorative way, the number of skills – elaborated in the competence model and specified in the Q-matrix – pre-defines the number of latent classes of a CDM. The maximum number of groups of respondents with different mastery/non-mastery skill profiles is 2 to the power of the number of latent skills assumed (confirmatory latent class approach).

(2) While CFA is based on classical test theory, CDMs are grounded in probabilistic test theory. Classical test theory represents a deterministic approach, by predicting item responses in an "absolute manner" via linear regression. In contrast, probabilistic test theory estimates the probability of an outcome employing a logistic regression. Hence, these models inform on the probability of a respondent with a given ability to solve an item correctly or wrongly – sometimes even controlling for guessing or inattention.

(3) While CFA usually assumes continuous (uni- or multidimensional) factors, CDMs assume categorical (multidimensional) factors indicating mastered/non-mastered skills.

(4) Finally, while CFA can involve categorical as well as continuous outcomes (i.e., item responses), CDMs use categorical, usually dichotomous, data like 0/1 responses (although ordinal data is possible, too).

The log-linear cognitive diagnosis framework

CDMs represent log-linear models (with latent classes) and thus fall under the framework of log-linear (cognitive diagnosis) models (LCDM, Henson et al. 2008). Within the LCDM framework, the most general notation of an item response function is as follows:

Equation 1:

$$P\left(Y_{ij} = 1 \,|\alpha_j\right) = \frac{exp(\lambda_0 + \lambda_{\alpha 1} + \lambda_{\alpha 2} + \lambda_{\alpha \ldots} + \lambda_{\alpha 1 * \alpha 2} + \ldots)}{1 + exp(\lambda_0 + \lambda_{\alpha 1} + \lambda_{\alpha 2} + \lambda_{\alpha \ldots} + \lambda_{\alpha 1 * \alpha 2} + \ldots)} \tag{1}$$

The left-hand side of the equation represents the probability *P* of an outcome *Y*, which is a person's *j* correct response (*1*) to item *i*, given that the person *j* belongs to a certain latent class *α* of item *i*. α_j represents a vector of mastered or non-mastered skills.

The right-hand side of the equation links *Y* in a probabilistic/log-linear fashion (*exp()*) to the skill pattern (α_j) that is required to solve item *i*. Thereby, an intercept coefficient *λ_0* and main effects *λ_α1* as well as interaction effects *λ_(α1 ∗ α2)* are defined and combined in an additive way (+). The intercept coefficient can be interpreted as the probability of solving an item correctly when none of the necessary skills is mastered. In this general framework, each mastered skill additionally contributes to the item response probability (main effects). Moreover, for each possible (second and higher order) interaction of mastered necessary skills, the probability of solving the item is raised additionally (interaction effects). For instance, if an item requires three skills α1, α2, and α3, three second order interactions (α1 ∗ α2, α1 ∗ α3, α2 ∗ α3) and a third order interaction (α1 ∗ α2 ∗ α3) are specified. In the literature and CDM package, this model is called "Reduced NC-RUM", if non-compensatory skills are assumed, or "GDINA", if compensatory skills are assumed. It represents the most complex CDM concerning the number of freely estimated parameters (skill class probability parameters as well as intercept, main effect, and interaction effect parameters for each item). To reduce model complexity, higher-order interactions might be constrained to 0. In other words, the GDINA model might be restricted such that only second-order interactions are allowed ("GDINA2").

In contrast to NC-RUM and GDINA models, the DINA model does not allow for main effects. Instead, for each item, only the intercept parameter and the highest order interaction term are estimated. While the intercept parameter is usually called the *guessing parameter* because it quantifies the probability of solving an item correctly when none of the necessary skills is mastered, the interaction parameter can be interpreted as the *slipping parameter* (i.e., the probability of solving an item incorrectly when all of the necessary skills are mastered). Equation 2 represents the item response function of the DINA model. Since the DINA model estimates only two parameters per item (guessing and slipping, which are restricted across skills) in addition to the skill class probability parameters, it represents the simplest CDM.

Equation 2:

$$P\left(Y_{ij} = 1 \,|\alpha_i\right) = \frac{exp(\lambda_0 + \lambda_{\alpha 1 * \alpha 2 * \alpha i})}{1 + exp(\lambda_0 + \lambda_{\alpha 1 * \alpha 2 * \alpha i})} \quad (2)$$

Rupp et al. (2010) categorize CDMs regarding their assumptions on (1) the scale type of item responses and the latent skills and (2) the compensability of latent skills. Concerning the scale types, CDMs may be grouped according to whether item responses are dichotomous or polytomous and whether the skills are assumed to represent dichotomous or polytomous latent variables. Concerning the second classification criterion, a model is characterized as compensatory if a particular skill (required for an item) can be compensated by another skill (required for the same item), leading to an item response probability (the probability of solving the item correctly) that is equal to the item response probability of a person who has mastered both skills. Hence, in compensatory models, the mastery of more than one of the required skills does not increase the probability of solving an item correctly. In a non-compensatory model, all skills that are required for solving an item must be mastered to obtain the maximum item response probability. If only one required skill is not mastered, the item response probability decreases to or near zero, equaling the value of a person who has mastered none of the required skills.

Constructing and validating CDMs for introductory accounting classes

Assessment goal

By developing CDMs as diagnostic tools in introductory accounting classes, we aim at delivering diagnostic information that offers insights into each student's conceptual understanding and operational proficiency. CDM-based assessments should equip teachers with fine-grained and valid information that can serve as a data-based decisional basis for designing appropriate instructional measures that foster subsequent learning processes. In line with the generic strategies of adaptive teaching set out in the first section, these assessments should reveal:

- each learner's particular skills and deficits in dealing with accounting tasks at different stages of his/her learning progress, which teachers may use to give elaborate and substantial feedback;
- sub-groups of learners with homogeneous or heterogeneous profiles of skills and deficits to facilitate ability grouping and differentiated task assignments;
- the total number of learners within given organizational entities (such as school classes) that can or cannot yet master distinct skills to set priorities for targeted exercises, further explanations, etc.

Competence model for solving typical tasks in introductory accounting classes

Review of reference models

Conceptual approaches to delineate cognitive operations that are to be mastered by learners in accounting largely rely on mathematical didactics. This is because the requirements for processing tasks in mathematics and accounting are largely comparable (Berding 2019, p. 90 et seq.). In both disciplines, it is necessary to encode and formalize reality, taking into account certain parameters, terminology, and procedures, and to transfer the meaning of the results obtained to real situations as well as to interpret them regarding specific effects (Berding 2019, p. 94; Phillips and Heiser 2011, p. 683 et seq.).

In the following, an approach by Seifried et al. (2010) will serve as a helpful initial categorization of mental operations and associated abilities that are needed to process task-inherent information in the domain of introductory accounting. Integrating mathematical and linguistic concepts, four phases for the cognitive processing of domain-specific tasks can be distinguished:

1. *Capturing economic reality:* recognizing economic issues of a real-world activity or occurrence, identifying and comprehending technical terms
2. *Encoding economic reality:* building or transforming verbal representations (case descriptions) and document-based representations
3. *Formalization and mathematics:* account assignment, posting, mathematical calculations
4. *Reflection and Assessment:* checking formal results and solution quality, interpreting results economically

The first phase refers to seizing key indicators to grasp the given economic situation, which necessitates interpretative abilities (Winther 2010). These abilities would be equally required for the encoding of the economic reality: for example, to be able to identify all relevant information on receipts or to assume certain roles or perspectives in a given situation, such as the role of the buyer of goods. Seifried et al.'s basic model does not (yet) deal deeply with the question if these logically separable phases require highly integrated cognitive processes for solving a given task effectively. They allocate, for example, a correct understanding of technical terms to the first phase and verbal representations (case descriptions) to the second phase. However, an understanding of economic terminology is of equal importance for being able to correctly interpret the context of verbally represented facts (e.g., "incoming and outgoing invoices") so that, for example, the perspective of the seller or buyer can be adopted correctly. Thus, it can be assumed that a suitable mental representation of economic reality relies on a thorough comprehension of its inherent domain-specific terms and details to such an extent that "Capturing" and "Encoding" empirically fall into one processing phase. This integrative phase could be referred to as the (mental) *modeling of the economic situation,* and its mastery would demonstrate a thorough conceptual understanding of the flow of goods and money as well as the participating entities that constitute a given task and affect the domain of accounting.

The transition between the mental modeling of economic facts and processes on the one side and their depiction in the formal accounting system on the other takes place in the third phase. Here, it is indispensable to formalize and mathematize the economically relevant information following the specialist logic of accounting from a particular role or perspective, which requires mainly operational proficiency. The third phase can thus be clearly distinguished from the previous phases due to this shift between reference systems. According to empirical findings on necessary cognitive operations in the creation of accounting records, "Formalization & mathematization" involves a total of three distinct processing steps (Helm 2016, p. 38 et seq.; Phillips and Heiser 2011, p. 684 et seq.). These relate to the selection of appropriate *accounts,* the decision on *account sides* (debit/credit) that must be addressed, and the *calculation of the amount* that must be posted. This separation is also theoretically plausible since the three processing steps correspond to three subtasks to be completed when registering any transaction or business event. Performing these steps requires specific knowledge about the function of each account, the accounting record logic, and the required mathematical operations. Therefore, it can be argued that "Formalization & mathematization" consists of at least three distinct cognitive operations and is primarily associated with operational proficiency in the accounting domain.

In the last phase, it is important to critically reflect on one's task solution regarding the chosen path and outcome but also to assess the produced formalized and mathematical information regarding their economic significance and impact. However, the present designation of this phase mingles processes from different spheres of mental activity. The *examination of one's task solution (path and outcome)* is *not necessarily* part of a competence model delineating essential cognitive operations for solving introductory accounting tasks since these steps represent *meta-cognitions.* According to Marzano and Kendall (2007, p. 53 et seq.), these steps serve to monitor, evaluate, and regulate (any) cognitive processes, which is why meta-cognitions are superordinate to task-specific mental operations. Nevertheless, the *economic interpretation of the produced formalized and mathematized information* is a cognitive task that immediately deals with purely domain-specific concepts as it crosses the lines between the accounting system and the economic reality once again. Analogous to the third phase, a further domain-specific processing step can thus be established. In tasks of introductory accounting, this step essentially relates to the assessment of the economic profit-/loss- effects resulting from the accounting records, which, like the first step, requires primarily conceptual understanding.

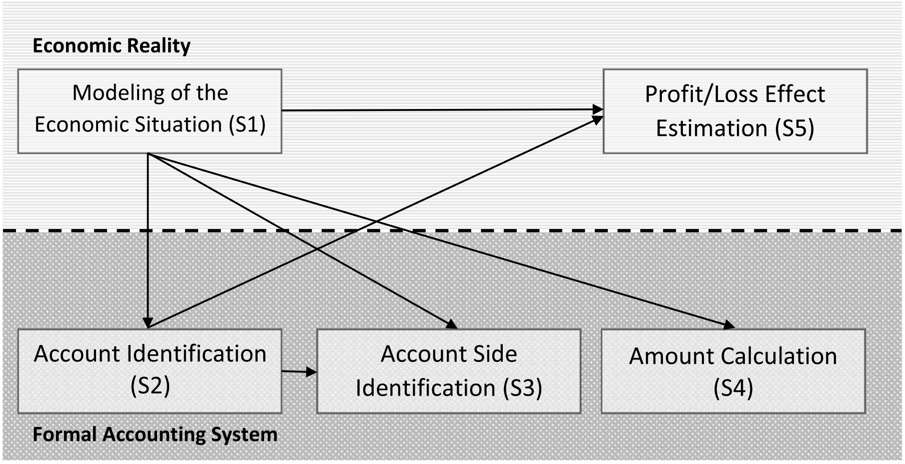
Economic Reality
Modeling of the Economic Situation (S1)
Profit/Loss Effect Estimation (S5)
Account Identification (S2)
Account Side Identification (S3)
Amount Calculation (S4)
Formal Accounting System

Fig. 2 Competence Model depicting cognitive skills (S1-S5) required for processing tasks in the domain of introductory accounting

Structural elements and basic assumptions of the proposed competence model

In consequence of the above considerations, five steps of cognitive processing that are necessary to solve tasks in the curricular field of introductory accounting can be distinguished. Their execution requires particular skills, which comprise Modeling the economic situation (Skill S1), Account Identification (S2), Account Side Identification (S3), Amount calculation (S4), and Profit/Loss Effect Estimation (S5). The relationships between these distinct but closely connected processing steps are depicted in a cognitive processing model (Fig. 2).

Processing steps on the economic reality and those on the formal accounting system are located in two distinct areas. The related skills are performed in sequential order (1–5) according to the procedure of task accomplishment. Furthermore, these skills cannot compensate each other. This means that once a skill, which is required for a particular task, is not mastered, the probability of completing the task correctly is greatly reduced, as shown in the following.

Relations among the models' structural elements

Following Seifried et al. (2010) and Phillips and Heiser (2011), the modeling of the economic situation (S1) provides the starting point for identifying the affected accounts (S2), choosing the right account sides (debit/credit) (S3), and making the correct calculations of relevant amounts (S4). The ability to model the economic situation (S1) is indispensable for translating flows of goods and money into the notations and procedures of accounting, following task requirements. Thus, it is a central skill that most obviously cannot be substituted by other skills located on the formal accounting level. Furthermore, S1 forms the basis for correctly estimating profit/loss effects (S5) that result from certain economic transactions and occurrences. Consequently, starting from S1 to all further processing steps, a directed connection is mapped. According to the non-compensability of the skills, the outgoing arrows imply that the subsequent processing steps are less likely to be mastered if S1 is lacking. In other words, cognitive operations of formalizing and mathematizing within the accounting-system logic are prone to error if there is little or no economic understanding, and they cannot offset economic misconceptions. Moreover, a critical, conclusive reflection of the accounting records would only be possible to a limited extent.

Within the area of the formal accounting system, only outgoing relationships have been considered for the correct selection of the affected accounts (S2), since S2 is the prerequisite for estimating the profit/loss impact of economic transactions (S5) based on accounting. To this end, the ability to correctly identify relevant account sides in terms of debit/credit (S3) plays a subordinate role, since the estimation of profit-/loss- effects only requires knowing whether the affected account increases or decreases due to the economic situation (S1), but not which account side (debit/credit) is affected. The ability to calculate relevant amounts (S4) is also of secondary importance for impact assessments, as the monetary value does not affect the nature of a transaction, such as paying a received invoice, but only the extent of its impact on business success. For the skills S3 and S4, no directed connections to S5 are considered. Further, S2 and S3 show a unidirectional relationship. This is because the identification of particular accounts (S2) is the basis for determining which account sides in terms of debit or credit (S3) must be addressed due to a given economic situation. As explained above, S1 would also be relevant here because, without this step, one cannot decide whether an account is increasing or decreasing. However, no directional relationship is assumed between S3 and S4 because S4 refers to capabilities of mathematizing economic issues as modeled in S1, which can be done independently of S3.

Construction of items to assess basic skills acquired during introductory accounting instruction

According to the assessment triangle, the assessment tasks should trigger skills that are described in the domain competence model (cognition corner; see Fig. 1). We therefore screened tasks in two common accounting textbooks that are certified based on the Austrian national competence-centered curriculum (BMUKK 2014) as well as the Austrian vocational education standards (http://www.bildungsstandards.berufsbildendeschulen.at) for their compatibility with the competence model described in Competence model for solving typical tasks in introductory accounting classes. Thus, from the range of available text books, we have only selected those that correspond to the national competence orientation framework, i.e., those that have been examined by domain experts and found to be in line with the national educational standards, while we have excluded all other textbooks. Next, we inspected the tasks in the selected textbooks ourselves and chose tasks that cover the dimensions of our cognitive diagnosis model in varying combinations. That way, we assembled textbook assignments that required all or particular sets of the postulated skills. For example, we purposefully collected items that do and don’t require S1 (“Modeling”), the latter items presenting the task not as an economic-situation stimulus but in an abstract and formalized wording. For instance, tasks that merely prompt students to identify the correct account side for the cash account (e.g., increase of revenues) fall into this latter category. This also applies to questions on which side the opening stock of a supplier account is kept and which account sides are involved when an expense or revenue arises. The number of selected items was limited by test time, which was one lesson (45 min). In total, 42 items were selected.

In a further step, we asked subject teachers to check whether the assembled test booklet is following the curriculum implemented at their school. They did not recommend any changes or amendments. However, statistical item analyses at a later time point showed that one category of items (8 tasks that asked for building a balance sheet) had to be removed due to low or even negative item discrimination. The final test booklet thus consisted of 34 items (tasks).

Specification, validation, and completeness check of the Q-matrix

In line with the approach described by Li and Suen (2013), an initial Q-matrix was constructed by coding each item (task) based on which skills (specified in the multidimensional competence model) are required to solve it. Drawing on the results from thinking-aloud studies (lead Helm 2016) and the researchers' expert knowledge of teaching and learning theories, the cognitive processes students should engage in when effectively solving the selected tasks were anticipated. In a stepwise procedure, we mapped an initial Q-matrix, which was then subjected to both theoretical and statistical validation checks. From a theoretical perspective, the initial Q-matrix was discussed with two domain experts, which led to the version presented in Table 1.

The Q-matrix reveals that the selected item pool from approved textbooks lacks tasks that require only one skill. Since such single-skill items are a prerequisite for the CDMs called DINA (non-compensatory skills assumed) and DINO (compensatory skills assumed), these two models were excluded from further analyses. For more complex CDMs like the NC-RUM and the C-RUM, we adopted the analytic procedure for testing model completeness proposed by Köhn and Chiu (2018), which yielded positive results.[1] Consequently, the selection of the most suited CDM outlined in the next section considers all models except DINO and DINA.

Selection of a psychometric model from the family of cognitive diagnosis models

Model selection should encompass multiple criteria (Dibello et al. 2007). We comply with this request by considering:

a) the conformity of different CDM specifications with the assumptions of the competence model,
b) statistical measures of model fit and classification reliability, indicating model alignment with the empirical data structure and allowing to test various hypotheses about skill relations and task-skill assignments,
c) the legitimacy and interpretability of CDM-output information, for instance concerning extant scientific evidence about domain-relevant competencies of the assessed learner group.

Before reporting the selection procedure and criteria in detail, we briefly outline the assessment context and sample.

[1] We thank Hans-Friedrich Köhn (University of Illinois) for determining the completeness of our Q-matrix.

Table 1 Q-Matrix used for cognitive diagnosis in introductory accounting

Item	Skill 1	Skill 2	Skill 3	Skill 4	Skill 5
Cost of sales (AR)	1	1	1	1	0
Return of goods (AR)	1	1	1	1	0
Rental expense (AR)	1	1	1	1	0
Cash contribution (AR)	1	1	1	0	0
Interest income (AR)	1	1	1	0	0
Debt conversion (AR)	1	1	1	1	0
Loans (AR)	1	1	1	0	0
Shipping charges (AR)	1	1	1	1	0
Internal consumption (AR)	1	1	1	1	0
Outgoing invoice (AR)	1	1	1	0	0
Customer discount (AR)	1	1	1	1	0
Cost of fuel (AR)	1	1	1	1	0
Cost of cleaning material (AR)	1	1	1	0	0
Initial balance of suppliers account	0	1	1	0	0
Income adjustments	1	1	1	0	0
Increase in cash account	0	1	1	0	0
Formation of expenses	0	0	1	0	0
Formation of income	0	0	1	0	0
Decrease in trade payables	1	1	1	0	0
Formation of cash discount	1	1	1	0	0
Closing balance of liability accounts	1	0	1	0	0
Cost of sales (P/L)	1	1	0	0	1
Return of goods (P/L)	1	1	0	0	1
Rental expense (P/L)	1	1	0	0	1
Cash contribution (P/L)	1	1	0	0	1
Interest income (P/L)	1	0	0	0	1
Debt conversion (P/L)	1	1	0	0	1
Loans (P/L)	1	1	0	0	1
Shipping charges (P/L)	1	1	0	0	1
Internal consumption (P/L)	1	1	0	0	1
Outgoing invoice (P/L)	1	1	0	0	1
Customer discount (P/L)	1	1	0	0	1
Cost of fuel (P/L)	1	1	0	0	1
Cost of cleaning material (P/L)	1	1	0	0	1

AR Accounting record, *P/L* Profit/loss effect estimation. *Skill 1* Modeling, *Skill 2* Account identification, *Skill 3* Account side identification, *Skill 4* Amount calculation, *Skill 5* Profit/loss effect estimation

Sample characteristics

We collected assessment data in 31 upper secondary vocational education classes in Austria (Helm 2016). 773 students (age: M = 14.5 years, SD = 9 months; 75% female) voluntarily took part in the assessment, which was carried out during accounting lessons at the end of Grade 9 (i.e., the first year of introductory accounting). The official national curriculum sets accounting records and the double entry method as central educational goals for Grade 9. Hence, at the time of assessment students already had sufficient opportunities to learn the assessed concepts.

Model selection based on the structure of skills and scale type of observed and latent variables

In line with Rupp et al. (2010, p. 98), we select CDMs that align with (1) the theoretically proposed structure of skills (i.e., skill hierarchy and skill compensation) and (2) the assumed scale type of the observed variables (student responses) as well as the latent variables (skills).

(1) The technical justifications for the domain competence model described in Competence model for solving typical tasks in introductory accounting classes characterize five skills for effectively solving tasks from introductory accounting curricula, which are supposed to be *non-compensatory*. Consequently, a student's lack of individual skills presumably reduces his/her chances to produce correct task solutions markedly. Students cannot easily offset, for example, deficient knowledge about how to calculate the exact amount of money to settle an invoice (skill 4) with knowledge about how this invoice settlement in principle affects a company's profit situation (skill 5). It has to be noted, however, that solution possibilities might not be reduced to zero because of two reasons: (1) Skill acquisition in accounting is often classified as the "rote learning" of standard procedures of bookkeeping (Muldoon et al. 2007), hence students may give a correct answer though they lack in-depth conceptual understanding. (2) Guessing: Different types of tasks such as the request to select the correct account side represent multiple-choice items with only 2 options.

(2) As further argued in Competence model for solving typical tasks in introductory accounting classes, we assume the following *skill hierarchy*: The mastery of skill 1 is a requirement for all other skills. Moreover, skill 2 precedes skill 3 and skill 5.

(3) Regarding the potential scale types of student responses, we chose dichotomous coding. It must be stressed that a 'response' in our test environment does not refer to varying solution qualities in highly complex tasks that could be solved more or less completely, accurately, efficiently, etc. From a domain expert's view, solutions to very narrowly defined assessment items on an *introductory level* of accounting education (standard and clearly delineated business transactions such as settling an invoice) are either right or wrong. Moreover, the test was purposefully constructed to include many tasks that merely *require specific skills* (in varying combinations) from our multidimensional model of accounting competence in order to draw safe conclusions from solutions of 'focus-tasks' to 'focus-skill' mastery (e.g., tasks that merely ask for the accounts that have to be addressed OR for the monetary sum that has to be transferred when settling an invoice). Given the pedagogical context and the intention to support learning, response coding aimed to locate exactly in which cognitive steps of processing accounting tasks a student repeatedly succeeds or fails (e.g., identifying the appropriate accounts), not on evaluating overall solution quality of complete tasks in many graduations as it would be done for summative performance assessments.

Following the above considerations and the completeness checks reported in Section CDMs of the NC-RUM type (with or without higher-order interactions) are suitable for the present diagnostic utilization and deserve closer inspection.

Model selection based on statistical evaluation criteria and tests of hypotheses inherent in the competence model and the Q-matrix

Measures of model fit and classification reliability serve to evaluate the quality and appropriateness of the selected model even further, that is, its accuracy of predicting students' response data. Included in these statistical examinations are checks of the empirical support for the theoretically justified skill relations and task-skill assignments.

Analytical procedure

After coding all student responses wrong or right (0/1), model-fit tests were carried out in 3 steps. In the first step, we examined the fit indices of the most complex possible model (NC-RUM). This step further includes testing the appropriateness of the assumed task-to-skill assignments (as defined in the Q-matrix). Model fit is evaluated using test-level (global) measures (MADcor, SRMSR, MADQ3, MADaQ3, maxχ^2 *p*-value) as well as measures on the item level (RMSEA, IDI). In the following, we briefly describe how these indices are determined, that is, how they work.

MADcor—Mean of absolute deviations in observed and expected correlations (e.g., DiBello et al. 2007). This index denotes the average absolute deviation between observed correlations (r_{ij}) and model predicted correlations ($\hat{}r_{ij}$) of item pairs ($_{i,j}$).

SRMSR—Standardized mean square root of squared residuals (e.g., Maydeu-Olivares 2013). Like the MADcor, the SRMSR is also based on comparing the observed and expected correlations of item pairs. While MADcor is based on absolute deviations SRMSR is based on squared deviations.

MADQ3 and MADaQ3—Mean of absolute values of pairwise correlations of residuals (e.g., Yen 1984). For calculating MADQ3 and MADaQ3, residuals of observed and expected responses for respondents $_n$ and items $_i$ are constructed. Then, the average of the absolute values of pairwise correlations of these residuals is computed for MADQ3. For MADaQ3, the average of the centered pairwise values (i.e., by subtracting the average Q3 statistic) is calculated.

maxχ^2 *p* value – Max Test of global absolute model fit using test statistics of all item pairs (Groß et al. 2016). The statistic max(χ^2) is the maximum of all χ^2_{ij} statistics accompanied with a *p*-value obtained by the Holm procedure.

IDI – Item discrimination index. The IDI reflects an item's ability to separate respondents who have mastered all measured attributes from respondents who have not mastered any of the attributes. According to Lee et al. (2012) IDI values larger than or equal to 0.35 indicate adequate fit.

RMSEA – root mean square error of approximation. RMSEA "essentially compares the model-predicted item response probabilities for a correct response for respondents in different latent classes with the observed proportions of correct responses by the responses weighted by the proportion of respondents in each latent class" (Kunina-Habenicht et al. 2009, p. 67).

For every of the listed fit statistics (except the IDI) it holds that smaller values (values near zero) indicate a better fit (Robitzsch et al. 2019). According to pertinent literature, the values presented in Table 2 attest to the model's adequacy.

Table 2 Central test-level and item-level fit indices used in CDM research

Fit measure	Values indicating adequate fit	References
Test-level		
MADcor	≤ 0.05	DiBello et al. (2007)
SRMSR	≤ 0.05	Maydeu-Olivares (2013)
MADQ3/MADaQ3	≤ 0.20 or .10	Yen (1984)
max χ^2 value	≥ 14.87[a]	Groß, Robitzsch, and George (2016)
Item-level		
RMSEA	≤ 0.10	Oliveri and Davier (2011)
IDI	approx. ≥ 0.35	Lee et al. (2012)
accuracy	≥ 0.70	Nunnally and Bernstein (1994)
consistency	≥ 0.70	Nunnally and Bernstein (1994)

MADcor Average absolute deviation between observed correlations and model predicted correlations, *SRMR* Standardized root mean square root of squared residuals, MADQ3 = mean of absolute values of Q3 statistic (Yen 1984), *MADaQ3* Mean of absolute values of centered Q3 statistic, *RMSEA* Root mean square error of approximation, IDI = item discrimination index

[a] Groß et al. (2015) propose a chi square value with a bonferroni corrected significance level ($\alpha^a = \alpha/((J(J-1))/2)$; α^a = bonferroni corrected significance level, α = original/desired significance level, J number of items) which equals 14.87 for 30 item

Table 3 Relative model fit indices

Model	# Par	N	loglike	Deviance	AIC	BIC	AIC3	AICc	CAIC	MADRC (SE)
NC-RUM34	324	773	− 13835.8	27671.7	28319.7	29826.4	28643.7	28789.8	30150.4	0.404 (0.050)
NC-RUM30	300	773	− 12069.4	24138.7	24738.7	26133.8	25038.7	25121.4	26433.8	0.591 (0.059)

NC-RUM30/34 non-compensatory reparameterized unified model with 30/34 tasks, *# Pa* Number of parameters, *N* Sample size, *loglike* log-likelihood, *AIC* Akaike information criterion, *BIC* Bayesian information criterion, *AIC3* AIC with # Par * 3, *AICc* AIC corrected, *CAIC* Consistent AIC, *MADRC* Mean absolute deviation residual covariance, *SE* Standard errors

In the second step, classification reliability serves to evaluate the quality of CDMs. Related literature offers two different measures of classification reliability on both skill profile level and skill level. Classification *consistency* refers to "the proportion of examinees that are classified into the same category (AMP) on parallel replications of the same test" (Wang et al. 2015, p. 462). Classification *accuracy* provides a measure of "the percentage of agreement between the observed and expected proportions of examinees in each of the attributes or AMPs under the CDA framework" (Wang et al. 2015, p. 460).

In the third step, we examined whether more restrictive models fit the data equally well. This step further allows testing hypotheses on the *non-compensatory* nature (H1) and *hierarchy* (H2) of skills that are required for solving assessment items in the domain of introductory accounting. These assumptions are tested via model comparisons using relative measurements of model fit (AIC, BIC, and derivatives thereof) as well as chi-square difference testing as implemented in the anova() command in R.

All analyses are done using the CDM package in R (George et al. 2016; Robitzsch et al. 2019).

Step 1: Examination of model fit indices. The inspection of statistical evaluation criteria starts with the most complex model, the NC-RUM – hereafter referred to as NC-RUM34, indicating the 34-item test length. The first lines in Tables 3 and 4 comprise the relative and absolute model fit indices at the test level after fitting an NC-RUM to the test data.

Table 4 Absolute model fit indices

Model	# Par	N	max χ^2	p max χ^2	MADcor (SE)	SRMSR (SE)	MADQ3 (SE)	MADaQ3 (SE)
NC-RUM34	324	773	68.240	0.001	0.018 (0.003)	0.030 (0.004)	0.022 (0.004)	0.023 (0.004)
NC-RUM30	288	773	46.503	0.001	0.029 (0.003)	0.046 (0.004)	0.039 (0.005)	0.040 (0.005)

NC-RUM30/34 Non-compensatory reparameterized unified model with 30/34 tasks. χ^2 Chi-square, *p* significance, *MADcor* Average absolute deviation between observed correlations and model predicted correlations, *SRMR* Standardized root mean square root of squared residuals, *MADQ3* Mean of absolute values of Q3 statistic (Yen 1984), *MADaQ3* Mean of absolute values of centered Q3 statistic, *SE* Standard errors

Table 5 Item pairs with contribution to model misfit

Item pair		χ^2
Debt conversion (AR)	Debt conversion (p/l)	46.503
Outgoing invoice (AR)	Outgoing invoice (p/l)	43.353
Customer discount (AR)	Customer discount (p/l)	41.366
Cash contribution (AR)	Cash contribution (p/l)	38.489
Loans (AR)	Loans (p/l)	32.358
Interest income (AR)	Interest income (p/l)	24.063
Internal consumption (AR)	Internal consumption (p/l)	21.619
Truck diesel (P/L)	Cleaning material (p/l)	19.859

χ^2 Chi square values above threshold of 15. *AR* accounting record, *P/L* Profit/Loss effect estimation

All absolute fit estimates are below the cut-off limits as described in Table 2 (MADcor 0.018 < 0.050; SRMSR 0.030 < 0.050; MADQ3 0.022 < 0.100; MADaQ3 0.023 < 0.100) with one exception: The maximum χ^2 value of the pairwise item comparisons appears to be statistically significant ($\chi^2 = 68.240$, $p < 0.001$).

Hence, we scrutinized those item pairs that showed a significant association – even after controlling for the five skills that are supposed to explain these associations (see Q-matrix in Table 1) – more closely. Those items that contributed most strongly to model misfit were excluded. Only one item was omitted at a time, namely the four items "profit/loss estimation for shipping charges," "formation of cash discount," "initial balance of suppliers account," and "increase in the cash account."

After excluding these four items, eight item pairs still showed significant misfits. However, this is most likely due to the common item stem of each pair, as each of the two items in an item pair refers to the same business case. Hence, the model misfit as indicated by the conservative max χ^2 criterion can be explained by the fact that the assessment comprises items that are independently solvable but have the same item stem in common.

Lines two of Tables 3 and 4 comprise the model fit criteria for the NC-RUM30 after deleting the four above-mentioned items. Although the relative model fit measures point to superior model fit compared to NC-RUM34, absolute model fit indices are slightly dropping but still below the cut-off values. Hence, for both models, NC-RUM34 and NC-RUM30, model evaluation *at the test level* indicate a good model fit.

In addition, model fit evaluation *at the item-level* points to good psychometric properties of the items. Table 5 lists those item pairs that contribute most to the model misfit. As documented in Table 6, the RMSEA values of all items are below or around the

Table 6 Item-level fit measures

	Min	Max	Mean	SD
NC-RUM34				
RMSEA	0.031	0.106	0.065	0.017
IDI	0.045	1.000	0.709	0.260
NC-RUM30				
RMSEA	0.019	0.107	0.065	0.022
IDI	0.038	1.000	0.737	0.237

NC-RUM30/34 Non-compensatory reparameterized unified model with 30/34 tasks. *RMSEA* Root mean square error of approximation, *IDI* Item discrimination index, *SD* Standard deviation

Table 7 Classification reliability of NC-RUM30

	Skill pattern	Skill 1: modeling	Skill 2: account	Skill 3: account side	Skill 4: amount calculation	Skill 5: profit/ loss
Accuracy	0.841	0.962	0.970	0.964	0.962	0.932
Consistency	0.740	0.954	0.967	0.954	0.955	0.901

MLE values (Cui et al. 2012; Wang et al. 2015; George et al. 2019, p. 95): "Note. The first column gives the probability of classifying a randomly selected student consistently. The other rows contain the same measure, but focused on the classification of the selected student in only one skill."

cut-off value. Moreover, for most items, the item discrimination index (IDI) values are above the desired value of 0.35. Only in 4 out of the 34 items (NC-RUM34) and in 1 out of 30 items (NC-RUM30) IDI is critical. On average, the test items strongly discriminate among students with different skill profiles (mean IDI = 0.71/0.74).

Step 2: Examination of classification reliability. To evaluate the reliability and validity of classification results produced by cognitive diagnostic assessments, Cui et al. (2012) developed two measures: classification consistency and classification accuracy. *Classification consistency* (often referred to as the reliability of classifications) states the probability of an examinee being classified into the same category on two test occasions (Cui et al. 2012). *Classification accuracy* examines the degree to which classifications based on observed scores match those based on true scores. Table 7 shows that all measures of classification reliability on both skill profile level and skill level are larger than the rule of thumb value of 0.7 (e.g., George et al. 2019). Hence, the assessment is highly accurate and consistent in classifying examinees into skill profiles and distinguishing those with mastered skills from those with non-mastered skills.

Taken together, the reported model evaluations corroborate that students' item scoring can be reproduced well enough by the NC-RUM and the specified Q-matrix. The model fit indices at the test and item level as well as the reliability measures support the assumptions we made when constructing the Q-matrix.

Step 3: Examination of the non-compensatory and skill-hierarchy hypotheses. When testing the assumed non-compensatory nature (H1) and hierarchical arrangement (H2) of basic accounting skills, NC-RUM30 provides the baseline model since all relative measures of model fit favor this model over the NC-RUM34 (see Table 3). Model comparisons based on relative model fit and χ^2 difference testing (see Table 8) indicate that—com-

Table 8 Hypotheses testing

Model		loglike	# Par	AIC	BIC	χ^2	df	*p*
	NC-RUM30	− 12069.4	300	24738.7	26133.8			
H1	C-RUM30	− 12468.6	138	25213.1	25854.9	798.4	162	< 0.001
H2	HIR-NC-RUM30	− 12512.9	300	25625.7	27020.8	887.0	0	–

H1/2 Hypothesis 1/2, *NC-RUM30* Non-compensatory reparameterized unified model with 30 tasks, *C-RUM30* Compensatory reparameterized unified model with 30 tasks, *HIR-NC-RUM30* Hierarchical non-compensatory reparameterized unified model with 30 tasks, *loglike* Log-likelihood, *# Par* Number of parameters, *AIC* Akaike information criterion, *BIC* Bayesian information criterion, χ^2 Chi-square, *df* Degrees of freedom, *p* Significance level

Table 9 Skill correlations

	Modeling	Accounts	Account side	Amount calculation	Profit/loss
Modeling	1.000	− 0.071	0.304	0.319	− 0.030
Accounts		1.000	0.058	0.130	0.220
Account side			1.000	0.070	− 0.108
Amount calculation				1.000	0.423
Profit/loss					1.000

pared to the baseline model (NC-RUM30)—all alternative models lead to a substantive drop in model fit as indicated by their significantly higher χ^2 values.

Regarding hypothesis 1, model comparison confirms that the investigated skills for introductory accounting tasks cannot substitute each other. To obtain the maximum item response probability, all skills that are required for solving an item (according to the Q-matrix) must be mastered. If only one required skill is not mastered, the item response probability substantially decreases. From a statistical point of view, the model fit significantly drops ($\chi^2 = 798.4$, $df = 162$, $p < 0.001$) when assuming a C-RUM model (main effects only) instead of an NC-RUM (main effects and interaction effects). This indicates that second (and higher) order interaction effects play an important role to predict students' response data. (However, this finding is by no means unambiguous, as the BIC improves).

To test hypothesis 2, a new model called HIR-NC-RUM30 is estimated. In contrast to the baseline model NC-RUM30, the Q-matrix of HIR-NC-RUM30 is specified in a way that hierarchy among the skills is assumed. That is, skill 1 is needed for all items since we assume that mastery of skill 1 is a requirement for all other skills. Moreover, as skill 2 precedes skill 3 and skill 5, the column "skill 2" of the Q-matrix of HIR-NC-RUM30 will always have entry 1, whenever the columns "skill 3" and "skill 5" have entry 1. Testing HIR-NC-RUM30 (with the new Q-matrix) against the baseline model NC-RUM30, shows, however, that the relative model misfit indices for non-nested models significantly increase (AIC: 24,738.7 vs. 25,625.7, BIC: 26,133.8 vs 27,020.8, χ^2 887.0). Accordingly, Table 9 shows that the five skills are not, or at best, moderately correlated with each other.

CDM output evaluation against the assessment context

From a domain expert point of view, finding the best fitting CDM to contradict the assumption of a *hierarchical* order of the investigated accounting skills might at first be surprising since competence models of disciplines like math and accounting postulate that modeling real-world situations represents a key initial step in problem-solving processes (see e.g., Artigue et al. 2007; Minnameier 2013). However, the missing *strong* links between "modeling" and all other skills are rather plausible against the background of the *instructional conditions* and the comparably low complexity of economic issues in *introductory* accounting education as opposed to advanced classes. Accounting education in Austrian as well as German schools have recurrently been criticized for concentrating rather mechanically on bookkeeping rules while neglecting the underlying economic processes (cf. lead Helm 2016; Bouley 2017; Reinisch, 1996; Seifried 2004). Instructional designs are considered to concentrate on routine applications of formal accounting operations, often in a drill-and-practice approach (Bouley 2017, p. 14), whereas insights into the functions and meaningfulness of accounting for business processes may only be gained at the end of a course using illustrative examples (Seifried 2004, p. 23). Thus, concerns arise that instructional processes might not prepare students optimally for cognitive steps like modeling business processes and events, which they have to undertake when dealing alone with tasks that align with approved domain standards (Berding et al. 2020; for similar discussions in the international literature see, for example, Bloemhof and Christensen Hughes 2013; Muldoon et al. 2007). However, while students' attempts to solve accounting tasks in a stimulus–response manner – drawing on standard bookkeeping records they learned by heart – certainly fall short when business processes or events are complex in nature, they might be sufficient for the processes and events that are dealt with in lessons (and related tasks) of *introductory* accounting. In other words: At least when dealing with tasks on the complexity level of introductory accounting, some signaling words like "sales costs" or "internal consumption" may be sufficient to trigger the reproduction of common, memorized procedures of bookkeeping and calculation, thereby producing correct accounting records (operational proficiency) without a *thorough* conceptual understanding of their economic significance. It must be stressed again that the tasks presented in the CDM-assessment do require cognitive processes of modeling, and these are positively related to the subsequent cognitive processes, but given the introductory level within the accounting curriculum, the economic issues are clearly delineated and the relational strength between modeling skills and others is weak.

In the following, we elucidate essential elements of CDM output by presenting exemplary findings from the investigated accounting classes in Austrian vocational schools. The closing section will revive these examples to illustrate how they might facilitate a teacher's attempts to deliver adaptive teaching.

Skill profiles

Skill profiles can be obtained for each student separately, revealing his/her strengths and weaknesses when dealing with introductory accounting tasks. The X-axis of Fig. 3 shows the student's probability of mastery and non-mastery of the investigated skills, which themselves are displayed on the Y-axis.

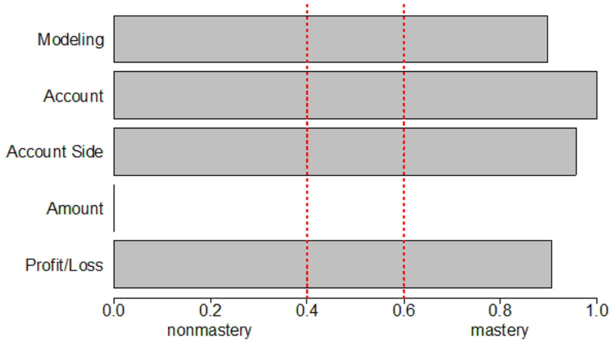

Fig. 3 Student with response pattern [000010111101101010110101111011]

More precisely, Fig. 3 shows the skill mastery profile of a student with the *item response pattern* [000010111101101010110101111011], indicating which of the 30 assessment tasks were solved (1) or not solved correctly (0). According to the task-skill-assignments of the Q-matrix, the student's skill profile reveals a consistent pattern of deficits in performing required calculations (skill Amount Calculation), whereas he/she mastered all other skills. Expressed in the sequence of the five skills defined in the competence model (see Fig. 2) and on the Y-axis, this *individual skill profile* is labeled 11101. Based on the individual skill profile and the individual total score, a differentiated picture of student competencies emerges. The profiles and total scores provide differentiated information for the teacher, e.g. whether someone has not mastered only one partial skill or almost all of them. Still, as with *any statistical* model, the parameter estimates for CDMs are subject to uncertainty. In our CDM output, dashed lines at 0.4 and 0.6 of the X-axis indicate a "region of uncertainty" about a student's mastery status. If a student falls within the region of uncertainty, the teacher could gather further information, for example by involving him/her targetedly in dialogic interactions to probe into the student's domain-specific concepts and operations (see chapter 1), or simply by having the student solve further tasks that require exactly the skills for which skill mastery is unclear. In contrast to high-stakes evaluative assessments that form the basis of giving grades and therefore must come to an unambiguous classification, it is well advisable in learning contexts, aiming at competence development, to assume that those learners who have not yet demonstrated skill mastery stably over diverse tasks (still) deserve support.

Skill class distribution in a given school class

Table 10 shows the skill class distribution for a given school class, that is, how often certain *patterns* of mastered and non-mastered skills occur among its members. Hence, each section of the table pools individual skill profiles that are identical. In other words, the different skill classes represent empirically distinguishable subgroups of students from this specific school class that differ in the particular *combination* of strengths and weaknesses that characterizes their skill profiles. The school class displayed in Table 10 includes 20 students. Their teacher receives the following diagnostic information:

None of the students mastered all skills.

Four students mastered all but the first skill (modeling).

Table 10 Skill class distribution for NC-RUM30 for a given school class

	Modeling	Account	Account side	Amount calculation	Profit/ Loss	Skill Profile	N Skills mastered
Skill class 1: profile 01111							
Student A	NO	YES	YES	YES	YES	01111	4
Student B	NO	YES	YES	YES	YES	01111	4
Student C	NO	YES	YES	YES	YES	01111	4
Student D	NO	YES	YES	YES	YES	01111	4
Skill class 2: profile 11011							
Student E	YES	YES	NO	YES	YES	11011	4
Skill class 3: profile 01011							
Student F	NO	YES	NO	YES	YES	01011	3
Student G	NO	YES	NO	YES	YES	01011	3
Student H	NO	YES	NO	YES	YES	01011	3
Student I	NO	YES	NO	YES	YES	01011	3
Skill class 4: profile 01101							
Student J	NO	YES	YES	NO	YES	01101	3
Student K	NO	YES	YES	NO	YES	01101	3
Student L	NO	YES	YES	NO	YES	01101	3
Skill class 5: profile 01110							
Student M	NO	YES	YES	YES	NO	01110	3
Student N	NO	YES	YES	YES	NO	01110	3
Student O	NO	YES	YES	YES	NO	01110	3
Skill class 6: profile 01001							
Student P	NO	YES	NO	NO	YES	01001	2
Student Q	NO	YES	NO	NO	YES	01001	2
Skill class 7: profile 00011							
Student R	NO	NO	NO	YES	YES	00011	2
Skill class 8: profile 01010							
Student S	NO	YES	NO	YES	NO	01010	2
Skill class 9: profile 01000							
Student T	NO	YES	NO	NO	NO	01000	1

One student mastered all but the third skill (account side).

Four students mastered all but the first (modeling) and the third skill (account side).

Three students mastered all but the first (modeling) and the fourth skill (amount calculation).

Three students mastered all but the first (modeling) and the fifth skill (profit/loss).

Two students mastered only the second (account) and the fifth skill (profit/loss).

One student mastered only the fourth (amount calculation) and the fifth skill (profit/loss).

One student mastered only the second (account) and the fourth skill (amount calculation).

One student mastered only the second skill (account).

For decisions on whether and how to adapt instructional strategies, the descriptive distribution of different *skill classes* delivers comparatively detailed information on the different domain-specific strengths and weaknesses of distinct subgroups of students as

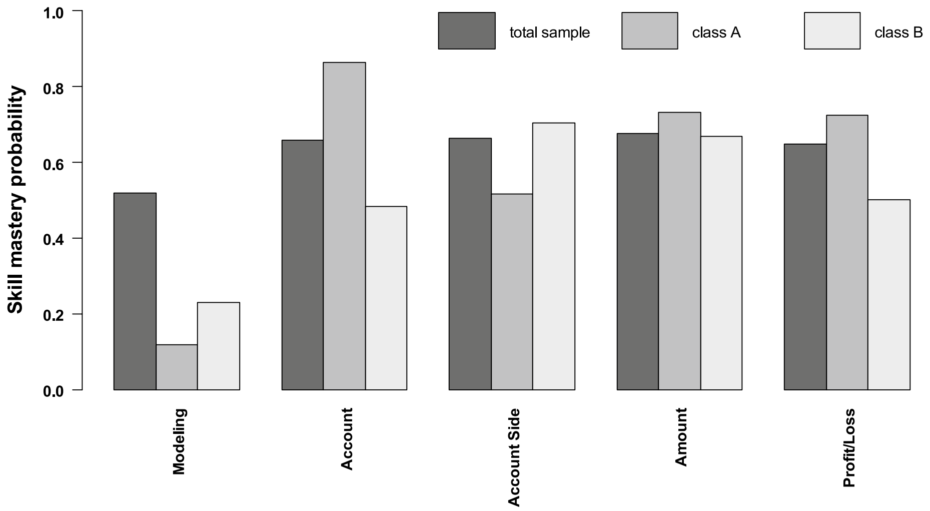

Fig. 4 Skill mastery in the entire sample and distinct school classes

well as on their prevalence: While a teacher might otherwise base these decisions on his/her vague impression that most students struggle more or less in dealing with tasks in their textbooks, he/she is informed, in the present case, that one-third of students have narrowly defined deficits (as they have mastered all skills except one), whereas about half of the students lack multiple skills. Moreover, he/she can locate the deficits of learners in a particular skill class within the domain-specific competence model (Fig. 1). The teacher may find, for example, one group to lack primarily conceptual understanding of the economic issues that should be recognized and interpreted, while another group predominantly lacks operational proficiency as they fail to translate these issues into the notations and procedures of the accounting system.

Skill mastery in the entire sample and in distinct school classes

To provide teachers with diagnostic information on how the focal school class performs in comparison with other classes or the entire sample that underwent the same diagnostic assessment, CDM outputs allow the teacher to contrast different (pre-defined) groupings of students. For instance, Fig. 4 contrasts the average probability of skill mastery in the entire sample against school class A and school class B. It displays the share of learners within each student entity that has developed or still lacks particular skills. Thus it provides reference points or even benchmarks to evaluate attained levels of conceptual understanding and operational proficiency in accounting for a specific class taught and to set priorities in the planning of subsequent instructional units.

Discussion

An interim balance: potentials and limits of CDMs for adaptive teaching in introductory accounting lessons

The present paper demonstrated how Cognitive Diagnosis Models can deliver fine-grained information on students' skills and deficits in dealing with domain-specific tasks, using introductory accounting as an exemplary field of application. Statistical analyses that incorporated several criteria for evaluating model fit corroborate theoretical assumptions on distinct skills as *multiple* dimensions of accounting competence, which have repeatedly been proposed in conceptual contributions in the literature on accounting education (Phillips and Heiser 2011; Seifried et al. 2010).

The (standard) CDM-method defines that single skills can be either mastered or not mastered. Accordingly, the statistical models estimate the probability of (non-)mastery of each skill for each test taker and assign each test taker to one of two latent groups – skill mastery or non-mastery. This is accompanied by the restriction that in the case of ambiguous results (=skill mastery probabilities around 50%), two similarly skilled test takers may be classified once as "mastered" and once as "non-mastered". CDM-users who do not consult further information (e.g., through further test items) in this area of uncertainty run the risk of misinterpretations.

The outlined problem also has a content- or domain-related side, i.e. the question of whether single *accounting skills* that together make up the multiple constituents of accounting competence are assumed to be dichotomous or continuous. We will briefly discuss this issue:

1. For narrowly defined skills such as "selecting the right account", it can well be argued that they can be either mastered or not. In contrast, for more complex skills such as "modeling economic reality", gradations are conceivable. However, complex skills tend to be the exception in the *introductory accounting* curriculum if we take the number and interactivity (Paas et al. 2003) of economic elements in a business process as an indicator of complexity. From a domain expert's view, the basic and common business transactions/events that belong to this curricular domain (e.g., calculation of 20% VAT) leave no or minimal space for deviating interpretations or alternative procedures.
2. While *single* skills are dichotomous, *skill profiles* are always ordinal. For instance, one student masters 2 out of 5 skills while another masters 4 out of 5 skills. Thus, skill profiles allow a more differentiated assessment and portrayal of a whole spectrum of relevant skills that together form domain-specific competence.
3. Empirically, the question can also be answered by model fit, which is sufficiently good in our case. In other words, the data support our assumption that the focal skills are dichotomous. Of course, this finding leaves open if models that assume continuous skills could show even better fit indices.
4. This leads to another argument in favor of the dichotomy assumption. Statistically, this assumption has the advantage that the statistical model has significantly less parameters than models assuming continuous skills (e.g., multidimensional IRT models) and can therefore be estimated at all, especially with relatively small samples.

Nevertheless, future research should explore the question of whether the dichotomy assumption could be an oversimplification even for introductory accounting skills and therefore a theoretically serious limitation.

From a *pedagogical* point of view, concerns may arise if grounding adaptive teaching strategies on detailed diagnostic information from continuous assessments might induce a teacher to concentrate primarily on remedial measures for observed deficits. Before elaborating how specific adaptive strategies (see chapter 1) can be implemented based on the available data of the examined accounting classes (see the following section), we discuss this issue more generally, thereby considering

- *The 'profiling' CDM outputs:* Individual skill profiles delineate not only the deficits but also the attained skills of each student. While the former can and should indeed reveal starting points for supportive measures, the latter can and should be used to formulate *reinforcing feedback*. Moreover, even feedback on deficits can take on an encouraging style by determining *proximal developmental steps* for precisely located difficulties of domain-specific conceptual understanding or operational proficiency, rather than giving a general evaluation of, for example, "satisfactory performance = grade C in a short test". Pooling students with heterogeneous profiles opens up another way of *valuing individual attainments* by involving very skilled learners with *responsible roles* in instructional rearrangements (e.g., as peer tutors). However, when forming homogeneous learner groups, it might indeed be easier for the teacher to conceive targeted and varied remedial tasks to overcome group-specific deficits than to design stimulating and further-reaching measures for learner groups who master *all* assessed skills already. Additionally, a teacher's own communicative and pedagogical skills (not the diagnostic information as such) will play a decisive role in avoiding motivational gaps between homogeneous learner groups.
- *The diagnostic content and goal:* Introductory accounting lessons aim to teach incommensurable *basic* skills, i.e. to reach clearly defined levels of conceptual understanding and operational proficiency in the domain. Thus, introductory accounting education has *convergent learning goals* in the sense of correct mental and formal representations of *elementary* business transactions and events that occur in any company. This delineated topical focus certainly implies a stronger diagnostic focus on detecting and overcoming deficits than domains that approve a wider range of viable ways to solve complex problems and thus, a wider spectrum of skills—such as intrapreneurship education (George et al. 2019). Whereas a teacher of *introductory* accounting primarily needs diagnostic information that helps him/her to guide students *towards desirable (basic) skill levels* in the run-up to more demanding and complex accounting tasks, in intrapreneurship education both the possibility and the desire to highlight individual strengths are more pronounced as these particular strengths might inform subsequent *career counseling*.

Practical implications

In line with the generic strategies of adaptive teaching set out in the first section, CDM-based assessments should deliver diagnostic information that helps teachers to (a) *give detailed feedback and individual support,* (b) *organize class members into homogeneous or heterogeneous learner groups,* and (c) *set priorities for deliberate practice in different school classes*. Based on the exemplary CDM outputs in CDM output evaluation against the assessment context, we now describe how available data on students' accounting skills might guide adaptive strategies in introductory accounting classes. Without profile-based diagnostic information, teachers are in danger of overlooking a student's lack of specific yet critical skills and consequently, of failing to create appropriate learning opportunities.

The *individual skill profile* of the student selected in Skill profiles shows a pronounced deficit regarding Amount Calculation (S4). In this case, learning opportunities are needed that promote the student's ability to carry out essential mathematical operations for determining the exact monetary values of business transactions and events, such as value-added tax. Since the skill profile locates individual strengths and weaknesses in dealing with domain-specific task requirements, the teacher has specific indications to formulate proximal learning goals for the student concerned (such as building up calculatory knowledge and procedures) and to provide tailored learning material to achieve these goals. Additionally, the individual skill profile helps to formulate positive, reinforcing feedback. A teacher who points out accomplished skills and highlights individual skill acquisition throughout multiple assessments can be more specific and attribute achievements to controllable causes more credibly (Henderlong and Lepper 2002) than a teacher who gives only global praise such as "You're quite a good bookkeeper!".

Beyond the individual level, skill profiles facilitate *ability-based differentiation* within a school class. With this aim, groups of learners with similar or dissimilar patterns of domain-specific strengths and weaknesses *(skill classes)* can be assembled to *address their deficits* and even to *utilize their developed abilities* in a targeted manner. In the case of students A-D in Skill class distribution in a given school class, Table 10, *common deficits* concern their domain-specific ability to model the economic situation (identical skill profile 01111). These learners might receive small case studies that prompt them to elaborate on how flows of goods and money are reflected in documents (Tramm 2003). Classmates who *possess this skill* (such as student E with Profile 11011) could actively support their learning process as tutors. If in the present example, all of the named students analyze the small case studies and construct the related book entries *cooperatively*, they might *mutually benefit* from each other, thus strengthening their economic expertise (students A-D) and their technical skills in the case-relevant formal bookkeeping rules (student E).

Finally, CDM output on shares of learners who have mastered particular skills in a direct comparison of "natural" student groupings such as different school classes can support teachers in (1) making *accurate evaluations of average skill levels* currently reached in a given group and (2) considering these levels when *planning subsequent lessons* for the whole class. Regarding Fig. 4 as an example, the teacher responsible for class A can conclude that many students' skills for modeling an economic situation (S1) and for choosing the correct account sides (debit/credit) (S3) are below the level of relevant reference groups. Based on this diagnostic information, the teacher may provide additional explanations (focusing on S3) and discuss small business cases (focusing on S1) during classroom dialogue. The latter strategy again has the potential not only to strengthen deficient skills but also to involve those learners with developed (modeling) skills actively, giving them the additional cognitive challenge of *explaining and justifying* their mental models to classmates.

Prospective developmental work

The present CDM should be a starting point for future research. The reported analyses don't attest to the assumption of hierarchical skills. It would be too early to reject the hierarchy assumption though, as it is compatible with the literature. It may be possible to confirm the hierarchy assumption with newly developed test items with more element interactivity to raise the complexity level of the transactions/events that have to be modeled mentally and thus, reduce chances that test takers 'produce' correct operations within the accounting system logic without a thorough conceptual understanding. At the same time, it must be remembered that CDMs require even 'economically reduced' tasks that only prompt single or few particular skills to precisely identify mastered and non-mastered skills. Therefore, a good mixture of simple and complex tasks is necessary.

Although we successfully demonstrated how multidimensional (basic) accounting skills can be assessed by employing Cognitive Diagnosis Models, we are aware that pedagogical practice will only profit from these tools if their usability increases. Laying the scientific groundings for reliable and valid assessments has to be recognized as an indispensable first step toward CDM applications that neither overstrain teachers with data collection and statistical details nor patronize their instructional decisions. Rather, the developmental goal should be to strengthen the evidence-based foundation of teachers' professional choices. We, therefore, differentiate two aspects of enhanced usability.

Firstly, teachers should be equipped with CDM output that is readily available and easily comprehensible. To this end, conceptual work from subject didactics and psychometric research is needed to develop web applications like platforms for Cognitive Diagnostic-Computerized Adaptive Testing (CD-CAT). Operating on fast and robust analytical procedures, such platforms administer tasks for students, integrate data from their task solutions, and deliver adaptive *real-time* feedback to teachers and students. This feedback can not only include reports of skill profiles and skill classes within the tested student group but also provide benchmarks for task performance regarding curricular standards or comparison groups. To optimize the user interface of these platforms, systematic evaluation studies should compare different options to label and graphically display diagnostic information on skill profiles, classes, and mastery (see CDM output evaluation against the assessment context). Evaluations may focus on perceived comprehensibility and helpfulness to support instructional decisions from the view of teachers as well as on the appropriateness of teachers' data interpretations and conclusions from the view of educational and psychometric scientists. Taken together, these developmental steps could eventually generate forms of CD-CAT that come close to on-the-fly assessments regarding their simple handling and short intervals of measurement but possess higher reliability and validity.

Secondly, teachers should receive training that makes them 'enlightened users' of CDMs. Training must therefore include application-oriented knowledge on how to select and interpret CDM outputs. Moreover, it has to equip users with expert pedagogical knowledge on adaptive teaching to foster students' skill acquisition. This training focus is a necessary complement because diagnostic information can facilitate a teacher's content- and person-related instructional decisions by answering questions like 'Which particular skills for solving domain-specific tasks deserve targeted support?'. Yet diagnostic information alone does not prescribe a particular strategy ('Which measure is most adequate'?). Thus, to ultimatively enhance teachers' informational basis *and* broaden their instructional repertoire, more empirical research is needed on the effectiveness of alternative strategies of task assignment, teacher feedback, and deliberate practice among students with differing patterns of skills and deficits. Intervention studies may examine if certain strategies of adaptive teaching (such as providing worked examples for homogeneous ability groups or establishing peer tutoring within heterogeneous ability groups) are generally preferable in the focal curricular domain or if they show differential effects, depending on the specific skills that are to be acquired (such as performing calculations vs. modeling economic processes). Nevertheless, teachers must stay in command of instructional designs, aligning scientific knowledge about essential, effective strategies with the specificities of the present classroom context.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s40461-022-00137-3.

Additional file 1.

Additional file 2.

Additional file 3.

Acknowledgements
Not applicable.

Author contributions
CH designed the study, performed the data analyses and was major contributor in writing the manuscript. All authors contributed to the article and approved the submitted version.

Authors' Information
An explanation of why your manuscript should be published in Empirical Research in Vocational Education and TrainingThe manuscript presents an empirical study in the field of cognitive student assessment in early accounting instruction. We thus think that it well fits the scope of the journal.An explanation of any issues relating to journal policiesNo issues related to the journal policies.Confirmation that all authors have approved the manuscript for submissionWe confirm that all authors have approved the manuscript for publication.Confirmation that the content of the manuscript has not been published, or submitted for publication elsewhere (see our Duplicate publication policy)We confirm that the content of the manuscript hast not been published, or submitted for publication elsewhere.If you are submitting a manuscript to a particular special issue, please refer to its specific name in your covering letterNot applicable.

Funding
Not applicable.

Availability of data and materials
We are happy to provide the data and any further material on request. Please contact the corresponding author.

Declarations

Competing interests
There is no competing interests.

Received: 1 February 2022 Accepted: 5 September 2022
Published online: 23 September 2022

References

Achtenhagen F, Winther E (2014) Workplace-based competence measurement: developing innovative assessment systems for tomorrow's VET programmes. J Voc Educ Train 66(3):281–295. https://doi.org/10.1080/13636820.2014.916740

Artigue M, Hodgson BR, Blum W, Galbraith PL, Henn H-W, Niss M (2007) Modelling and applications in mathematics education, vol 10. Springer, US, Boston. https://doi.org/10.1007/978-0-387-29822-1

Bennett RE (2011) Formative assessment: a critical review. Assess Educ Princ Policy Pract 18(1):5–25. https://doi.org/10.1080/0969594X.2010.513678

Berding F, Jahncke H, Slopinski A (2020) Moderner rechnungswesenunterricht 2020. Springer Fachmedien, Wiesbaden. https://doi.org/10.1007/978-3-658-31146-9

Berding F (2019) Rechnungswesenunterricht: Grundvorstellungen und ihre Diagnose. Rainer Hampp

Black P, Wiliam D (2009) Developing the theory of formative assessment. Educ Assess Eval Account 21(1):5–31. https://doi.org/10.1007/s11092-008-9068-5

Blayney P, Kalyuga S, Sweller J (2015) Using cognitive load theory to tailor instruction to levels of accounting students' expertise. J Educ Technol Soc 18(4):199–210

Bley S (2017) Developing and validating a technology-based diagnostic assessment using the evidence-centered game design approach: an example of intrapreneurship competence. Empir Res Vocat Educ Train 9(1):281. https://doi.org/10.1186/s40461-017-0049-0

Bloemhof B, Christensen Hughes J. (2013) Active learning strategies in introductory financial accounting classes. Higher education quality council of Ontario; Canadian Electronic Library

Bloom BS (1968) Learning for Mastery. Eval Comm 1(2):1–5

BMUKK (2014) *Lehrplan der Handelsakademie* [BGBl. II -Ausgegeben am 27. August 2014—Nr. 209]. https://www.hak.cc/files/syllabus/Lehrplan_HAK_2014.pdf

Bouley F (2017) Kompetenzerwerb im Rechnungswesenunterricht. Springer Fachmedien, Wiesbaden

Cui Y, Gierl MJ, Chang H-H (2012) Estimating classification consistency and accuracy for cognitive diagnostic assessment. J Educ Meas 49(1):19–38. https://doi.org/10.1111/j.1745-3984.2011.00158.x

de La Torre J, Minchen N (2014) Cognitively diagnostic assessments and the cognitive diagnosis model framework. Psicol Educ 20(2):89–97. https://doi.org/10.1016/j.pse.2014.11.001

DiBello L, Roussos L, Stout W (2007) Review of cognitively diagnostic assessment and a summary of psychometric models. In: Rao CR, Sinharay S (eds) Handbook of statistics. Elsevier, pp 979–1030

Dillard JF, Bhaskar R, Stephens RG (1982) Using first-order cognitive analysis to understand problem solving behavior: an example from accounting. Instr Sci 11(1):71–92. https://doi.org/10.1007/BF00120982

Dioso-Henson L (2012) The effect of reciprocal peer tutoring and non-reciprocal peer tutoring on the performance of students in college physics. Res Educ 87(1):34–49. https://doi.org/10.7227/RIE.87.1.3

Fuchs LS, Powell SR, Seethaler PM, Cirino PT, Fletcher JM, Fuchs D, Hamlett CL (2010) The effects of strategic counting instruction, with and without deliberate practice, on number combination skill among students with mathematics difficulties. Learn Individ Differ 20(2):89–100. https://doi.org/10.1016/j.lindif.2009.09.003

George AC, Bley S, Pellegrino J (2019) Characterizing and diagnosing complex professional competencies—an example of intrapreneurship. Educ Meas Issues Pract 38(2):89–100. https://doi.org/10.1111/emip.12257

George AC, Robitzsch A, Kiefer T, Groß J, Ünlü A (2016) The R package CDM for cognitive diagnosis models. J Stat Softw. https://doi.org/10.18637/jss.v074.i02

Groß J, Robitzsch A, George AC (2016) Cognitive diagnosis models for baseline testing of educational standards in math. J Appl Stat 43(1):229–243. https://doi.org/10.1080/02664763.2014.1000841

Hattie J, Gan M (2011) Instruction based on feedback. In: Mayer RE, Alexander PA (eds) Educational psychology handbook series. Handbook of research on learning and instruction. Routledge, New York, pp 263–285

Helm C (2016) Welche denkschritte durchlaufen Schüler/innen beim erstellen von buchungssätzen? Wissenplus 15/16(1):38–41

Henderlong J, Lepper MR (2002) The effects of praise on children's intrinsic motivation: a review and synthesis. Psychol Bull 128(5):774–795. https://doi.org/10.1037/0033-2909.128.5.774

Henson RA, Templin JL, Willse JT (2008) Defining a family of cognitive diagnosis models using log-linear models with latent variables. Psychometrika 74(2):191. https://doi.org/10.1007/s11336-008-9089-5

Hopster-den Otter D, Wools S, Eggen TJHM, Veldkamp BP (2019) A general framework for the validation of embedded formative assessment. J Educ Meas 56(4):715–732. https://doi.org/10.1111/jedm.12234

Klotz VK, Winther E (2017) Assessing Tomorrow's Potential: a competence measuring approach in vocational education and training. In: Leutner D, Fleischer J, Grünkorn J, Klieme E (eds) Methodology of educational measurement and assessment. competence assessment in education: research, models and instruments. Springer, Cham. https://doi.org/10.1007/978-3-319-50030-0_14

Köhn H-F, Chiu C-Y (2018) How to build a complete q-matrix for a cognitively diagnostic test. J Classif 35(2):273–299. https://doi.org/10.1007/s00357-018-9255-0

Kunina-Habenicht O, Rupp AA, Wilhelm O (2009) A practical illustration of multidimensional diagnostic skills profiling: comparing results from confirmatory factor analysis and diagnostic classification models. Stud Educ Eval 35(2–3):64–70. https://doi.org/10.1016/j.stueduc.2009.10.003

Lee Y-S, de La Torre J, Park YS (2012) Relationships between cognitive diagnosis, CTT, and IRT indices: an empirical investigation. Asia Pac Educ Rev 13(2):333–345. https://doi.org/10.1007/s12564-011-9196-3

Lehtinen E, Hannula-Sormunen M, McMullen J, Gruber H (2017) Cultivating mathematical skills: from drill-and-practice to deliberate practice. ZDM 49(4):625–636. https://doi.org/10.1007/s11858-017-0856-6

Li H, Suen HK (2013) Constructing and validating a q-matrix for cognitive diagnostic analyses of a reading test. Educ Assess 18(1):1–25. https://doi.org/10.1080/10627197.2013.761522

Lyon CJ, Nabors Oláh L, Caroline Wylie E (2019) Working toward integrated practice: Understanding the interaction among formative assessment strategies. J Educ Res 112(3):301–314. https://doi.org/10.1080/00220671.2018.1514359

Marzano RJ, Kendall JS (2007) The new taxonomy of educational objectives, 2nd edn. Corwin Press, California

Maydeu-Olivares A (2013) Goodness-of-fit assessment of item response theory models. Meas Interdiscip Res Perspect 11(3):71–101. https://doi.org/10.1080/15366367.2013.831680

Minnameier G (2013) The inferential construction of knowledge in the business and economics domain. In: Beck K, Zlatkin-Troitschanskaia O (eds) From diagnostics to learning success: Proceedings in vocational education and training. Sense, Rotterdam, pp 141–156

Mislevy RJ, Haertel GD (2006) Implications of evidence-centered design for educational testing. Educ Meas Issues Pract 25(4):6–20. https://doi.org/10.1111/j.1745-3992.2006.00075.x

Muldoon N, Pawsey N, Palm CT (2007) An investigation into the use of a blended model of learning in a first year accounting subject. Proceedings 2007 AFAANZ Conference Accounting & Finance Association of Australia and New Zealand, Gold Coast, Queensland, Australia

Paas F, van Gog T (2006) Optimising worked example instruction: different ways to increase germane cognitive load. Learn Instr 16(2):87–91. https://doi.org/10.1016/j.learninstruc.2006.02.004

Paas F, Renkl A, Sweller J (2003) Cognitive load theory and instructional design: recent developments. Edu Psychol 38(1):1–4. https://doi.org/10.1207/S15326985EP3801_1

Park V, Datnow A (2017) Ability Grouping and differentiated instruction in an era of data-driven decision making. Am J Educ 123(2):281–306. https://doi.org/10.1086/689930

Pellegrino JW, Chudowsky N, Glaser R (2001) Knowing what students know: The science and design of educational assessment. National Academy Press. http://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=87034

Pellegrino JW (2010) The design of an assessment system for the race to the top: a learning sciences perspective on issues of growth and measurement. Educational Testing Service

Phillips F, Heiser L (2011) A field Experiment examining the effects of accounting equation emphasis and transaction scope on students learning to journalize. Issues Account Educ 26(4):681–699. https://doi.org/10.2308/iace-50051

Robitzsch A., Kiefer T, George AC, Ünlü A (2019) CDM: Cognitive diagnosis modeling. https://CRAN.R-project.org/package=CDM

Roscoe RD, Chi MTH (2007) Understanding tutor learning: knowledge-building and knowledge-telling in peer tutors' explanations and questions. Rev Educ Res 77(4):534–574. https://doi.org/10.3102/0034654307309920

Ruiz-Primo MA (2011) Informal formative assessment: the role of instructional dialogues in assessing students' learning. Stud Educ Eval 37(1):15–24. https://doi.org/10.1016/j.stueduc.2011.04.003

Rupp AA, Templin J, Henson RA (2010) Diagnostic measurement: theory, methods, and applications Methodology in the social sciences. Guilford Press, New York

Schneider M, Preckel F (2017) Variables associated with achievement in higher education: a systematic review of meta-analyses. Psychol Bull 143(6):565–600. https://doi.org/10.1037/bul0000098

Seifried J (2004) Fachdidaktische Variationen in einer selbstorganisationsoffenen Lernumgebung. Eine empirische Untersuchung im Rechnungswesenunterricht. Deutscher Universitätsverlag

Seifried J, Türling JM, Wuttke E (2010) Professionelles Lehrerhandeln. Schülerfehler erkennen und für Lernprozesse nutzen. In: Warwas J, Sembill D (Hrsg.) Schule zwischen Effizienzkriterien und Sinnfragen (S. 137–156). Baltmannsweiler: Schneider-Verl. Hohengehren

Shepard LA (2005) Linking formative assessment to scaffolding. Educ Leadersh 63(3):66–70

Sithole STM, Abeysekera I (2017) Accounting education: a cognitive load theory perspective Routledge studies in accounting, vol 20. Routledge, New York

Steenbergen-Hu S, Makel MC, Olszewski-Kubilius P (2016) what one hundred years of research says about the effects of ability grouping and acceleration on K–12 students' academic achievement. Rev Educ Res 86(4):849–899. https://doi.org/10.3102/0034654316675417

Tatsuoka KK (1983) Rule Space: An Approach for Dealing with Misconceptions Based on Item Response Theory. J Educ Meas 20(4):345–354

Tomlinson CA, Jarvis J (2009) Differentiation: Making curriculum work for all students through responsive planning and instruction. In: Renzulli JS (ed) Systems and models for developing programs for the gifted and talented, 2nd edn. Prufrock Press, Texas, pp 559–628

Tramm T (2003) Wirtschaftsinstrumentelle Rechnungswesen und die Modellierungsmethode—eine fachdidaktische Einführung. In: Joost D, Kripke G, Tramm PT (eds) Wirtschaftsinstrumentelles rechnungswesen: gültig ab der 1. auflage des lehrbuches, 1st edn. Bildungsverl. EINS, Braunschweig, pp 4–10

Valiandes S (2015) Evaluating the impact of differentiated instruction on literacy and reading in mixed ability classrooms: quality and equity dimensions of education effectiveness. Stud Educ Eval 45:17–26. https://doi.org/10.1016/j.stueduc.2015.02.005

Vogt F, Rogalla M (2009) Developing adaptive teaching competency through coaching. Teach Teach Educ 25(8):1051–1060. https://doi.org/10.1016/j.tate.2009.04.002

Wang W, Song L, Chen P, Meng Y, Ding S (2015) Attribute-level and pattern-level classification consistency and accuracy indices for cognitive diagnostic assessment. J Educ Meas 52(4):457–476. https://doi.org/10.1111/jedm.12096

Warwas J, Kärner T, Golyszny K (2015) Diagnostische sensibilität von lehrpersonen im berufsschulunterricht: explorative prozessanalysen mittels continuous-state-sampling. Zeitschrift Für Berufs Und Wirtschaftspädagogik 111(3):437–454

Waxman HC, Alford BL, Brown DB (2012) Individualized instruction. In: Hattie J, Anderman EM (eds) International guide to student achievement. Routledge, pp 405–407

Winther E (2010) Kompetenzmessung in der beruflichen Bildung. Habilitation, Bertelsmann

Yen WM (1984) Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. Appl Psychol Meas 8(2):125–145

Publisher's Note