72 Competence-Based Tests: Measurement Challenges of Competence Development in Vocational Education and Training

Christian Michaelis and Susan Seeber

Contents

Abstract

In the current vocational education and training (VET) research, many empirical studies focus primarily on the development of competence-based test instruments that aim to measure the cognitive dispositions of a domain-specific competence construct. While these studies mainly focus on the analysis of competence structures and levels, the valid measurement of competence development represents a research gap. Studies with this focus differ significantly in their theoretical-conceptual foundations and their methodological approaches. Moreover, different challenges arise in the methodological aspects of a valid change measurement. For this purpose, the first chapter reviews the research on the measurement of competence development in the

C. Michaelis · S. Seeber (✉)
Faculty of Economic Sciences, Georg-August-University of Göttingen, Göttingen, Germany
e-mail: christian.michaelis@wiwi.uni-goettingen.de; susan.seeber@wiwi.uni-goettingen.de

S. McGrath et al. (eds.), *Handbook of Vocational Education and Training*,
https://doi.org/10.1007/978-3-319-94532-3_83

VET field. Given the increasing significance of sustainable development-related competencies in VET, an empirical analysis is performed on commercial trainees' development of declarative knowledge related to sustainability from a societal perspective and sustainability in business processes. The in-depth analyses are based on data sets from cross-sectional and longitudinal studies. The differences and similarities between these research designs are compared with regard to their analysis potential, analysis methods, and results.

Keywords

Competence measurement · Competence development · Cross-sectional research design · Panel studies

Research on the Measurement of Competence Development in VET: Theoretical Assumptions and Methodological Designs

The discourse on measuring competencies in vocational education and training has received increasing attention in recent years (see Beck et al. 2016; Seeber 2017; Klotz and Winther 2017; Holtsch et al. 2016; Winther et al. 2016). Nevertheless, the understanding of competence as a latent construct and its associated causality is not always clear (Klieme et al. 2008). Currently, individual cognitive and affective-motivational dispositions (Blömeke et al. 2015) serve to explain performance related to the requirements in a domain-specific context; these dispositions are understood to be the mental resource potentials of an individual. Due to the assumption that competencies refer to clusters of knowledge, abilities, beliefs, and emotional and motivational resources and considered as a result of previous learning processes (Weinert 1999), theoretical foundations of studies aiming to measure competence development have to take into account research on learning, teaching, learning environments, and processes as well as on specific characteristics of the domain. Moreover, psychometric research has to be considered.

In principle, competence development is a lifelong process (Blossfeld and Schneider 2011). Nevertheless, the research on competence development in VET is often restricted to observations in the life stage of a respective training program and goes seldom beyond. The theoretical frameworks in this context are often based on approaches to expertise research, particularly the five-stage skill acquisition model by Dreyfus and Dreyfus (1980), which describes the development from a novice to an expert or master level. However, corresponding (more deterministic) models of competence development, such as that of Dreyfus and Dreyfus (1980), are very general, and due to the abstract level differentiations and the lack of explicit explanation-relevant predictors, such models are limited to describing the development of professional competence in VET (Michaelis 2017, p. 45f.). In any case, it is still questionable whether competence development in terms of temporal development is to be understood as a continuous or discontinuous process (Fleischer et al. 2013). Because of the high degree of abstraction in Dreyfus and Dreyfus' (1980) model, competence development, which results from specific intervening learning arrangements, can be explained only to a limited extent by this theory. Even the offer-and-use models discussed in school effectivity research (e.g., Helmke and Weinert 1997) are of limited use in VET because these models do not refer to the context of workplace learning and its prevailing norms (see Lempert 2009; Rausch 2011; Michaelis 2017, p. 45ff.; Sembill 2008).

Up to this day, the diagnosis of domain-specific competence in different educational areas has been predominantly based on curricular and therefore formal learning processes. For such research projects, and particularly for the previously noted intervention designs, the curriculum instruction assessment triad (Mislevy and Riconscente 2005; Mislevy and Haertel 2006) can be used as a theoretical framework (Winther 2010; Seeber 2017). This means that the development of an assessment should be convergent with the curricular requirements (intended curriculum) but also with the teaching process (implemented curriculum/intervention). Moreover, assessments in VET have to take into account different learning contexts: school-based theoretical learning and practical-oriented learning in companies or workplaces. It can be assumed that the two learning contexts help to develop different knowledge representations; both are necessary to cope with workplace requirements and to act as an expert in a respective domain. Subsequently, the assessment of the development of occupation-related competencies must refer to both learning contexts and their specific outcomes (see Winther 2010, p. 93ff.; Seeber 2017). Furthermore, assessments of the development of learning outcomes must not only consider different learning opportunities in vocational schools and training companies but also nonformal and informal learning opportunities outside of these education or training institutions. Nonformal and informal learning possibilities, e.g., extracurricular activities or learning possibilities in the private context, represent interesting and important explanatory factors. However, explanatory factors of competence development are not considered further in this chapter. This chapter is more concentrated on the methodological challenges due to the underestimation in research on the measurement of change.

Methodological differences can be attributed to general fundamental questions of the measurement of change. Studies with cross-sectional (survey of different samples at one occasion of measurement, e.g., different years of training), trend (survey of different, but comparable, samples at different occasions of measurement), as well as panel designs (longitudinal surveys in which a sample is repeatedly questioned at different occasions of measurement) are suitable for analyzing competence development. Independent of the research design, it must be ensured that the construct validity is the central condition. Consequently, the comparability of the outcome variable or a set of outcome variables across different samples (cross-sectional and trend) or at different occasions of measurement (trend and panel) is of central importance (von Davier et al. 2008). As a general rule, no change measurement methodology is free of measurement errors. Rost (2004) calls this the “validity problem of measuring change.” Each procedure has its own advantages and disadvantages, which have to be weighed against the objectives of a study. Table 1 presents major challenges associated with cross-sectional and longitudinal research designs.

In the past mean comparisons based on sum scores were common for the observation of competence development; however, this approach does not meet the expectation of a valid psychometric foundation. Recent research with validated psychometric procedures primarily focus on the German VET system with its specific characteristics of the so-called dual system. On the one hand, there are studies that analyze the development

Table 1 Challenges of different methodical designs for measuring competence development

	Cross-sectional design/trend design	Longitudinal design
Analytic potential	Interindividual differences between subgroups (e.g., differentiated by years of training)	Interindividual differences between occasions of measurement and intraindividual development
Sample requirements	Comparable samples in difficult accessible test fields: In addition to the usual comparison of subgroups (especially by gender, level of education, migration background, socioeconomic status), VET faces the challenge of cyclical economic fluctuations that can lead to rapid compensation effects in the training market (Seeber et al. 2017). Therefore, regional aspects of the training market have to be compared across the subgroups	Constant sample over time or appropriate methods for use with incomplete data sets
Test instruments	Ensuring the measurement invariance/construct validity	Ensuring the measurement invariance/construct validity over time Risk of memory effects if subjects answer an item multiple times at all occasions of measurement within short intervals of time

of basic competencies in prevocational training programs (e.g., mathematics, linguistics, or natural sciences) (e.g., Weißeno et al. 2016; Behrend et al. 2017). On the other hand, there are studies that aim to measure the development of domain-specific competencies in selected occupations. Table 2 provides an overview of selected longitudinal studies measuring the change in domain-specific vocational competence during the course of training. These recent studies are methodologically acceptable, based on item response theory (IRT), a probabilistic approach where the models and underlying assumptions are verified in the process of scaling.

Methodological Differences in the Use of Cross-Sectional and Longitudinal Study Designs for the Measurement of Competence Development

In scientific psychometrics, different approaches exist for measuring competence development. First, it should be noted that change can be observed and discussed on two levels: the group level and the individual level. The specific research interest and the study design determine which level should be addressed. For large-scale assessments on the effectiveness of VET as a whole or the effectiveness of selected training programs, a group-level observation may be sufficient; however, in most diagnostic processes, it is indispensable to obtain precise information on the individual change in competences. Both methodical designs can be analyzed

Table 2 Selected studies of the "research field" diagnosis of competence development in VET

Domain reference	Authors	Sample/target group	Focus	Methodical approach
Commercial/ business	Rosendahl and Straka (2011a, b)	Bank clerks	Domain-specific competence	Longitudinal study: IRT and structural equation modeling
	Klotz (2015) and Klotz et al. (2015)	Industrial clerks	Domain-specific competence	Cross-sectional study: IRT
	Michaelis (2017)	Freight forwarding and logistics services clerks	Competencies toward sustainable operational management	Longitudinal study: IRT and structural equation modeling
Industrial-technical	Nickolaus et al. (2008, 2011)	Electronic technicians – specializing in energy and building technology Motor vehicle mechatronics technicians (only in the study of 2008)	Domain-specific competence	Longitudinal study: structural equation modeling
	Atik and Nickolaus (2016)	Plant mechanics	Domain-specific competence	Longitudinal study: IRT
	Abele (2014)	Motor vehicle mechatronics technicians Production mechanics Electronic technicians – specializing in automation technology Mechatronic fitters	Domain-specific competence	Longitudinal study: IRT and structural equation modeling

by using item response theory models (e.g., von Davier et al. 2011; Meiser 2007). Already the one-parameter logistic (1-PL) IRT model is suitable to analyze competence developments (see Rasch 1960; Fischer and Molenaar 2012). This chapter follows this approach. A crucial feature of this method is the representation of person skills and item difficulties on a scale based on the response behavior of a random sample.

To be able to evaluate competence-based test instruments with IRT models, an essential condition is the use of suitable linking designs (see von Davier et al. 2008). A common approach is to use anchor items. Studies with a cross-sectional design use identical subgroup-overlapping items, and longitudinal studies use identical occasion-overlapping items. These approaches are adopted in this chapter and are explained in more detail below.

An Approach to Analyzing Competence Development in Cross-Sectional Designs

If anchor items exist between the considered subgroups, differential item functioning (DIF) analyses (as a method for testing measurement invariance) can be used to check subgroup differences both globally and at the item level. The basic assumption of DIF analyses is to calculate item parameters specifically for the considered subgroups within the Rasch analysis, such as trainees with different years of training. By comparing these subgroup-specific parameters, differences in the difficulty degree of a test instrument (global) or of items (local) can be derived for the subgroups. At the item level, the DIF parameter generally indicates the deviation from the average item difficulty in logit units. To calculate the item difficulty, the estimate average and the subgroup-specific estimate must be added. A positive item DIF parameter therefore means a higher item difficulty for the subgroup, whereas a negative item DIF parameter indicates a lower item difficulty (Wu et al. 2007).

To assess the DIF effect, the recommendations of the National Educational Panel Study (NEPS) will be considered. An absolute difference in the average item difficulty greater than 0.4 and less than 0.6 is a weak effect, a difference greater than 0.6 and less than 1.0 is a medium effect, and a difference greater than 1.0 is a strong effect (Pohl and Carstensen 2012). A DIF parameter is considered significant if it is at least twice as large as the associated standard error.

In accordance with expertise research (e.g., Dreyfus and Dreyfus 1980), it can be assumed that apprentices will find it easier to answer items as the number of years of training increases if the content of the items is implemented in curricula. Accordingly, for this case, lower DIF parameters are expected during the training program and can be interpreted as a positive development of competence. Klotz et al. (2015) used this approach in cross-sectional study for commercial trainees. Nevertheless, a high uncertainty remains with regard to the usability of these items as anchor items in longitudinal designs, because the link assumption assumes the item's measurement invariance over time.

An Approach to Analyzing Competence Development in Longitudinal Designs

However, the 1-PL Rasch model faces a challenge in the analysis of longitudinal data with regard to the data structure. Due to the different occasions of measurement, the classical structure, which consists of items and cases, is extended by a time-specific third dimension. However, the three-dimensional structure can be resolved by a data restructuring with virtual test persons or virtual items. In this particular context, "virtual" means that the data of survey times n + 1 are added to the data set of t1 as additional variables or test participants. However, the two procedures should be viewed not as alternatives but as complementary methods. Data restructuring via virtual persons enables analyses of temporal measurement invariance, which is referred to in the literature as item parameter drift. In principle, the data set is treated with a cross-sectional design, and interindividual differences between the occasions of measurement can be calculated via DIF analyses, as described above.

The advantage of longitudinal designs, however, lies in the ability to analyze intraindividual competence development. For this purpose, the method of virtual items has been developed. This approach requires the specification of a multi-dimensional Rasch model. The items of each occasion of measurement form a separate dimension. Put simply, items of the first occasion of measurement form the first dimension, and items of the subsequent occasion of measurement n + 1 form the n + first dimension. Corresponding models are based primarily on the methods of Andersen (1985) and Embretson (1991). The main difference between the two methods lies in the test booklet design. While in Anderson's design, identical items are used at all occasions of measurement (anchor items), Embretson's method allows occasion-specific items as well as anchor items. Embretson (1991) recommends using all items at each test time but distributing them to different test booklets so that a test participant answers each item only once. To implement a corresponding procedure, IT-based test assessments should be used.

Based on the assumption of construct validity, anchor items should still measure the same construct at later occasions of measurement in multidimensional models. For this purpose, it is important to keep the item parameters of the anchor items constant over time. This means that the anchor items receive the same item parameters for all occasions of measurement (parameter fixation). An essential aspect, however, is that competence development is based solely on anchor items. Additional occasion-specific items are used to improve the estimation quality of personal parameters. However, in this method, it is important to select for the item parameter fixation only anchor items whose stability over time is given. Therefore, the analysis of item parameter drift should be advanced. A significant item parameter drift of the anchor items can lead to inappropriate item fit values of the subsequent Rasch scaling. Whether an anchor item is finally usable can be checked via the subsequent Rasch scaling with item parameter fixations. The anchor items should have acceptable item fit values after scaling.

Exemplary Presentation of Challenges of the Measurement of Competence Development in VET

Methodology

The following example for the measurement of competence development in VET refers to the latent construct of sustainability competencies in business processes. The empirical analysis of competence development is focused on the specific competence construct "acting in the sense of sustainability in the context of operational management." "Competence in sustainability management are defined as a complex ability to act adequately in business contexts, in particular to be able to take into account the medium- and long-term economical, ecological and social – intra-company as well as external – consequences of (strategic and operational – the authors) management decisions" (Seeber et al. 2016). Sustainability aspects are becoming increasingly important in sustainable development education (Barth and Rieckmann 2016) as well as in VET. However, curricular analyses show that in many German training programs, the implementation of sustainability aspects is insufficient (Brötz et al. 2014). For the purpose of this chapter, however, the complexity of this construct (Seeber and Michaelis 2014) is limited to the declarative knowledge (see Shavelson et al. 2005) or conceptual knowledge (see Anderson and Krathwohl 2001) regarding sustainability and sustainability in business administration. In terms of content, the declarative knowledge items asked about general principles of sustainability, i.e., theoretical and widespread normative concepts, facts and their significance, the impacts of currently discussed sustainability examples, and operational application possibilities in companies (Michaelis 2017).

The final test instrument consists of 40 items. In terms of item design, the test includes single- and multiple-choice items. The scoring is dichotomous. Accordingly, subjects earn one point if they answer the item correctly. A total of 698 apprentices in a 3-year freight forward and logistic clerk training program were tested in summer 2013. This training program provides a reasonable supply of training offerings in Germany (to ensure an adequate sample size); additionally, the issue of sustainability is becoming more and more part of the business models in the logistic sector (Handfield et al. 2013). Therefore, it could be assumed that the apprentices are familiar with topics of sustainability.

The survey was administered at the beginning of the school year. Trainees of all 3 training years were interviewed, so the data represented approximately 2 years of training. In the sense of a cross-sectional analysis, differences in competence levels between different training periods can be interpreted as competence development within the training. In addition, 185 of 326 apprentices in the first year of training were tested on two further occasions (summer 2014 and summer 2015). This sample was used for the longitudinal analysis. Table 3 shows the main characteristics of the tested sample.

Table 3 Main characteristics of the tested sample

Methodological design	Year of training	Sample size	Age (mean/ median)	Female	University entrance qualification	Number of declarative knowledge items
Cross-sectional	1	326	20.9/20 (28 missing)	116/35.6%	249/71.1% (3 missing)	20
	2	164	21.9/21 (3 missing)	61/37.4% (1 missing)	118/72.4% (1 missing)	23
	3	208	23.0/22 (2 missing)	82/39.6% (1 missing)	145/70.4% (2 missing)	23
Longitudinal	1–3	185	20.9/20 (4 missing)	71/38.4%	143/77.7% (1 missing)	20 per occasion of measurement

Note: The characteristics of the longitudinal analysis refer to the first year of training

The tests took place in four different vocational schools. Because the IT equipment in the vocational school was limited, it was not possible to use IT-based test software. Therefore, item rotations in the test booklets were not possible. For economic reasons, not all trainees received all items. Trainees in the cross-sectional sample answered 20–23 items. Trainees in the longitudinal sample received 20 items per occasion of measurement, with 10 items held constant as anchor items between the occasions of measurement.

To analyze the data, the 1-PL Rasch model by the ConQuest software program was used. This program offers different coefficients for analyzing reliability and construct validity, which are explained in the following:

- The estimate of item difficulty is given in logit units; in addition, the standard error is specified.
- The expected a posteriori/plausible value reliability (EAP/PV reliability) "is [the] explained variance according to the estimated model divided by total person variance" (Draney and Wilson 2008, p. 425).
- The weighted means square (wMNSQ) indicates how accurately an item fit the model. The expected value is 1.0. For multiple-choice tests, a wMNSQ between 0.8 and 1.2 is recommended (Bond and Fox 2007, p. 243).
- The T-value informs about the significance of the deviation from the perfect fit per item. Values between −2.0 and 2.0 are recommended (Bond and Fox 2007, p. 43).
- For the discriminatory power of an item, ConQuest reported the item-total correlation. In this study, values of 0.2 are sought. In some cases, such as when an item represents meaningful conceptual content, lower values are acceptable.

Results

Cross-Sectional Scaling

For the cross-sectional analysis, three Rasch scaling procedures were performed. To check the construct validity of the test instrument, a general Rasch scaling over all items was carried out first. Overall, the reliability of this test instrument is still reasonable (EAP/PV reliability, 0.630). The item fit values (see Table 4, left columns) also have good item properties. Only three items (1.1, 1.6, and 1.9) are

Table 4 Global DIF analysis for comparing the level of declarative knowledge regarding sustainability and sustainability in business administration between the years of training

Items	Year of training	Estimate	Error
1.1–1.12	1	−0.151	0.053
	2	0.081	0.062
	3	0.170	0.059
2.1–2.11	2	0.026	0.049
	3	−0.026	0.049

conspicuous, as their item-total correlation lies under 0.2. However, these values are still acceptable due to the content-related focus of the items.

In the second step, two DIF analyses were calculated for the items that were answered by trainees with different years of training. Accordingly, subgroup differences can only be calculated between the groups that received the same anchor items. Apprentices of all training years answered items 1.1–1.12, and apprentices from the second and third years of training additionally answered items 2.1–2.11. Table 4 displays the results on the global level, and Table 5 shows those on the item level in Table 5 (right columns). In contrast to the interpretation of the local DIF estimates, the lower the DIF parameter at the global level, the worse the individuals of the associated subgroup performed in comparison to the other subgroups (Wu et al. 2007). With regard to items 1.1–1.12, a decrease in test difficulty is observed with an increase in training courses. At first glance, this finding is in line with the expectation. With regard to items 2.1–2.11, there is only a marginal difference in sustainability-related knowledge between apprentices in the second and third years of training.

However, the global subgroup difference can be assumed only if the associated items show systematic discrimination between the subgroups. Accordingly, the level of item difficulty would have to decrease with increasing training process. In conclusion, negative DIF estimates are to be expected for higher years of training.

The DIF parameters show that the majority of item difficulties do not change significantly between the years of training. All DIF parameters (except item 1.10 toward the third year of training) are below the weak level of 0.4 in terms of the NEPS categorization described above. Only three items (1.1, 1.6, and 1.10) change significantly as the number of years of training increases. However, two items have an unexpected effect direction. Items 1.1 and 1.6 are significantly easier for trainees in the first years of training than for the other trainees. Certainly, the local item analyses only partially confirm the global findings. The global effect of the weaker performance of the first year of training is understandable with regard to the absolute number of stronger item parameters. Only 8 of the 12 items indicate a slightly higher item parameter value for the first year of training. While there were only marginal global differences for items 2.1–2.11 between the years of training, the DIF parameters show unsystematic drifts in the item parameters. The global result (a stable, declarative knowledge achievement) arises because of mutually decreasing item parameter differences between the subgroups.

Longitudinal Scaling

In the first step, the data of each occasion of measurement were individually scaled to identify items containing measurement errors (see more differentiated information in Michaelis 2017). Of the original 40 items, 5 items had to be excluded because of the inappropriate fit of the values. One anchor item was below the adequate value, so the number of anchor items was reduced to nine. Second, the data set was structured such that the test participants of later occasions of measurement were treated as virtual persons. To analyze the item parameter drifts between the occasions of measurement, a DIF analysis was applied. The global-level results are shown in Table 6. Again, it can be observed that the test difficulty decreases over the course of training.

Table 5 Cross-sectional scaling results

1-PL Rasch model						DIF analysis							
Item	Estimate	Error	wMNSQ	T	Item-total cor.	Average		Training year 1		Training year 2		Training year 3	
						Estimate	Error	Estimate	Error	Estimate	Error	Estimate	Error
1.1	2.494	0.126	1.04	0.4	**0.15**	2.697	0.130	−0.380	0.161	0.059	0.192	0.321	0.190
1.2	0.204	0.081	1.05	1.9	0.28	0.298	0.082	0.044	0.107	−0.128	0.125	0.084	0.117
1.3	1.572	0.096	1.03	0.6	0.26	1.722	0.097	−0.136	0.125	0.129	0.146	0.007	0.136
1.4	-0.218	0.082	1.05	1.7	0.28	−0.133	0.084	0.080	0.108	−0.093	0.128	0.013	0.120
1.5	-0.946	0.091	0.92	−1.8	0.47	−0.861	0.093	0.069	0.118	0.058	0.140	−0.127	0.134
1.6	1.806	0.101	1.06	0.9	**0.19**	1.983	0.104	−0.291	0.131	0.120	0.156	0.171	0.148
1.7	−1.238	0.097	1.00	0.1	0.31	−1.166	0.100	0.099	0.126	0.017	0.151	−0.116	0.144
1.8	0.237	0.081	0.97	−1.2	0.41	0.338	0.082	0.092	0.107	0.079	0.125	−0.171	0.117
1.9	−1.857	0.116	1.04	0.5	**0.19**	−1.789	0.119	0.094	0.148	0.042	0.179	−0.136	0.173
1.10	−1.173	0.096	0.97	−0.5	0.35	−1.170	0.102	0.360	0.125	0.062	0.151	−0.422	0.152
1.11	−1.083	0.094	0.91	−1.9	0.52	−0.973	0.095	−0.132	0.122	−0.055	0.145	0.187	0.133
1.12	−0.996	0.092	0.91	−2.1	0.50	−0.946	0.096	0.101	0.121	−0.290	0.150	0.189	0.133
2.1	1.338	0.121	1.04	0.8	0.22	0.935	0.117			−0.204	0.116	0.204	0.116
2.2	0.534	0.109	0.97	−1.1	0.44	0.138	0.106			−0.224	0.106	0.224	0.106
2.3	2.070	0.145	0.98	−0.1	0.29	1.663	0.139			−0.077	0.138	0.077	0.138
2.4	0.274	0.108	1.00	0.0	0.36	−0.103	0.106			−0.088	0.105	0.088	0.105
2.5	−1.191	0.130	0.98	−0.2	0.36	−1.517	0.126			0.046	0.125	−0.046	0.125
2.6	1.322	0.120	1.06	1.1	0.20	0.925	0.116			−0.103	0.116	0.103	0.116
2.7	1.045	0.115	1.02	0.5	0.29	0.649	0.111			−0.148	0.111	0.148	0.111
2.8	−1.211	0.132	1.04	0.5	0.25	−1.525	0.128			0.165	0.127	−0.165	0.127
2.9	1.307	0.120	1.04	0.8	0.26	0.983	0.120			0.353	0.120	−0.353	0.120
2.10	0.979	0.114	1.01	0.3	0.30	0.593	0.110			−0.037	0.110	0.037	0.110
2.11	−2.435	0.195	0.99	−0.0	0.26	−2.741	0.188			0.315	0.186	−0.315	0.186
3.1	−0.929	0.133	0.93	−1.1	0.50								
3.2	−1.002	0.135	0.96	−0.7	0.44								
3.3	0.711	0.124	1.04	0.9	0.28								

(*continued*)

Table 5 (continued)

1-PL Rasch model						DIF analysis							
Item	Estimate	Error	wMNSQ	T	Item-total cor.	Average		Training year 1		Training year 2		Training year 3	
						Estimate	Error	Estimate	Error	Estimate	Error	Estimate	Error
3.4	−1.057	0.137	0.95	−0.7	0.45								
3.5	−1.020	0.136	0.95	−0.7	0.42								
3.6	−0.326	0.123	1.03	0.8	0.32								
3.7	0.790	0.125	0.99	−0.2	0.40								

Note: One item of the 1-PL Rasch model has been deleted for better scaling results. Conspicuous item fit values are highlighted in bold. Significant DIF parameters are grayed out

Table 6 DIF analysis for comparing the development of item parameter drift on a global level (Michaelis 2017, p. 205)

Occasion of measurement	Estimate	Error
t1	−0.304	0.058
t2	0.127	0.058
t3	0.177	0.058

The anchor items of the declarative knowledge test have a globally significant subgroup difference and show a significant decrease in difficulty over the course of training. This decrease is most pronounced between occasion times t1 and t2. At the item level (Table 7), three of the nine anchor items show a significant item parameter drift. These are items A7, A8, and A10 at t1 and t3. At this time, these items should not be excluded but should be observed within the longitudinal scaling. As mentioned before, a significant item parameter drift is merely an indicator of potential measurement errors of item parameter fixation. Therefore, whether a potential impairment of the measurement model arises is examined below via a longitudinal scaling with item parameter fixation.

Third, the longitudinal scaling is performed by modeling each occasion of measurement as a separate dimension (virtual item structure) and using the item parameter fixation of anchor items. After the deviance (an indicator for assessing the quality of the Rasch model; see Wu et al. 2007) was compared, the best model fit was obtained when the anchor item parameters of the second occasion of measurement from the time-specific individual scaling were used as the item parameter fixation. In the following, only the results of the final scaling are discussed (for the final results of the scaling, see Michaelis 2017, p. 344f). In addition to the abovementioned exclusion of five items, nine additional items were excluded from the scaling due to insufficient item fit values during the longitudinal scaling. A few item parameter fixations were also removed as they resulted in inappropriate item fit values. This mainly concerns the items that were observed in

Table 7 DIF analysis for the identification of items in the declarative knowledge test, which have an item parameter drift between the occasions of measurement. (Based on Michaelis 2017, S. 206)

Item	Estimate (average)	Error (average)	Estimate (t1)	Error (t1)	Estimate (t2)	Error (t2)	Estimate (t3)	Error (t3)
A2 (1.3)	1.266	0.094	−0.107	0.134	−0.088	0.129	0.195	0.132
A3 (1.4)	−0.328	0.087	−0.131	0.122	0.081	0.123	0.050	0.123
A4 (1.5)	−1.142	0.100	−0.022	0.136	−0.151	0.145	0.173	0.140
A5 (1.6)	1.651	0.102	−0.258	0.143	0.167	0.143	0.091	0.141
A6 (1.8)	0.025	0.085	−0.084	0.120	0.066	0.120	0.018	0.120
A7 (1.12)	−0.886	0.094	−0.367	0.133	−0.056	0.135	0.423	0.130
A8 (3.3)	−0.448	0.094	0.817	0.127	−0.085	0.130	−0.733	0.139
A9 (3.2)	−0.369	0.087	−0.186	0.122	−0.030	0.124	0.217	0.123
A10 (3.7)	0.230	0.086	0.339	0.123	−0.095	0.121	−0.333	0.123

Note: Significant DIF parameters are grayed out

the item parameter drift analysis. The final scaling is reasonable to good scores in terms of item fit values and reliabilities (EAP/PV reliability, 0.689 (t1), 0.708 (t2), 0.733 (t3)). Only two items stand out with an item-total correlation under the recommendations of 0.2. For content-related reasons, however, these items are kept in the analysis.

The longitudinal survey design allows the calculation of the latent variances and correlations as an indicator of the stability of declarative knowledge about sustainability over time. The results, including those in Table 8, show a strong correlation between the occasions of measurement. Accordingly, interindividual differences in intraindividual development are less pronounced.

Comparison of Scaling Results of Cross-Sectional and Longitudinal Research Design

Based on the previous results, the competence developments of both research designs are compared in Table 9. For this purpose, Warm's mean weighted likelihood estimates (WLE) as an estimator for a persons' ability (Warm 1989) were z-standardized per research design (mean value = 500, standard deviation = 100), and the effect sizes of the change (Cohen's/Hedge's d) were calculated. The effect sizes were calculated as the relative mean difference in the z-standardized WLE person ability scores between the higher and the lower occasions of measurement on the mean total standard deviation (Howell 2013). Hedge's d is similar to Cohen's d, but it calculates the effect size for subgroups of different sizes. The comparison shows that with a continuous training process, the performance on the declarative knowledge test regarding sustainability and sustainability in business administration

Table 8 Latent covariances (top right) and correlations (bottom left) between the occasions of measurement of the declarative knowledge test (Michaelis 2017, p. 218)

Occasion of measurement	t1	t2	t3
t1		0.393	0.423
t2	0.702		0.412
t3	0.626	0.797	

Table 9 Comparison of competence development between cross-sectional and longitudinal research designs

Subgroup	N	WLF (mean)	SD	Cohen’s/Hedge’s d to previous year of training
LS (t1)	185	466.497	107.233	
LS (t2)	185	511.292	86.246	0.460
LS (t3)	185	522.211	97.049	0.119
CS (first year)	306	488.057	109.456	
CS (second year)	164	513.507	93.548	0.244
CS (third year)	208	506.920	88.080	−0.073

Note: CS, cross-sectional study; LS, longitudinal study

increases. The greatest increase in knowledge is measured between the first and second years of the training program in both research designs; however, there is a stronger effect in the longitudinal study. One explanation may lie in the larger number of anchor items in the cross-sectional study design and the stronger unsystematic developments of their item parameters (see Table 4, left columns), which could lead to a reduced knowledge development in the cross-sectional design. The above-average improvement during the first year of the training program is congruent to curricular analyses that suggest the importance of developing competencies regarding sustainability in the freight forwarding and logistics clerk occupation (Michaelis 2017). Between the second and third years of training, the longitudinal study shows a weak but positive effect. In the cross-sectional data set, there is even a weak deterioration in performance between the second and third years of training.

In addition to the comparison of competence development using different research designs, it is possible to consider item parameter differences as well as the development of the items that were used in both study designs (cross-sectional and longitudinal) at all occasions of measurement. Five items fulfill this condition, for which a new DIF analysis is calculated. Therefore, the data of the second and third occasions of measurement of the longitudinal study were again structured as virtual persons and matched with the data of the cross-sectional study. Table 10 includes the results of the DIF analysis.

The DIF effects of the five-item twins from the different study designs show only slightly pronounced item parameter drifts between the individual occasions of measurement or years of training. Nevertheless, the comparison of the item

Table 10 Comparison of item parameter development between cross-sectional and longitudinal research designs (DIF analysis)

Item twin	Design	Average		Year of training 1/t1		Year of training 2/t2		Year of training 3/t3	
		Estimate	Error	Estimate	Error	Estimate	Error	Estimate	Error
1.3	CS	1.657	0.067	−0.091	0.123	0.247	0.159	0.121	0.141
A2	LS			−0.053	0.147	−0.157	0.140	−0.067	0.143
1.4	CS	−0.143	0.058	0.021	0.109	−0.092	0.142	0.012	0.128
A3	LS			0.032	0.132	0.136	0.134	−0.109	0.135
1.5	CS	−0.946	0.066	0.063	0.117	0.120	0.154	−0.079	0.142
A4	LS			0.089	0.143	−0.160	0.156	−0.033	0.149
1.8	CS	0.287	0.058	0.089	0.108	0.139	0.139	−0.122	0.126
A6	LS			0.028	0.131	0.064	0.132	−0.198	0.132
1.12	CS	−0.855	0.065	−0.082	0.117	−0.414	0.164	0.067	0.138
A7	LS			−0.096	0.144	0.116	0.147	0.408	0.139

Note: CS, cross-sectional study; LS, longitudinal study

parameters makes it clear that depending on the study design, different (but usually only slightly pronounced) item parameters and developments can be measured. The most noticeable difference between the study designs is the development of the item parameter of the five-item twin, which compares items 1.12 and A7. However, the absolute item parameter difference of only 0.341 logits between items 1.12 and A7 toward the third year of training is rated as a low effect.

Discussion

As this chapter emphasized, the diagnosis of competence development in VET is a research gap. As clarified, the challenge of a valid change measurement is one of the main reasons for this research gap. Although there are comprehensive recommendations for measuring competence developments (e.g., von Davier et al. 2008, 2011), concrete measurement standards could not prevail in VET so far. Previous studies in VET differ mainly in their methodological design. The majority of VET studies aiming to measure competence development use longitudinal designs (e.g., Abele 2014; Atik and Nickolaus 2016; Rosendahl and Straka 2011a; Michaelis 2017; Nickolaus et al. 2008, 2011). In contrast, there are cross-sectional designs (Klotz 2015; Klotz et al. 2015), which are less common so far. However, both approaches have their own potential and advantages as well as aspects that promote measurement errors (see Table 1). The assets and drawbacks of the measurement approach have to be weighed against the objectives of a study. Hence, in the present analysis, psychometric issues of cross-sectional and longitudinal research designs were considered by the example of an instrument for measuring declarative knowledge regarding sustainability and sustainability in business administration. The analysis revealed the following findings:

- The greatest increase in declarative knowledge arises between the first and second years of the training program. However, the effect in the longitudinal research design is stronger than that in the cross-sectional research design.
- The comparison of the development of the items, which were all used in both survey designs for all occasions/training years, produced comparable results, with some individual occasion-specific deviations.

These differences can have a variety of methodological explanatory backgrounds. Therefore, causes are identified, and recommendations for the measurement of competence development in VET are briefly outlined in the following.

A substantial methodological difference is that the selection of anchor items in the two study designs differs. The cross-sectional research design includes 12 anchor items between all 3 subgroups and another 11 anchor items between the second and third years of training. The anchoring in the longitudinal study, however, is based on only five to seven items per measurement occasion. For both longitudinal and cross-sectional research designs, it is questionable whether the selection of anchor items represents a representative selection of items for measuring the declarative knowledge development regarding sustainability and sustainability in business administration. Using an IT-supported test procedure with the function of randomly distributing items on test booklets could improve the quality of the measurement for both research designs.

The sample can also be determined critically. In the longitudinal section, only persons who participated in all occasions of measurement were considered in the analysis. The representativeness of the sample can be limited by this (risk of selection effects). Here, however, imputation methods could be used to work with incomplete data sets. The cross-sectional sample is also problematic in terms of subgroup comparability. Particularly in the case of changing training market situations, the comparison of subgroups may be limited. In the case of changing critical comparison features, weighting methods would be recommended.

It is conspicuous for both research designs that unsystematic developments of the item parameters were measurable. To a certain degree, curricular analyses could explain item-specific difficulty-reducing developments (especially with regard to environmental aspects, which are anchored in the curriculum during the first year of training; Michaelis 2017). The finding that items become more difficult over the course of the training program must be attributed to memory barriers or oblivion processes. It could be assumed that corresponding concepts were acquired during previous educational processes (especially in general education) or perhaps informal learning processes (such as media contributions) but are less present during the training process. On the other hand, this result is compatible with critiques that training programs in Germany insufficiently support holistic sustainability aspects (Brötz et al. 2014). Despite partial curricular anchors, the promotion of sustainability-related competencies is more facilitated by the voluntary initiatives of training institutions. This makes it much more difficult to use cross-sectional study designs, as it is not possible to ensure comparability with regard to the content and quality of learning processes.

Conclusion

The explanations in this chapter are concentrated on aspects on the psychometric challenges to measure development in domain-specific knowledge of future business administration clerks. Considering the previous aspects, it becomes clear that both methods can generate data that could contain measurement errors. However, the choice of the survey design also depends on the analysis objective. It is possible that the analysis of intraindividual developments is only possible via longitudinal research designs. Additionally, with regard to the analysis of explanatory contexts (e.g., relationships to affective-motivational dispositions), longitudinal studies are superior to cross-sectional studies. Complex structural relationship models can be computed if explanatory features have been tested with regard to the respective occasions of measurement.

A second challenge was mentioned in the introduction: the multiple influencing factors. Domain-specific competence development is not only a result of structured and intended learning processes in vocational schools and training companies but also of informal learning opportunities at workplaces and in private life, especially the discussed competence development in the domain of sustainable business processes. Competence measurement in VET is not an end in itself; rather the aim is to improve learning and training opportunities to support the individual competence development. Therefore, individual prerequisites at the beginning of training in particular knowledge, attitudes, motivation, prevocational experiences, social and cultural backgrounds, etc. and structured learning opportunities as well as nonformal and informal learning during the course of training have to be considered in research designs. A second aspect has to be taken into account in the future research on competence development: the measurement of noncognitive competence facets. Most studies avoid the integration of motivational and emotional competence facets, beliefs, and attitudes. The main reason for this specific research lack lies in the difficulty of their simultaneous measurement and in the lack of appropriate psychometric methods for such complex research designs. First research efforts can be observed by using the experience sampling method (see Rausch et al. 2016), albeit not yet in a longitudinal design.

References

Abele S (2014) Modellierung und Entwicklung berufsfachlicher Kompetenz in der gewerblich-technischen Ausbildung. Steiner, Stuttgart

Andersen EB (1985) Estimating latent correlations between repeated testings. Psychometrika 50(1):3–16. https://doi.org/10.1007/BF02294143

Anderson LW, Krathwohl DR (2001) A taxonomy for learning teaching and assessing: a revision of Bloom's taxonomy of educational objectives. Longman Publishing, New York

Atik D, Nickolaus R (2016) Die Entwicklung berufsfachlicher Kompetenzen von Anlagenmechanikern m ersten Ausbildungsjahr. Z Berufs- Wirtschaftspädagogik 112(2):243–269

Barth M, Rieckmann M (2016) State of the art in research on higher education for sustainable development. In: Barth M, Michelsen G, Rieckmann M, Thomas I (eds) Routledge handbook of higher education for sustainable development. Routledge, London, pp 100–113

Beck K, Landenberger M, Oser F (eds) (2016) Technologiebasierte Kompetenzmessung in der beruflichen Bildung: Ergebnisse aus der BMBF-Förderinitiative ASCOT. Bertelsmann, Gütersloh. https://doi.org/10.3278/6004436w

Behrend S, Nickolaus R, Seeber S (2017) Die Entwicklung der Basiskompetenzen im Übergangssystem. Unterrichtswissenschaft 1:51–66. https://doi.org/10.3262/UW1701051

Blömeke S, Gustafsson JE, Shavelson RJ (2015) Beyond dichotomies: competence viewed as a continuum. Z Psychol 223(1):3–13. https://doi.org/10.1027/2151-2604/a000194

Blossfeld H-P, Schneider T (2011) Data on educational processes: national and international comparisons. In: Blossfeld H-P, Roßbach H-G, von Maurice J (eds) Education as a lifelong process: the German National Educational Panel Study, Sonderheft 14 – Zeitschrift für Erziehungswissenschaft. VS Verlag für Sozialwissenschaften, Wiesbaden, pp 35–50. https://doi.org/10.1007/s11618-011-0180-9

Bond TG, Fox CM (2007) Applying the Rasch model: fundamental measurement in the human sciences, 2nd edn. Lawrence Erlbaum Associates, Mahwah

Brötz R, Annen S, Kaiser F, Kock A, Krieger A, Noack A, . . . Tiemann M (2014) Gemeinsamkeiten und Unterschiede kaufmännisch-betriebswirtschaftlicher Aus- und Fortbildungsberufe (GUK) Abschlussbericht (ergänzte Fassung von 2014). BIBB, Bonn

Draney K, Wilson M (2008) A LLTM approach to the examination of teachers' ratings of classroom assessment tasks. Psychol Sci 50(3):417–432

Dreyfus SE, Dreyfus HL (1980) A five-stage model of the mental activities involved in directed skill acquisition. University of California Berkeley, Berkley

Embretson SE (1991) A multidimensional latent trait model for measuring learning and change. Psychometrika 56(3):495–515. https://doi.org/10.1007/BF02294487

Fischer GH, Molenaar IW (eds) (2012) Rasch models: foundations recent developments and applications. Springer, New York

Fleischer DPJ, Koeppen K, Kenk DPM, Klieme E, Leutner D (2013) Kompetenzmodellierung: Struktur Konzepte und Forschungszugänge des DFG-Schwerpunktprogramms. Z Erzieh 16(1):5–22. https://doi.org/10.1007/s11618-013-0379-z

Handfield R, Straube F, Pfohl H-C, Wieland A (2013) Trends and strategies in logistics and supply chain management – embracing global logistics complexity to drive market advantage. DVV Media Group GmbH, Hamburg

Helmke A, Weinert FE (1997) Bedingungsfaktoren schulischer Leistungen. In: Weinert FE (ed) Psychologie des Unterrichts und der Schule. Pädagogische Psychologie, 3rd edn. Hogrefe, Göttingen, pp 71–175

Holtsch D, Rohr-Mentele S, Wenger E, Eberle F, Shavelson RJ (2016) Challenges of a cross-national computer-based test adaptation. Empir Res Vocat Educ Train 8(18):1. https://doi.org/10.1186/s40461-016-0043-y

Howell DC (2013) Statistical methods for psychology, 8th edn. Wadsworth, Belmont

Klieme E, Hartig J, Rauch D (2008) The concept of competence in educational contexts. In: Hartig J, Klieme E, Leutner D (eds) Assessment of competencies in educational contexts. Hogrefe, Göttingen, pp 3–22

Klotz VK (2015) Diagnostik beruflicher Kompetenzentwicklung: Eine wirtschaftsdidaktische Modellierung für die kaufmännische Domäne. Springer Gabler, Wiesbaden. https://doi.org/10.1007/978-3-658-10681-2

Klotz VK, Winther E (2017) On improving current assessment practices – competence measurement in the domain of business and commerce. In: Leutner D, Fleischer J, Grünkorn J, Klieme E (eds) Competence assessment in education: research, models and instruments. Springer, Heidelberg, pp 221–243

Klotz VK, Winther E, Festner D (2015) Modeling the development of vocational competence: a psychometric model for economic domains. Vocat Learn 8(3):247–268. https://doi.org/10.1007/s12186-015-9139-y

Lempert W (2009) Berufliche Sozialisation: Persönlichkeitsentwicklung in der betrieblichen Ausbildung und Arbeit, 2nd edn. Schneider-Verlag Hohengehren, Baltmannsweiler

Meiser T (2007) Rasch models for longitudinal data. In: von Davier M, Carstensen CH (eds) Multivariate and mixture distribution Rasch models extensions and applications. Springer, New York, pp 191–199

Michaelis C (2017) Kompetenzentwicklung zum nachhaltigen Wirtschaften: Eine Längsschnittstudie in der kaufmännischen Ausbildung. Peter-Lang, Frankfurt am Main. https://doi.org/10.3726/b10896

Mislevy RJ, Haertel GD (2006) Implications of evidence-centered design for educational testing SRI International and University of Maryland Ravenswood. https://padi.sri.com/downloads/TR17_EMIP.pdf. Accessed 14 Dec 2017

Mislevy RJ, Riconscente MM (eds) (2005) Evidence-centered assessment design: layers structures and terminology, PADI technical report 9. SRI International, Menlo Park

Nickolaus R, Gschwendtner T, Geißel B (2008) Entwicklung und Modellierung beruflicher Fachkompetenz in der gewerblich-technischen Grundbildung. Z Berufs- Wirtschaftspädagogik 104(1):48–73

Nickolaus R, Geißel B, Abele S, Nitzschke A (2011) Fachkompetenzmodellierung und Fachkompetenzentwicklung bei Elektronikern für Energie- und Gebäudetechnik im Verlauf der Ausbildung – Ausgewählte Ergebnisse einer Längsschnittstudie. In: Nickolaus R, Pätzold G (eds) Lehr-Lernforschung in der gewerblich-technischen Berufsbildung, Zeitschrift für Berufs- und Wirtschaftspädagogik – Beiheft 2. Franz Steiner Verlag, Stuttgart, pp 78–94

Pohl S, Carstensen CH (2012) NEPS technical report. Scaling the data of the competence test. NEPS working paper no 14. Nationales Bildungspanel, Universität Bamberg. https://www.neps-data.de/Portals/0/Working%20Papers/WP_XIV.pdf. Accessed 14 Dec 2017

Rasch G (1960) Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests. Nielsen & Lydiche, Oxford

Rausch A (2011) Erleben und Lernen am Arbeitsplatz in der betrieblichen Ausbildung. VS Verlag für Sozialwissenschaften, Wiesbaden

Rausch A, Seifried J, Wuttke E, Kögler K, Brandt S (2016) Reliability and validity of a computer-based assessment of cognitive and non-cognitive facets of problem-solving competence in the business domain. Empir Res Vocat Educ Train 8:9. https://doi.org/10.1186/s40461-016-0035-y

Rosendahl J, Straka GA (2011a) Kompetenzmodellierungen zur wirtschaftlichen Fachkompetenz angehender Bankkaufleute. Z Berufs- Wirtschaftspädagogik 107(2):190–217

Rosendahl J, Straka GA (2011b) Effekte personaler schulischer und betrieblicher Bedingungen auf berufliche Kompetenzen von Bankkaufleuten während der dualen Ausbildung Ergebnisse einer dreijährigen Längsschnittstudie. https://elib.suub.uni-bremen.de/edocs/00102039-1.pdf. Accessed 14 Dec 2017

Rost J (2004) Lehrbuch Testtheorie – Testkonstruktion, 2nd edn. Hans Huber, Bern

Seeber S (2017) Economic competencies and situation specific commercial competencies: reflections on conceptualization and measurement. J Citizenship Soc Econ Educ 15(3):162–182. https://doi.org/10.1177/2047173417695275

Seeber S, Michaelis C (2014) Development of a model of competencies required for sustainable economic performance among apprentices in business education. AERA, Washington, DC. http://www.aera.net/Publications/Online-Paper-Repository. Accessed 14 Dec 2017

Seeber S, Hartig J, Dierkes S, Schumann M (2016) Ko-NaMa – simulation-based measurement and validation of a competence model for sustainability management. In: Pant HA, Zlatkin-Troitschanskaia O, Lautenbach C, Toepper, M, Molerov, D (eds) Modeling and measuring competencies in higher education – validation and methodological innovations (KoKoHs) – overview of the research projects (KoKoHs working papers, 11). Humboldt University, Berlin; Johannes Gutenberg University, Mainz, pp 53–56. https://www.kompetenzen-im-hochschulsektor.de/Dateien/KoKoHs_Working_Papers_No.11_Final_04.07.pdf. Accessed 14 Dec 2017

Seeber S, Baethge M, Baas M, Richter M, Busse R, Michaelis C (2017) Ländermonitor berufliche Bildung 2017: Leistungsfähigkeit und Chancengerechtigkeit – ein Vergleich zwischen den Bundesländern. wbv, Bielefeld. https://doi.org/10.3278/6004634w

Sembill D (2008) Führung und Zeit – institutionelle und unterrichtliche Perspektiven. In: Warwas J, Sembill D (eds) Zeitgemäße Führung – zeitgemäßer Unterricht. Schneider, Hohengehren, pp 81–98

Shavelson RJ, Ruiz-Primo MA, Wiley EW (2005) Windows into the mind. High Educ 49(4):413–430. https://doi.org/10.1007/s10734-004-9448-9

von Davier AA, Carstensen CH, von Davier M (2008) Linking competencies in horizontal vertical and longitudinal settings and measuring growth. In: Hartig J, Klieme E, Leutner D (eds) Assessment of competencies in educational contexts. Hogrefe, Göttingen, pp 53–80

von Davier M, Xu X, Carstensen CH (2011) Measuring growth in a longitudinal large-scale assessment with a general latent variable model. Psychometrika 76(2):318–336. https://doi.org/10.1007/s11336-011-9202-z

Warm TA (1989) Weighted likelihood estimation of ability in item response theory. Psychometrika 54:427–450. https://doi.org/10.1007/BF02294627

Weinert FE (1999) Concepts of competence. Definition and selection of competencies: theoretical and conceptual foundations (DeSeCo). OECD, Paris. https://pdfs.semanticscholar.org/8b88/efa9dd5e0a4b605aea6e5e3b9ec640beb089.pdf. Accessed 14 Dec 2017

Weißeno S, Seeber S, Kosanke J, Stange C (2016) Development of mathematical competency in different German pre-vocational training programmes of the transition system. Empir Res Vocat Educ Train 8(1):14. https://doi.org/10.1177/2047173417695275

Winther E (2010) Kompetenzmessung in der beruflichen Bildung. Bertelsmann, Bielefeld. https://doi.org/10.3278/6004148w

Winther E, Festner D, Sangmeister J, Klotz VK (2016) Facing commercial competence: modeling domain-linked and domain-specific competence as key elements of vocational development. In: Wuttke E, Schumann S, Seifried J (eds) Economic competence and financial literacy of young adults – status and challenges. Barbara Budrich, Opladen, pp 149–164

Wu ML, Adams RJ, Wilson MR, Haldane SA (2007) ACERConQuest Version 2: Generalised item response modeling software. Australian Council for Educational Research, Camberwell