Pretest-Posttest-Posttest Multilevel IRT Modeling of Competence Growth of Students in Higher Education in Germany

Susanne Schmidt, Jean-Paul Fox & Olga Zlatkin-Troitschanskaia

Longitudinal research in higher education faces several challenges. Appropriate methods of analyzing competence growth in higher education are needed to deal with those challenges and thereby obtain generalizable and valid results. In this paper, a newly developed pretest-posttest-posttest multivariate multilevel item response theory (IRT) model for repeated measures is introduced which is designed to address educational research questions according to a German research project with genuine data. In this model, dependencies between repeated observations of the same students are considered not, as usual, by clustering observations within participants but rather by clustering observations within semesters. Estimation of the model is conducted within a Bayesian framework with a Markov Chain Monte Charlo (MCMC) algorithm.

Introduction and Research Background

Questions about learning processes such as how competences are acquired are related to individual change and growth (Willett, 1988). To answer such questions, studies in which data is gathered to describe and explain changes in individual students' learning outcomes over time are needed (Singer & Willett, 2003). Longitudinal research allows more precise analyses to be conducted of the causal relationships between learning process variables and the growth in learning outcome variables than cross-sectional research (Bijleveld & Kamp, 1998, p. 2). However, the benefits of longitudinal research come at the cost of complications in its realization in field studies. Especially longitudinal research in higher education faces several challenges.

First, data usually is collected on volunteers, who often have little incentive to participate (repeatedly) and sometimes miss measurement sessions or drop out of the panel completely. As in cross-sectional research, missing data poses challenges to statistical analysis, for example, by reducing the power of statistical inferences or because of biased estimates of standard errors. Second, large numbers of college or university students are difficult to reach (repeatedly), rendering repeated assessment of the same students difficult especially if their anonymity has to be preserved. Consequently, samples tend to be small. Third, unlike in primary school and secondary school, where students usually are grouped according to age, have had the same amount of schooling, and have similar prior knowledge, students in higher education – at least in Germany – do not belong to fixed classes and therefore may attend courses with students who are in different semesters and have different prior knowledge and competences. This means that on one occasion there may be students in different semesters of study. Consequently, conducting studies in higher education may mean having a combined cohort-longitudinal design (cf. Happ, Zlatkin-Troitschanskaia, Beck, & Förster, in press (b); Hof, Roede, & Kowalski, 1977) where students of different semesters are assessed repeatedly. This requires an appropriate statistical model. Also, longitudinal research is expensive and time-consuming; it involves complex theoretical and methodological decisions. Finally, there is no guarantee that change will occur simply because of repeated measurements (Ployhart & Vandenberg, 2009, p. 95). These probably are the main reasons longitudinal studies are less common than cross-sectional studies in empirical research on education. Furthermore, since there is less research on higher education than on other sectors of education (Blömeke, Zlatkin-Troitschanskaia, Kuhn, & Fege, 2013), there are very few longitudinal studies addressing questions about learning processes over time in higher education (e.g., Schaap, Schmidt, & Verkoeijen, 2012; Coertjens, Donche, Maeyer, Vanthournout, & Petegem, 2013; Zlatkin-Troitschanskaia, Shavelson, & Kuhn, 2015).

In Germany, the number of students in higher education is rising steadily (Federal Statistical Office, 2014), and therefore it is becoming more important to understand students' competence acquisition in the tertiary sector. Especially in the major field of study of business and economics, with approximately 15% of all students in Germany (Federal Statistical Office, 2013), it is becoming increasingly important to examine the individual growth of domain-specific competence, since the acquisition of business and economic competences is one of the most important learning outcomes of studying in this field. Along with the assessment of competence growth, it is important to assess its relevant predictors. In many studies of higher education, predictors of academic success are not adequately defined or measured (Robbins et al, 2004, p. 262). In terms of individual learning processes of students in higher education, the most relevant predictors of growth of competence are prior knowledge, motivational orientations, and general cognitive abilities (see Alexander, Jetton, & Kulikowich, 1995; Hambrick et al., 2008; Shulman, 1970). To determine how the initial level and rate of growth of competence are related to motivational and cognitive variables, it is necessary to assess all variables (outcomes and predictors) together at all measurement occasions. Only in this way is it possible to describe growth of competence and explain differences in competence among students in different semesters by the relevant predictors. Furthermore, students' competences often are assessed using performance tests with different tasks, such as items in multiple choice (MC) format, which can be answered correctly or incorrectly only (cf. Nusche, 2008; Shavelson, 2013; Zlatkin-Troitschanskaia et al., 2015). Usually, the correct responses are calculated to obtain a sum score to interpret the level or growth of the competence construct of interest. However, for interpretations of such data in longitudinal studies it would be inappropriate to analyze the growth of competence based on manifest scale scores such as sum scores. This would ignore measurement errors associated with items and item scales, which often are on a dichotomous scale due to their MC format, and therefore could lead to biased standard errors and inconsistent parameter estimates (cf. Coertjens et al., 2013; Skrondal & Laake, 2001). In addition, the hierarchical structure of the data needs to be taken into account. For example, higher education students are nested in different degree courses and, within these, in different semesters. Therefore, multilevel modeling is needed to determine the amount of variance at the student level and at the semester level that can be explained by covariates (Snijders & Bosker, 2012). In multilevel modeling, the total variance in students' competence can be partitioned into variance within and variance between semesters of higher education in order to examine separately the influence of personal and semester-specific factors (independent variables) on students' competence (dependent variable) (Rabe-Hesketh, Skrondal, & Gjessing, 2008).

In this paper these challenges and deficits are taken into account within the context of valid assessment of the growth of competence of students in higher education. The aim of this paper is to analyze appropriately hierarchical and longitudinal data gathered within a research project conducted in Germany marked by the problems mentioned above. The theoretical and conceptual framework of the growth of competence of business and economics students as well as the research design of the corresponding study are presented in Section "Theoretical Foundation and Study Design". Taking the aforementioned challenges into account an innovative method of analysis which could enhance assessment, evaluation, testing, and measurement practices in higher education is presented and discussed. A newly developed pretest-posttest-posttest multivariate multilevel item response theory (IRT) model for repeated measures is introduced which is designed to address educational research questions according to the German research project with genuine data. Estimation of the model is conducted within a Bayesian framework with a Markov Chain Monte Charlo (MCMC) algorithm. The model with its potential and restrictions are explained in Section "Method and Results". Results of the application of this model to genuine data are also shown in Section "Method and Results". Implications and perspectives for future methodological and psychometrical developments needed to be able to assess growth of students' competence in higher education are discussed in Section "Discussion and Conclusion".

Theoretical Foundation and Study Design

Conceptual Model of the Growth of Competence of Semester Cohorts within the Multilevel IRT Framework and Research Questions

Competence is a theoretical construct which is not directly observable; it must be inferred from performance on tasks related to the specific competence of interest (Shavelson, 2013). To measure domain-specific competence in business and economics, performance tests comprise tasks or questions about business and economic situations (Zlatkin-Troitschanskaia, Förster, Brückner, & Happ, 2014), meaning the competence construct within this study domain has at least two sub-dimensions: business and economics. A one-dimensional measurement model was used as starting point to develop the new pretest-posttest-posttest model, and the sub-dimension of economic competence was randomly chosen as the focus of analysis in this paper. Although there is no uniform definition of *competence* in the literature, in many large-scale studies of education such as PISA[1] or NEPS[2] (see Artelt, Weinert, & Carstensen, 2013; Zlatkin-Troitschanskaia et al., 2014) Weinert's (2001) definition of competence is followed, and therefore cognitive and non-cognitive components of competence are distinguished but non-cognitive components often are excluded from modeling and measuring. We follow this tradition. Narrowing competence to cognitive components means that the in assessing economic competence focus is on measuring the specialized understanding and knowledge of economic principles and situations (cf. Walstad, Watts, & Rebeck, 2007; 2013; Zlatkin-Troitschanskaia et al., 2014). Growth of economic competence is modeled as gain and acquisition over the course of studies. This usually results in measuring individuals' initial amount and growth rate of economic competence. However, as described in Section 1, students in higher education do not belong to fixed classes and when assessments take place in lectures and courses, the test takers most likely are students in different semesters. This means, within one particular occasion, various cohorts of students grouped according to the number of semester they have completed are surveyed. This fact frames the approach presented in this paper and growth of competence is modeled as growth over semester cohorts (the cohorts are assembled as shown in Table 2). In other words, students are clustered according to the semester they are in, which is in line with a traditional multilevel framework. Dependencies between students' observations due to repeated measures are considered not, as usual, by clustering observations within individuals, but rather by clustering observations within semesters. By doing this, it should be possible to answer the following questions:

1) Is economic competence at one measurement occasion greater when students are in higher semesters?
2) Does economic competence grow over the three measurement points and with the number of semesters of study completed?
3) What is the amount of unobserved heterogeneity in the economic competence of students within semester-specific clusters?
4) What is the amount of unobserved heterogeneity in the economic competence of students between semesters?
5) What are the predictors of this unobserved potential heterogeneity?

6) Which effects of which predictor variables are stable over time?

The aim of this paper is not to measure individuals' growth of competence in economics but rather to measure the growth of economic competence within and across semester cohorts. In this way it is possible to identify (a) differences in economic competence between students in different semesters and (b) overall growth of economic competence across all semesters. Growth is represented via pairwise correlation between two measurement occasions as growth over cohorts and not as one individual's overall growth rate (see Figure 1; a detailed formulation of the model is described in Section "Method and Results"; the aim of the present Section is to describe how the conceptual model was built in terms of the dependent variable and its predictors as well as in terms of the research questions). In the present study assessment of economic competence took place at three occasions (T1, T2, and T3) to determine the growth of competence and its relevant predictors (for a detailed description see Section 2.2).

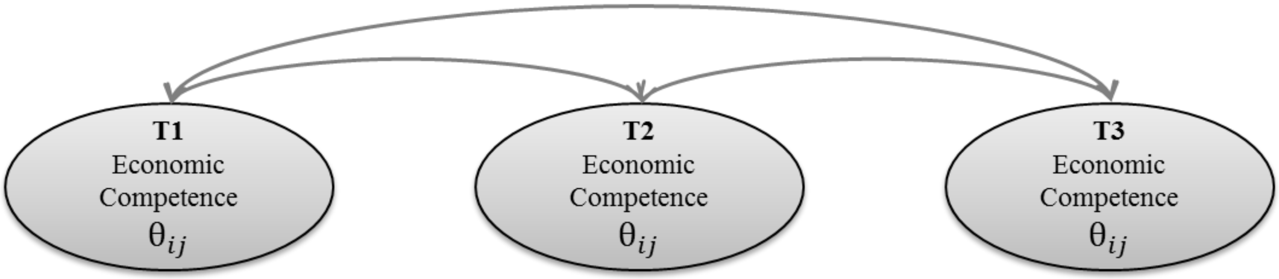

Figure 1. Conceptual model of the growth of economic competence of semester cohorts.

Figure 2 represents the conceptual model for one measurement occasion whereas T1, T2, and T3 are correlated as shown in Figure 1. Figure 2 illustrates that students (level 1) are clustered in semester-cohorts (level 2) as described above. Circles represent latent constructs whereas rectangles represent direct observations.

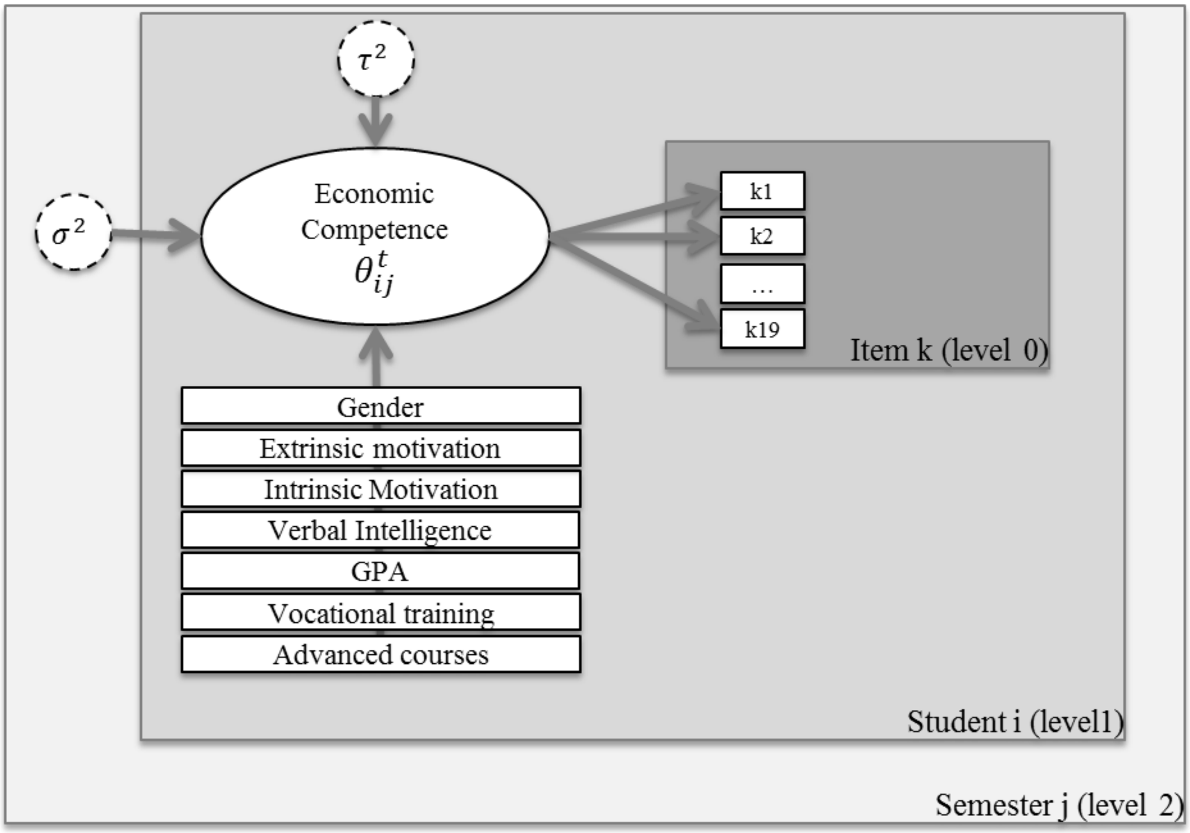

Figure 2. Conceptual model of the multilevel concept of economic competence.

The students' competence is denoted as θ_{ij}^t with $t = 1, 2, 3$ representing the measurement occasions, $j = 1, \ldots, J$ the number of semesters, and $i = 1, .., n_j$ the number of students nested within semesters j. As shown in Figure 2, items k with $k = 1, .., 19$ represent economic competence, which is located at level 0. By including level 0 in the conceptual framework, the measurement model of economic competence with its item discriminations and difficulties is considered along with the analysis model to answer research questions 1 to 5. Economic competence θ_{ij}^t estimated for each occasion T1, T2, and T3 as well as all predictors (motivational orientations, etc.) are located at level 1. At level 2, there are no predictors because variables that are constant for students within one semester but differ between students in different semesters are difficult to define without a fixed curriculum.

Various predictors were considered in this paper. First, gender was taken into account, as it almost always is relevant to the level of economic competence (Brückner, et al., 2015a). Female students tend to perform worse than male students on economic tasks, at least in western countries such as Germany or the United States of America (e.g., Beck & Krumm, 1992; Owen, 2012; Walstad & Robson, 1997; Williams, Waldauer, & Duggal, 1992). Second, prior education was examined because in cross-sectional analyses the completion of vocational training prior to studying often has a significant impact on economic competence in higher education (Brückner et al., 2015b). Similarly, attending advanced business and economics courses in secondary school can influence the level of economic competence of students in higher education (Gill & Gratton-Lavoie, 2011). Third, in addition to sociodemographic factors and personal factors, individual factors, in particular, such as motivational orientations and intelligence were taken into consideration, as they can have an impact on the level and growth of competence (cf. Shulman, 1970; Alexander et al., 1995; Hambrick et al., 2008). Intelligence as general cognitive ability usually is associated with competence measures because of the g-factor and the general relationship among learning, cognitive abilities (such as speed of information processing and transferring available knowledge to solve new problems), and the acquisition of competence (cf. Ackermann, 1988; Coyle & Pillow, 2008; Glaser, 1991). Motivational orientations influence the growth of competence in terms of the degree of self-determination during the process of acquisition as described by Ryan and Deci (2000). They distinguish between extrinsic motivation and intrinsic motivation. Extrinsic motivation can be described as a construct that pertains to an activity being is done to attain some separable outcome (e.g., better grades on exams). Extrinsic motivation is outcome-oriented and remains rather stable throughout the course of studies. Intrinsic motivation is defined as the doing of an activity for inherent satisfaction rather than for some separable consequence. Several studies have shown that both types of motivation play a role in higher education. Intrinsic motivation as the drive to study because the content and activity of studying are interesting is rather dynamic in nature (Ryan & Deci, 2000).

In addition to fixed effects, there are random effects which express the unobserved heterogeneity in economic competence among students within the same semester and among students in different semesters. The level of unobserved heterogeneity of students within a semester (as asked in question 3) is represented by τ^2 and of unobserved heterogeneity of students in different semesters (as asked in question 4) is represented by σ^2. If τ^2 is greater than zero, it can be assumed that the level of competence of students within the same semester is different and the goal is to explain some of the variance τ^2 with the aforementioned relevant predictors. If σ^2 is greater than zero, it can be assumed that the average level of competence differs between semesters. These differences also could be explained by level 1 predictors.

Instruments and Sample Size

To determine the growth of competence of students in business and economics over the course of their studies and within and between semesters, data from three repeated assessments over six semesters within the Innovative Teach-Study Network in Academic Higher Education[3] (ILLEV) research project were analyzed. In ILLEV longitudinal surveys were conducted in three one-year intervals: autumn 2009, 2010, and 2011 (cf. Happ et al., 2013; Happ et al., in press (b)). From a cross-sectional view, the sample consisted of 770 students in 2009 (T1), 1,279 students in 2010 (T2), and 1,239 students in 2011 (T3). To do longitudinal analysis it is important to have large numbers of students and to assess them repeatedly. However, due to the above-mentioned challenges of conducting longitudinal research in higher education (see Section 1), wave-nonresponses also were an issue in the current study. Only approximately 20% of all students in the sample could be assessed more than once (see Table 1).

Table 1.
Sample Sizes of Students Assessed at All Three Measurement Occasions

Freq.	**Percent**	**Cum.**	**Pattern**
795	30.11	30.11	010
787	29.81	59.92	001
484	18.33	78.26	100
288	10.91	89.17	011
122	4.62	93.79	110
90	3.41	97.20	101
74	2.80	100.00	111
2640	100.00		

From this 20%, it was possible to gather data at all three measurement occasions on 74 students only. In addition to having to handle a large amount of missing data, identifying the kind of missing mechanism (i.e., missing completely at random (MCAR), missing at random (MAR), or missing not at random (MNAR)) and determining how best to deal with it (cf. Rubin, 1976; Little & Rubin, 2002) were issues in this study. Data on students who participated only once were assumed to be MNAR for the first analyses such as the one presented in this paper and they could not be used to measure growth of competence. Rather, they were used to determine the level of competence at one measurement occasion only. Data on students who were assessed at least twice should be included in the analysis of competence growth but the growth model (presented in Section "Methods and Results") was first formulated and tested on students without wave-nonresponses. Developing a model that can handle wave-nonresponses is beyond the scope of this paper but will be the aim of further analyses based on the pretest-posttest-posttest model (see Section "Discussion and Conclusion").

The following analysis is of data on the 74 students whose economic competence was assessed at three occasions during various semesters (see Table 2). The distribution of T1 should actually just be shifted two semesters because all the students in their first semester at T1 should have been in their third semester at T2, and in their fifth semester at T3, and so on. Table 2 shows that from T1 to T2 a student in his/her second semester in T1 was assigned to semester 3 in T2, which was possible if he/she did not study for a semester, for example, due to an internship. In addition, there is unequal distribution of students in even and in odd semesters because students in Germany usually start their studies in autumn and the assessments always took place in autumn. Therefore, students who started their studies in spring were not underrepresented; they merely were less common.

Table 2
Distribution over semesters at T1, T2, and T3

Semester	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**
T1	43	2	14	3	6	2	2	1	1	0	0	0	0
T2	0	0	44	1	15	2	6	2	2	1	1	0	0
T3	0	0	0	0	44	1	15	2	6	2	2	1	1

To assess economic competence and its growth items from the validated standardized Wirtschaftskundlicher Bildungstest (WBT) by Beck, Krum, and Dubs (1998), which is the German adaption of the Test of Economic Literacy (TEL) by Soper and Walstad (1987), were administered. Soper and Walstad (1987) developed the TEL to permit differentiation between relatively low and high development levels of economic knowledge and understanding. As described in Happ et al. (2013), the measurement features and quality factors of the WBT have been researched and validated for both the English and German versions of the test (Beck & Krumm, 1989, 1992; Beck, Krumm, & Dubs, 2001; Soper & Brenneke, 1981; Soper & Walstad, 1987). Although the WBT was designed to assess economic knowledge and understanding of students in vocational business training (Beck et al., 2001), several of the items on the test are appropriate for measuring economic knowledge of university-level students, which has been proven in ILLEV (cf. Beck et al., 1998; Happ et al., 2013). Therefore, particular attention was paid to the possible occurrence of ceiling effects and item selectivity index (Happ et al., 2013). Furthermore, to guarantee curricular validity of the WBT items, the curricula of the participants' business and economics programs were analyzed and lecturers of the relevant classes were surveyed (Happ & Zlatkin-Troitschanskaia, 2014). Consequently, only the WBT items with curricular relevance and appropriate difficulty were employed (Happ et al., 2013). Thus, these selected WBT items are adequate to use in the assessment of economic knowledge and understanding in higher education.

There are two parallel versions of the WBT each consisting of 46 items (which includes 15 anchor items allowing comparison of the two versions). The data analyzed in this paper were responses to 19 items on the original WBT. Each item had one correct answer and three distractors. The items selected for assessments within the ILLEV study were those appropriate for university-level students. Over the three measurement occasions, the same 19 items were administered to 2,640 students of business and economics to determine the level and trace the growth of their economic competence (Cronbach's alpha = .65). In addition, the effects of a number of personal traits as well as structural and individual factors influencing the level and growth of economic competence (see Happ et al., in press (a)) were examined on the questionnaire. In line with current research and prior analyses (cf. Happ et al., 2013; Happ, et al., in press (b); Schmidt, in prep.) the seven predictors described in Section 2.1 were explored in the present analysis to explain differences in economic competence over time. Of the 74 students that could be assessed repeatedly 32 were male (43%) and 42 female (57%), and 22 (30%) had completed vocational training and 33 (45%) had attended an advanced course in economics at school before starting their studies. On the ILLEV questionnaire students were asked about their extrinsic motivation and intrinsic motivation with four items for each dimension adapted from a standardized and validated instrument by Schiefele, Moschner, and Hustegge (2002). Students' extrinsic motivation was assessed by analyzing their responses to questions about expectations in terms of having a good job after completing their studies. Students' intrinsic motivation was assessed by analyzing their responses to questions about how interesting and enjoyable they found the content of their business and economics studies. Cronbach's alpha for extrinsic motivation and intrinsic motivation was approximately .87 (based on the whole sample with n=2,640 as with the WBT reliability). To measure general cognitive abilities, the students responded on the ILLEV questionnaire to items on two subscales (analogies and numerical series) taken from the Intelligence Structure Test (IST 2000) by Liepmann, Beauducel, Brocke, and Amthauer (2007), which is a test commonly used in Germany to assess intelligence. However, only the analogy scale measuring verbal intelligence could be used because the scale on numeric intelligence showed ceiling effects. Therefore, at each measurement occasion 20 items on analogies were administered (Crobach's alpha=.65 for n=2,640). As a further indicator of cognitive abilities, on the ILLEV questionnaire the students were asked about their average school leaving grade (GPA[4]). GPA ranged from 1 (highest level of ability) to 6 (lowest level of ability). The average leaving grade of the 74 students was approximately 2.3 (see Table 3).

Table 3
Descriptive Statistics for Individual Motivational and Cognitive Factors (n=74)

		Mean	S.d.*	Minimum	Maximum
GPA		2.272	0.523	1.2	3.3
Verbal intelligence		0.045	0.679	-1.398	1.617
Intrinsic motivation	**T1**	0.344	1.034	-2.080	1.983
	T2	0.036	0.971	-2.194	1.965
	T3	0.118	0.861	-2.472	1.981
Extrinsic motivation		-0.008	0.697	-1.623	1.536

*S.d. = standard deviation.

Because verbal intelligence and both motivational facets are not directly observable, as underlying constructs which are measured with several questions in standardized test instruments, they must be inferred from test scores. In the present paper, the test scores for intelligence as well as for extrinsic motivation and intrinsic motivation were employed as factor scores as expected a posteriori (EAP) estimates (for a detailed description see next Section “Method and Results”, as this kind of EAP is estimated within the newly developed pretest-posttest-posttest model presented in this paper). These EAPs as empirical Bayes estimates (cf. Bock & Aitkin, 1981; Skrondal & Rabe-Hesketh, 2004) are generated with the multilevel IRT (MLIRT) model by Fox (2007) whereby students are clustered within occasions to take into account the longitudinal structure of the data and, accordingly, the dependencies of observations. The model estimates EAPs for each student across all measurement occasions as well as for each student at each separate measurement occasion. For extrinsic motivation and intelligence, which are assumed not change over the course of studies, the overall EAPs for each student were used. In contrast, because intrinsic motivation is assumed to change over time, for each student the EAPs from each measurement occasion were used. As shown in Table 3, the mean verbal intelligence of the sample of 74 students was approximately 0.05 on the latent ability scale. It ranged from -1.4 to 1.6 and had a standard deviation of approximately 0.7. The mean intrinsic motivation varied over time and was highest at T1 with a mean of 0.3 and lowest at T2 with a mean of 0.04. At T3 intrinsic motivation was on average approximately 0.1 on the latent ability scale (for corresponding standard deviations and ranges see Table 3). Extrinsic motivation had a mean of -0.008 with a standard deviation of 0.7 ranging from -1.6 to 1.5.

Method and Results

A Pretest-Posttest-Posttest Multilevel IRT Model to Assess the Growth of Competence of Semester Cohorts

Following the pretest-posttest multivariate multilevel model for two measurement occasions by Keuning, van Geel, Visscher, and Fox (2015) a further extended model was developed as an extension of the multilevel IRT model by Fox and Glas (2001) and Fox (2010) to analyze the growth of competence over more than two measurement occasions. This extended model can be employed to analyze hierarchical and longitudinal data for any number of occasions. As students were assessed at three occasions in this study, the extended model was applied and is presented for three waves of data (pretest-posttest-posttest model).

As shown in Figure 2, a multilevel model was needed because observations of students were clustered in semesters. By doing this, students within one semester were treated as one cohort and dependencies between student scores within one cohort were considered by the multilevel framework. Furthermore, the multilevel modeling approach can be applied to the measurement model, where the ability such as economic competence is treated as a latent variable rather than an observed variable (Fox, 2001). By doing this an IRT model is used to describe the relationship between the latent variable and test items (Fox, 2007). In the present study, economic competence was assessed using items from the WBT, each of which had four possible responses but only one correct response. This led to binary responses with value 0 for an incorrect response and 1 for a correct response. According to Fox (2007), who described the IRT model formulation regardless of the number of measurement occasions, the probability of responding correctly at a given occasion *t* to item *k* by student *i* in semester *j* is given by

$$P\left(y_{ijk}^{t} = 1 \,\middle|\, \theta_{ij}^{t}, a_{k}^{t}, b_{k}^{t}\right) = \Phi\left(a_{k}^{t}\theta_{ij}^{t} - b_{k}^{t}\right),$$

where a_k^t and b_k^t represent item discrimination and difficulty parameter for item k at occasion *t*, respectively, and $\Phi(\cdot)$ represents the cumulative normal distribution function. One requirement for repeated measurement models is measurement invariance: items are supposed to measure the same construct over time (Horn & McArdle, 1992) and item discrimination and difficulty parameters must be constant over time. In the present study, all items were assumed to be invariant.[5] However, the newly developed pretest-posttest-posttest multivariate multilevel IRT model could handle items showing item drift (see Bock, Murakl, & Pfeiffenberger, 1988) over time as well as an incomplete test design with a few anchor items and different items at measurement occasions. Items with different discrimination and difficulty parameters could be specified across measurement occasions in the pretest-posttest-posttest model.

The structural part of the model, where the relationship between the latent variable economic competence θ_{ij}^t and other observed predictor variables X_{ij}^t were considered, was given by

$$\theta_{ij}^t = \beta_{0j}^t + X_{ij}^t \beta^t + e_{ij}^t$$

as level 1 equation and with $e_{ij}^t \sim \mathcal{N}(0, \tau^2)$, whereas τ^2 represented the residual variance at level 1. The parameter β_{0j}^t represented the semester-specific mean of economic competence in semester j and occasion *t*. As usual in multilevel models, the β_{0j}^t as random intercept was formulated as a level 2 equation by

$$\beta_{0j}^t = \gamma_{00}^t + u_{0j}^t \text{ with } u_{0j}^t \sim \mathcal{N}(0, \begin{pmatrix} \sigma_{t1t1}^2 & \sigma_{t1t2}^2 & \sigma_{t1t3}^2 \\ \sigma_{t1t2}^2 & \sigma_{t2t2}^2 & \sigma_{t2t3}^2 \\ \sigma_{t1t3}^2 & \sigma_{t2t3}^2 & \sigma_{t3t3}^2 \end{pmatrix}),$$

whereas the covariance matrix u_{0j}^{t} represented the unobserved heterogeneity at level 2. The covariance matrix u_{0j}^{t} included the variance parameter of the error term at level 2 for each occasion (the diagonal of the matrix). At each measurement occasion, variance represented the variability in semester-average student scores. When variance increased over time, the semester-average scores differed more over time. Each covariance parameter represented the covariance between semester-average student scores measured on two occasions. The multivariate multilevel model defined a common correlation between the semester scores over occasions. Therefore, each covariance parameter defined the assumed common correlation between average-semester scores at two measurement occasions. The modeled covariance between semester-average scores over time also was used to link the occasion-specific scales to each other.

The parameter γ_{00}^{t} represented the grand-mean of economic competence of the population at each occasion *t*. So, the occasion-specific average and the overall change in economic competence could be inferred from differences between γ_{00}^{1}, γ_{00}^{2} and γ_{00}^{3}.

The level 1 equation showed that explanatory variables X_{ij}^{t} could be considered. At level 2, (time-specific) explanatory variables differentiating between semesters could be included, but they were not available for the present study. Only (time-specific) student variables were available and were included as level-1 explanatory variables. These student-level predictors could explain variability between semesters.

The model could be identified by restricting the mean and variance of the latent scale at occasion T1. One or more anchor items were needed to link the latent scales over time.

The joint estimation of the multivariate multilevel IRT model with correlations modeled at level 2 between occasions was implemented in the statistical software R (R Core Team, 2014) as an extension of the mlirt package (Fox, 2007) and of the package by Keuning et al. (2015). In this implementation, an MCMC algorithm was conducted. In this Bayesian approach all sources of uncertainty were taken into account in the estimation of the model parameters, and all parameters at all occasions were estimated simultaneously.[6] To make inferences about the effects of the estimated parameters, highest posterior density (HPD) intervals were calculated. The HPD intervals are a way to test for significance. The region of the interval in the present paper represented the 95% highest posterior probability of the parameter.

Results from the ILLEV Study

In Table 4 the results for the empty model, meaning the model without predictors, are presented. The intercept represents the overall score in economic competence for each occasion. The intercept increased constantly from 0.3 at T1 (with a standard deviation of 0.3 and an HDP interval from -0.3 to 1) to over 0.7 at T2 (with a standard deviation of 0.5 and an HDP interval from -0.3 to 1.6) to 1.6 at T3 (with a standard deviation of 0.4 and an HDP interval from 0.7 to 2.4). The economic competence of all students over all semesters increased on average over time.

Table 4

Results of the Empty Model – Development of Economic Competence of Students Clustered According to Semester

	T1			**T2**			**T3**		
	Est.	S.D.	HPD	Est.	S.D.	HPD	Est.	S.D.	HPD
Fixed Effects									
Intercept	0.305	0.340	[-0.3,1.0]	0.691	0.478	[-0.3,1.6]	1.592	0.433	[0.7,2.4]
Random Effects									
Var. Level 2	0.541	0.459	[0.1,1.7]	1.506	1.312	[0.3,4.9]	1.263	1.164	[0.2,4.3]
Var. Level 1	0.855	0.201	[0.5,1.3]	2.443	0.743	[1.1,3.9]	2.282	0.739	[1.0,3.7]
ICC	0.387			0.381			0.356		
Covariance									
T1T2	0.248	0.485	[-0.5,1.4]						
T1T3	0.267	0.437	[-0.4,1.3]						
T2T3	0.642	0.826	[-0.4,2.6]						

Est. = Estimated coefficient, S.d.= Standard deviation, HPD= Highest posterior density interval

The intraclass correlation (ICC), which represents the proportion of variance due to clustering students according to semester, shows that at all occasions approximately 36% to 39% of the variance was explained by clustering students in semesters. However, despite similar proportions of variance, as we can see in the ICC, level 1 variance and level 2 variance increased over time and were greatest at T2. This means that the students' responses to the WBT items between semesters as well as within semesters were less heterogeneous at T1 than at T2 and at T3. The covariance between two measurement occasions represents the covariance between the average semester scores at the two occasions. The positive covariance means that the average semester scores correlated over time. However, the sample size was too small to make inferences about the correlations between average scores over time.

Table 5
Results of the Final Model - Development of Economic Competence with Predictors

	T1			**T2**			**T3**		
	Est.	S.D.	HPD	Est.	S.D.	HPD	Est.	S.D.	HPD
Fixed Effects									
Intercept	0.545	0.450	[-0.4, 1,4]	0.811	0.825	[-0.8,2.6]	2.568	0.693	[1.2,3.8]
Female	-0.436	0.267	[-0.9,0.1]	-0.633	0.461	[-1.5,0.3]	-0.955	0.405	[-1.7,-0.1]
Verbal intel.	0.207	0.206	[-0.2,0.6]	0.754	0.320	[0.1,1.4]	1.017	0.225	[0.6,1.4]
GPA	-0.040	0.139	[-0.3,0.2]	-0.633	0.212	[-1.1,-0.2]	-0.152	0.164	[-0.5,0.2]
Intrinsic Mot.	0.210	0.128	[0.0,0.5]	0.311	0.216	[-0.1,0.7]	0.018	0.174	[-0.4,0.3]
Extrinsic Mot.	0.125	0.200	[-0.3,0.5]	0.036	0.307	[-0.6,0.6]	-0.075	0.241	[-0.5,0.4]
Voc. Training	0.024	0.281	[-0.5,0.6]	0.228	0.459	[-0.7,1.1]	1.098	0.401	[0.2,1.8]
Adv. courses	0.285	0.275	[-0.3,0.8]	0.192	0.467	[-0.7,1.1]	0.148	0.317	[-0.50.8]
Random Effects									
Var. Level 2	0.552	0.465	[0.1,1.7]	0.955	0.941	[0.2,3.4]	0.941	0.950	[0.2,3.5]
Var. Level 1	0.793	0.218	[0.4,1.3]	1.958	0.570	[1.0,3.2]	0.804	0.338	[0.3,1.5]
Covariance									
T1T2	0.183	0.421	[-0.4,1.2]						
T1T3	0.158	0.380	[-0.5,1.0]						
T2T3	0.283	0.564	[-0.5,1.6]						

Est. = Estimated coefficient, S.d.= Standard deviation, HPD= Highest posterior density interval

Next, a model that included all relevant predictors as mentioned in Section 2 was employed. Table 5 shows the results of the model with all predictors. Here, the intercept represented the average economic competence of male students across all semesters, with the value 0 in the latent ability scores for verbal intelligence, extrinsic motivation and intrinsic motivation, average GPA (the predictor of which was mean-centered), and no prior economic education. The intercept scores were approximately 0.25 higher on the latent ability scale at T1, 0.12 higher at T2, and 1 higher at T3 in this model (see Table 5) than in the empty model (see Table 4). Further, female students had a lower level of economic competence than male students with estimated differences between 0.4 at T1, 0.6 at T2 and approximately 1.0 at T3. This means that as competence increased, the difference in the amount of competence between male students and female students also increased. This was the case for verbal intelligence and for students who had completed vocational training. So, students who exhibited greater abilities on the intelligence test and students who had completed vocational training showed a higher level of economic competence at all occasions. Moreover, competence grows more quickly for less intelligent students and for those who had not completed vocational training.

The effect of the GPA was negative because higher grades, which are not *better* grades in the German education system, reflect less cognitive ability: Students with grades higher than 2.3 exhibited less economic competence than students with average or above average grades. However, the development of this effect over time was not steady as it was for verbal intelligence. Students with better GPAs developed their competence more quickly between T1 and T2 and between T2 and T3.

Concerning effects of motivation, high levels of intrinsic motivation resulted in greater economic competence at all measurement occasions whereas high levels of extrinsic motivation led to higher scores in ability in the construct of economic competence at T1 and T2 only. At T3, when competence was already at a very high level, greater extrinsic motivation accompanied lower levels of competence.

Finally, attending advanced courses in economics at school improved higher education students’ economic competence; however, the improvement decreased over time while economic competence grew. Students who had completed these advanced courses had a higher level of competence at T1 of approximately 0.3, at T2 of approximately 0.2, and at T3 of approximately 0.1 on the latent ability scale. Students who had not attended such courses at school were able to compensate for this drawback over time as their competence grew while studying.

Discussion and Conclusion

In this paper, various challenges of longitudinal research in higher education have been described. In addition, the need for appropriate models for analyzing students’ growth of competence in higher education and obtaining valid and generalizable results thereof has been argued. Valid assessment of students’ growth of competence in higher education requires sophisticated methodological designs and statistical methods for analyzing hierarchical longitudinal data. Traditional methods such as multilevel analysis (e.g., Singer & Willett, 2003; Snijders & Bosker, 2012) and latent growth curve analysis based on structural equation models (SEM; e.g., Bollen & Curran, 2006) cannot be employed to address questions about the growth of competence if the data is affected by missing cases due to panel dropouts (wave nonresponses) and if the measured constructs are not directly observable, as with students’ competences (Little, Lindenberger, & Maier, 2008). However, both of these traditional methods have their strengths which can be combined in a multivariate multilevel structural equation (or IRT resp.) model for repeated measures.

In this paper, a newly developed multivariate multilevel IRT model for repeated measures was introduced which was designed to address those typical challenges of longitudinal research in higher education and was tested with genuine data. The model was estimated within a

Bayesian framework with an MCMC algorithm. In this model, dependencies between repeated observations of the same students are considered not, as usual, by clustering observations within participants but rather by clustering observations within semesters. This way, missing values from single students at one or more measurement occasions presented no challenge (as long as they were missing at random), because growth of competence was measured as growth of semester cohorts. This means that focus was placed on growth patterns of different semester cohorts without conditioning on unreliable individual growth trajectories based on a small number of measurements. The multivariate modeling component accounted for dependencies between scores of cohort members over time. The multilevel component accounted for the nesting of students in semesters at each measurement occasion. Using real data from ILLEV project results, the multivariate multilevel modeling approach was shown to be particularly relevant to analyzing longitudinal data collected over a limited number of measurements occasions but with many students clustered in higher-level units measured on each occasion.

Some limitations of this study should be discussed. The new model presented in this paper was applied to data without wave nonresponses. Therefore, the results should be interpreted with caution because the sample and the number of clusters were small. A model is needed that can represent students with missing observations in terms of, for example, panel dropouts. Such a model exists but needs to be tested more thoroughly before it can be applied to genuine data. With such a model, missing values from single students at one or more measurement occasions will not present a challenge (as long as they are missing at random), because growth in competence still will be measured as growth of semester cohorts.

In follow-up studies, the limitations of the approach highlighted in this paper should be explored in more detail, especially taking into account sampling procedures, the test instrument, and the operationalization of students' competence. Many challenges remain, including more appropriate modeling and measuring of students' learning outcomes and learning gains in higher education based on a broad concept of competence in higher education which includes not only content knowledge but also motivational orientations, epistemological beliefs, and so on (see section "Theoretical Foundation and Study Design"). This concept invites an equally broad range of assessment approaches focusing on students' and/or graduates' knowledge, skills, and motivational, volitional, and social dispositions and using innovative methods such as computer-based adaptive testing.

Further significant challenges lie in the test instrument. The multidimensional and context-specific characteristics of students' competence complicate the development of measurement instruments. Internationally, there are few reliable and valid instruments to assess students' competence in higher education. Therefore, in spite of the challenges of assessing students' learning outcomes and learning gains, specifically in higher education, more research as well as objective, reliable, and valid models and instruments are needed to assess students' knowledge and skills and, thus, improve educational measurement practices in the respective domains of higher education.

To provide more reliable and generalizable findings on the higher education system in Germany and its institutions, future research should address the challenge of drawing random samples of institutions and students. Further validation criteria should be examined for in-depth validation. Analyses can focus on the five validation criteria laid out in the international Standards for Educational and Psychological Testing (AERA, APA, & NCME 2015).

Based on the above limitations and perspectives, various implications and avenues are worth considering in future research. For instance, future analyses may involve exploring further potential influence factors. Greater focus may be placed on the content of studies and on instructional practices in economics in higher education. Brückner et al. (2015a) have found systematic effects of the content of economic items based on their verbal and mathematical components. Thus, another focus for future research could be in-depth analysis of differences in the effects of personal or study-related characteristics on economic numeracy and economic literacy. Furthermore, comparative analyses with additional educational institutions and types of study models are needed. Zlatkin-Troitschanskaia et al. (in press) and Brückner et al. (2015b) have found systematic effects of the type of institution, for example universities and universities of applied sciences in Germany. Future research on students' growth of competence in higher education should have experimental research designs, for example, with several instructional formats, additional comprehension tests, and qualitative explanatory methods such as think aloud (see Brückner et al., in press). Such studies and their results would offer a valuable basis for drawing very important practical conclusions about teaching and assessment practices in economics in higher education.

Notes

[1] Programme for International Student Assessment (for further information, see http://www.oecd.org/pisa/).

[2] German National Educational Panel Study (for further information, see https://www.neps-data.de/en-us/home.aspx)

[3] The ILLEV project was financed by Germany's Federal Ministry of Education and Research (BMBF) (grant no. 01PH08013).

[4] The average school leaving grade is similar to grade point average in the United States of America and therefore is referred to hereafter as GPA.

[5] In Happ et al. (in press (b)) the longitudinal properties of the WBT items were verified and documented.

[6] For further details on the MCMC algorithm, see Fox (2007, 2010).

References

Ackerman, P. L. (1988). Determinants of individual differences during skill acquisition: cognitive abilities and information processing. *Journal of experimental psychology: General, 117*(3), 288.

AERA, APA, & NCME (2014). *Standards for Educational and Psychological Testing.* Washington DC: American Educational Research Association.

Alexander, P. A., Jetton, T. L. & Kulikowich, J. M. (1995). Interrelationship of knowledge, interest, and recall: Assessing a model of domain learning. *Journal of Educational Psychology, 87*(4), 559–575.

Artelt, C., Weinert, S., Carstensen, C. H. (2013). Assessing competencies across the lifespan within the German National Educational Panel Study (NEPS) – Editorial. *Journal for educational research online 5*(2), 5-14.

Beck, K. & Krumm, V. (1989). Economic Literacy in German Speaking Countries and the United States. First Steps to a Comparative Study. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA, USA, March 27-31, 1989. *ERIC, Microfiche No. ED 310 027*, 1-23.

Beck, K. & Krumm, V. (1992). *Economic Literacy in the United States, Germany, and Austria: Results of cross national studies.* Paper presented at the Annual Meeting of the Joint Council on Economic Education and the National Association on Economic Education JCEE/NAEE, Los Angeles, USA. Nov. 09, 1990. ERIC. Microfiche No. ED 340 629, 1-65. (Abstract in Resources in Education, May 1992).

Beck, K., Krumm, V. & Dubs, R. (1998). *Wirtschaftskundlicher Bildungs-Test* (WBT). Göttingen: Hogrefe.

Beck, K., Krumm, V. & Dubs, R. (2001). WBT - Wirtschaftskundlicher Bildungstest. In W. Sarges & H. Wottawa (Eds.), *Handbuch wirtschaftspsychologischer Testverfahren* (pp. 559–562). Lengerich: Pabst.

Bijleveld, C. J. H & Kamp, L. J. T. van der (1998). *Longitudinal data analysis. Designs, models and methods*. London: Sage.

Blömeke, S., Zlatkin-Troitschanskaia, O., Kuhn, C. & Fege, J. (2013). Modeling and measuring competencies in higher education: Tasks and challenges. In S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Eds.), *Modeling and measuring competencies in higher education* (pp. 1-12). Rotterdam: Sense Publishers.

Bock, R. D. & Aitkin, M. (1981). Marginal maximun likelihood estimation of item parameters: An application of a EM algorithm, *Psychometrika*, *46*, 179-197

Bock, R. D., Murakl, E. & Pfeiffenberger, W. (1988). Item pool maintenance in the presence of item parameter drift. *Journal of Educational Measurement*, *25*(4), 275–285.

Bollen, K. A. & Curran, P. J. (2006). *Latent curve models. A structural equation perspective*. Hoboken, N.J.: Wiley-Interscience.

Brückner, S., Förster, M., Zlatkin-Troitschanskaia, O., Happ, R., Walstad, W. B., Yamaoka, M. et al. (2015a). Gender effects in assessment of economic knowledge and understanding: Differences among undergraduate Business and Economics students in Germany, Japan, and the United States. *Peabody Journal of Education*, *90*(4), 503–518.

Brückner, S., Förster, M., Zlatkin-Troitschanskaia, O. & Walstad, W. B. (2015b). Effects of prior economic education, native language, and gender on economic knowledge of first-year students in higher education. A comparative study between Germany and the USA. *Studies in Higher Education*, *40*(3), 437–453.

Brückner , S. (in press). Die Analyse mentaler Prozesse von Studierenden beim Lösen wirtschaftswissenschaftlicher Testaufgaben in einem prozessbezogenen Validierungsansatz. Landau: Verlag Empirische Pädagogik.(Doctoral Dissertation).

Coertjens, L., Donche, V., Maeyer, S. de, Vanthournout, G. & Petegem, P. van (2013). Modeling change in learning strategies throughout higher education: a multi-indicator latent growth perspective. *PloS one*, *8*(7), e67854.

Coyle, T. R. & Pillow, D. R. (2008). SAT and ACT predict college GPA after removing g. *Intelligence*, *36*(6), 719–729.

Federal Statistical Office (2013). *Bildung und Kultur - Studierende an Hochschulen* [Education and culture – students in higher education]. Wiesbaden: Statistisches Bundesamt [Federal Statistical Office]. Available at https://www.destatis.de/DE/Publikationen/Thematisch/BildungForschungKultur/Hochschulen/StudierendeHochschulenEndg2110410137004.pdf?__blob=publicationFile

Federal Statistical Office (2014). *Long term series of the total number of students in Germany*. Available at https://www.destatis.de/EN/FactsFigures/Indicators/LongTermSeries/Education/lrbil01.html.

Fox, J.-P. (2001). *Multilevel IRT: A Bayesian perspective on estimating parameters and testing statistical hypotheses*. Enschede: University of Twente.

Fox, J.-P. & Glas, C. A. W. (2001). Bayesian estimation of a multilevel IRT model using Gibbs sampling. *Psychometrika, 66*(2), 271-288.

Fox, J.-P. (2007). Multilevel IRT modeling in practice with the package mlirt. *Journal of Statistical Software, 20*(5), 1 - 16.

Fox, J.-P. (2010). *Bayesian Item Resonse Modeling – Theory and applications*. New York: Springer.

Gill, A., & C. Gratton-Lavoie, C. (2011). Retention of high school economics knowledge and the effect of the California state mandate. *The Journal of Economic Education 42(*4). 319–37.

Glaser, R. (1991). The maturing of the relationship between the science of learning and cognition and educational practice. *Learning and instruction, 1*(2), 129-144.

Hambrick, D. Z., Pink, J. E., Meinz, E. J., Pettibone, J. C. & Oswald, F. L. (2008). The roles of ability, personality, and interests in acquiring current events knowledge: A longitudinal study. *Intelligence, 36*(3), 261–278.

Happ, R., Kuhn, C. & Zlatkin-Troitschanskaia, O. (in press (a)). Effects of the structural and curricular changes following the bologna reform in Germany on the content knowledge and pedagogical content knowledge of student teachers of business and economics. In A. Aelterman, B. De Wever, M. Tuytens & R. Vanderlinde (Eds.), *Professional learning of teacher educators, teachers, and student teachers: Research-based practices.* Academia Press, Ginkgo-imprint.

Happ, R. & Zlatkin-Troitschanskaia, O. (2014). Assessing the Quality of the Bachelor Degree Courses: An Analysis of their Effects on Students' Acquisition of Economic Content Knowledge. Zeitschrift für Hochschulentwicklung, 9(2), 64-77.

Happ, R., Zlatkin-Troitschanskaia, O., Beck, K. & Förster, M. (in press (b)). Increasing Heterogeneity in Students' Prior Economic Content Knowledge – Impact on and Implications for Teaching in Higher Education. In E. Wuttke, J. Seifried & S. Schumann (Eds.), Economic Competence of Young Adults in European Countries. Opladen: Barbara Budrich Publishers.

Happ, R., Zlatkin-Troitschanskaia, O., Förster, M., Preuße, D., Kuhn, C. & Schmidt, S. (2013): *The research project ILLEV (Innovative Teach-Study-Network in Academic Higher Education) – A Short Summary of the Main Results.* Mainz: Johannes Gutenberg-Universität. Available at: http://www.wipaed.uni-mainz.de/ls/ArbeitspapiereWP/gr_Nr._65.pdf

Hof, M. A. van't, Roede, M. J. & Kowalski, C. J. (1977). A mixed longitudinal data analysis model. *Human Biology, 49(*2), 165-179.

Horn, J. L. & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental Aging Research: An International Journal Devoted to the Scientific Study of the Aging Process, 18*(3), 117-144.

Keuning, T., van Geel, M., Visscher, A. J. & Fox, J.-P. (2015). *The development of teaching quality during a DBDM-Intervention: A pre-post multilevel IRT approach Trynke Keuning, Marieke van Geel, Adrie Visscher, Jean-Paul Fox. Submitted to AERJ (2015).*

Liepmann, D., Beauducel, A., Brocke, B., & Amthauer, R. (2007). I-*S-T 2000 R: Intelligenz-Struktur-Test 2000R.* Göttingen: Hogrefe.

Little, T. D., Lindenberger, U. & Maier, H. (2008). Selectivity and generalizability in longitudinal research: On the effects of continuers and dropouts. In T. D. Little, K.-U. Schnabel, J. Baumert (Eds.), *Modeling longitudinal and multilevel data. Practical issues, applied approaches and specific examples* (pp. 187–200). Mahwah: Erlbaum.

Little, R. J. A. & Rubin, D. B. (2002). *Statistical analysis with missing data.* New York: Wiley.

Nusche, D. (2008). Assessment of learning outcomes in higher education: A comparative review of selected practices. OECD Education Working Paper No. 15. Available at http://www.oecd.org/edu/skills-beyond-school/40256023.pdf

Owen, A. L. (2012). Student characteristics, behavior, and performance in economics classes. In G. M. Hoyt & K. McGoldrick (Eds.), *International handbook on teaching and learning economics* (pp. 341-350). Cheltenham: Edward Elgar.

Ployhart, R. E. & Vandenberg, R. J. (2009). Longitudinal research: The theory, design, and analysis of change. *Journal of Management, 36*(1), 94–120.

R Core Team (2014). *R: A language and environment for statistical computing.* Vienna, Austria: R foundation for statistical computing.

Rabe-Hesketh, S. Skrondal, A. and Gjessing, H.K. (2008). Biometrical modeling of twin and family data using standard mixed model software. *Biometrics 64*(1), 280-288.

Robbins, S. B., Lauver, K., Le, H., Davis, D., Langley, R. & Carlstrom, A. (2004). Do Psychosocial and Study Skill Factors Predict College Outcomes? A Meta-Analysis. *Psychological Bulletin, 130*(2), 261–288.

Rubin, D. B. (1976). Inference and missing data. *Biometrika, 63*(3), 582–592.

Ryan, R. M. & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology, 25* (1), 54–67.

Schaap, L., Schmidt, H. G. & Verkoeijen, P. P. (2012). Assessing knowledge growth in a psychology curriculum: which students improve most? *Assessment & Evaluation in Higher Education, 37*(7), 875–887.

Schiefele, U., Moschner, B. & Husstegge, R. (2002). *Skalenhandbuch SMILE-Projekt.* Bielefeld: Universität Bielefeld, Abteilung für Psychologie.

Schmidt, S. (in prep.). *Eine Längsschnittanalyse der Veränderung von betriebswirtschaftlichem Fachwissen bei Studierenden der Wirtschaftswissenschaften und der Wirtschaftspädagogik [A longitudinal analysis of change of business administration knowledge among students in business education and business administration and economics]*. (Doctoral dissertation in preparation.)

Shavelson, R. J. (2013). An approach to testing & modeling competence. In S. Blömeke, O. Zlatkin-Troitschanskaia, C. Kuhn & J. Fege (Eds.), *Modeling and measuring competencies in higher education* (pp. 19-34). Rotterdam: Sense Publishers.

Shulman, L. S. (1970). Cognitive Learning and the Educational Process. *The Journal of Medical Education, 45*(11), 90–100.

Singer, J. D. & Willett, J. B. (2003). *Applied longitudinal data analysis. Modeling change and event occurrence*. Oxford: Oxford Univ. Press.

Skrondal, A. & Laake, P. (2001). Regression among factor scores. *Psychometrika, 66*(4), 563–575.

Skrondal, A. & Rabe-Hesketh, S. (2004). *Generalized latent variable modeling. Multilevel, longitudinal, and structural equation models*. Boca Raton: Chapman & Hall.

Snijders, T. A. B. & Bosker, R. J. (2012). *Multilevel analysis. An introduction to basic and advanced multilevel modeling* (2. ed.). London: Sage.

Soper, J. C. & Brenneke, J. S. (1981). The test of economic literacy and an evaluation of the DEEP system. *Journal of Economic Education*, *12*(2), 1-14.

Soper, J. C. & Walstad, W. B. (1987). *Test of Economic Literacy: Second Edition Examiner's Manual.* New York: Joint Council on Economic Education.

Walstad, W. B. & Robson, D. (1997). Differential item functioning and male-female differences on multiple-choice tests in economics. *Journal of Economic Education*, *28*, 155-171.

Walstad, W. B., Watts, M. & Rebeck, K. (2007). *Test of understanding in college economics: Examiner's manual* (4th ed.). New York NY: National Council on Economic Education.

Weinert, F. E. (2001). Concept of Competence: A Conceptual Clarification. In D. S. Rychen & L. H. Salganik (Eds.), *Defining and selecting key competencies* (pp. 45–65). Seattle: Hogrefe & Huber.

Willett, J. B. (1988). Questions and answers in the measurement of change. *Review of Research in Education*, *15*(1), 345–422.

Williams, M. L., Waldauer, C., & Duggal, V. G. (1992). Gender Differences in Economic Knowledge: An Extension of the Analysis. *The Journal of Economic Education*, *23*(3), 219-231.

Zlatkin-Troitschanskaia, O., Förster, M., Brückner, S., & Happ, R. (2014). Insights from a German assessment of business and economics competence. In H. Coates (Ed.), *Higher education learning outcomes assessment – International perspectives* (pp. 175-197). Frankfurt am Main: Peter Lang.

Zlatkin-Troitschanskaia, O., Schmidt, S., Brückner, S., Förster, M., Yamaoka, M. & Asano, T. (in press). Macroeconomic Knowledge of Higher Education Students in Germany and Japan – A Multilevel Analysis of Contextual and Personal Effects. *Assessment & Evaluation in Higher Education.*

Zlatkin-Troitschanskaia, O., Shavelson, R. J., & Kuhn, C. (2015). The International state of research on measurement of competency in higher education. *Studies in Higher Education*, *40*(3), 393-411.