The Improvement of Student Teachers' Instructional Quality during a 15 Week Field Experience: A Latent Multimethod Change Analysis

Peter Holtz

Knowledge Media Research Center Tübingen

Timo Gnambs

Leibniz Institute for Educational Trajectories

Author Note

Correspondence concerning this article should be addressed to Peter Holtz, Knowledge Media Research Center Tübingen, Schleichstraße 8, 72076 Tübingen, Germany, E-mail: p.holtz@iwm-tuebingen.de

Accepted for publication in *Higher Education*.

Abstract

Most studies evaluating the effectiveness of school internships have relied on self-assessments that are prone to self-presentational distortions. Therefore, the present study analyzed the improvement in the instructional quality of 102 student teachers (46 women) from a German university during a 15 week internship at a local secondary school across three rating sources: the student teachers themselves, their students, and their mentors (experienced teachers). A latent multimethod change analysis identified a significant increase in instructional quality during the practice semester. However, ratings from the three informant groups only marginally converged.

The Improvement of Student Teachers' Instructional Quality during a 15 Week Field Experience: A Latent Multimethod Change Analysis

A continuing challenge in teacher education is coping with the 'theory-practice gap' (cf. Allen & Wright 2014; Cohen, Hoz, & Kaplan 2013): how can abstract academic learning be transformed into applied instructional practice Whereas in 'early entry' programs student teachers are exposed to classroom teaching with relatively little prior training, in many European countries prospective teachers require an academic degree before starting their on-the-job training at school (e.g., Buchberger, Campos, Kallos, & Stephenson 2000; Korthagen, Loughran, & Russell 2006). The optimal balance between academic input and practical experience is still a prevalent source of dispute. This controversy is further complicated by ambiguities in the evaluation of student teachers' field experiences that make empirical findings difficult to compare. So far, most studies on field experiences in teacher education have focused on self-ratings of instructional quality (Cohen, Hoz, & Kaplan 2013). However, increasing evidence suggests that self-reports are prone to various forms of self-presentational distortions (e.g., Gnambs & Kaspar 2015, 2016; Nasser & Fresko 2006; Podsakoff, MacKenzie, & Podsakoff 2012) and, thus, represent biased indicators of learning outcomes. Therefore, the present study answers repeated calls for more multi-methodical approaches in educational research (e.g., Keller-Margulis 2012) and evaluated changes in the instructional quality of student teachers after a 15 week field experience across different rating sources.

1 Field Experiences in Teacher Education

1.1. Previous findings on the efficacy of field experiences

Field experiences in the form of internships at schools are an integral part of many college-based teacher education programs, meant to give student teachers the opportunity to translate their formal learning experiences from the university context into practical competencies

(Zeichner, Payne, & Brayko 2012). However, early on—particularly in the 1970s and 1980s—ambiguous findings on the effectiveness of field experiences threw into question the generalized, 'naïve' idea of universally positive effects of practical experience (e.g., Sandgreen & Smith 1956; Iannacone & Button 1964; Sorenson 1967; Tabachnik & Zeichner 1984). Since then, research has shifted towards identifying more specific activities and settings that contribute to successful theory-practice integration (cf. Zeichner 2010). In the early 2000s, Wilson and Floden (2003) criticized the descriptive nature of most empirical studies conducted on field experiences: most studies only described how field experiences were integrated into teacher education curricula, but did not evaluate the outcome of these attempts. Thus, they seemed inadequate for deriving any substantiated conclusions regarding the factors contributing to successful field experiences in teacher education.

More recently, a review of 113 studies on field experiences in teacher education published between 1996 and 2009 corroborated these findings and found that the majority of the studies had (107/113) adopted a descriptive focus, meaning that they had focused on a single internship program (Cohen, Hoz, & Kaplan 2013, p. 349). Of these 107 studies, 51 were categorized as 'descriptive-neutral' (no evaluative aspects) and 56 as 'descriptive-evaluative'; here, "experimental or qualitative comparisons" (p. 349) were used to assess the consequences of a given internship program. Overall, the results were "generally favorable" (p. 368). Of the 37 articles that evaluated the internship's efficacy in any way, only twelve studies reported unfavorable outcomes at all (p. 366 f.): For example, teachers failed to implement the intended teaching approaches, such as constructivist problem solving or the integration of English and native languages (seven studies, p. 367). Three studies found that preservice teachers were not able to translate improvements in lab teaching to actual classroom teaching (p. 366). And another three studies reported a lack of teachers' willingness to engage with students' critical thinking (p.

366). Among the favorable outcomes, 17 studies found improvements regarding the teachers' 'cognitive and emotional development', such as an improvement of the teachers' observational capabilities (p. 366). Eleven studies reported 'improvement in instruction competences and skills' over the course of the internship (p. 367). Seven studies reported positive effects on 'efficacy and self-confidence in teaching, views, opinions, and appreciation' (p. 366), and another seven studies found improvements in the domain of 'professionalism: implementation of teaching approaches' (p. 367).

Over the last decade, a number of studies (see below) in Germany took up the question of the overall effectiveness of field experiences anew and examined self-rated pedagogical competencies before and after field experiences, comprising several weeks or even months of in-school training. Overall, these studies concordantly showed that student teachers perceived themselves as more competent after field experiences (e.g., Bodensohn & Schneider 2008; Müller 2010; Gröschner, Schmitt, & Seidel 2013; for an English language review see Besa & Büdcher 2014).

Despite the substantial number of empirical studies on field experiences in teacher education, one limitation of most studies is their reliance on self-reports. Typically, student teachers evaluated their own competencies or rated their own motivations and attitudes before and after field experiences. However, self-reports are suspect to various forms of distortion (e.g., Gnambs & Kaspar 2015, 2016; Nasser & Fresko 2006; Podsakoff et al. 2012) that might have obfuscated previous findings. For example, most individuals tend to exhibit a positivity bias (cf. Gnambs 2013; Paulhus & John 1998) leading them to evaluate themselves rather favorably. Thus, frequently people overestimate their own performance as compared to a more objective standard (e.g., Janssen & Van der Vegt 2011; Oeberst, Haberstroh, & Gnambs 2015). Moreover, subjective ratings are typically also susceptible to various response styles, such as acquiescence

or extreme responding, that have been shown to inflate observed statistics by up to 54% (Baumgartner & Steenkamp 2001). In order to remedy these biasing influences on empirical findings, there have been repeated calls for more multi-methodical approaches in educational and psychological research (e.g., Keller-Margulis 2012; Podsakoff et al. 2012). However, so far there are no studies on changes in instructional quality after field experiences that validate their findings across different rating sources.

1.2 Multi-perspective ratings of instructional quality

Whereas multi-perspective evaluations of teaching competence after a field experience are scarce, there is a large body of research on ratings of instructional quality in general. A substantial part of these studies is based on college students' ratings of their instructors' teaching. In general, college student ratings are regarded as reliable and valid indicators of teachers' learning and achievements (e.g., Davis 2009; Feldman 1989a; Marsh & Roche 1997; McKeachie 1997). Although research on younger students is still limited, there is increasing evidence that ratings of instructional quality are comparably reliable and valid in younger age cohorts (e.g., Kyriakides 2005; Peterson, Wahlquist, & Bone 2000; Wagner, Göllner, Helmke, Trautwein, & Lüdtke 2013); even elementary students' ratings seem to be almost as reliable as those of students in higher grades (e.g., Follman 1995; Strong & Ostrander 1997).

However, instructional quality can be assessed in a number of different ways (Berk 2005, distinguishes no less than 12 approaches) that can provide information additional to student ratings. For example, teaching performance is sometimes evaluated using peer reports from external observers such as colleagues or mentors. These evaluations are able to capture aspects that students are typically unable to fully evaluate (e.g., instructional expertise). Frequently, teachers are also asked to evaluate themselves. Although these self-ratings might be subject to self-presentational biases (Nasser & Fresko 2006; Paulhus & John 1998), they give access to

aspects of teacher behavior that might be difficult to observe by students (e.g., motivational states). Therefore, different rating sources do not necessarily yield identical results. Correlations between student ratings of instructional quality, teacher's self-ratings, and peer ratings typically range from small to moderate. In his review, Feldman (1989b) found an average correlation between student ratings and teachers' self-ratings of instructional effectiveness of $r = .29$ (based on 19 studies), an average correlation of $r = .22$ (six studies) between self-ratings and external observers' ratings, and an average correlation of $r = .50$ (14 studies) between students' ratings and external observers' ratings.

These results highlight that different informants seem to capture related, but by no means identical, concepts. The rather small to moderate correlations indicate that most of the variance in instructional ratings is unique to the specific rating source. Each informant (i.e., student teacher, students, or external observer) takes a unique perspective and, thus, seems to evaluate different facets of instructional quality. Therefore, the unimodal evaluations of teaching quality that dominate research on field experiences in teacher education so far might yield biased conclusions if some aspects of instructional quality are not adequately observed by the specific rater. The present study overcomes this limitation by analyzing changes in instructional quality from different perspectives, including the student teachers themselves, their students, and the student teachers' mentors.

2 Study Overview and Predictions

Although a number of studies have evaluated the efficacy of field experiences in teacher education, most of them are based on self-ratings of competence alone. In light of evidence that the assessment of instructional quality varies by rating source, the present study examined the effect of a mandatory school internship on student teachers' instructional quality from multiple perspectives.

2.1 The 'Jena Practice Semester'

This study draws on student teachers at Jena University in Germany (State of Thuringia). There, student teachers have to complete a five-year curriculum subdivided into ten semesters to be allowed to complete their 'first state exam' ("Erstes Staatsexamen"; comparable to MA in education). The students attend university classes in two main subjects and attend classes in pedagogy, educational psychology and other related disciplines as well. After another 18 months of preparatory service, student teachers can take the second state exam and qualify to work as state-employed school teachers in the State of Thuringia. In the 5th or 6th semester of their university curriculum, student teachers complete the 'practice semester,' comprising an approximately 15 week internship at a regional secondary school. Every two weeks, for one day, the students attend classes at the university on educational psychology, research methods, and didactics of their two subjects. Apart from that, the students are at their schools five days a week, six hours a day, mentored by teachers from the respective schools. During the first weeks of the practice semester, student teachers passively observe the experienced teachers in class and later become more active in the classroom. All in all, they are supposed to take an active teaching role in 40 to 80 lessons (depending on the specific regulations for their school subjects) – always under the supervision of an experienced teacher (for more details see Kleinespel, 2014).

2.2 Hypotheses

The majority of studies based on self-reports of instructional quality showed that student teachers and pre-service teachers perceive a substantial increase in their teaching abilities during field experiences (e.g., Cohen, Hoz, & Kaplan 2013; Besa & Büdcher 2014). Although there is still no consensus on the overall effectiveness of internships in teacher education (cf. Wilson & Floden 2003) and a strong need for multi-methodical approaches to evaluating it (e.g., Keller-Margulis 2012), our first hypothesis will be:

H1: Student teachers, their students, and their mentors perceive an increase in instructional quality during the 'practice semester'.

Prior research has identified moderate correlations among ratings of instructional quality from different perspectives (e.g., Feldman, 1989b). Hence, to some degree, students and mentors are likely to perceive similar changes in instructional quality during the practice semester as the student teachers themselves perceived. However, other studies (e.g., Nesser & Fresko 2006) highlighted that different raters seem to focus on different aspects when evaluating the instructional quality of teachers. So far, there has not been a systematic comparison of student teachers', their mentors', and their students' perceptions of a change in student teachers' instructional quality during a long field experience. We hypothesize:

H2: Changes in student teachers' instructional quality differ by rating source.

There is broad agreement that ratings of instructional qualities are multidimensional, that is, they reflect different aspects of teaching (e.g., Cohen 1981, Feldman 2007). Nevertheless, these dimensions are usually substantially correlated. Therefore, this study focuses on two central dimensions of instructional quality, motivation and structure, that are focal determinants of students' achievements: Among a total of 28 dimensions of instructional quality reviewed by Feldman (2007), factors relating to the structuring of the course (e.g., teacher preparation, course organization) and teacher's ability to motivate students (e.g., stimulate interest in the subject or motivate students to do their best) were the most important predictors of students' learning success explaining between 14 to 32 percent of the variance in student achievement. Not surprisingly, nine of the eleven instruments for student evaluation of teaching (SET) that were compared in a recent literature review (Spooren, Brockx, & Mortelmans, 2013) featured at least one scale that corresponded to aspects of 'motivation' (e.g., 'motivation', 'instructor helpfulness', 'caring and supportive') and one scale that captured aspects of the 'structure'-dimension (e.g.,

'organization', 'course rigor', 'instructor's delivery of course information'). Whereas the majority of SET studies focused on university students' course evaluations, there is some empirical evidence that structure and motivation are important dimensions of school students' evaluations of teaching as well. For example, 'motivation' and 'structure' were among the five most important factors in school students' course evaluations in a recent study with 6,909 German ninth-grade students (Wagner et al., 2013). Because of the established impact of these instructional dimensions on educational outcomes (see also Cohen 1981), their widespread use in SET instruments, their applicability to school students' evaluation of teaching, and due to the fact that they can be measured reliably by focusing on just one single lesson (Praetorius, Pauli, Reusser, Rakoczy, & Klieme 2014), the present study focuses on changes in motivation and structure as proxies of instructional quality.

3. Method

3.1 Participants and Procedure

3.1.1 Procedure. During the winter semester 2013/2014, the teaching activities of 181 student teachers participating in the practice semester were subjected to a multi-perspective evaluation. At the beginning and the end of the semester, one lesson was evaluated by the student teachers, their students, and one experienced teacher (N = 181) who mentored the student teacher and served as an external observer. Participation in the study was voluntary, but the student teachers were repeatedly encouraged to collect and provide data. Information on the school subjects was not obtained because this would have (together with information on age and sex) exposed the identities of some of the participants. Each class (5^{th} to 12^{th} grade) included up to 30 students (M = 18.79, SD = 5.06) from different secondary schools in Thuringia, Germany. The average time between the two assessments was 80 days (SD = 22.05). Because the regulations for

the different school subjects varied slightly, the student teachers taught approximately 30 to 50 lessons during this time period.

3.1.2 Participants and attrition analysis. The 181 student teachers (46 women) had a mean age of 23.39 years (*SD* = 3.48). About two thirds (120 student teachers) participated at both measurement occasions, whereas one third only provided responses at the beginning of the semester. To rule out a systematic bias due to nonresponse, the respondents participating twice were compared to those that participated only once with regard to student teachers' sex, age, and initial levels of instructional quality (motivation and structure). These analyses revealed no significant effects (all *p*s > .10). Therefore, dropout is unlikely to have introduced a systematic bias.

3.2 Measures

Instructional quality at the two measurement points was operationalized with two subscales, each including six items from an inventory for the evidence-based multi-perspective assessment of instructional quality (EMU; Helmke 2010). One subscale referred to the degree teachers managed to motivate their students to engage with the subject matter at hand ('motivation'; (1) treating students in a respectful way; (2) being friendly towards students; (3) letting the students finish their answers; (4) giving the students enough time to think about answers; (5) ability to liven up the lesson; (6) lauding the students for helpful contributions; the wording was different for student teachers, students, and mentors/observers), whereas the other subscale measured the degree to which the lesson was well organized and followed a clear structure ('structuring'; (1) the students knew what they were supposed to do; (2) content of previous lessons was taken up; (3) the teacher recapped key contents; (4) the students were asked to express themselves in a clear and understandable way; (5) the students were aware of what the

lesson was about; (6) the lesson was interesting for the students; again, the wording was different for the three groups of informants). All items were rated on 4-point response scales ranging from 1 (strongly disagree) to 4 (strongly agree). Student teachers provided self-ratings on these items, whereas students and mentors provided observer reports. For all three groups, the formulation of the items was kept as similar as possible. For example, the second item of the motivation subscale would be "I was friendly towards the students" for the student teachers; "the teacher was friendly towards the students" for the mentors/observers; and "the teacher was friendly towards me" for the students. No individual responses were available for students; therefore, the analyses are limited to the mean responses in each class. On the two measurement occasions, the motivation scale resulted in satisfactory coefficient alpha reliabilities from .71 to .86, whereas the structuring scale yielded reliabilities between .69 and .83 (see Table 1).

The EMU-instrument is in many regards similar to established English-language instruments for classroom observation such as the Protocol for Language Arts Teaching Observation (PLATO; Grossman, Loeb, Cohen, & Wyckoff 2013), the Classroom Assessment Scoring System (CLASS; Pianta, Karen, Paro, & Hamre 2008), the UTeach Observation Protocol (UTOP; Walkington, Arora, Ihorn, Gordon, Walker, Abraham, & Marder 2012), and the Framework for Teaching Evaluation (FfTE; Danielson 2013). The 'motivation' subscale we used in this study bears similarities to the PLATO subscales 'feedback' and 'behavior management', the domain 'emotional support' in the CLASS instruments, several subscales (e.g., classroom engagement and classroom management) of UTOP's 'classroom environment' domain and the sub-domain 'implementation questioning', and the subscales 3a, 3b, 3c, and 3e of the FfTE's domain 'instruction'. 'Structuring' resembles PLATO's subscales 'representation of content', 'connection to prior knowledge' and 'connections to personal and cultural experience', and 'strategy use and instruction'; it is related as well to the 'classroom organization' domain of

CLASS, UTOP's 'lesson structure' domain, and the FfTE's subscales 1d, and 1e of the domain 'preparation and planning' as well as subscale 2c ('management of classroom procedures').

In contrast to the aforementioned instruments, which are created for evaluations by trained observers, EMU complements the observer perspective ('mentors') with ratings from teachers and students themselves. Compared to the conceptual richness of tools such as PLATO or the FfTE, EMU captures a rather narrow array of instructional features.

3.3 Statistical Analyses

Changes in instructional quality during the practice semester were examined using latent change modelling (cf. McArdle 2009). At each measurement occasion, one latent factor (T_l) was specified that represented a common instructional quality dimension across all raters (see Figure 1). In order to create more parsimonious measurement models, we did not analyze individual item scores but created two test halves (item parcels) for each rater following the item-to-construct balance technique (Little, Cunningham, Shahar, & Widaman 2002). Thus, at each measurement occasion, the latent instructional quality factor was represented by six manifest indicators with uncorrelated residuals. To account for test halves that were not strictly parallel, we also included indicator-specific teaching skill factors (IS_k) that capture unshared variance unique to the second test half (cf. Geiser & Lockhart 2012). Change in instructional quality across the two measurement occasions was represented by a latent difference variable (McArdle 2001; Steyer, Eid, & Schwenkmezger 1997). This approach decomposes the second latent teaching skill factor (T_2) into the initial skill factor (T_1) and a latent difference factor (T_2-T_1). The latter represents the portion of T_2 not identical to T_1 (McArdle, 2009), that is, the change between the two time points.

Because instructional quality was evaluated by different raters (student teacher, students, and mentors/observers) the model was extended to a multimethod change model (Geiser, Eid,

Nussbeck, Courvoisier, & Cole 2010) that contrasts change in different methods (i.e., raters). To this effect, a reference method (here: student teachers) was selected against which the remaining methods (here: students and mentors/observers) were compared. The multimethod change model adopted in the present study is presented in Figure 1. In this model the latent difference factor T_2-T_1 represents the change in teaching skills as observed by student teachers (i.e., the reference method), whereas the latent difference factors M_{12}-M_{11} and M_{32}-M_{31} represent the *residual* change for students and external observer ratings that is not captured by student teacher ratings. Thus, the latter represent change that is unique to these raters.

All change models were estimated in Mplus 7 (Muthén & Muthén 1998-2012) with a robust full maximum likelihood estimator that has been shown to yield unbiased parameter estimates in covariance analyses when responses are missing at random (Enders & Bandalos, 2001; Newman, 2003). Departures from multivariate normality were acknowledged by adopting the Yuan-Bentler test statistic (Yuan & Bentler, 2000) and estimating heteroskedasticity-robust standard errors (cf. Hays & Cai, 2007). The goodness of fit of these models was evaluated using the *Comparative Fit Index* (CFI; Bentler 1990), *Non-Normed Fit Index* (NNFI; Bentler, 1990), and the *Root Mean Square Error of Approximation* (RMSEA; Steiger 1990). In line with conventional standards (e.g., Hu & Bentler 1999; Schermelleh-Engel, Moosbrugger, & Müller 2003), fit for models with a CFI ≤ .90, NNFI ≤ .90, or a RMSEA ≥ .10 is considered ‘bad’, those with .90 > CFI < .95, .90 > NNFI < .95, and .05 > RMSEA < .10 as ‘acceptable’, and CFI ≥ .95, NNFI ≥ .95, and RMSEA ≤ .05 as ‘good’.

4. Results

Descriptive statistics, bivariate correlations, and coefficient alpha reliabilities between all measures are presented in Table 1. The two instructional quality dimensions, motivation and structure, were moderately correlated, $M(r) = .58$, indicating that the two scales measured, albeit

related, but by no means identical concepts. Moreover, the mean stability coefficient was $M(r)$ = .54; that is, instructional quality changed during the practice semester.

4.1 Goodness of Fit

The multimethod change models (see Figure 1) for the motivation and structure subscales resulted in acceptable fits, $\chi^2(41) = 67.45$, $p = .01$, CFI = .96, NNFI = .93, RMSEA = .06, RMSEA 90% CI = [.03, .09], and $\chi^2(41) = 49.96$, $p = .16$, CFI = .99, NNFI = .98 RMSEA = .04, RMSEA 90% CI = [.00, .07], respectively. Meaningful interpretations of the model parameters require strong measurement invariance across time (Little, 2013). Therefore, we refitted models to the data that held the factor loadings and intercepts of the indicators equal over time. These invariance models fit the data well, $\chi^2(49) = 78.06$, $p = .01$, CFI = .95, NNFI = .93, RMSEA = .06, RMSEA 90% CI = [.03, .08], and $\chi^2(49) = 54.19$, $p = .28$, CFI = .99, NNFI = .99, RMSEA = .02, RMSEA 90% CI = [.00, .06]. Moreover, difference tests indicated that the constrained models did not fit worse than the unconstrained models, $\Delta\chi^2(8) = 10.89$, $p = .21$ and $\Delta\chi^2(8) = 4.86$, $p = .77$, respectively. Thus, the assumption of invariant measurement structures across time was supported. Therefore, all subsequent analyses are based on the more parsimonious models with invariance of factor loadings and intercepts.

4.2 Consistency and Specificity of Change

From the variance estimates in the latent change models, several coefficients can be derived that reflect the consistency (convergent validity), specificity, and reliability of the difference scores for each observed indicator (Geiser et al. 2010). Reliability reflects the amount of variance in difference scores that is not attributable to measurement error. Overall, reliability was rather low (see Table 2): the Spearman-Brown corrected reliability coefficients for the difference scores fell between .31 and .81 for the motivation subscale and between .26 and .82 for the structure subscale. More importantly, the observed difference scores based on students and

external observer ratings showed very low consistencies and very high specificities: only about 2% to 9% of the variance in observed difference scores was determined by respective changes in student teacher ratings, whereas the majority of variance was unique to the two raters. Thus, change in teaching skills as measured by the three rating sources only marginally converged.

4.3 Mean Change

The estimated latent means of the change scores (see Table 3) indicated that on average, teachers increased their instructional skills during the practice semester. However, there were notable differences in change scores for the three raters. Change scores were smallest for student ratings and in the case of the motivation subscale, not even significantly different from zero, p = .12. The change scores for student teacher and external observer ratings that reflect a unique change of the two rating sources independent of the student ratings were considerably larger. The mean differences for the motivation and structure subscales are plotted in Figure 2. Students tended to report less change in instructional quality than the student teachers themselves or their mentors.

5. Discussion

5.1 Key Findings

Field experiences are generally regarded as an essential part of teacher education programs. However, empirical findings on the effectiveness of internships have been rather inconclusive because most previous studies were limited to descriptive analyses (Cohen et al., 2013) and relied on self-assessments of teaching competencies that can be attenuated by self-presentational styles (Paulhus & John 1998; Podsakoff et al. 2012). Therefore, the present study evaluated student teachers' field experiences using multiple perspectives. Overall, the study provided three central findings. First, student teachers' instructional quality as captured by the two dimensions 'motivation' and 'structuring' *did* in fact improve during the practice semester

according to the ratings of the student teachers themselves, their mentors, and—at least as far as 'structuring' is concerned—also according to their students. This is an important finding in light of the frequent criticism regarding naïve beliefs in the mythical power of practice (cf. Hascher 2011). Hence, there is empirical support for our hypothesis 1: Student teachers, their students (at least with regard to the structuring dimension), and their mentors perceive an increase in instructional quality during the practice semester.

Second, observed change scores in instructional quality exhibited rather poor reliabilities (cf. Kopp 2011; Peter, Churchill, & Brown 1993). Thus, empirical analyses using observed statistics, typically, would be unable to identify significant improvements of teaching quality because measurement error attenuated true changes (Gnambs 2014, 2015). In contrast, the latent variable modelling approach (Geiser et al. 2010) adopted in the present study has allowed for the examination of true change independent of measurement error. Third, ratings from the different informant groups converged to a rather limited degree. In line with previous studies (e.g., Feldman 1989b), the three perspectives were moderately correlated cross-sectionally (see Table 1). In contrast, the convergent validity of change in teaching skills was rather low; that is, interindividual differences in intraindividual change estimated from students' ratings did not correspond well with interindividual differences in intraindividual change estimated from self- or external observer ratings. Hence, in line with our hypothesis 2, we found considerable differences between the student teachers, their students, and their mentors with regard to the magnitude of the perceived changes during the practice semester.

5.2 Implications

Conclusions about change in teaching skills from student ratings might be rather different from conclusions drawn on the basis of self- or observer ratings. Researchers and educators would be ill advised to base assessments of the effectiveness of field experiences on self-ratings

of competence gains alone, as is the common practice. However, given the overwhelming evidence on the external validity of student ratings (e.g., Benton, Duchon, & Palett 2013), the students' perspective should not be ignored when discussing good practice in teacher education. As, to our knowledge, the different perceptions of change in instructional quality during field experiences have not been studied systematically before, these results substantially add to the existing knowledge on the effectiveness of school internships.

The study also extends prior empirical evidence on the efficacy of field experiences in teaching experience insofar as it opens up hypotheses about the reasons for the discrepancies between student teachers and mentors/observers on the one side and students on the other. In this study, we tried to use the same items for all three groups of informants in order to gather information on the overall effectiveness of the practice semester. Still, it is possible that different foci in the evaluation of the lesson and, for example, different degrees of background knowledge about pedagogical principles and teaching techniques may lead the students to base their judgments in part on other criteria than the mentors/observers and the student teachers.

Moreover, cognitive-motivational processes, such as the reduction of dissonance (Festinger, 1957), could also have contributed to these results: the student teachers as well as their mentors had invested a substantial amount of time and resources into the practice semesters. The student teachers had to adapt their lifestyles during the internship to those of professional school teachers and had to integrate into a new environment and pick up many new social and professional skills in a relatively short period of time. In general, the practice semester is perceived by the student teachers as an exhausting but overall positive and enriching experience (Holtz, 2014). The mentors as well have to allocate resources to the training of the student teachers in addition to their own teaching workload. They have to give feedback, provide instruction, and – in some cases – give emotional support to the student teachers as well. It might

be speculated that these investments may influence the ratings of change in instructional quality insofar as a lack of success could lead to the unpleasant feeling of cognitive dissonance ('I invested a lot but got nothing in return'). Hence, there may be a tendency to overestimate increases in competencies on the side of the student teachers and the mentors. Future studies are needed to test these specific hypotheses based on our explorative findings.

There is also another practical reason for us to encourage other researchers to more strongly focus on the students' perspective when discussing the effectiveness of educational measures in teacher education: In Hattie's (2009) well known synthesis of meta-analyses on the influences on achievement of school-aged students, student orientation, student feedback, and multi-perspective feedback were among the most effective measures in teacher training. Hence, introducing multi-perspective evaluation procedures might have an overall effect on student teachers instructional quality in its own right over and above getting reliable data on the overall effectiveness of the respective programs.

5.3 Limitations and Directions for Future Research

Some limitations of the present study should be acknowledged. First, we were unable to obtain individual responses from the students. Therefore, the analyses presented were based on the aggregated responses within each class. More precise estimates of the model parameters could be obtained if individual responses were available and the change models were extended by an appropriate multilevel structure (e.g., Koch, Schultze, Eid, & Geiser 2014). Moreover, because different clusters of students were recruited at the beginning and the end of the term, changes in instructional quality might have been underestimated to some degree. Therefore, future studies should replicate these findings using the same students throughout the entire course of the longitudinal design. Second, only slightly more than half of the student teachers participating in the practice semester volunteered for this study. Although we did not find any significant

differences in instructional quality between participants and those who decided not to participate, it cannot be ruled out that selective dropout might have biased the results to some degree. Moreover, due to our non-randomized sampling procedure, the validity of our results might be compromised to some degree, if respondents and non-respondents differed in certain key characteristics (e.g., competence or motivation). Third, our sample size was somewhat small for precise parameter estimates in covariance structure analyses; as a consequence, our models achieved only a power of .71 (McCallum, Browne, & Sugawara, 1996). Fourth, because of privacy concerns, we did not have any information about the student teachers' school subjects. It is possible that the practice semester might have different effects in different academic subjects. We also could not control for contextual factors, such as characteristics of the respective schools and classrooms that could influence the efficacy of the field experience (Ronfeldt 2012). Moreover, the sample included students from grades 5 to 12 and, thus was considerably younger than samples in most of the previous research on teacher effectiveness. Although several studies supported the reliability and validity of student ratings in this age range (e.g., Follman 1995; Kyriakides 2005; Peterson et al. 2000; Strong & Ostrander 1997; Wagner et al. 2013) further research is needed to determine potential moderating effects. Fifth, it would be worthwhile to replicate the findings presented here with other established English-language instruments for classroom observation such as PLATO, UTOP, CLASS, and the FfTE (see above). In light of the multidimensional nature of instructional quality (see Feldman 2007), it could also be of interest to contrast our results on motivation and structure with other instructional facets, such as class climate or even teacher personality. Finally, models of teacher education vary substantially between countries and sometimes even within countries (Darling-Hammond, & Liberman 2013; Campos, Kallos, & Stephenson 2000). Thus, the 'Jena practice semester' is only one among many approaches to implementing field experiences in teacher education. Therefore, researchers

at other institutions are encouraged to replicate these results by incorporating different perspectives in different environments to identify potential causal factors underlying competence gains during field experiences and to advance our understanding of effective models of teacher education.

5.4 Conclusion

A 15 week field experience at a local school increased student teachers' instructional skills. However, the perceived change was significantly smaller (in case of the 'structure' dimension, not even significant) for student ratings as compared to ratings by student teachers or their mentors. To the best of our knowledge, this was the first study that systematically compared differences in the perception of the acquisition of competencies in field experiences in teacher education using latent multimethod change analyses. Although the study focused on a very specific set-up of a field experience in teacher education, the 'Jena practice semester', we hope that our findings will encourage researchers interested in the effectivity of school internships to take into account the students' perspective as well.

References

Allen, J. M., & Wright, S. E. (2014). Integrating theory and practice in the pre-service teacher education practicum. *Teachers and Teaching, 20*, 136-151. doi:10.1080/13540602.2013.848568

Baumgartner, H., & Steenkamp, J. B. E. M. (2001). Response styles in marketing research: a cross-national investigation. *Journal of Marketing Research, 38*, 143-56. doi:10.1509/jmkr.38.2.143.18840

Bentler, P. M. (1990). Comparative fit indexes in structural models. *Psychological Bulletin, 107*, 238-46. doi:10.1037/0033-2909.107.2.238

Benton, S. L., Cashin, W. E., & Kansas, E. (2012). *IDEA PAPER# 50 Student Ratings of Teaching: A Summary of Research and Literature*. Manhattan, KS: The IDEA Center.

Benton, S. L., Duchon, D., & Pallett, W. H. (2013). Validity of student self-reported ratings of learning. *Assessment & Evaluation in Higher Education, 38*, 377-388. doi:10.1080/02602938.2011.636799

Berk, R. A. (2005). Survey of 12 strategies to measure teaching effectiveness. *International Journal of Teaching and Learning in Higher Education, 17*, 48-62.

Besa, S., & Büdcher, M. (2014). Empirical evidence on field experiences in teacher education. In K.-H. Arnold, A. Gröschner, & T. Hascher (Eds.), *Schulpraktika in der Lehrerbildung: Theoretische Konzeptionen, Grundlagen und Effekte* (pp. 129-145). Münster, Germany: Waxmann.

Buchberger, F., Campos, B. P., Kallos, D., & Stephenson, J. (2000). *Green paper on teacher education in Europe*. Umeå, Sweden: Thematic Network on Teacher Education in Europe.

Cohen, P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies. *Review of Educational Research, 51*, 281-309. doi:10.3102/00346543051003281

Cohen, E., Hoz, R., & Kaplan, H. (2013). The practicum in preservice teacher education: a review of empirical studies. *Teaching Education, 24*, 345-380. doi:10.1080/10476210.2012.711815

Danielson, C. (2013). *The Framework for Teaching Evaluation Instrument*. Princeton, NJ: Danielson Group.

Darling-Hammond, L., & Lieberman, A. (2013). *Teacher education around the world: Changing policies and practices*. London, England: Routledge.

Enders, C. K., & Bandalos, D. L. (2001). The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural Equation Modeling, 8*, 430-457. doi:10.1207/S15328007SEM0803_5

Feldman, K. A. (1989a). The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies. *Research in Higher Education, 30*, 583-645. doi:10.1007/BF00992392

Feldman, K. A. (1989b). Instructional effectiveness of college teachers as judged by teachers themselves, current and former students, colleagues, administrators, and external (neutral) observers. *Research in Higher Education, 30*, 137-194. doi:10.1007/BF00992716

Feldman, K. A. (2007). Identifying exemplary teachers and teaching: Evidence from student ratings. In R. P. Perry & J. C. Smart (Eds.), *The scholarship of teaching and learning in*

higher education: An evidence-based perspective (pp. 93-143). New York, NY: Springer.

Festinger, L. (1957). *A theory of cognitive dissonance.* Stanford, CA: Stanford university press.

Follman, J. (1995). Elementary public school pupil rating of teacher effectiveness. *Child Study Journal, 25*, 57-78.

Geiser, C., & Lockhart, G. (2012). A comparison of four approaches to account for method effects in latent state-trait analyses. *Psychological Methods, 17*, 255-283. doi:10.1037/a0026977

Geiser, C., Eid, M., Nussbeck, F. W., Courvoisier, D. S., & Cole, D. A. (2010). Multitrait-multimethod change modelling. *Advances in Statistical Analysis, 94*, 185-201. doi:10.1007/s10182-010-0127-0

Gnambs, T. (2013). The elusive general factor of personality: The acquaintance effect. *European Journal of Personality, 27*, 507-520. doi:10.1002/per.1933

Gnambs, T. (2014). A meta-analysis of dependability coefficients (test-retest reliabilities) for measures of the Big Five. *Journal of Research in Personality, 52*, 20-28. doi:10.1016/j.jrp.2014.06.003

Gnambs, T. (2015). Facets of measurement error for scores of the Big Five: Three reliability generalizations. *Personality and Individual Differences, 84*, 84-89. doi:10.1016/j.paid.2014.08.019

Gnambs, T., & Kaspar, K. (2015). Disclosure of sensitive behaviors across self-administered survey modes: A meta-analysis. *Behavior Research Methods, 47*, 1237-1259. doi:10.1177/1073191115624547

Gnambs, T., & Kaspar, K. (2016). Socially desirable responding in web-based questionnaires: A meta-analytic review of the candor hypothesis. *Assessment*. Advance online publication. doi:10.3758/s13428-014-0533-4

Gröschner, A., Schmitt, C. & Seidel, T (2013). Veränderung subjektiver Kompetenzeinschätzungen von Lehramtsstudierenden im Praxissemester [Changes in student teachers‘ competence self-ratings during the practice semester]. *Zeitschrift für Pädagogische Psychologie, 27*, 77-86. doi:10.1024/1010-0652/a000090

Grossman, P., Loeb, S., Cohen, J., & Wyckoff, J. (2013). Measure for measure: The relationship between measures of instructional practice in middle school English language arts and teachers’ value-added scores. *American Journal of Education, 119*, 445-470. doi:10.1086/669901

Hascher, T. (2011). Vom «Mythos Praktikum» … und der Gefahr verpasster Lerngelegenheiten [On the mythical power of field experience … and the danger of missing out on educational opportunities]. *Journal für Lehrerinnen- und Lehrerbildung, 3*, 8-16.

Hattie, J. A. C. (2009). Visible learning: A synthesis of 800+ meta-analyses on achievement. Abingdon, England: Routledge.

Hayes, A. F., & Cai, L. (2007). Using heteroskedasticity-consistent standard error estimators in OLS regression: An introduction and software implementation. *Behavior Research Methods*, *39*, 709-722. doi:10.3758/BF03192961

Helmke, A. (2010). *Unterrichtsqualität und Lehrerprofessionalität. Diagnose, Evaluation und Verbesserung des Unterrichts* [Instructional quality and teacher professionalism] (3rd edition). Seelze, Germany: Klett-Kallmeyer.

Holtz, P. (2014). „Es heißt ja auch Praxissemester und nicht Theoriesemester“: Quantitative und qualitative Befunde zum Spannungsfeld zwischen Theorie und Praxis im Jenaer Praxissemester [“it is called ‘practice semester’ and not ‘theory semester’”: Quantitative and qualitative findings on the tension between theory and practice in the Jena practice semester]. In: A. K. Kleinespel (eds.), *Ein Praxissemester in der Lehrerbildung: Konzepte, Befunde und Entwicklungsprozesse im Jenaer Modell der Lehrerbildung* [a practice semester in teacher education …], 97-118. Bad Heilbrunn: Klinkhardt.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55. doi:10.1080/10705519909540118

Janssen, O., & Van der Vegt, G. S. (2011). Positivity bias in employees' self-ratings of performance relative to supervisor ratings: The roles of performance type, performance-approach goal orientation, and perceived influence. *European Journal of Work and Organizational Psychology, 20*, 524-552. doi:10.1080/1359432X.2010.485736

Keller-Margulis, M. A. (2012). Fidelity of implementation framework: A critical need for response to intervention models. *Psychology in the Schools, 49*, 342-352. doi:10.1002/pits.21602

Kleinespel, K. (2014). *Ein Praxissemester in der Lehrerbildung: Konzepte, Befunde und Entwicklungsperspektiven am Beispiel des Jenaer Modells der Lehrerbildung* [The Jena practice semester: ideas, results, and perspectives]. Bad Heilbrunn, Germany: Klinkhardt.

Koch, T., Schultze, M., Eid, M., & Geiser, C. (2014). A longitudinal multilevel CFA-MTMM model for interchangeable and structurally different methods. *Frontiers in Psychology, 5*, Article 311. doi:10.3389/fpsyg.2014.00311

Kopp, B. (2011). Neuropsychologists must keep their eyes on the reliability of difference measures. *Journal of the International Neuropsychological Society, 17*, 562-563. doi:10.1017/S1355617711000361

Kyriakides, L. (2005). Drawing from teacher effectivess research and research into teacher interpersonal behaviour to establish a teacher evaluation system: A study on the use of student ratings to evaluate teacher behaviour. *Journal of Classroom Interaction, 40*, 44-66.

Little, T. D. (2013). Longitudinal Structural Equation Modeling. New York, NY: Guilford Press.

Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question and weighing the merits. *Structural Equation Modeling, 9*, 151-173. doi:10.1207/S15328007SEM0902_1

Marsh, H. W., & Roche, L. A. (1997). Making students' evaluations of teaching effectiveness effective: The critical issues of validity, bias, and utility. *American Psychologist, 52*, 1187-1197. doi:10.1037/0003-066X.52.11.1187

McArdle, J. J. (2001). A latent difference score approach to longitudinal dynamic structural analyses. In R. Cudeck, S. du Toit, & D. Sorbom (Eds.), *Structural Equation Modeling: Present and Future* (pp. 342–380). Lincolnwood, IL: Scientific Software International.

McArdle, J. J. (2009). Latent variable modeling of differences and changes with longitudinal data. *Annual Review of Psychology, 60*, 577–605. doi:10.1146/annurev.psych.60.110707.163612

MacCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods, 1*, 130-149. doi:10.1037//1082-989X.1.2.130

McKeachie, W. J. (1997). Student ratings: The validity of use. *American Psychologist, 52*, 1218-1225. doi:10.1037/0003-066X.52.11.1218

Müller, K. (2010). *Das Praxisjahr in der Lehrerbildung* [A practice year in teacher education]. Bad Heilbrunn, Germany: Klinkhardt.

Muthén, L. K., & Muthén, B. O. (1998-2012). *Mplus user's guide (7th edition).* Los Angeles, CA: Muthén & Muthén.

Nasser, F., & Fresko, B. (2006). Predicting student ratings: the relationship between actual student ratings and instructors' predictions. *Assessment & Evaluation in Higher Education, 31*, 1-18. doi:10.1080/02602930500262338

Newman, D. A. (2003). Longitudinal modeling with randomly and systematically missing data: A simulation of ad hoc, maximum likelihood, and multiple imputation techniques. *Organizational Research Methods, 6*, 328-362. doi:10.1177/1094428103254673

Oeberst, A., Haberstroh, S., & Gnambs, T. (2015). Not really the same: Computerized and real lotteries in decision making research. *Computers in Human Behavior, 44*, 250-257. doi:10.1016/j.chb.2014.10.060

Paulhus, D. L., & John, O. P. (1998). Egoistic and moralistic biases in self-perception: The interplay of self-deceptive styles with basic traits and motives. *Journal of Personality, 66*, 1025-1060. doi:10.1111/1467-6494.00041

Peter, J. P., Churchill Jr, G. A., & Brown, T. J. (1993). Caution in the use of difference scores in consumer research. *Journal of Consumer Research, 19*, 655-662. doi:10.1086/209329

Peterson, K. D., Wahlquist, C., & Bone, K. (2000). Student surveys for school teacher evaluation. *Journal of Personnel Evaluation in Education, 14*, 135-153. doi:10.1023/A:1008102519702

Pianta, R. C., Karen, M., Paro, L., & Hamre, B. K. (2008). *Classroom Assessment Scoring System (CLASS) Manual: K-3*. Baltimore, MD: Paul H. Brookes Publishing Company.

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology, 63*, 539-569. doi:10.1146/annurev-psych-120710-100452

Praetorius, A. K., Pauli, C., Reusser, K., Rakoczy, K., & Klieme, E. (2014). One lesson is all you need? Stability of instructional quality across lessons. *Learning and Instruction, 31*, 2-12. doi:10.1016/j.learninstruc.2013.12.002

Ronfeldt, M. (2012). Where should student teachers learn to teach? Effects of field placement school characteristics on teacher retention and effectiveness. *Educational Evaluation and Policy Analysis, 34(1)*, 3-26. doi: 10.3102/0162373711420865.

Schermelleh-Engel, K., Moosbrugger, H., & Müller, H. (2003). Evaluating the fit of structural equation models: Test of significance and descriptive goodness-of-fit measures. *Methods of Psychological Research Online, 8*, 23-74.

Spooren, P., Brockx, B., & Mortelmans, D. (2013). On the validity of student evaluation of teaching the state of the art. *Review of Educational Research, 83*(4), 598-642. doi:10.3102/0034654313496870

Steiger, J. H. (1990). Structural model evaluation and modification: an interval estimation approach. *Multivariate Behavior Research, 25*, 173-180. doi:10.1207/s15327906mbr2502_4

Steyer, R., Eid, M., & Schwenkmezger, P. (1998). Modeling true intraindividual change: true change as a latent variable. *Methods of Psychological Research Online, 2*, 21-33.

Stronge, J. H., & Ostrander, L. P. (2006). Client surveys in teacher evaluation. In J. H. Stronge (Ed.), *Evaluating Teaching* (2nd edition, pp. 125-152). Thousand Oaks, CA: Sage.

Tabachnick, B. R., & Zeichner, K. M. (1984). The impact of the student teaching experience on the development of teacher perspectives. *Journal of Teacher Education, 35*, 28-36. doi:10.1177/002248718403500608

Wagner, W., Göllner, R., Helmke, A., Trautwein, U., & Lüdtke, O. (2013). Construct validity of student perceptions of instructional quality is high, but not perfect: Dimensionality and generalizability of domain-independent assessments. *Learning and Instruction, 28*, 1-11. doi:10.1016/j.learninstruc.2013.03.003

Walkington, C., Arora, P., Ihorn, S., Gordon, J., Walker, M., Abraham, L., & Marder, M. (2012). *Development of the UTeach observation protocol: A classroom observation instrument to evaluate mathematics and science teachers from the UTeach preparation program.* Retrieved from https://utop.uteach.utexas.edu/

Wilson, S. M., & Floden, R. E. (2003). *Creating effective teachers: Concise answers for hard questions*. New York, NY: AACTE Publications.

Zeichner, K. (2010). Rethinking the connections between campus courses and field experiences in college- and university-based teacher education. *Journal of Teacher Education, 61*, 88-99. doi:10.1177/0022487109347671

Zeichner, K., Payne, K., & Brayko, K. (2012). *Democratizing knowledge in university teacher education through practice-based methods teaching and mediated field experience in schools and communities (Issue Paper 12-1).* Seattle, WA: University of Washington.

Yuan, K., & Bentler, P. M. (2000). Three likelihood-based methods for mean and covariance structure analysis with nonnormal missing data. *Sociological Methodology, 30*, 167-202. doi:10.1111/0081-1750.00078

Table 1.

Descriptive Statistics and Correlations

							First measurement occasion						*Second measurement occasion*					
							Student		Teacher		Mentor/ observer		Student		Teacher		Mentor/ observer	
		M	*SD*	γ	κ	1.	2.	3.	4.	5.	6.	7.	8.	9.	10.	11.	12.
	First measurement occasion																
S	1. Motivation	3.41	0.25	-0.44	-0.14	(.85)											
	2. Structure	3.32	0.28	-0.69	0.28	.74*	(.71)										
T	3. Motivation	3.38	0.37	-0.74	0.85	.31*	.21*	(.81)									
	4. Structure	3.10	0.47	-0.28	-0.40	.15	.27*	.36*	(.85)								
M	5. Motivation	3.58	0.35	-0.99	0.81	.37*	.32*	.39*	.19*	(.69)							
	6. Structure	3.31	0.44	-0.46	-0.28	.27*	.33*	.24*	.49*	.58*	(.78)						
	Second measurement occasion																
S	7. Motivation	3.42	0.26	-0.38	-0.59	.46*	.39*	.21*	.17	.30*	.23*	(.84)					
	8. Structure	3.36	0.27	-.067	0.03	.36*	.49*	.21*	.29*	.22*	.22*	.77*	(.73)				
T	9. Motivation	3.46	0.33	-0.97	2.38	.26*	.21*	.62*	.44*	.28*	.35*	.36*	.23	(.79)			
	10. Structure	3.28	0.41	-0.59	0.17	.11	.19*	.31*	.60*	.34*	.29*	.12	.25*	.53*	(.85)		
M	11. Motivation	3.68	0.28	-1.20	1.60	.28*	.29*	.16	.11*	.29*	.50*	.40*	.27	.30*	.11	(.72)	
	12. Structure	3.51	0.35	-0.59	-0.09	.03	.21*	.21*	.41*	.54*	.39*	.32*	.34*	.37*	.42*	.52*	(.78)

Note. S = Student, T = Teacher, M = Mentor; γ = Skewness, κ = Excess kurtosis. Coefficient alpha reliabilities are presented in diagonal.

* $p < .05$

Table 2.

Estimates of Variance Components for the Difference Scores

	Difference	Consistency	Specificity	Reliability
Motivation				
S	$H_{122} - H_{121}$	.08	.52	.60 (.75)
	$H_{222} - H_{221}$	.07	.61	.68 (.81)
T	$H_{112} - H_{111}$	.22	-	.22 (.37)
	$H_{212} - H_{211}$	.18	-	.18 (.31)
M	$H_{132} - H_{131}$	.08	.29	.37 (.54)
	$H_{232} - H_{231}$	.04	.23	.26 (.42)
Structure				
S	$H_{112} - H_{111}$	.02	.62	.64 (.78)
	$H_{212} - H_{211}$	.03	.68	.70 (.82)
T	$H_{122} - H_{121}$	.20	-	.20 (.33)
	$H_{222} - H_{221}$	.15	-	.15 (.26)
M	$H_{132} - H_{131}$	.09	.32	.41 (.58)
	$H_{232} - H_{231}$	.04	.28	.31 (.48)

Note. S = Student, T = Teacher, M = Mentor; H_{ikl} = Test half with i = test half, k = rater, and l = measurement occasion. The first number refers to the test half and the second number refers to the measurement occasion. In some cases the consistency and specificity coefficients do not add up to the reliability coefficient due to rounding errors. Spearman-Brown corrected reliabilities for full test length are in parenthesis.

Table 3.

Estimated Means and Standard Deviations of Latent Change Scores

		M	*SD*
Motivation			
S	$M_{12} - M_{11}$	-0.02	0.05*
T	$T_2 - T_1$	0.05*	0.01
M	M_{22} - M_{21}	0.05	0.04
Structure			
S	$M_{12} - M_{11}$	0.02	0.06*
T	$T_2 - T_1$	0.04*	0.00
M	M_{22} - M_{21}	0.10*	0.06*

Note. S = Student, T = Teacher, M = Mentor

* $p < .0$;

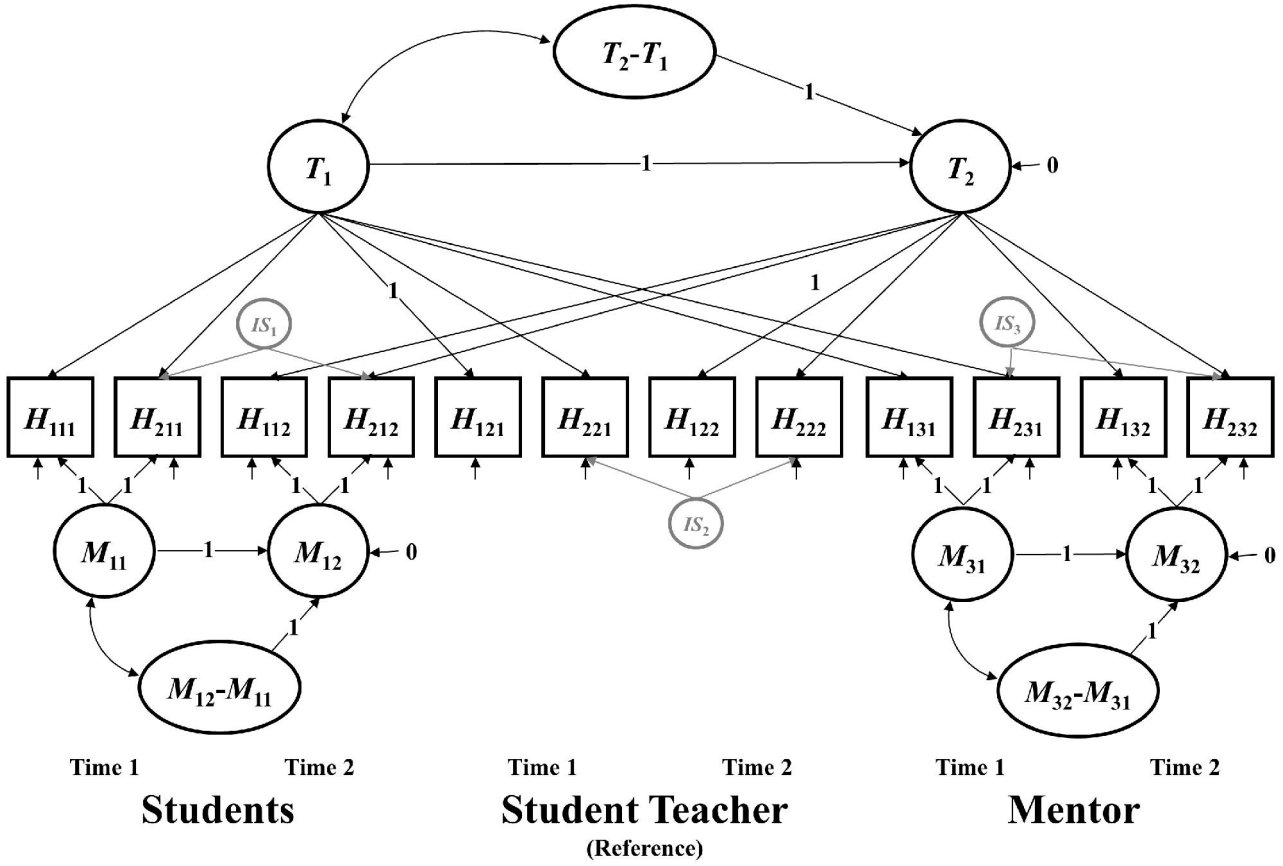

Figure 1. Latent multimethod change model. H_{ikl} = Test half, T_l = Teaching skill factor, IS_k = Indicator-specific factor, and M_{kl} = Rater-specific method factor with i = test half, k = rater, and l = measurement occasion. Correlations between indicator-specific factors are not shown.

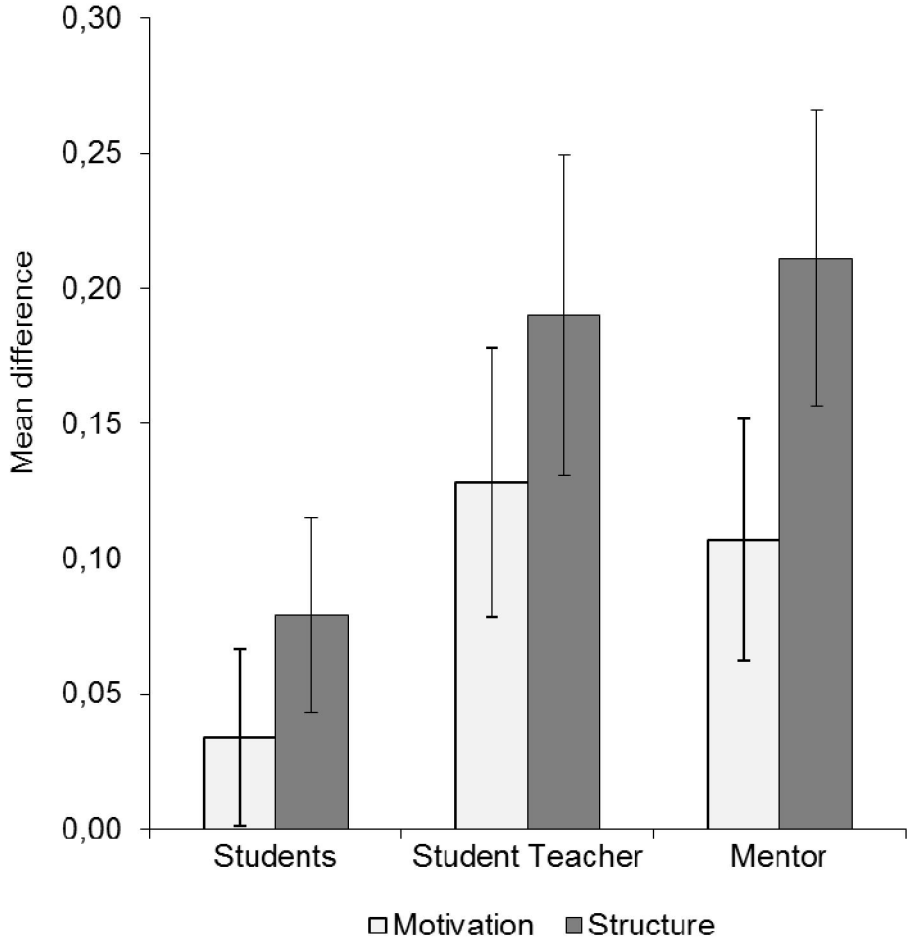

Figure 2. Mean differences with standard errors for teaching skills by rater