

Article

Expansion of the Field Practice Experience Scales (FPE-DK) to a Long Form to Improve Targeting to the Student Teacher Population

Tine Nielsen ^{1,*} , Beate Tommerup Headley ²  and Gitte Gorm Larsen ³

¹ Department of Applied Research in Education and Social Sciences, UCL University College, DK-5230 Odense, Denmark

² Independent Researcher

³ Department of Teacher Education, UCL University College, DK-5230 Odense, Denmark

* Correspondence: tini@ucl.dk

Abstract

Targeting measurement scales to the intended population is important to maximize the information gained. The Field Practice Experience scales (FPE-DK) measure student teachers' opportunities to learn through the observation of other teachers, own practice, and receiving feedback on 12 key teaching activities while in field practice placements during teacher education. While the three scales have previously been shown to fit pure Rasch models, they were not well targeted to the population. The present study aimed to improve the targeting of the three scales of the FPE-DK by adding more difficult items. The study used Rasch and graphical log-linear Rasch models for item analysis of field practice experience data of 307 Danish student teachers collected with the FPE-DK-long immediately after having completed a field practice placement. The three expanded scales did not fit pure Rasch models, as had been the case in previous research. Instead, each of the three proposed long scales had to be divided into two subscales: one set of seventeen items measuring opportunities to learn through observation and own practice, respectively, receiving feedback with regard to instruction-related practices; and one set of eight items measuring opportunities to learn through observation and own practice, respectively, receiving feedback with regard to collaboration-related practices. The six subscales fitted graphical log-linear Rasch models with some locally dependent items and some items functioning differentially, and targeting was improved compared with the FPE-DK short form. The findings provided an improved measurement instrument and advanced our knowledge of differences in opportunities to learn through observation, own practice, and receiving feedback over the course of teacher education.

Keywords: opportunities to learn; teacher education; field practice placement; field practice experience; observation; own practice; receiving feedback; Rasch model; graphical log-linear Rasch model



Academic Editor: Antonio
Hernández-Mendo

Received: 24 August 2026

Revised: 25 September 2026

Accepted: 1 October 2026

Published: 10 October 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Teacher education differs across cultures and countries, though with the common feature that it contains both academic components, usually acquired through on-campus teaching, and core skills, usually acquired through modeling or simulation on campus and/or through supervised own practice in field practice placements or practicum in schools. The organization of these two overall components of teacher education differs across countries and within countries across teacher education programs. The most common organization consists of a subject-matter bachelor's degree followed by an education

degree at the master's level, with the latter including in-school training of varying durations (Weisdorf, 2020). In Denmark and several other Nordic countries, teacher education is a coherent program with subject matter and educational training combined. Over several decades, the field practice part(s) of teacher education (in some contexts named practicum or clinical teaching experience) have been a subject of research (Menter, 2022). Prevalent foci of teacher education research have included the coherence of the academic and practical parts (e.g., Canrinus et al., 2017; Cavanna et al., 2021; Grossman et al., 2008; Sayed et al., 2026; Youngs et al., 2022), the practical part(s) separately (e.g., More, 2003; Nielsen et al., 2024, 2025, 2026b; Ulvik et al., 2021; Zeichner, 2021), and within the practical part(s), specific opportunities to learn (e.g., Cohen & Berlin, 2020; Fuentes-Abeledo et al., 2020; Nielsen, 2021, 2024).

Within professional education programs such as the teacher education program—but also nurse, social work, doctor, etc.—aimed at a specific profession, core skills and practices are predominantly acquired through one or more field placements in the intended profession. In teacher education, the weight and timewise placement of these field placements vary across countries and even within countries (Weisdorf, 2020). For example, in Denmark¹ and Sweden, field practice placement(s) amounts to 30 ECTS (European Credit Transfer System²); in Finland, 20 ECTS; in Norway, it amounts to a minimum of 115 days; while in Iceland, there is no minimum duration for field practice set by law, but all teacher education programs include field practice placement (Source: e-mail correspondence with the Icelandic Ministry of Education, Science and Culture). In Germany, teacher education has two distinct phases, an academic phase and a teacher start phase, where the latter is field practice with a duration of 16–24 months (Weisdorf, 2020). Such differences in the weight of field practice in teacher education are bound to have an impact on the student teachers' possibilities for training and acquiring core skills and practices for the teacher profession. Fuentes-Abeledo et al. (2020), in the Spanish context, suggested that field practice *“should be an opportunity for the students to come close to the teachers' role, so that they can directly discover and experience the multiple tasks facing teachers in their daily lives.”* (p. 347). Similarly, Gaikhorst et al. (2020) pointed to the need for student teachers to be provided with opportunities to learn through field practice in urban schools, and Nielsen (2024) suggested that teacher education programs could benefit from placing focus on opportunities to learn through observation in field practice placement late in the program.

Opportunities to learn could consist of enactment of core skills and activities within the on-campus parts of teacher education (Klette et al., 2017; Schutz et al., 2018), i.e., role-playing, teaching of fellow students, training micro-activities, and so on. However, the enactment approach does provide students with an opportunity to practice activities approximating real-world teaching activities, as the setting is controlled and artificial, and all unexpected situations and disturbances occurring in the real-world school context do not occur. Another more ecologically valid opportunity to practice and learn core skills is through field practice placement(s), where students, under supervision and guidance, can practice in the real-world context and learn to cope with unexpected situations. Although the ecological validity of such training is important, it is insufficient to ensure the best outcome, as it also depends on the field practice teachers/mentors in the schools (Ronfeldt, 2012). Cohen and Berlin (2020) suggested dividing opportunities to learn aimed at training skills into opportunities to engage with representations and decompositions of practice and opportunities to approximate or enact teaching practices. Nielsen (2021) attempted to measure both of these aspects of opportunities to learn, but only as they take place in field practice placements.

While most research into opportunities to learn for student teachers (pre-service teachers) is concerned with academic opportunities to learn (Schmidt et al., 2011), some research

into the opportunities to learn the core skills and practices of the teaching profession while in field practice in schools is available.

In a document- and interview-based study in the Netherlands teacher education context, [Gaikhorst et al. \(2020\)](#) found that different teacher education programs prepared teachers for urban teaching in different ways, for example, by mandatory field practice at urban schools.

[Fuentes-Abeledo et al. \(2020\)](#) conducted a survey on 57 opportunities to learn through tasks in field practice with 248 Spanish final-year student teachers. The tasks were chosen on the basis of focus-group interviews and consisted of six planning tasks, four evaluation tasks, seven psychoeducational guidance tasks, ten execution tasks, four training tasks, four reflection tasks, and twenty-two tasks of participating in school activities. Results were reported as mean scores for these groups of tasks, despite these not being established as measurement scales, and no tests of difference between the levels of opportunities to learn through the various categories of tasks were conducted. At the descriptive level, it was reported that the most frequent tasks students did in field practice were reflective, planning, evaluation, and execution tasks, while the least frequent tasks were training and educational guidance tasks as well as tasks related to participation in school activities. Further, the study tested whether there were differences in the types of tasks offered to students dependent on whether they were doing field practice in primary or infant education, or in rural or urban schools, and they found some differences. The authors concluded that scant participation in school life and culture during field practice limits the necessary and comprehensive training that should be provided.

The questionnaire—the elementary teaching candidate survey—used in the Development of Ambitious Instruction (DAI) Study (Youngs, www.daiproject.weebly.com accessed on 13 May 2024) included a section asking student teachers about 11 experienced opportunities to learn essential teacher activities related to teaching mathematics or English language while in field practice. Students could mark if they had “observed other teachers use this practice with students”, “used this practice with students once or twice”, “used this practice with students three or more times”, “received feedback on my attempts to use this practice with students”, and “none of the above”. While students could mark more than one option, the lack of response options “yes” and “no” meant that a missing response could indeed be missing, or it could signify a “no, I did not experience this”. In order to accommodate precise interpretation and develop measurement scales rather than single questions, [Nielsen \(2021\)](#) changed this format in the development of the Field Practice Experience scale (FPE-DK), so each item would be included in three separate scales measuring opportunities to learn across the same item, but through different modes of learning: observation of other teachers, own practice, and receiving feedback. Thus, also taking into account [Cohen and Berlin’s \(2020\)](#) division into opportunities to engage with representations and decompositions of practice and opportunities to approximate or enact teaching practices by measuring representations of practice through opportunities to learn through observation of other teachers, enactment of teaching practice through own practice, and decompositions of practice through feedback on this practice. For each mode of opportunity to learn and each item, students could respond “yes” or “no”. In this process, items were also generalized so that they did not specifically refer to teaching mathematics or English, and to include the relational side of classroom management, a 12th item was added.

In the Danish teacher education context, [Nielsen \(2021\)](#)—using the FPE-DK, which contains three validated scales measuring opportunities to learn through observation of the teacher, opportunities to learn through own practice, and opportunities to learn through receiving feedback on a set of 12 teacher practices³—found significant differences in the

mean scores on all three scales dependent on the level of field practice students had just completed⁴. More specifically, it was found that students who had just completed field practice level 1 or level 2 scored significantly and substantially higher ($M = 8.86$) on the observed scale than did students who had just completed level 3 ($M = 6.38$) ($p < 0.001$). Students who had just completed field practice level 2 or level 3 scored significantly but only slightly higher ($M = 10.87$) on the practiced scale than did students who had only completed level 1 ($M = 10.09$) ($p < 0.01$). Lastly, students who had just completed field practice level 1 or 2 scored significantly but only slightly higher ($M = 8.69$) than students who had just completed level 3 ($M = 9.54$) ($p < 0.05$).

Nielsen (2024), also using the FPE-DK in the Danish teacher education context, used Chain graph models to analyze the data of 560 Danish student teachers who had just completed field practice at three different levels. The results showed that opportunities to learn through observation of other teachers were negatively associated with the level of field practice and thus were reported less and less the further along students were in the program, while the opposite was the case for opportunities to learn through own practice.

The three scales in the FPE-DK measuring opportunities to learn through observation of other teachers, opportunities to learn through own practice, and opportunities to learn through receiving feedback each fitted a Rasch model (Nielsen, 2021). However, targeting of the three scales in the FPE to the student teacher population was not satisfactory.

The Current Study

The current study was aimed at expanding the Field Practice Experiences Questionnaire for teacher students—Danish version (FPE-DK; Nielsen, 2021, 2025) in order to improve its targeting to the population of student teachers. To achieve this aim, new items were developed, and item analyses using Rasch models were conducted to investigate the psychometric properties of the long-form FPE-DK. In addition, the research question originally posed by Nielsen (2021) with the short-form FPE-DK was investigated in the current study: How are students' opportunities to learn associated with the level of their latest field practice placement?

2. Methods

2.1. Instrument

The FPE-DK consists of three scales measuring student teachers' opportunities to learn 12 teacher practices through observation of other teachers, through own practice, and through receiving feedback while in field practice placement during their teacher education (Nielsen, 2021, 2025). The FPE-DK thus consist of three scales drawing on the same set of 12 items: the Observed scale, which reflects opportunities to learn through observation of teachers, the Practiced scale, which reflects opportunities to learn through practicing yourself, and the received Feedback scale, which reflects opportunities to learn through receiving feedback on their practice.

Students were asked, "Which opportunities did you have for each of the following during your latest field practice placement?" They were instructed to respond three times for each practice item, once for observation, once for own practice, and once for feedback, and were provided the opportunity to respond with a yes (1) or a no (0) in each case (see Supplementary File S1 for the item content in Danish, as well as a non-validated English version, and Supplementary File S2 for the matrix design of the questionnaire). Accordingly, each of the three scales has a range of 0–12.

In the initial validity study of the FPE-DK, the three scales were found to each fit a Rasch model (Rasch, 1960) with no issues of local dependence between items, differential item functioning, lack of invariance, or lack of overall homogeneity (Nielsen, 2021).

Thus, the summed scale scores could be used to assess the degree to which the students experienced opportunities to learn through the three modes of observation, own practice, and receiving feedback (Kreiner, 2013; Nielsen & Dyhrberg O'Neill, 2025). This made the FPE readily accessible to practitioners wanting to use the instrument to gauge student teachers' opportunities to learn while in field practice. Reliability of both the Observed and received Feedback scales was excellent across subgroups of students defined by field practice level and gender (above 0.85), while the Practiced scale showed varied reliability across the subgroups (from 0.23 for male students who had just completed field practice level 2 to 0.77 for female students who had just completed field practice level 1). However, targeting of all three scales was poor, and so for all subgroups, the received Feedback scale was best targeted to male students who had just completed level 1 field practice, with 56% of the maximum information obtained for this group. While this was the most information obtained for a subgroup across the three scales, this was not considered good targeting (Nielsen, 2021).

Expanding the FPE-DK

The expansion of the 12-item FPE-DK (Nielsen, 2021) with additional items was aimed at improving the targeting of the three scales. Thus, the goal was to obtain scales that were better targeted to the student teacher population. The main reason for the poor targeting appeared to be that most opportunities to learn reflected in the 12 items were commonly experienced by students during field practice and thus very easy to endorse for all students (Tables 3 and 4 in Nielsen, 2021). Thus, in Rasch model terms, we needed to include items that were more difficult—in this case, that meant opportunities to learn that were not as commonly experienced by the student population, whether the experience was by observation, own practice, or receiving feedback.

We proceeded to develop additional items based on (1) the national skills objectives for field practice in previous teacher education (*Act on the Education of Teachers*, 2015). These items would reflect practices that should be encountered in one way or another during field practice in order for students to train and acquire these skills. Since we developed the items, Danish teacher education has been reformed, and the objectives for field practice are now described in a more overall manner, but the objectives still cover the same areas as the previous teacher education, as the skills needed for a teacher remain the same (*Act on the Education of Teachers*, 2023; *Lærerruddannelsen*, 2026). (2) Input from the field practice coordinator, who had 14 years' experience with what student teachers report as occurring very rarely during their field practice placements, as the field practice coordinator organizes and analyzes both students' and field practice teachers' evaluations of field practice placements each year as soon as they are completed. By combining the two sources of information, we developed 13 new items. There were three experts in teacher education with first-hand teaching experience in the teacher education program, one with additional field practice coordinator experience, two with additional research experience of more than 20 years, and one of the latter additionally had more than 20 years' experience in the development and validation of questionnaires and measurement scales within educational psychology and related fields.

The three skills objectives on communication with parents within the competence area of relational work (c.f. the English translation of the national skills objectives for the field practice in Nielsen (2021)) was an area that students consistently reported as one they did not get acquainted with within field practice. We therefore developed three items reflecting ordinary communication practices with parents (items 18–20).

From another two skills objectives within the competence area of relational work, we developed two items on taking part in teamwork with other teachers and with other professionals (items 21 and 22).

Two items on formative and summative evaluation were developed from the two evaluation skills objectives within the competence area of didactics, as these also rarely occurred during field practice due to timing issues (items 23 and 24). We did, however, also develop three items from the three planning-oriented skills objectives within the competence area of didactics, which were not all rare in occurrence. Rather, these were increasingly rare in occurrence depending on their complexity (items 13–15). In addition, we developed a single item on the inclusion of physical activities in classes (item 25), as this was part of one of the skills objectives within the didactics competence, and one that potentially would be hard to include in all teaching subjects.

Two items were developed from the three skills objectives making up the competence area of classroom management, even though these were known to be more commonly occurring in field practice (items 16 and 17). We did this because we wanted to include this competence area as well.

We did not conduct cognitive interviews or pilot testing among students, as they had detailed knowledge of the skills objectives for the field practice levels and the linguistic level of the items was simple rather than complex. Thus, the students should readily be able to understand the items.

Supplementary File S1 contains the new items in Danish as well as in a non-validated English version, and the Danish long-form FPE-DK is available, with administration and scoring guidelines, in [Nielsen et al. \(2026a\)](#).

2.2. Participants

Data were collected among student teachers at one Danish university college just after they had completed a field practice placement in the first, second, or third year of teacher education ($n = 307$).

The majority of the students were enrolled in the ordinary teacher program (85.7%), and most of the students attended campus A (71%) (Table 1). The latter reflects the relative intake of students at the two campuses. The distribution of students in terms of the level of their last completed field practice was almost equal, with approximately a third of the sample at each level. In Danish teacher education, the major teaching subject, counting 50 ECTS, must be chosen from three subjects: Danish language, mathematics, or English language. However, university colleges are not required to offer English as a major teaching subject. In the study sample, the Danish language was chosen by half the students (49.8%), mathematics by a little more than a third (37.5%), and English language by the minority of students (11.4%) (Table 1). In the sample, most identified as female (69.4%), and the mean age in the sample was 27.3 years (Table 1).

Table 1. Characteristics of the sample.

	Total ($n = 307$)	
	n	%
Teacher program		
Ordinary	263	85.7
Other ^a	44	14.3
Campus		
Campus A	218	71
Campus B	89	29

Table 1. Cont.

	Total (n = 307)	
	n	%
Level of most recent field practice		
Level 1	111	36.2
Level 2	97	31.6
Level 3	99	32.2
Major ^b		
Danish language	153	49.8
Mathematics	115	37.5
English language	35	11.4
Gender ^c		
Female	213	69.4
Male	92	30
Age	Mean	SD
20–65	27.26	7.34

Notes. Percentages not summing to 100 within variable are due to missing responses. ^a Includes trainee program, distance (online) program, and STEM program. ^b Major regardless of grade. ^c Two students identified as other gender and were excluded from further analysis due to their small representation.

2.3. Item Analyses by Rasch Models

Item analyses were conducted to investigate, in detail, the psychometric properties and validity of the expanded 25-item versions of the Observed, Practiced, and Feedback scales, intended to measure opportunities to learn through observation of other teachers, opportunities to learn through own practice, and opportunities to learn through feedback, while in field practice placements in schools. As the short 12-item version of the scales in the FPE-DK has previously been shown to fit the Rasch model (RM; Rasch, 1960), and scales fitting the RM have particularly desirable properties, such as sufficiency of the summed score for the person parameter estimates, the RM was our model of choice in the current study.

The RM assumes that items in a scale adhere to the following five requirements (Kreiner, 2013):

Unidimensionality: The items of a scale assess a single underlying latent construct. In this study, the Observed, Practiced, and Feedback scales would each measure a single construct.

Monotonicity: The expected item scores increase with increasing values of the latent variable. In this study, the probability of having observed another teacher engaged in any of the field practice activities described in the items should increase with increasing scores on the Observed scale, and likewise for the Practiced and the Feedback scales.

Conditional independence of items (no local response dependence; No LD): Responses to a single item should be conditionally independent of responses to any other item given the latent variable. In this study, this means that responses to any one of the items in the scale should depend only on the level on the Observed, Practiced, and Feedback scales and not also on responses to the other items in the scale. This was tested for all item pairs.

Absence of differential item functioning (no DIF): Items and exogenous (i.e., background variables) should be conditionally independent given the latent variable. In this study, this would entail that responses to any one item in a scale should only depend on the level on the Observed, Practiced, and Feedback scales and not also on subgroup membership (e.g., level of field practice or gender). This was tested for each item and each exogenous variable.

Overall invariance: The entire set of items in a scale should function invariantly across subgroups (c.f. DIF, which was tested for a single item at a time; invariance was tested for the set of items and each exogenous variable).

Overall homogeneity: The rank order of the item locations (i.e., the item difficulties) should be the same across all persons, regardless of their level on the latent variable. In this study, this means that the order of the items in each of the three field practice experience scales should be the same in terms of difficulty, no matter the levels on the Observed, Practiced, and Feedback scales. Homogeneity was tested for low and high scores across the entire set of item parameters.

Fulfilment of the above requirements ensures criterion-related construct validity as defined by Rosenbaum (1989), as well as sufficiency of the sum score for the person parameter estimates in the sense that the sum score contains the necessary information to evaluate a person's level on the construct in question (Kreiner, 2013; Rasch, 1960, 1961).

2.3.1. Graphical Log-Linear Rasch Models

It is not unusual for measurement scales to suffer from DIF and/or LD between items, and thus not achieve fit with what could be termed the pure RM. One way to move forward in such cases is to use extended graphical log-linear Rasch models (GLLRMs) (Kreiner & Christensen, 2002, 2004, 2007), which can adjust for departures from the RM in the form of uniform⁵ DIF and/or uniform LD, while retaining the essential validity of the scale. The reliability of scales fitting a GLLRM locally dependent item cannot be assessed using Cronbach's alpha, as alpha represents the lower bound of reliability for scales with conditionally independent items. Thus, reliability should be calculated in a manner that takes into account any local dependence, or it will be spuriously inflated to some degree. Local dependence in a GLLRM does, however, not affect the sufficiency of the sum score (Kreiner & Christensen, 2007). GLLRMs that include uniform DIF interaction terms—between items and one or more exogenous variables—indicate that item responses fit a RM for each of the subgroups for which one or more items function differentially. Since these estimates are on the same logit scale (theta), they are still comparable and can be used in subsequent statistical analyses and for individual assessment, and essential validity is maintained (Kreiner & Christensen, 2007). Scales fitting a GLLRM retain a sufficient score if these are equated for DIF via the person parameter estimates (Kreiner, 2007). Thus, whether we want to use the person parameter estimates or the DIF-equated sum score for subsequent group comparisons or for individual assessment, this is possible.

GLLRMs can be illustrated using chain graph models, where nodes represent variables and edges represent associations between variables. In these graphs, missing edges or arrows between nodes mean that the variables in question are conditionally independent, given the remaining variables in the model, while arrows or undirected connections between variables refer to causal or non-causal relationships between these (Lauritzen, 1996). In terms of the GLLRM, items that are not connected by an edge are conditionally independent given the latent variable (i.e., items are locally independent). Items that are not connected to an exogenous variable mean that items and exogenous variables are conditionally independent given the latent variable (i.e., there is no DIF). The causal relationship between the latent variable and items is illustrated by arrows from the former to the latter, as are associations between exogenous variables and the latent variable.

2.3.2. Strategy of Analysis

The same strategy of analysis was used for all three field practice experience scales. The strategy was based on RM being the most parsimonious amongst the RM and different

GLLRMs with LD and/or DIF; thus, fit to the RM was tested first. The strategy of analyses for each of the three field practice experience scales was as follows:

First, the fit of the set of item responses for a scale to the RM was tested as rigorously as possible, thus trying our best to reject the model. If fit to the RM was rejected, the departures from the RM in the form of LD and DIF were catalogued in order to define a GLLRM with log-linear interaction terms for uniform LD and/or DIF. Subsequently, the fit of the set of item responses to the GLLRM was tested in the same manner as was the fit to the RM.

If there was no fit to either a RM or a GLLRM, the item deemed most problematic was eliminated from the scale (using the criteria proposed by [Nielsen & Kreiner, 2013](#)), and the iterative process of testing fit to the RM, cataloguing departures, and testing fit to a specified GLLRM was repeated. The order of the various tests of fit was not necessarily the same in each iteration, but instead, the order was determined by the intermediate results. If no fit was achieved, even if a substantial number of items had been removed, item content was scrutinized, and a suggestion for division into subscales was made based on content considerations. For each subscale, the above strategy of analysis was conducted. Once fit to final subscale models within each of the original three field practice experience scales was achieved, unidimensionality was tested across subscales. The reason why unidimensionality would be tested at the end of the analysis rather than at the start is that unrecognized local dependence within subscales can create spurious evidence against unidimensionality.

In addition, reliability and targeting were assessed for the final subscale models.

2.3.3. Statistics

The RM is characterized by the sufficiency of the raw score for the person parameter estimates, a property adhering only to Rasch models. Therefore, the estimation of parameters and inference in relation to the fit of items to the RM can be conditional on the score. Conditional inference eliminates person parameters by using the conditional distributions of items that do not depend on the person parameters, and thus, conditional inference does not require assumptions on the distribution or sampling of persons ([Christensen, 2013](#); [Kreiner, 2013, 2025](#); [Rasch, 1961](#)).

Item parameters (locations) were calculated using conditional maximum likelihood (CML) estimates, as they are known to be consistent ([Christensen, 2013](#); [Kreiner, 2025](#); [Kreiner & Nielsen, 2023](#)).

Two overall tests of fit were conducted using Andersen's conditional likelihood ratio test (CLR; [Andersen, 1973](#); [Kreiner & Christensen, 2013a](#)). Firstly, a test of global homogeneity was conducted by the comparison of item parameters in low- and high-scoring groups, defined as being as close to two equal-sized groups as possible. This tests the hypothesis that item parameters are the same for persons with low or high scores on the scale in question. Secondly, a test of invariance across the entire set of items in the scale was tested in relation to each background variable. This tests the hypothesis that all item parameters across all background variables are the same.

The fit of individual items to the RM was tested by conditional infit and outfit statistics ([Christensen & Kreiner, 2013](#); [Kreiner & Nielsen, 2023](#)) and by comparing observed item-rest score correlations with expected item-rest score correlations⁶ under the model ([Christensen & Kreiner, 2013](#); [Kreiner, 2011](#); [Kreiner & Nielsen, 2023](#)). Too-strong item discrimination will result in an item-rest score correlation that is stronger than expected under the RM, while too-low item discrimination results in the opposite ([Christensen & Kreiner, 2013](#)).

Conditional independence of items (i.e., no local response dependence) and absence of DIF were both tested using Kelderman's likelihood ratio test (Christensen & Kreiner, 2013; Kelderman, 1984; Kreiner & Christensen, 2013c). As items and exogenous variables used for testing DIF were binary or ordinal categorical variables, Goodman & Kruskal's Gamma (γ) (Goodman & Wallis, 1954) was used to assess the partial correlations among items and between items and exogenous variables. Both DIF and invariance were tested in relation to five central exogenous variables: level of field practice (level 1, level 2, level 3), teacher education program (ordinary, other), campus (A, B), gender (female, male), and age groups (24 years and younger, 25 years and older).

After fit of the various sets of item responses (i.e., the items in each scale) to either a RM or a GLLRM had been established, person parameters were estimated using weighted maximum likelihood estimation, as these have been shown to be less biased than maximum likelihood estimates (Kreiner & Christensen, 2013b; Warm, 1989). Unidimensionality was tested, DIF-equating of the sum scores was performed, and reliability and targeting were assessed.

Unidimensionality was tested across subscales within each of the three modes of field practice experience (Observed, Practiced, and received Feedback) by comparisons of the observed and expected gamma (γ) correlations under the unidimensional model (Horton et al., 2013). Scales measuring different constructs will be less strongly correlated than what is expected under the common unidimensional model.

Reliability was assessed using Hamon and Mesbah's (2002) Monte Carlo method for the reliability of Rasch scales, as this takes into account any LD between items and adjusts reliability accordingly.

Targeting of a population by a measure is good if information is high (or SEM is low) for the majority of the persons, and if items are not out of target when comparing item thresholds to the person parameter estimates or the scores (Kreiner & Christensen, 2013b). Targeting of the final theta scales was assessed numerically with two indices (Kreiner & Christensen, 2013b): the test information target index (the mean test information divided by the maximum test information) and the root mean squared error target index (the minimum standard error of measurement divided by the mean standard error of measurement). Both indices should have a value close to one. The targeting of each scale was illustrated graphically by plotting item maps showing the distribution of item threshold locations against weighted maximum likelihood estimations of the person parameter locations, the person parameters for the population assuming a normal distribution, and the information function. The target of the observed scores and the standard error of measurement of the observed scores were calculated to facilitate (a simpler) evaluation of targeting of these as well.

In GLLRMs with evidence of DIF, person parameters were estimated separately for each subgroup in order to obtain measurement on the same (logit) scale, and subsequently these were utilized for the purpose of DIF-equating the sum scores for each scale to bring the sum scores back on the same scale for unbiased subsequent statistical comparisons (Kreiner & Nielsen, 2023). DIF-equating of the sum scores is possible, as one of the properties of the RM, and only the RM, is the sufficiency of the sum scores for the person parameter estimates, as there is a one-to-one relationship between the person parameter estimates and the sum scores (Rasch, 1960).

All tests tested the null hypothesis that item responses fit the RM against different alternatives, thus effectively testing whether the data complied with the assumptions of the RM or subsequently a GLLRM, with significant p -values providing evidence against the respective model. p -values were evaluated as a continuous measure of evidence distinguishing between weak ($p < 0.05$), moderate ($p < 0.01$), and strong ($p < 0.001$) evidence

against the model (Cox et al., 1977). The Benjamini–Hochberg procedure was applied to adjust the false discovery rate (FDR), taking multiple testing into account when appropriate (Benjamini & Hochberg, 1995).

2.4. Differences in Scale Scores Dependent on Level of Field Practice

The relationship between the students' scores on the six field practice experience scales and the level of the students' latest completed field practice placement was assessed by testing whether students' mean scores on each of the scales differed significantly depending on the level of field practice placement. For this, a stepwise analysis of pairwise collapsibility of the mean scores across the three ordinal levels of field practice was used. This procedure is essentially a stepwise and pairwise t-test procedure, where categories with the largest p -values are collapsed into a single category if the p -value is larger than the set critical p -value (i.e., it is accepted that the mean scores in two categories are statistically equal). Thus, for three categories, it might be that two categories are collapsed in the first step, and then this new category is compared to the remaining third category, which may or may not be collapsed. This was done while controlling the false detection rate for multiple testing at the 5% critical level using the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995), also used in the item analyses using Rasch models.

2.5. Software

Item analyses, as well as a comparison of mean scale scores, were performed using the Digram software package (Kreiner, 2003; Kreiner & Nielsen, 2013, 2023), and item maps were created using R (R Core Team, 2021), with code developed by Pedro Ribeiro Santiago.

3. Results

Prior to engaging in item analysis using Rasch models, we wanted to gauge whether the developed items indeed reflected rarely occurring practices and activities for students to observe, practice on their own, or receive feedback on while in field practice. Thus, we looked at the prevalence of endorsement of each item in the three different modes of learning: observation, own practice, and receiving feedback (Table 2).

Table 2. Prevalence of endorsement of long-form FPE-DK items when observed, practiced, or received feedback in the study sample.

Item	n	Observed	Practiced	Feedback
1	307	39.4	91.2	76.8
2	306	31.4	66.0	53.6
3	306	53.6	93.2	76.8
4	306	47.4	88.6	69.6
5	306	35.6	68.0	53.3
6	306	39.9	82.4	62.5
7	306	53.6	76.9	57.5
8	285	67.7	94.7	66.9
9	285	68.4	81.7	68.7
10	284	74.3	96.5	83.4
11	284	58.8	77.4	58.0
12	283	52.3	75.6	51.9

Table 2. *Cont.*

Item	n	Observed	Practiced	Feedback
13	307	37.1	94.1	81.7
14	307	37.5	96.1	84.0
15	306	28.4	82.7	72.3
16	306	61.4	97.7	84.3
17	307	73.9	72.5	52.6
18	283	15.5	4.6	4.9
19	283	31.1	6.0	4.9
20	283	44.2	46.3	27.6
21	283	73.9	54.8	29.7
22	283	53.0	34.3	21.6
23	283	36.4	73.5	52.7
24	283	34.3	65.0	46.6
25	283	42.4	90.1	73.9

Notes. Italicized percentages above the bold line represent the original 12 items, while the 13 new items are below the bold line. Bottom tertile prevalence marked in green, middle tertile prevalence marked in yellow, and top tertile prevalences marked in red.

For the 13 newly developed items (items 13–25), the tertile distribution was 25.6% in the bottom tertile (i.e., the least commonly occurring opportunities to learn), 33.3% in the middle tertile, and 41.0% in the top tertile (i.e., the most commonly occurring opportunities to learn). Compared with the prevalence of the original 12 items, this is a marked improvement in terms of adding items that reflect more rarely occurring opportunities to learn, as the tertile distribution in the original items was 2.7% in the bottom tertile, 38.9% in the middle, and 58.3% in the top tertile.

The least commonly encountered practices among the newly generated items were item 18, Conduct a parents' evening/parents-teacher meetings, and item 19, Conduct a parent-teacher conference, independent of whether this was encountered as an opportunity to learn through observation, own practice, or receiving feedback. For more details, see Supplementary File S1 for the items' content.

The question is whether the new items represent enough of an improvement to improve targeting to an adequate level in terms of information achieved?

3.1. Initial Item Analysis

The initial item analyses of the extended 25-item version of the Observed, Practiced, and received Feedback scale of the FPE-DK showed that none of the extended scales fit the RM, as had been the case with the original 12-item scales. Further stepwise analysis to determine whether fit to log-linear Rasch models, which included uniform local dependence among items and/or uniform DIF, proved unsuccessful, and quite a few items had to be eliminated during analysis. The pool of items that was eliminated from each scale was quite similar⁷. Subsequent scrutinization of the item content across the eliminated and the remaining item sets led us to propose the division of the 25 items into two sets of items making up two subscales within each of the modes of opportunities to learn: 1 set of 17 items, including the original 12, in which content was focused on practices and activities directly connected with teaching students in the classroom, as well as 1 set of 8 items, in which content was focused on interpersonal practices surrounding or supporting teaching, such as collaboration with other professionals or parents, conflict resolution, and assessment. The first proposed dimension within all three modes of opportunities to learn was thus named "instruction", and the second was named "collaboration". Subsequently,

the six resulting subscales were analyzed using Rasch and graphical log-linear Rasch models, and after resolving the final models, dimensionality was tested within each of the three modes of opportunities to learn. Section 3.2 reports the results on the three instruction subscales, while Section 3.3 reports the results on the three collaboration subscales.

The central results for the final models for the six subscales are shown in Section 3.2 (the Observed-instruction, Practice-instruction, and received Feedback-instruction scales) as well as Section 3.3 (the Observed-collaboration, Practiced-collaboration, and received Feedback-collaboration scales). This is followed by results on dimensionality in Section 3.4. Additional results are provided in Supplementary Files S3–S8 (one for each scale), including results concerning item fit, item thresholds, and a table for converting summed scores to person parameter estimates on the logit scale. For scales with DIF, the Supplementary Materials Files S3–S8 also contain information for the DIF-equation of the summed scores. In addition, SPSS (v32.0.0) syntax facilitating easy conversion of scores and easy DIF-equating is available at zenodo.org: <https://doi.org/10.5281/zenodo.22832423>.

3.2. The Three Instruction Subscales

Neither of the instruction subscales fitted pure RMs. However, all three subscales fitted GLLRMs, which included interaction terms for locally dependent items in all three scales, and in the Practiced-instruction subscale also an interaction term for DIF of item o7 relative to gender (Figure 1). After including the local dependence and DIF interaction terms in the final GLLRMs, there were no issues with item fit (Supplementary Tables S3a, S5a, and S7a in Supplementary Materials Files S3, S5, and S7, respectively) or with global homogeneity or invariance across the included background variables (Table 3).

Table 3. Global tests of homogeneity and invariance for the final GLLRMs of the Observed-instruction, Practiced-instruction, and Feedback-instruction subscales.

Tests of Fit	Observed-I ^a			Practiced-I ^b			Feedback-I ^c		
	CLR	df	p	CLR	df	p	CLR	df	p
Global homogeneity	36.5	20	0.013 ⁺	34.7	18	0.010 ⁺	25.5	17	0.135
Invariance relative to:									
Field practice level	13.3	18	0.775	54.3	36	0.026 ⁺	38.9	34	0.259
Teacher program	15.5	12	0.214	26.2	18	0.096	29.9	17	0.027 ⁺
Campus	19.9	12	0.068	20.1	18	0.328	14.7	17	0.620
Gender	13.5	12	0.336	29.4	16	0.021 ⁺	31.7	17	0.016 ⁺
Age	17.7	12	0.125	40.9	18	0.002 ⁺	25.8	17	0.078

Notes. GLLRM: graphical log-linear Rasch model; Observed-I: Observed-instruction subscale; Practiced-I: Practiced-instruction subscale; Feedback-I: Feedback-instruction subscale; CLR: conditional likelihood ratio; df: degrees of freedom; p: p-value. Global homogeneity test compares item parameters in approximately equal-sized groups of high- and low-scoring students. ⁺: The critical limits for p-values after adjusting for false discovery rate were 5% limit $p = 0.0500$, and 1% limit $p = 0.0017$. ^a Model includes local dependence between items o3 and o8, items o4 and o5, items o12 and o14, and items o14 and o15. ^b Model includes local dependence between items p4 and p5, as well as DIF for item p7 relative to gender. ^c Model includes local dependence between items f13 and f14.

3.2.1. Local Dependence Among Items

The analyses found evidence of local dependence between some items in each of the three instruction subscales (Figure 1 and Table 3). Related to the observation of other teachers in the Observed-instruction subscale, four item pairs were strongly locally dependent (γ from 0.55 to 0.66, Supplementary Table S3b, Supplementary Materials File S3): items 3 (*Differentiate instruction*) and 8 (*Identify and respond to student thinking*), items 4 (*Connect subject specific content to students' prior knowledge*) and 5 (*Connect subject specific content to students' personal/cultural background*), items 12 (*Provide students subject specific feedback*) and 14 (*Plan short teaching modules*), and items 14 and 15 (*Plan long teaching modules*).

When it came to practicing yourself, Practiced-instruction subscale items 4 and 5 were also strongly locally dependent ($\gamma = 0.46$, Supplementary Table S5b, Supplementary Materials File S5), while items 13 (*Make plans for single lessons*) and 14 were strongly locally dependent when it came to receiving feedback on your practice in the received Feedback subscale ($\gamma = 0.80$, Supplementary Table S7b, Supplementary Materials File S7).

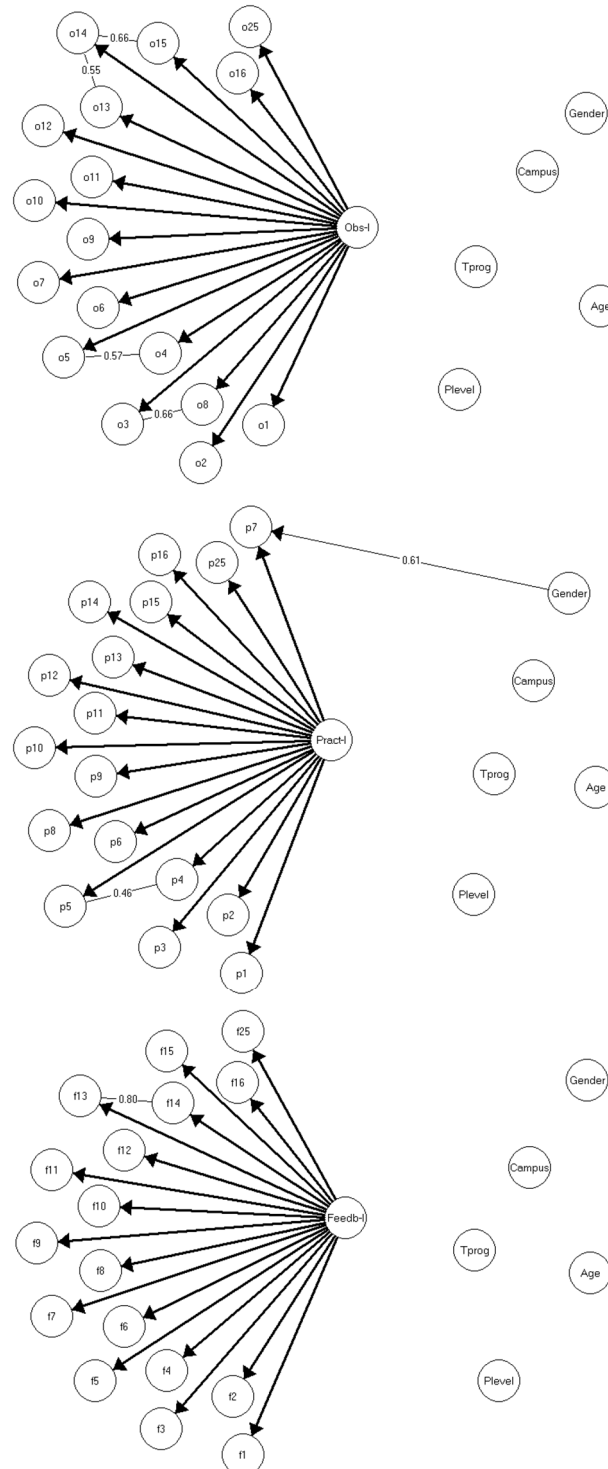


Figure 1. The final graphical log-linear Rasch models for the Observed-instruction (**top**), Practiced-instruction (**middle**), and Feedback-instruction (**bottom**) subscales. *Notes.* γ -correlations are partial Goodman and Kruskal's rank correlations for ordinal data. Campus (A vs. B), Tprog (ordinary teacher education program vs. other), Plevel (field practice level 1, level 2, level 3) Gender (female vs. male), Age (24 years and younger vs. 25 years and older).

3.2.2. Effect of Differential Item Functioning

For the three instruction subscales, evidence of DIF was only found within the Practiced-instruction subscale (Figure 1). Here, item 7 (*Facilitate students' use of technology*) functioned differentially relative to gender, so that male students systematically reported that they had had the opportunity to practice this in their field practice more often than female students, no matter their level on the scale. In other words, the item was easier to endorse for male students (Supplementary Table S5b, Supplementary Materials File S5). The effect of this DIF was slight in terms of bias (Table 4), and failure to adjust the scores for the DIF would not change the conclusions when comparing subgroups.

Table 4. Comparison of observed and DIF-equated scores for the Practiced-instruction subscale.

Subgroups of Students (n)	Observed Scores		DIF-Equated Scores		Bias
	Mean	SE	Mean	SE	
<i>Teacher program</i>					
Ordinary (243)	14.50	0.14	14.45	0.14	0.05
Other teaching program (38)	14.74	0.27	14.69	0.28	0.05
<i>p</i>	0.442		0.450		
<i>Campus</i>					
Campus A (194)	14.34	0.16	14.28	0.16	0.05
Campus B (87)	14.98	0.20	14.94	0.21	0.04
<i>p</i>	0.013		0.012		
<i>Field practice level</i>					
Level 1 (99)	13.87	0.22	13.80	0.23	0.06
Level 2 (93)	14.51	0.23	14.46	0.24	0.04
Level 3 (89)	15.30	0.17	15.27	0.17	0.03
<i>p</i>	<0.001		<0.001		
<i>Gender (source of DIF)</i>					
Female (200)	14.50	0.15	14.50	0.15	0.00
Male (81)	14.63	0.23	14.46	0.24	0.17
<i>p</i>	0.627		0.909		
<i>Age groups</i>					
24 years and younger (144)	14.51	0.18	14.46	0.18	0.05
25 years and older (137)	14.56	0.18	14.51	0.18	0.05
<i>p</i>	0.828		0.834		

3.2.3. Targeting and Reliability

Targeting was acceptable for the Observed-instruction and Feedback-instruction subscales, with more than 60% of the maximum information obtained (Table 5). The item maps (Figure 2) illustrate both the level of information along the scales, the point of maximum information (the target), and the degree of alignment of the person estimates and item difficulties (i.e., item thresholds). Figure 2 (top) shows that for the Observed-instruction subscale, the student measures are located quite evenly across the scale, though with a spike at the lower end, while the item difficulties are located within a narrower range around the middle of the scale. The item map for the Feedback-instruction subscale (Figure 2, middle) shows a poorer alignment of person estimates and item difficulties than for the Observed-instruction subscale, but better than the Practiced-instruction subscale. For the Practiced-instruction subscale, targeting was less good, as only approximately 50% of the maximum information was obtained, and least for the male students. In Figure 2, this poorer targeting of the Practiced-instruction subscale is explained by item difficulties being in the lower half of the scale, while the person estimates are distributed along the entire scale for both male and female students.

Table 5. Targeting and reliability for the Observed-instruction, Practiced-instruction, and Feedback-instruction subscales (n = 281).

Subscale and DIF Groups (n)	θ	Target		Person Parameter (θ) Population							Person Raw Score Population			
		Target Info	Target RMSE	Average θ	SD	Average Test Info	Average RMSE	Target Info Index	Target RMSE Index	PSI	Target	Average Score	Average SEM	r^a
Obs-I (281)	−0.01	4.063	0.496	−0.38	1.76	2.719	0.637	0.669	0.779	0.995	9.57	8.27	1.60	0.90
Pract-I														
Female (200)	0.11	3.416	0.547	2.27	0.92	1.754	0.732	0.513	0.740	0.735	8.98	14.50	1.28	0.62
Male (81)	−0.04	3.469	0.487	2.23	0.91	1.689	0.739	0.487	0.726	0.732	8.78	14.63	1.26	0.61
Feedb-I (281)	−0.36	3.853	0.509	1.13	1.73	2.346	0.687	0.609	0.742	1.067	7.77	11.75	1.46	0.89

Notes. Targeting and reliability results are only shown for subgroups if DIF was present. Obs-I: Observed-instruction subscale; Pract-I: Practiced-instruction subscale; Feedb-I: Feedback-instruction subscale; FPL: Field Practice level. The target info index = the ratio between the average test information and the target information. The target RMSE index = the ratio between the average RMSE and the RMSE at target. PSI = Person Separation Index = the ratio between the variance of θ in the population and the variance of the estimates of θ . SEM = the standard error of measurement of the observed score. r = reliability estimated as the ratio between the variance of the expected true score and the observed mean score. ^a Weighted mean reliability for the Pract-I subscale across DIF-subgroups = 0.62.

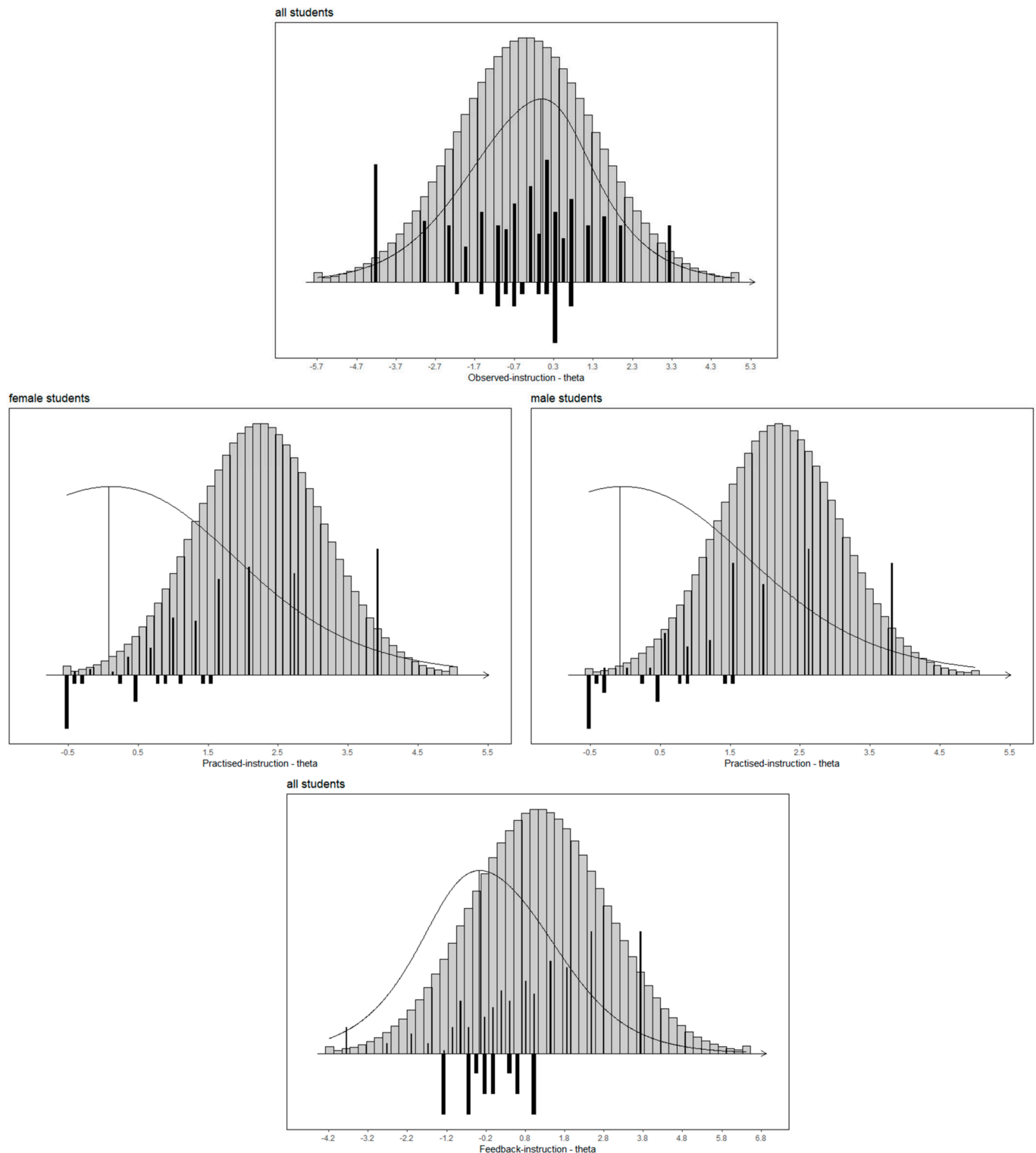


Figure 2. Item maps for the Observed-instruction (**top**), Practised-instruction (**middle**), and Feedback-instruction (**bottom**) subscales with distributions of person parameter estimates and information curve above item difficulties. *Notes.* Person parameter estimates are weighted maximum likelihood estimates (theta) and illustrate the distribution of these for the study sample (black bars above the line) and for the population under the assumption of normality (grey bars above the line), as well as the information curve, relative to the distribution of the item difficulties (black bars below the line).

The same pattern was found with regard to the reliability of the three instruction subscales. It was excellent for the Observed-instruction and Feedback-instruction subscales (Table 5), but poor for the Practised-instruction subscale for both male and female

students. The item maps illustrate this by showing that the variation is less for the Practiced-instruction subscales compared with the other two subscales.

3.3. The Three Collaboration Subscales

Neither of the collaboration subscales fitted pure Rasch models, but each fitted a GLLRM, as was the case with the three instruction subscales. The GLLRMs for the three collaboration subscales all included interaction terms for the same three pairs of locally dependent items, as well as interaction terms for DIF for one or two items in each scale and in relation to level of field practice in all three cases (Figure 3). After including the local dependence and DIF interaction terms in the final GLLRMs, there were no issues with item fit (Supplementary Tables S4a, S6a, and S8a in Supplementary Files S4, S6, and S8, respectively) or with global homogeneity or invariance across the included background variables (Table 6).

Table 6. Global tests of homogeneity and invariance for the final GLLRMs of the Observed-collaboration, Practiced-Collaboration, and Feedback-collaboration subscales.

Tests of Fit	Observed-C ^a			Practiced-C ^b			Feedback-C ^c		
	CLR	df	p	CLR	df	p	CLR	df	p
Global homogeneity	13.2	12	0.353	15.5	14	0.347	10.1	12	0.610
Invariance relative to:									
Field practice level	13.3	18	0.775	31.1	16	0.013 ⁺	26.6	18	0.087
Teacher program	15.5	12	0.214	9.0	14	0.831	21.8	12	0.040 ⁺
Campus	19.9	12	0.068	25.1	14	0.033 ⁺	17.6	12	0.129
Gender	13.5	12	0.336	15.5	14	0.343	12.4	12	0.411
Age	17.7	12	0.125	14.3	14	0.429	9.5	12	0.663

Notes. GLLRM: graphical log-linear Rasch model; CLR: conditional likelihood ratio; df: degrees of freedom; p: p-value; Observed-C: Observed-collaboration subscale; Practiced-C: Practiced-collaboration subscale; Feedback-C: Feedback-collaboration subscale. Global homogeneity test compares item parameters in approximately equal-sized groups of high- and low-scoring students. ⁺: The critical limit for the p-values at the 5% level after adjusting for false discovery rate was $p = 0.0083$. ^a Model includes local dependence between items o18 and o19, items o21 and o22, and items o23 and o24, as well as DIF for item o19 relative to level of field practice. ^b Model includes local dependence between items p18 and p19, items p21 and p22, and items p23 and p24, as well as DIF for items p23 and p24 relative to level of field practice. ^c Model includes local dependence between items f18 and f19, items f21 and f22, and items f23 and f24, as well as DIF for item f20 relative to level of field practice.

3.3.1. Local Dependence

There was evidence of local dependence between some items in each of the three collaboration subscales (Figure 3 and Table 6). In relation to the observation of other teachers in the Observed-instruction subscale, three item pairs were strongly locally dependent (Supplementary Table S4b, Supplementary Materials File S4): items 18 (*Conduct a parents' evening/parent-teacher meetings*) and 19 (*Conduct a parent-teacher conference*) ($\gamma = 0.62$), items 21 (*Take part in teamwork among teachers*) and 22 (*Take part in teamwork with professionals from other professions concerning students*) ($\gamma = 0.62$), as well as items 23 (*Do formative assessment with students*) and 24 (*Do summative assessment with students*) ($\gamma = 0.93$).

When it came to practicing yourself in the Practiced-collaboration subscale, the same three item pairs were strongly locally dependent (Supplementary Table S6b, Supplementary Materials File S6): items 18 and 19 ($\gamma = 0.93$), items 21 and 22 ($\gamma = 0.52$), as well as items 23 and 24 ($\gamma = 0.66$).

Finally, in relation to receiving feedback on your practice in the received Feedback-collaboration subscale, the three item pairs were again strongly locally dependent (Supplementary Table S8b, Supplementary Materials File S8): items 18 and 19 ($\gamma = 0.94$), items 21 and 22 ($\gamma = 0.73$), as well as items 23 and 24 ($\gamma = 0.81$).

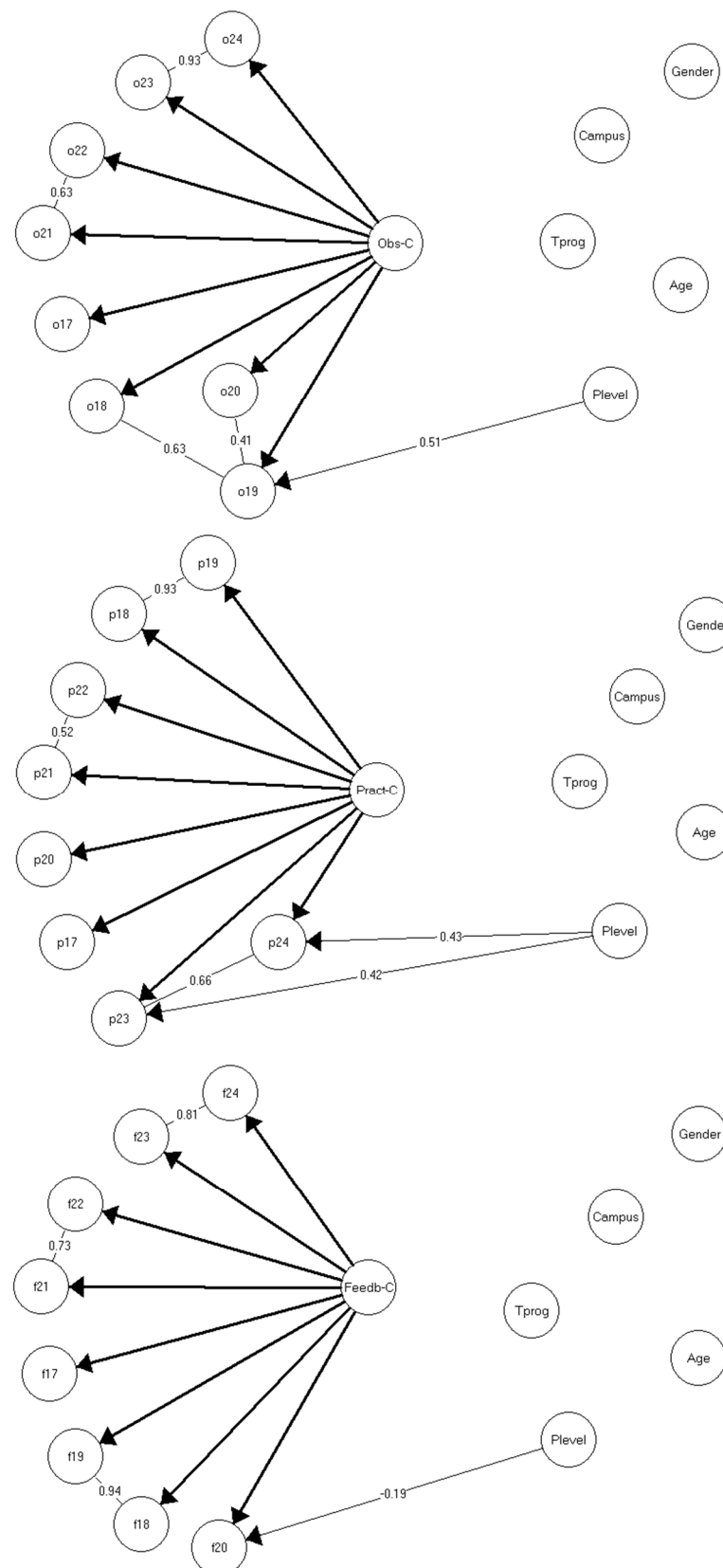


Figure 3. The final graphical log-linear Rasch models for the Observed-collaboration (**top**), Practiced-collaboration (**middle**), and Feedback-collaboration (**bottom**) subscales. *Notes.* γ -correlations are partial Goodman and Kruskal's rank correlations for ordinal data. Campus (A vs. B), Tprog (ordinary teacher education program vs. other), Plevel (field practice level 1, level 2, level 3) Gender (female vs. male), Age (24 years and younger vs. 25 years and older).

3.3.2. Effect of Differential Item Functioning

Evidence of DIF relative to the level of field practice just completed was found for all three collaboration subscales (Figure 3, Supplementary Tables S4b, S6b, and S8b in Supplementary Materials Files S4, S6, and S8, respectively).

In the Observed-collaboration subscale, it was item 19 (*Conduct a parent–teacher conference*) that functioned differentially relative to field practice level ($\gamma = 0.49$), so that the more advanced the field practice that had just been completed was, the more students were systematically inclined to report they had had the opportunity to observe this in their field practice, no matter their level on the Observed-collaboration scale (Supplementary Table S4b, Supplementary Materials File S4). The effect of this DIF was slight in terms of bias, except for the subgroups defined by field practice level (Table 7). Failure to adjust the scores for the DIF would not change the conclusions when comparing subgroups, but would change the strength of the evidence from weak to very strong for a difference between the joint group of students who had just completed field practice level 1 or 2 and the groups who had completed field practice level 3.

Table 7. Comparison of observed and DIF-equated scores for the Observed-collaboration subscale.

Subgroups of Students (n)	Observed Scores		DIF-Equated Scores		Bias
	Mean	SE	Mean	SE	
Teacher program					
Ordinary (243)	3.80	0.13	3.68	0.13	0.12
Other teaching program (38)	2.63	0.39	2.54	0.38	0.10
<i>p</i>	0.004		0.004		
Campus					
Campus A (194)	3.56	0.14	3.46	0.14	0.10
Campus B (87)	3.84	0.25	3.69	0.24	0.15
<i>p</i>	0.324		0.411		
Field practice level (<i>DIF source</i>)					
Level 1 (99)	3.82	0.19	3.82	0.19	0.00
Level 2 (93)	3.92	0.21	3.83	0.21	0.10
Level 3 (89)	3.16	0.24	2.90	0.23	0.26
<i>p</i>	0.036		0.002		
Gender					
Female (200)	3.55	0.15	3.43	0.15	0.12
Male (81)	3.89	0.22	3.78	0.21	0.11
<i>p</i>	0.191		0.183		
Age groups					
24 years and younger (144)	3.92	0.17	3.84	0.16	0.09
25 years and older (137)	3.35	0.18	3.21	0.18	0.14
<i>p</i>	0.021		0.010		

In the Practiced-collaboration subscale, items 23 (*Do formative assessment with students*) and 24 (*Do summative assessment with students*) functioned differentially relative to field practice level ($\gamma = 0.42$ and 0.43). In both cases, the DIF signified that the more advanced the just-completed field practice was, the more students were systematically inclined to report they had had the opportunity to practice this in their field practice, no matter their level on the Practiced-collaboration scale (Supplementary Table S6b, Supplementary Materials File S6). The effect of this DIF was large in terms of bias (Table 8), in fact so large that failure to adjust the scores for the DIF would change the conclusion when comparing subgroups defined by field practice level and lead to a type I error claiming that there was a difference between subgroups, when in fact there was not, when scores had been equated for DIF.

In the Feedback-collaboration subscale, item 20 (*Communicate in writing with parents*) functioned differentially relative to field practice level, but only moderately so and in the opposite direction than was the case for the Observed-collaboration and Practiced-collaboration

subscales ($\gamma = -0.19$). Thus, the DIF signified that the less advanced the just-completed field practice was, the more students were systematically inclined to report that they had the opportunity to receive feedback on this in their field practice, no matter their level on the Feedback-collaboration scale (Supplementary Table S7b, Supplementary Materials File S7). The effect of this DIF was slight in terms of bias (Table 9), and failure to adjust the scores for the DIF would not change the conclusions when comparing subgroups.

Table 8. Comparison of observed and DIF-equated scores for the Practiced-collaboration subscale.

Subgroups of Students (n)	Observed Scores		DIF-Equated Scores		Bias
	Mean	SE	Mean	SE	
<i>Teacher program</i>					
Ordinary (243)	3.62	0.11	3.30	0.12	0.31
Other teaching program (38)	3.32	0.21	2.84	0.22	0.48
<i>p</i>	0.197		0.057		
<i>Campus</i>					
Campus A (194)	3.43	0.11	3.11	0.12	0.32
Campus B (87)	3.90	0.19	3.52	0.20	0.37
<i>p</i>	0.034		0.083		
<i>Field practice level (DIF source)</i>					
Level 1 (99)	3.25	0.16	3.25	0.16	0.00
Level 2 (93)	3.56	0.18	3.07	0.20	0.48
Level 3 (89)	3.96	0.16	3.40	0.18	0.55
<i>p</i>	0.008		0.472		
<i>Gender</i>					
Female (200)	3.57	0.12	3.22	0.13	0.34
Male (81)	3.60	0.16	3.28	0.17	0.32
<i>p</i>	0.845		0.793		
<i>Age groups</i>					
24 years and younger (144)	3.68	0.14	3.39	0.15	0.29
25 years and older (137)	3.47	0.14	3.08	0.15	0.38
<i>p</i>	0.278		0.140		

Table 9. Comparison of observed and DIF-equated scores for the Feedback-collaboration subscale.

Subgroups of Students (n)	Observed Scores		DIF-Equated Scores		Bias
	Mean	SE	Mean	SE	
<i>Teacher program</i>					
Ordinary (243)	2.49	0.13	2.58	0.13	−0.09
Other teaching program (38)	1.95	0.28	2.03	0.29	−0.08
<i>p</i>	0.083		0.082		
<i>Campus</i>					
Campus A (194)	2.19	0.13	2.28	0.14	−0.09
Campus B (87)	2.91	0.23	3.00	0.23	−0.10
<i>p</i>	0.007		0.007		
<i>Field practice level (DIF source)</i>					
Level 1 (99)	2.11	0.17	2.11	0.17	0.00
Level 2 (93)	2.33	0.22	2.53	0.23	−0.20
Level 3 (89)	2.83	0.22	2.91	0.22	−0.07
<i>p</i>	0.033		0.016		
<i>Gender</i>					
Female (200)	2.28	0.15	2.36	0.15	−0.08
Male (81)	2.75	0.19	2.86	0.20	−0.11
<i>p</i>	0.048		0.043		
<i>Age groups</i>					
24 years and younger (144)	2.45	0.17	2.53	0.17	−0.08
25 years and older (137)	2.37	0.17	2.47	0.17	−0.10
<i>p</i>	0.738		0.789		

3.3.3. Targeting and Reliability

Targeting was good to excellent for the Observed-collaboration, the Practiced-collaboration, and the received Feedback-collaboration subscales, as between 70% and 89% of the maximum information was obtained across student subgroups defined by level of field practice (Table 10). The item maps in Figure 4 illustrate this, as the degree of alignment of the person estimates and the item difficulties appears balanced, and the test information is high along most of the scales. Targeting was poorest for the Observed-collaboration scale for students who had completed field practice level 3, as well as for the received Feedback-collaboration subscale for students who had just completed field practice level 2 (Table 10). The latter is illustrated in the item maps, where there are relatively many students towards the lower end of the scales, where there is little information (Figure 4).

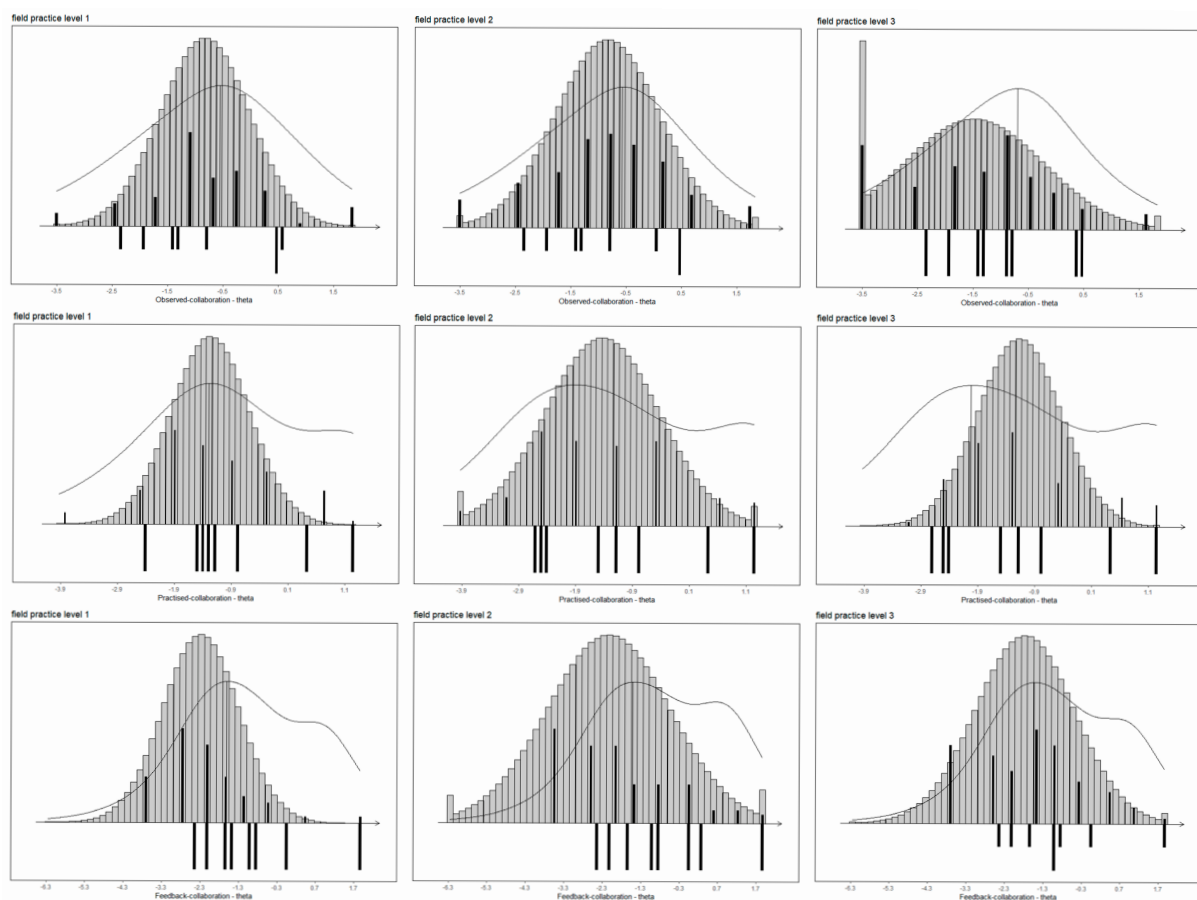


Figure 4. Item maps for the Observed-collaboration (**top**), Practiced-collaboration (**middle**), and Feedback-collaboration (**bottom**) subscales with distributions of person parameter estimates and information curve above item difficulties. *Notes.* Person parameter estimates are weighted maximum likelihood estimates (theta) and illustrate the distribution of these for the study sample (black bars above the line) and for the population under the assumption of normality (grey bars above the line), as well as the information curve, relative to the distribution of the item difficulties (black bars below the line).

A similar pattern was found with regard to the reliability of the three collaboration subscales. It was highest for the Observed-collaboration and received Feedback-collaboration subscales and lower for the Practiced-collaboration subscale (Table 10). However, the reliability was, by conventional standards, poor for most subgroups, as only the reliability of the Observed-collaboration scale for students who had completed field practice level 3, as well as of the received Feedback-collaboration subscale for students who had completed field practice level 2, was above 0.70—and it is notable that these are the same groups for which targeting was the poorest.

Table 10. Targeting and reliability for the Observed-collaboration subscale, Practiced-collaboration subscale, and Feedback-collaboration subscale (n = 281).

Subscale and DIF Groups (n)	θ	Target		Person Parameter (θ)							Person Raw Score			
		Target Info	Target RMSE	Population							Population			
				Average θ	SD	Average Test Info	Average RMSE	Target Info Index	Target RMSE Index	PSI	Target	Average Score	Average SEM	r^a
Obs-C														
FPL 1 (99)	−0.53	1.987	0.709	−0.84	0.80	1.710	0.765	0.860	0.927	0.54	4.34	3.82	1.30	0.53
FPL 2 (93)	−0.56	2.069	0.695	−0.81	0.94	1.696	0.759	0.820	0.916	0.63	4.44	3.92	1.29	0.61
FPL 3 (89)	−0.69	2.128	0.686	−1.50	1.24	1.486	0.790	0.699	0.867	0.84	4.46	3.16	1.19	0.71
Pract-C														
FPL 1 (99)	−1.27	1.752	0.756	−1.28	0.68	1.564	0.812	0.893	0.930	0.408	3.24	3.25	1.25	0.42
FPL 2 (93)	−1.91	1.500	0.816	−1.44	1.00	1.307	0.864	0.871	0.945	0.600	2.93	3.56	1.14	0.57
FPL 3 (89)	−2.00	1.493	0.818	−1.18	0.73	1.314	0.878	0.881	0.932	0.413	2.84	3.96	1.14	0.40
Feedb-C														
FPL 1 (99)	−1.59	1.716	0.763	−2.30	0.98	1.284	0.790	0.748	0.966	0.760	3.08	2.11	1.11	0.58
FPL 2 (93)	−1.47	1.588	0.793	−2.13	1.60	1.100	0.867	0.693	0.916	1.220	2.94	2.33	1.01	0.75
FPL 3 (89)	−1.52	1.697	0.768	−1.77	1.24	1.329	0.797	0.783	0.963	0.862	3.08	2.83	1.13	0.69

Notes. Targeting and reliability results are only shown for subgroups if DIF was present. Obs-C: Observed-collaboration subscale; Pract-C: Practiced-collaboration subscale; Feedb-C: Feedback-collaboration subscale; FPL: Field Practice level. The target info index = the ratio between the average test information and the target information. The target RMSE index = the ratio between the average RMSE and the RMSE at target. PSI = Person Separation Index = the ratio between the variance of θ in the population and the variance of the estimates of θ . SEM = the standard error of measurement of the observed score. r = reliability estimated as the ratio between the variance of the expected true score and the observed mean score. ^a Weighted mean reliability for subscales with DIF: Obs-C = 0.61, Pract-C = 0.46, and Feedb-C = 0.67.

3.4. Unidimensionality

The tests of unidimensionality of each of the pairs of subscales, i.e., the instruction subscales and the collaboration subscales within each of the modes of opportunities to learn, showed varied results (Table 11). For the Observed subscales and the Feedback subscales, unidimensionality was clearly rejected ($p < 0.0001$ in both cases), and thus, within each of these modes of opportunities to learn, two subscales are necessary. For the two subscales in the practiced mode of opportunities to learn, unidimensionality was, however, accepted. As the preliminary analysis (c.f. Section 3) had shown that it was not possible to reach fit to a model with 25 items, the finding was not a simple one.

Table 11. Tests of unidimensionality.

Subscales	Observed γ	Expected γ	SE Exp γ	Asymptotic p	Exact p^a
Observed-instruction vs. Observed-collaboration	0.50	0.68	0.028	<0.0001	<0.0001
Practiced-instruction vs. Practiced-collaboration	0.45	0.48	0.043	0.524	0.165
Feedback-instruction vs. Feedback-collaboration	0.59	0.69	0.028	<0.001	<0.0001

Notes. γ -correlations are Goodman and Kruskal's rank correlations for ordinal data. Observed γ = observed correlations between subscales. Expected γ = expected correlations under the model.
^a. Parametric bootstrapping with 1000 samples.

We thus proceeded to test the fit of the joint model of 25 items to a GLLRM with the DIF and LD discovered within each subscale. Initially, fit was rejected for this joint model, as invariance was rejected for level of field practice and age, and more LD terms and DIF terms were suggested for inclusion, though item fit was accepted. Further iterations, where we included one additional interaction term for LD and three additional interaction terms for DIF, were conducted, with strong evidence of negative local dependence between items from the two subscales, as well as items within the Practiced-collaboration subscale (results not included). These results collectively cast enough doubt on the unidimensionality of the two practiced subscales, as we could not obtain fit for a model with 25 items, but we could obtain fit for each of the two subscales. We therefore concluded that Practiced-instruction and Practiced-collaboration were two separate subscales, as was the case for the Observed subscales and the Feedback subscales, even though the evidence was more complex.

3.5. Differences in Scale Scores Dependent on Level of Field Practice

As the majority of the students are within target of the items in the six scales (c.f. Figures 3 and 4), and the sum scores are counting scales, where the scores signify the number of opportunities to learn, the summed scale score facilitates direct interpretation of the number of opportunities to learn related to instruction and collaboration that are experienced in each of the three modes of experience (Observed, Practiced, and received Feedback). We therefore used the summed scale scores for comparison of mean scores across levels of field practice.

Table 12 shows the mean scores on each of the three instruction subscales stratified by field practice level and after the stepwise analysis of pairwise collapsibility of the mean scores across levels of field practice. With regard to the Observed-instruction scale, equality of the mean scores for students who had completed field practice level 1 and level 2 was accepted ($p = 0.480$), and thus the two groups were collapsed to a single group with a common mean score on the Observed scale. Equality of the common mean score for students who had completed field practice levels 1 + 2 ($M = 9.32$) and the mean score

for students who had completed level 3 field practice ($M = 6.06$) was strongly rejected ($p < 0.001$). The received Feedback-instruction scores were also found to be equal for students who had completed levels 1 + 2 field practice ($p = 0.574$), and the common mean score for this combined group ($M = 12.82$) was not equal to the mean score for students who had completed level 3 field practice ($M = 11.29$) ($p = 0.006$). Lastly, for the Practiced-instruction scale score, equality of the mean scores for students who had completed levels 1, 2, and 3 field practice was rejected ($p < 0.001$).

Table 12. Common mean instruction scale scores for students in collapsed groups of field practice levels ($n = 281$).

Field Practice	Observed-Instruction		Practiced-Instruction ^a		Feedback-Instruction	
	Mean	SE	Mean	SE	Mean	SE
Level 1			13.80	0.23		
Level 2	9.32	0.35	14.46	0.24	12.82	0.33
Level 3	6.06	0.54	15.27	0.17	11.29	0.45

Notes. ^a DIF-adjusted scores were used.

Students who had completed field practice level 3 scored significantly and substantially lower on Observed-instruction than did students who had completed field practice levels 1 + 2 ($M\text{-diff} = -3.36$). Thus, the more advanced and timewise later field practice afforded students fewer opportunities to learn through observation of other teachers within the dimension of instruction than was the case for earlier field practices. The same was the case, but to a lesser degree, with regard to received Feedback ($M\text{-diff} = 1.53$). Conversely, the mean Practiced-instruction score for students who had completed field practice level 3 was significantly but not substantially higher than for students who had completed level 2 ($M\text{-diff} = 0.81$), and the same was the case between students who had completed level 2 compared with those who had completed level 1 ($M\text{-diff} = 0.66$). Thus, with increasing complexity in field practice, more students appear to have been afforded opportunities to learn through own practice within the instruction dimension.

Turning to the collaboration subscales (Table 13), we found that equality of the mean Observed-collaboration scores was accepted for students who had completed field practice level 1 and level 2 ($p = 0.989$), and the two groups were collapsed to a single group with a common mean score. Equality of the common mean score for students who had completed field practice levels 1 + 2 ($M = 3.82$) and the mean score for students who had completed level 3 field practice ($M = 2.85$) was strongly rejected ($p < 0.001$).

Table 13. Common mean collaboration scale scores for students in collapsed groups of field practice levels ($n = 281$).

Field Practice	Observed-Collaboration		Practiced-Collaboration		Feedback-Collaboration	
	Mean	SE	Mean	SE	Mean	SE
Level 1						
Level 2	3.82	0.14	3.24	0.26	2.11	0.17
Level 3	2.85	0.22			2.73	0.16

Notes. DIF-adjusted scores were used for all three collaboration scales.

The received Feedback-collaboration scores were also found to be equal for students who had completed levels 1 + 2 field practice ($p = 0.244$), and the common mean score for this combined group ($M = 2.11$) was not equal to the mean score for students who

had completed level 3 field practice ($M = 2.73$) ($p = 0.008$). Lastly, with the Practiced-collaboration scores, equality of the mean scores for students who had completed levels 1, 2 and 3 field practice was accepted ($p = 0.314$).

Students who had completed field practice level 3 scored significantly and substantially lower on Observed-collaboration than did students who had completed field practice levels 1 + 2 ($M\text{-diff} = -0.97$). Thus, again, the more advanced and timewise later field practice afforded students fewer opportunities to learn through observation of other teachers within the collaboration dimension than was the case for earlier field practices. The opposite was the case, but to a lesser degree, with regard to received Feedback-collaboration ($M\text{-diff} = 0.62$). Quite differently, no differences were found in the mean Practiced-collaboration scores across students having completed field practice at levels 1, 2, or 3. Thus, increasing complexity in field practice is not associated with differences in the opportunities to learn through own practice afforded to students within the collaboration dimension.

4. Discussion and Implications

As this is the first validity study of the expanded long-form FPE-DK, we will discuss the psychometric properties briefly in relation to the original short-form scales, while focusing on the objective of the study: to improve targeting to the study population, and the implications the results have for use of the scales. We will also discuss the differences in scores dependent on the level of field practice in the current study against previous findings.

4.1. The Psychometric Properties of the Six Scales

The results showed that none of the six scales in the extended long-form FPE-DK fitted pure RMs, as had been the case with the three field practice experience scales in the short-form FPE-DK (Nielsen, 2021). However, all six scales fitted GLLRMs with local dependence between some items, and four of the scales also included items that functioned differentially. The scale that was the closest fit to the pure RM was the Feedback-instruction scale, where the only departure from the model was in the form of a single pair of locally dependent items. This was followed by the Practiced-instruction scale, which also included a single pair of locally dependent items, as well as an item functioning differentially relative to gender. The scales with the most departures from the pure RM were the three collaboration scales, as they all included three pairs of locally dependent items, as well as one or two items functioning differentially relative to the just completed level of field practice.

While fit to the RM is desirable in the teacher education context in which the scales are intended, fit to GLLRMs with uniformly locally dependent items and/or uniform DIF across some subgroupings of students complicates matters somewhat, but still allows practitioners (and researchers) to use the summed scale scores for assessment. Thus, the summed and, in case of DIF, DIF-equated sum scores can be used to assess the degree to which students experience opportunities to learn instruction-related practices and collaboration-related practices through observation of other teachers, own practice, and receiving feedback, as these scores are sufficient for the estimated Rasch scores resulting from the psychometric models. The Rasch scores (i.e., the person parameter estimates) can, of course, also be used if researchers or practitioners prefer those.

4.1.1. Local Dependence

As mentioned, all six scales included items that were locally dependent, which is in contrast to the findings of Nielsen (2021) with the short-form FPE-DK. However, two of the instruction scales, which included the 12 items from the FPE-DK short form, only had

a single pair of locally dependent items. For the received Feedback-instruction scale, the locally dependent items were items added in the current study (items 13 and 14), while the locally dependent items in the Practiced-instruction scale were original items from the short-form FPE-DK (items 4 and 5). These are considered minor differences from the previous finding by Nielsen (2021). The third instruction scale (i.e., Observed-instruction) included four pairs of locally dependent items, with only two of these being items that were added in the current study (items 13 and 14, and items 14 and 15), while the remaining local dependence was with items from the short-form FPE-DK (items 4 and 5, as in the Practiced-instruction scale, and items 3 and 8).

In addition, in the three collaboration scales, which consist only of items added in the current study, some items were locally dependent—interestingly, it is the same three pairs of items across the three scales. The first pair of locally dependent items were item 18 (*Conduct a parents' evening/parent–teacher meetings*) and 19 (*Conduct a parent–teacher conference*). The second pair was item 21 (*Take part in teamwork among teachers*) and 22 (*Take part in teamwork with professionals from other professions concerning students*). The third pair was item 23 (*Do formative assessment with students*) and 24 (*Do summative assessment with students*). While the strengths of the associations varied somewhat across the three scales, all were very strong. This means that the occurrence of opportunities to learn one teacher practice (e.g., formative assessment), in a locally dependent pair of practices, follows the occurrence of the opportunity to learn the other teacher practice (e.g., summative assessment), and vice versa. In addition, this is the case whether the opportunities to learn are via observation of other teachers, via own practice, or via receiving feedback. These opportunities to learn items have not previously been investigated. However, a recent study on changes in the self-efficacy for single teacher practices across field practices in the Danish teacher education context found that student teachers' self-efficacy for writing to/with parents, speaking with parents, and participating in parents' evenings decreased substantially in the first year, and still decreased slightly in the second year (Nielsen et al., 2026). These findings may be connected to the generally low scores on the collaboration scales in the current study, indicating that only 2–4 practices presented opportunities to learn across modes (c.f. the mean scores in Table 13). Future research could successfully combine measurement of changes in practice self-efficacy across field practice with retrospective measurement of opportunities to learn while in field practice, using items addressing similar practices in the two instruments (Nielsen et al., 2024, 2026a, 2026; Nielsen & Pettersson, 2025).

Conditional independence of items is a requirement in all IRT models, but a requirement that is rarely met in practice. Thus, we consider the limited local dependence discovered in the current study as a positive finding.

4.1.2. Differential Item Functioning

A total of five items across four of six scales functioned differentially. With the instruction scales, only the Practiced-instruction scale included an item that functioned differentially. This was item 7 (*Facilitate students' use of technology*), which functioned differentially relative to gender, so that male students systematically reported that they had the opportunity to practice this in their field practice more often than female students, no matter their level on the scale. Previous research found that male teachers were more involved in class use of computers; however, when considering other relevant predictors, gender differences had no predictive power for class use of computers (van Braak, 2001). A more recent study found that male teachers had higher computer self-efficacy in relation to operational skills, but also that there were gender differences in teachers' computer self-efficacy in relation to use of computers for instructional purposes (Scherer & Siddiq, 2015). However, within the current study, we cannot say whether this is because male

students in this case are more efficacious towards using technology in classrooms, if they seek and gain these opportunities more often, or if it is an equity issue. Future research should seek to replicate or dispute this finding with larger and more diverse samples of student teachers. However, qualitative observational studies might also shed some light on the mechanisms at play.

The four remaining items that functioned differentially were items in the collaboration scales that were all added in the current study.

In the Observed-collaboration scale, item 19 (*conduct a parent–teacher conference*) functioned differentially relative to field practice level, so that the more advanced the just-completed field practice was, the more inclined students were to report they had the opportunity to observe a parent–teacher conference, independent of the level of Observed-collaboration. This finding reflects the field practice skills objectives, as the objectives that include collaboration with parents become more advanced as the level of field practice progresses. Thus, at level 1, students were required to be able to communicate with parents about the teaching and the aims of the school; at level 2, to communicate with parents in writing or orally about the content and objectives of planned teaching; and at level 3, to communicate with parents about the children’s participation and their well-being at school (Læreruddannelsen, 2026). In addition, level 3 includes a knowledge objective on forms of collaboration in relation to meeting with parents, which could include parent–teacher conferences.

Also related to collaboration with parents, item 20 (*Communicate in writing with parents*) in the Feedback-collaboration scale functioned differentially relative to field practice level, but only moderately so and in the opposite direction than was the case for the Observed-collaboration scale. Here, the less advanced the just-completed field practice was, the more inclined students were to report they had the opportunity to receive feedback on written communication with parents, independent of the level of Feedback-collaboration. Thus, while written communication with parents is a skills objective at level 2, it appears that receiving feedback on this is most commonly occurring at level 1, and systematically so. This is an interesting finding, as it is somewhat contrary to the objectives for field practices, and thus indicates that the reality of field practice does not always follow the prescribed progression.

In the Practiced-collaboration scale, items 23 (*Do formative assessment with students*) and 24 (*Do summative assessment with students*) also functioned differentially relative to field practice level, so that the more advanced the just-completed field practice was, the more inclined students were to report they had the opportunity to practice formative and summative assessment on their own, independent of the level of Practiced-collaboration. Again, the finding reflects the field practice skills objectives, as the objectives that include evaluation—and implicitly assessment—become more advanced as the level of field practice progresses (Læreruddannelsen, 2026). Thus, at level 1, students were required to be able to analyze teaching sequences for the purpose of developing the teaching; at level 2, to observe the individual pupil’s learning in order to develop the teaching; and at level 3, to evaluate pupils’ learning outcomes and the overall effect of teaching. Furthermore, at level 3, a knowledge objective on formative and summative evaluation methods was also included.

Thus, while DIF is not desirable in measurement scales, in this case the DIF is very informative when considered in relation to the field practice objectives, as the DIF points towards a systematic and structural progression taking place in accordance with the skills and knowledge objectives for the field practice placements when it comes to observation of parent–teacher collaboration and own practice of assessment.

4.1.3. Unidimensionality

The formal tests of unidimensionality across each pair of instruction and collaboration subscales rejected unidimensionality for the Practiced-instruction and the Practiced-collaboration subscales. However, as further analysis proved efforts to define a unidimensional GLLRM futile, we, at this point in time, can only suggest that these two subscales are in fact different scales, as was the case with the instruction and collaboration subscales for Observation and Feedback. Future research should test the dimensionality of the Practiced-instruction and Practiced-collaboration subscales again, preferably with larger and more diverse samples of student teachers.

4.1.4. Targeting and Reliability

The key objective of the current study was to improve targeting of the three field practice experiences scales in the short-form FPE-DK (Nielsen, 2021) by expanding the instrument with additional items. The issue is thus whether this was achieved.

The targeting of the original three field practice experiences scales was indisputably poor to just acceptable, with between 12% and 59% of the maximum information obtained for subgroups defined by level of field practice and gender (Nielsen, 2021). The longer instruction scales in the current study, which included the 12 original items along with 5 new items, showed improved targeting compared with the short-form FPE-DK. For the Observed-instruction scale, 67% of the information was obtained with the expanded scale compared with between 34% and 50% information obtained with the original 12-item Observed scale. For the Practiced-instruction scale, 51% of the maximum information was obtained for male students and 49% for female students, compared with between 12% and 47% with the original 12-item Practiced scale. For the received Feedback-instruction scale, an improvement to 61% of the maximum information was obtained on average, compared to between 20% and 59% with the original 12-item received Feedback scale. Thus, we succeeded in improving targeting of the expanded scales in the current study compared with the short-form scales in Nielsen (2021). However, targeting of the Practiced-instruction scale remains poor. Whether this can be improved with additional items is hard to say, as it appears from the item maps in Figure 3 that this would require items reflecting opportunities to learn that are extremely rare to encounter in field practice, and this may make the items appear irrelevant and de-motivate students to complete the questionnaire. It is an issue to be considered in future research.

It may be argued that targeting of the 12-item scales in Nielsen (2021) cannot be compared with the targeting of the 17-item instruction scales, which included the original 12 items, as targeting depends not only on item difficulty but also on the ability distribution of the participants, and thus, comparisons across samples cannot provide sufficient evidence that the improvement in targeting resulted from the newly added items. The samples are, however, highly comparable: both samples were student teachers who had just completed field practice at one of three levels at the same university college. The distributions of gender and type of teacher education program were identical across samples, and the mean age varied by less than 1 year. The distribution of campuses varied slightly, and only the distributions of field practice levels and chosen major teaching subject varied substantially. The latter is to be expected, as the number of students admitted varied across year cohorts, and students' preferences for teaching subjects fluctuated. We would be hard-pressed to find a more comparable sample for the present study. When comparing the "ability" distributions of the participants across the two studies, we found that there was virtually no difference when it came to the Observed scale and the longer Observed-instruction scale, showing us that the added items are indeed less commonly experienced in field practice. With the Practiced and the Practiced-instruction scales, there were some differences in the

level of the scores, with the latter having the highest mean scores, showing us that some of the newly added items were also commonly experienced in field practice. Lastly, with the Feedback and the Feedback-instruction scales, we again saw higher scores for the longer version. We thus find it plausible that improvements in targeting were due to the newly added items and not just differences in score distributions.

Turning to the three collaboration scales, which only contained newly developed items, targeting was good to excellent, as between 70% and 89% of the maximum information was obtained on average across subgroups defined by level of field practice. The best targeting was for the Practiced-collaboration scale, as it was excellent, targeting students at all levels of field practice.

Reliability of the original 12-item Observed and received Feedback scales in Nielsen (2021) was excellent and high enough (or close to) for individual assessment (Hale & Astolfi, 2014, Well & Wollack, 2003), while reliability of the Practiced scale varied from unacceptable to appropriate for statistical group analyses. In the current study, reliability of the three instruction subscales, which contained all the 12 original items from Nielsen (2021) as well as five newly developed items, followed the same pattern. As reliability is sample-dependent, the reason for the very similar reliability results across the current study and Nielsen (2021) may, at least partially, be found in the high degree of comparability of the student samples (c.f. the above).

In the current study, the reliability of the new collaboration scales was less than satisfactory by conventional standards.

Thus, for the FPE-DK long form, only two subscales were considered suitable for individual assessment in the classical test theory sense—the same two scales that lived up to this level of reliability in Nielsen (2021).

4.2. Differences in Scale Scores Dependent on Level of Field Practice

When comparing the mean scale scores of the three instruction scales for students who had just completed one of three levels of field practice, not all three groups differed significantly. With regard to observation of other teachers in instruction-related practices, students who had just completed the most advanced field practice scored 3.26 points lower than the combined groups of students who had completed level 1 or level 2 field practice. This finding is a replication of Nielsen's (2021) results with the short-form FPE-DK, where all items were instruction-related, and it supports the direct negative association between field practice level and the short-form Observed scale in Nielsen (2024). Thus, it appears to be a consistent finding that the Danish student teachers had fewer opportunities to learn through observation of instruction-related practices in their last and most advanced field practice. As suggested by Nielsen (2021), this may feel intuitively right, but it may also be viewed as problematic from a learning progression point of view, as the ability to learn from observation increases with knowledge, reflectional capacity, and professional maturity; thus, vicarious learning beyond mere imitation and conditioning is possible. In addition, the level of experiences through observation is quite low considering that the scores ranged from 0 to 17 on the instruction scales.

With regard to having practiced the instruction-related practices yourself, students who had completed each of the three levels of field practice all differed significantly, with mean scores increasing with the level of field practice, and by a little more than 0.5 points per level. This finding differs slightly from Nielsen (2021), where students who had just completed field practice level 2 or 3 scored a little more than 0.5 points higher on the 12 instructional items than students who had completed level 1 field practice. Both findings are supported by the positive and direct association between level of field practice and the short-form Practiced scale in Nielsen (2024). It thus appears that the core purpose of

field practice, namely, to practice yourself, increases with the level of field practice, so that students are afforded more experiences of instruction-related practice as they progress through teacher education, though with many experiences at all levels.

With regard to having received feedback on the instruction-related practices, students who had just completed field practice level 1 or 2 scored approximately 1.5 points higher than students who had completed level 3, while all students scored relatively high on the scale. This result is in contrast to the results reported by Nielsen (2021) on the 12-item Feedback scale. In the 2021 study, students who had just completed field practice level 3 scored just under 1 point higher than students who had completed level 1 or 2. Whether this “reversal” of findings is due to the addition of five items in the Feedback-instruction scale in the current study may be worth looking into in future studies. As it stands, it is disconcerting, as feedback is an essential element in students’ development towards being fully qualified teachers, and thus, such opportunities to learn should not decrease as students gain knowledge and skills to benefit the most from feedback.

Results concerning the three collaboration subscales cannot be readily discussed against previous research by Nielsen (2021, 2024), as all the collaboration items are new. However, they can be discussed in relation to the results of the three instruction scales in the current study, and more broadly in relation to previous research. With regard to observing other teachers in collaboration-related practices, the differences showed the same patterns as for the instruction-related practices, namely that students who had just completed level 3 field practice scored lower than students who had completed level 1 or 2. The difference was smaller (approximately 1 point), but the collaboration scales only had the range 0–8 compared with 0–17 for the instruction scales. This further supports the findings that opportunities to learn through observation decline in commonality as students progress through teacher education.

With regard to having practiced the collaboration-related practices yourself, there were no differences in the mean scores dependent on the level of field practice. Thus, there was no progression in the opportunities to learn through own practice when it came to collaboration-related practices, as was the case with instruction-related practices.

With regard to having received feedback on collaboration-related practices, the results were opposite to the results for instruction-related practices. Thus, students who had just completed field practice level 3 scored approximately 0.5 points higher when it came to collaboration-related practices than did students who had completed level 1 or 2, while they scored lower on instruction-related practices.

5. Conclusions

Improved targeting in terms of test information along the scales and better alignment to the intended student population was obtained for all six expanded scales compared with the original shorter scales. However, targeting of the Practiced-instruction scale remained less than satisfactory. Whether this can be improved further with more items that occur even more rarely than the ones included in the instruction scales remains an open question. Reliability in the classical test theory sense remained too low for some scales and subgroups. However, as test information is high along most of the scales, we consider the measurement precision of the scales to be very good from the IRT perspective. With improved targeting and meaningful results concerning differences in mean scores dependent on the level of field practice that had just been completed, the scales show promise for both research and practical use.

Supplementary Materials: The following supporting information can be downloaded at: <https://www.mdpi.com/article/10.3390/psycholint8040066/s1>. File S1: Items and subscales; File S2: Survey design; File S3: Additional results for the Observed-instruction subscale; File S4: Additional results for the Observed-collaboration subscale; File S5: Additional results for the Practiced-instruction subscale;

File S6: Additional results for the Practiced-collaboration subscale; File S7: Additional results for the Feedback-instruction subscale; File S8: Additional results for the Feedback-collaboration subscale; Table S1: The Danish item texts of the three expanded field practice experience scales and their English counterparts; Table S2: Matrix design of the expanded Field Practice Experience scales; Table S3a: Item fit statistics for the Observed-instruction subscale under the final GLLRM; Table S3b: Conditional likelihood ratio tests of local dependence under the final GLLRM for the Observed-instruction subscale; Table S3c: Item and partial credit thresholds, locations, midpoints and targets for conditionally independent items and component items made up of locally dependent items in the Observed-instruction subscale; Table S3d: Conversion of Observed-instruction subscale sum scores to weighted maximum likelihood estimates of person parameters (theta) resulting from the final subscale GLLRM; Table S4a: Item fit statistics for the Observed-collaboration subscale under the final GLLRM; Table S4b: Conditional likelihood ratio tests of local dependence and DIF under the final GLLRM for the Observed subscale 2; Table S4c: Item and partial credit thresholds, locations, midpoints and targets for conditionally independent items and component items made up of locally dependent items in the Observed-collaboration subscale; Table S4d: Conversion of Observed-collaboration subscale sum scores to weighted maximum likelihood estimates of person parameters (theta) and DIF-equated sum scores resulting from the final subscale GLLRM; Table S5a: Item fit statistics for the Practised-instruction subscale under the final GLLRM; Table S5b: Conditional likelihood ratio tests of local dependence and DIF under the final GLLRM for the Practised-instruction subscale; Table S5c: Item and partial credit thresholds, locations, midpoints and targets for conditionally independent items and component items made up of locally dependent items in the Practised-instruction subscale; Table S5d: Conversion of Practised-instruction subscale sum scores to weighted maximum likelihood estimates of person parameters (theta) and DIF-equated sum scores resulting from the final subscale GLLRM; Table S6a: Item fit statistics for the Practised-collaboration subscale under the final GLLRM; Table S6b: Conditional likelihood ratio tests of local dependence under the final GLLRM for the Practised-collaboration subscale; Table S6c: Item and partial credit thresholds, locations, midpoints and targets for conditionally independent items and component items made up of locally dependent items in the Practised subscale 2; Table S6d: Conversion of Practised-collaboration subscale sum scores to weighted maximum likelihood estimates of person parameters (theta) and DIF-equated sum scores resulting from the final subscale GLLRM; Table S7a: Item fit statistics for the Feedback-instruction subscale under the final GLLRM; Table S7b: Conditional likelihood ratio tests of local dependence under the final GLLRM for the Feedback-instruction subscale; Table S7c: Item and partial credit thresholds, locations, midpoints and targets for conditionally independent items and component items made up of locally dependent items in the Feedback-instruction subscale; Table S7d: Conversion of Feedback-instruction subscale sum scores to weighted maximum likelihood estimates of person parameters (theta) resulting from the final subscale GLLRM; Table S8a: Item fit statistics for the Feedback-collaboration subscale under the final GLLRM; Table S8b: Conditional likelihood ratio tests of local dependence and DIF under the final GLLRM for the Feedback subscale 2; Table S8c: Item and partial credit thresholds, locations, midpoints and targets for conditionally independent items and component items made up of locally dependent items in the Feedback-collaboration subscale; Table S8d: Conversion of Feedback-collaboration subscale sum scores to weighted maximum likelihood estimates of person parameters (theta) and DIF-equated sum scores resulting from the final subscale GLLRM.

Author Contributions: Conceptualization, T.N. and G.G.L.; methodology, T.N.; validation: T.N.; formal analysis, T.N.; investigation, T.N., B.T.H., and G.G.L.; data curation, T.N.; writing—original draft, T.N. and B.T.H.; writing—review and editing, T.N., B.T.H., and G.G.L.; project administration, T.N.; supervision, T.N.; resources, T.N. All authors have read and agreed to the published version of the manuscript.

Funding: The authors received no specific funding for this work from any funding agencies.

Institutional Review Board Statement: The study was conducted in accordance with the ethical standards of the Helsinki Declaration. All participants received written information about the project

aims, including details on the voluntary nature of participation, confidentiality of their data, and compliance with GDPR for data storage and use. Ethical review and approval were waived for this study due to the non-interventional nature of the survey.

Informed Consent Statement: All participants received written information about the project aims, including details on the voluntary nature of participation, confidentiality of their data, and compliance with GDPR for data storage and use.

Data Availability Statement: The data set used is publicly available at zenodo.org: <https://doi.org/10.5281/zenodo.22079932>.

Acknowledgments: We thank the participating students for their time and effort, as well as Pedro Ribeiro Santiago for providing the R code used to produce the item maps.

Conflicts of Interest: The authors declare no conflicts of interest.

Notes

- ¹ In Denmark, field practice has since been increased to 40 ECTS (*Act on the Education of Teachers*, 2023).
- ² One ETCS compares to approximately 28 hours of work for a student.
- ³ Each scale has a score range of 0–12.
- ⁴ At that point in time, the Danish teacher education included three field practice placements throughout the 4-year-long program.
- ⁵ *Uniform DIF exists when the statistical relationship between the item response and group is constant for all levels of a matching variable”* (Hanson, 1998, p. 244). Similarly, uniform LD exists when the statistical relationship between responses to two items is constant for all levels of a matching variable.
- ⁶ Item–rest score correlations are correlations between an item and the score with the item excluded.
- ⁷ Results from these preliminary analyses are not included, as they encompass several hundred pages, but data are openly available.

References

- Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, 38, 123–140. [CrossRef]
- Bekendtgørelse om uddannelsen til professionsbachelor som lærer i folkeskolen [Act on the Education of Teachers], BEK nr. 1068 af 08/09/2015. (2015). Available online: <https://www.retsinformation.dk/eli/lt/2015/1068> (accessed on 16 April 2024).
- Bekendtgørelse om uddannelsen til professionsbachelor som lærer i folkeskolen [Act on the Education of Teachers], BEK nr. 374 af 29/3/2023. (2023). Available online: <https://www.retsinformation.dk/eli/lt/2023/374> (accessed on 16 April 2024).
- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57, 289–300. [CrossRef]
- Canrinus, E. T., Bergem, O. K., Klette, K., & Hammerness, K. (2017). Coherent teacher education programmes: Taking a student perspective. *Journal of Curriculum Studies*, 49(3), 313–333. [CrossRef]
- Cavanna, J., Molloy Elreda, L., Youngs, P., & Pippin, J. (2021). How methods instructors and program administrators promote teacher education program coherence. *Journal of Teacher Education*, 72(1), 27–41. [CrossRef]
- Christensen, K. B. (2013). Estimation of item parameters. In K. B. Christensen, S. Kreiner, & M. Mesbah (Eds.), *Rasch models in health* (pp. 49–62). ISTE. John Wiley & Sons, Inc. [CrossRef]
- Christensen, K. B., & Kreiner, S. (2013). Item fit statistics. In K. B. Christensen, S. Kreiner, & M. Mesbah (Eds.), *Rasch models in health* (pp. 83–103). ISTE. John Wiley & Sons, Inc. [CrossRef]
- Cohen, J., & Berlin, R. (2020). What constitutes an “opportunity to learn” in teacher preparation? *Journal of Teacher Education*, 71(4), 434–448. [CrossRef]
- Cox, D. R., Spjøtvoll, E., Johansen, S., van Zwet, W. R., Bithell, J. F., Barndoff-Nielsen, O., & Keuls, M. (1977). The role of significance tests [with discussion and reply]. *Scandinavian Journal of Statistics*, 4(2), 49–70.
- Fuentes-Abeledo, E.-J., González-Sanmamed, M., Muñoz-Carril, P.-C., & Veiga-Rio, E.-J. (2020). Teacher training and learning to teach: An analysis of tasks in the practicum. *European Journal of Teacher Education*, 43(3), 333–351. [CrossRef]
- Gaikhurst, L., Post, J., März, V., & Soeterik, I. (2020). Teacher preparation for urban teaching: A multiple case study of three primary teacher education programmes. *European Journal of Teacher Education*, 43(3), 301–317. [CrossRef]
- Goodman, L. A., & Wallis, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association*, 49, 732–764. [CrossRef]
- Grossman, P., Hammerness, K. M., McDonald, M., & Ronfeldt, M. (2008). Constructing coherence: Structural predictors of perceptions of coherence in NYC teacher education programs. *Journal of Teacher Education*, 59(4), 273–287. [CrossRef]
- Hale, C. D., & Astolfi, D. (2014). *Measuring learning and performance: A primer*. Saint Leo University.

- Nielsen, T., Headley, B. T., & Larsen, G. G. (2026a). *Muligheder for læring i lærerpraktik spørgeskema, lang version (FPE-DK-long)—Scoringssguide—Administrationsguide* [The long form Field Practice Experiences questionnaire for students in teacher education (FPE-DK-long)]. UCL Erhvervsakademi og Professionshøjskole.
- Nielsen, T., Jensen, L. S., Toft, L., & Pettersson, M. (2026b). Differences between first- and second-year student teachers' practice self-efficacy: A cross-sectional study. *Psychology International*, 8, 24. [CrossRef]
- Nielsen, T., & Kreiner, S. (2013). Improving items that do not fit the Rasch model. In K. B. Christensen, S. Kreiner, & M. Mesbah (Eds.), *Rasch models in health* (pp. 317–334). ISTE. John Wiley & Sons, Inc. [CrossRef]
- Nielsen, T., Lindemann, D. M., Toft, L., Pettersson, M., Jensen, L. S., Nielsen, E. H., & Larsen, G. G. (2024). Teacher and teaching-related practices student teachers may have difficulty believing they can enact/engage in while in field practice: Perspectives from student teachers and teachers in field-practice schools. *Nordic Studies in Education*, 44(2), 103–121. [CrossRef]
- Nielsen, T., & Pettersson, M. (2025). Praksis self-efficacy for lærerstuderende (PSEQ). In *Spørgeskema, scoringssguide og administration*. UCL Erhvervsakademi og Professionshøjskole.
- Nielsen, T., Pettersson, M., & Toft, L. (2025). Student teachers' practice self-efficacy prior to their first field practice in schools: Interrelatedness of subconstructs within three domains of practice. *Psychology International*, 7(3), 59. [CrossRef]
- Nielsen, T., Wisniewski, A. H., & Larsen, G. G. (2026). *Ændringer i første-og andetårs lærerstuderendes self-efficacy for specifikke praksisser henover praktikforløb*. UCL Erhvervsakademi og Professionshøjskole.
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research.
- Rasch, G. (1961). *On general laws and the meaning of measurement in psychology*. The Regents of the University of California. Available online: <http://projecteuclid.org/euclid.bsm/1200512895> (accessed on 7 June 2008).
- R Core Team. (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.
- 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, 3–26. [CrossRef]
- Rosenbaum, P. R. (1989). Criterion-related construct validity. *Psychometrika*, 54(4), 625–633. [CrossRef]
- Sayed, Y., Ramchand, M., Mosito, C. P., Milani, R., & Bertini, L. F. (2026). School-university partnership for effective teaching practicum experiences: A study of an initial teacher education programme from Brazil, India and South Africa. *International Journal of Educational Research*, 137, 102956. [CrossRef]
- Scherer, R., & Siddiq, F. (2015). Revisiting teachers' computer self-efficacy: A differentiated view on gender differences. *Computers in Human Behavior*, 53, 48–57. [CrossRef]
- Schmidt, W., Cogan, L., & Houang, R. (2011). The role of opportunity to learn in teacher preparation: An international context. *Journal of Teacher Education*, 62(2), 138–153. [CrossRef]
- Schutz, K. M., Grossman, P., & Shaughnessy, M. (2018). Approximations of practice in teacher education. In P. Grossman (Ed.), *Teaching core practices in teacher education* (pp. 57–83). Harvard Education Press.
- Ulvik, M., Eide, L., Helleve, I., & Kvam, E. K. (2021). Praksisopplæringens oppfattende og erfarte formål sett fra ulike aktørperspektiv. *Nordisk Tidsskrift for Utdanning og Praksis*, 15(3), 19–37. [CrossRef]
- van Braak, J. (2001). Individual characteristics influencing teachers' class use of computers. *Journal of Educational Computing Research*, 25(2), 141–157. [CrossRef]
- Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, 54, 427–450. [CrossRef]
- Weisdorf, A. K. (2020). *Læreruddannelsen i globalt perspektiv—Et komparativt studie af læreruddannelsen i Danmark, England, Finland, Holland, New Zealand, Norge, Ontario, Singapore, Sverige og Tyskland* [Teacher education in a global perspective—A comparative study of teacher education in Denmark, the UK, Finland, the Netherlands, New Zealand, Norway, Ontario, Singapore, Sweden and Germany]. Danske Professionshøjskoler. Available online: https://www.ucviden.dk/ws/portalfiles/portal/183471468/Weisdorf_2020_L_reruddannelsen_i_globalt_perspektiv_Et_komparativt_studie.pdf (accessed on 2 April 2020).
- Well, C. S., & Wollack, J. A. (2003). An instructor's guide to understanding test reliability. In *Testing and evaluation services*. University of Wisconsin.
- Youngs, P., Elreda, L. M., Anagnostopoulos, D., Cohen, J., Drake, C., & Konstantopoulos, S. (2022). The development of ambitious instruction: How beginning elementary teachers' preparation experiences are associated with their mathematics and English language arts instructional practices. *Teaching and Teacher Education*, 110, 103576. [CrossRef]
- Zeichner, K. (2021). Critical unresolved and understudied issues in clinical teacher education. *Peabody Journal of Education*, 96(1), 1–7. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.