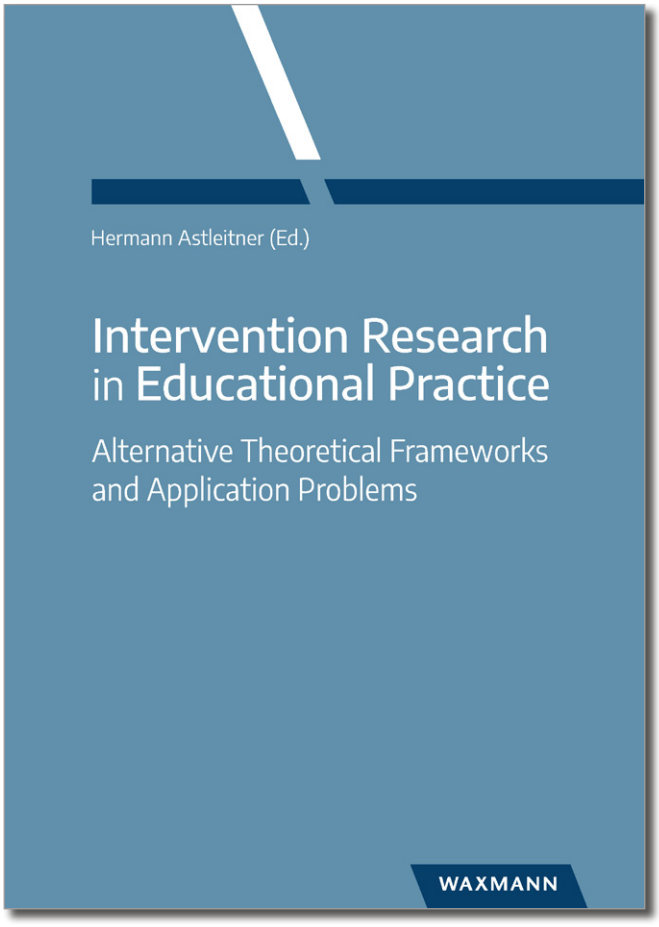
Alexander Naumann, Stephanie Musow & Michaela Katstaller

Instructional Sensitivity as a Prerequisite for Determining the Effectiveness of Interventions in Educational Research



Hermann Astleitner (Ed.)

Intervention Research in Educational Practice

Alternative Theoretical Frameworks and Application Problems

2020, 188 pages, br., € 29,90,
ISBN 978-3-8309-4197-2

E-Book: € 26,99,
ISBN 978-3-8309-9197-7

WAXMANN
Steinfurter Str. 555
48159 Münster
Fon +49 (0)2 51 – 2 65 04-0
Fax +49 (0)2 51 – 2 65 04-26
info@waxmann.com
order@waxmann.com
www.waxmann.com
Further book information here.

7. Instructional Sensitivity as a Prerequisite for Determining the Effectiveness of Interventions in Educational Research

Alexander Naumann, Stephanie Musow & Michaela Katstaller

Abstract: Student achievement has become a major criterion for evaluating the effectiveness of schooling and teaching. However, valid interpretation and use of test scores in educational contexts require more detailed information about the degree to which the applied test instruments are appropriate to evaluate the intended educational and interventional effects. Instructional sensitivity is the psychometric property of tests or single items to capture effects of classroom instruction. Although instructional sensitivity is a prerequisite for valid inferences on teaching effectiveness, sensitivity is rather assumed than verified in practice. The aim of this chapter is to improve the understanding of instructional sensitivity and its measurement in educational intervention research. Specifically, it first provides an overview of the theoretical framework and relevance of instructional sensitivity. Then, different approaches of measuring instructional sensitivity are outlined and procedures of implementing instructional sensitivity in educational intervention studies are introduced and contrasted with each other. Finally, the role of time spans is discussed and modelling change for short-time and long-time effects in pretest-posttest-follow-up designs is addressed.

Introduction

This chapter aims at embedding instructional sensitivity in the scientific discourse of educational intervention research. Educational intervention research is expected to provide evidence-based insights into the effectiveness of educational measures (e.g., Hascher & Schmitz, 2010). However, evidence-based insights necessitate the availability of instructionally sensitive test instruments for drawing valid conclusions on the effectiveness of educational interventions in schools, higher education, or out-of-school learning activities. Yet, fulfilling such methodological requirements like instructional sensitivity may be challenging in a practice-oriented field like educational intervention research. To foster the methodological foundation of educational intervention studies, we will address the following three issues: (a) the theoretical background and relevance of instructional sensitivity, (b) its measurement, and (c) ways of practical implementation in educational intervention studies.

Throughout the chapter, we will discuss particularities of intervention studies with respect to instructional sensitivity.

Theoretical Background and Relevance of Instructional Sensitivity

While instructional sensitivity received little attention in European countries until recently (Deutscher & Winther, 2018; Naumann, Musow, Aichele, Hochweber, & Hartig, 2019a), the concept has been discussed in the U.S. since the mid-1960s (e.g., Cox & Vargas, 1966). Back then, researchers argued whether traditional item statistics like item difficulty or discrimination were appropriate for selecting items in criterion-referenced testing (e.g., Kosecoff & Klein, 1974). However, the concept has been exposed to essential changes since then. By the end of the 1970s, the main focus shifted from item selection in criterion-referenced testing to issues of validity and test fairness in educational assessments (e.g., Linn & Harnisch, 1981). Essentially, there were two concepts of instructional sensitivity, namely instructional validity and instructional bias. Instructional validity referred to the question to what degree classroom instruction contributes to students' test scores (e.g., Schmidt, Porter, Schwille, Floden, & Freeman, 1983). In contrast, instructional bias referred to differential item functioning for students when they were exposed to different kinds of schooling (e.g., Linn & Harnisch, 1981). Both were seen as essential for drawing inferences on instruction (e.g., Burstein, 1989; Linn, 1983), and consequently, both strands merged in the concept of instructional sensitivity (D'Agostino, Welsh, & Corson, 2007). In 2010, Polikoff defined instructional sensitivity as the psychometric capacity of a test or a single test item of capturing effects of teaching. That is, instructional sensitivity (a) can be seen as a necessary prerequisite for valid test use and interpretation if tests are used for drawing inferences on teaching (Burstein, 1989; Popham, 2007) and (b) can be quantified as a psychometric property of an assessment (Polikoff, 2010). While some researchers have expressed their preferences on whether assessments should be sensitive to the content or to the quality of teaching (e.g., Popham, 2007), today's understanding of instructional sensitivity equally encompasses both aspects of teaching (D'Agostino et al., 2007).

In line with the current *Standards for Educational and Psychological Testing* (AERA, APA, & NCME, 2014), instructional sensitivity can be seen as a necessary validity aspect when detecting effects of schooling and teaching. Already in the 1980s, Airasian and Madaus (1983) emphasized the role of instructional sensitivity as an important aspect of construct validity. More specifically, instructional sensitivity was seen as a necessary, though not sufficient requirement for consequential validity (Messick, 1989). Following today's argument-based approach to validity (Kane, 2013), the evaluation of instructional sensitivity provides empirical evidence for a valid use and interpretation of test scores. Unlike other validity aspects such as content or curricular validity aiming at the linkage of tests and items and the intended curriculum, instructional sensitivity refers to the alignment of assessments with the implemented curriculum (Naumann et al., 2019a). Nevertheless, in con-

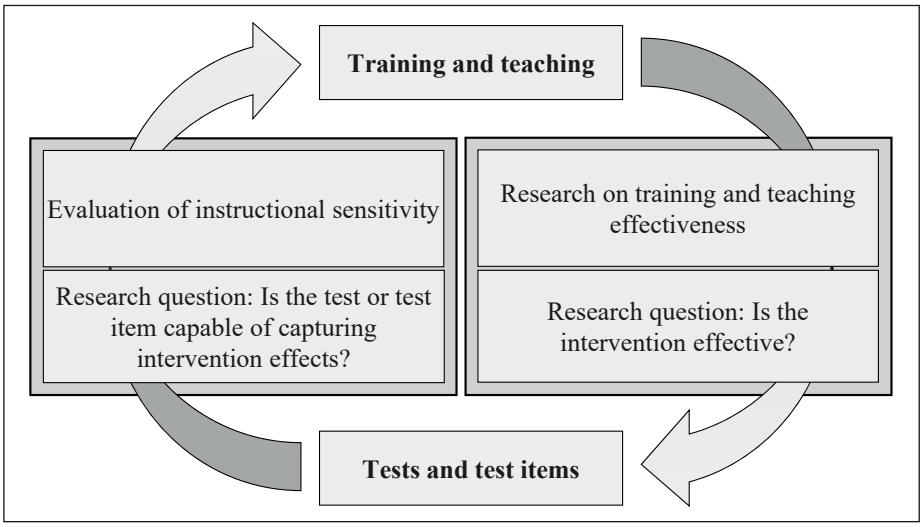


Fig. 1: Relationship of the evaluation of instructional sensitivity and research on education and teaching effectiveness.

trast to the U.S., where discussions of instructional sensitivity have mainly focused on accountability issues (e.g., Popham, 2007), the European discussion puts more emphasis on instructional sensitivity as a central validity aspect in research on educational effectiveness (Naumann, Rieser, Musow, Hochweber, & Hartig, 2019b).

In educational effectiveness research, measures of students' achievement and competencies are the most widespread criteria for evaluating whether or not teaching has been effective (Klieme, 2019). The usual strategy is to use student test scores as dependent variable in (multilevel) regression analyses (Marsh et al., 2012). Applying this strategy requires that the test in principle needs to be instructionally sensitive, that is, capable of capturing effects of teaching. Otherwise, if instructional sensitivity is unclear when evaluating teaching effectiveness, a lack of effects might be either due to ineffective teaching or insensitive assessments (Naumann, Hochweber, & Hartig, 2014; Naumann, Hochweber, & Klieme, 2016). Accordingly, instructional sensitivity needs to be ensured during test development prior to the main effectiveness studies as both explanations remain inextricably confounded otherwise (Naumann et al., 2019a). That is, studies on educational effectiveness have to rely on instruments that are instructionally sensitive to check the degree to which teaching is effective (see right-hand side of Figure 1). However, instructional sensitivity oftentimes is rather assumed than actually investigated empirically (D'Agostino et al., 2007; Naumann et al., 2016).

Two recent studies emphasize practical consequences for inferences on teaching effectiveness stemming from a varying degree of instructional sensitivity. First, Grossman, Cohen, Ronfeldt and Brown (2014) found that tests that operationalize the same construct such as students' achievement in English language arts may show a different extent of instructional sensitivity. Consequently, the test matters whether

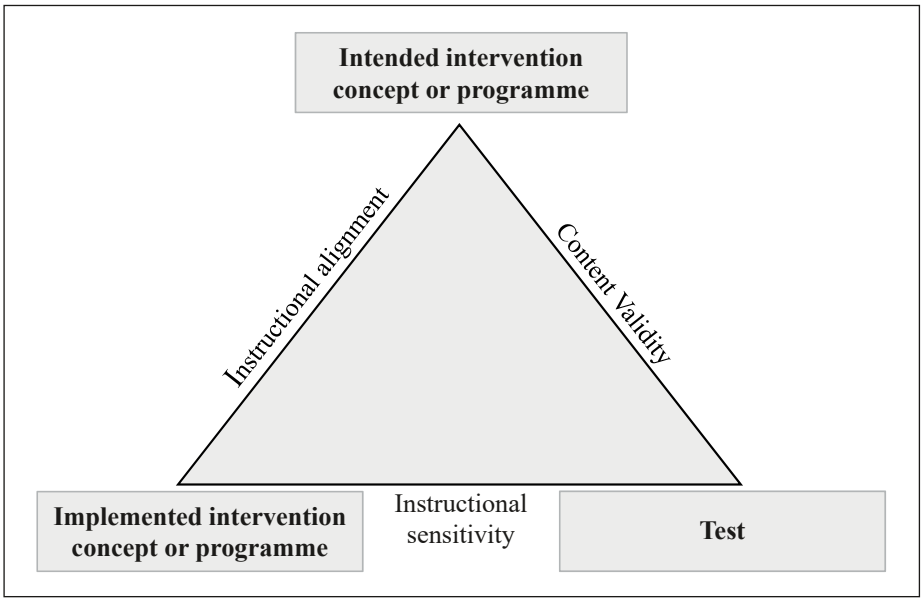


Fig. 2: Ratio of intended intervention concept or programme, implemented intervention concept or programme and test (adapted from Naumann et al., 2019a).

or not certain aspects of teaching are identified to be effective. Second, Naumann and colleagues (2019b) illustrated that a test's instructional sensitivity varies concordantly with the degree of its items' instructional sensitivity. Their results suggest that even slight changes in a test's composition may lead to different conclusions on teaching effectiveness even when the sampled items originate from the same item pool. Taken together, both studies provide empirical evidence that the associations of test scores and the construct(s) of interest, and thus the effect sizes, in an educational effectiveness study depend on the assessment's instructional sensitivity.

The previous considerations correspond to educational intervention research. Students' test scores are also a major criterion for assessing an intervention's success (Hascher & Schmitz, 2010). Accordingly, only if a test and its items are instructionally sensitive, intervention effects – or the lack thereof – can be validly interpreted. Thus, instructional sensitivity may be seen as an essential validity criterion for evaluating an educational intervention to ensure that results are validly interpretable with respect to the intervention's effectiveness. Similar to educational effectiveness research, validly detecting the intervention's effectiveness depends on the extent to which (1) the test itself, (2) the intended intervention concept or programme, and (3) the implemented intervention concept or programme are aligned with each other (Naumann et al., 2019b). Figure 2 depicts the relationship of these three elements as a triad, with each of its sides describing the alignment between two out of the three elements (adapted from Anderson, 2002; Naumann et al., 2019a; Pellegrino, 2002).

A particularity of intervention studies is that both the intended and thus also the implemented curriculum may differ between experimental and control group. The

degree depends on the overall intervention concept. While in some interventions the content may differ completely between experimental and control group (e.g., uninstructed vs. instructed students), there may be an overlap in the intervention content across groups (e.g., same content, but different teaching method; Decristan et al., 2015a). Accordingly, the alignment of intended and implemented intervention concept or programme can either be seen as indicating instructional alignment or treatment adherence within each of the intervention or control conditions, respectively. Treatment adherence is necessary for valid test score interpretation to avoid that the causes of potential effects remain unclear.

Analogously, the alignment of test and intended intervention concept or programme provides arguments for content validity (Hartig, Frey, & Jude, 2012). Empirical evidence for content validity may be given, for instance, by content reconciliation of the test material and formal documents of the intervention concept or programme (Naumann et al., 2019b). If the degree of content validity differs substantially between the experimental and control conditions, valid interpretation of results may be impaired, for example, due to a lack of test fairness. Finally, the actual implemented intervention concept or programme is crucial for the intervention's contribution to students' performance on the test. Accordingly, the alignment of the test and the implemented intervention concept or programme is of special interest, that is, instructional sensitivity. Only if the test is capable of capturing potential intervention effects, results can be validly interpreted. Yet, while the test has to be sensitive to the intervention, it should not favor the intervention conditions compared to the control group. Thus, researchers are required to investigate instructional sensitivity prior to the intervention, for example, in an intervention's pilot study. To provide an understanding of how to achieve this requirement in practice, we will first provide an overview on the measurement of instructional sensitivity hereafter and then propose ways of implementation in educational interventions.

Measuring Instructional Sensitivity

In the course of the last decades, different approaches have been developed to profoundly evaluate instructional sensitivity: (1) item statistics (e.g., Cox & Vargas, 1966; Linn & Harnisch, 1981; Robitzsch, 2009), (2) approaches relating test scores and item responses to instructional measures (e.g., Ing, 2018; Muthén et al., 1995; Ruiz-Primo et al., 2012), and (3) expert ratings (Chen, 2012; Popham, 2007; Popham & Ryan, 2012). Although expert ratings on instructional sensitivity appear beneficial due to economic reasons, they have not been sufficiently evaluated yet. Thus, we will focus on approaches based on actual student tests and item response data in the following section and discuss expert ratings later.

As mentioned before, evaluation of instructional sensitivity should take place prior to the main study to prevent confounding of effectiveness and sensitivity. When evaluating instructional sensitivity, the underlying assumption is that teaching is effective to check whether an instrument is sensitive or not (left-hand side of

Figure 1). That is, instructional sensitivity can be seen as a relational concept that describes the psychometric capacity of a test or a single item of capturing effects of classroom instruction under the condition that teaching is effective (Naumann et al., 2019b).

There are different procedures available for empirical investigation of instructional sensitivity. These procedures can be classified based upon (a) whether they address absolute or relative sensitivity (Naumann, Hartig, & Hochweber, 2017) and (b) their perspective on instructional sensitivity (Naumann et al., 2016). In the following, we will give a brief overview of the resulting framework for measuring instructional sensitivity and how it relates to commonly applied research designs in educational intervention studies.

The Framework for Measuring Instructional Sensitivity

When evaluating the instructional sensitivity of test items, we can distinguish two kinds of sensitivity measures: absolute and relative measures (Naumann et al., 2017). Absolute measures capture an item's overall sensitivity, while relative measures capture the degree to which an item's sensitivity deviates from the test's sensitivity. Absolute and relative sensitivity can be evaluated from each of the three perspectives on instructional sensitivity within the framework. In educational intervention practice, however, absolute measures are usually of more interest.

The three perspectives relate to different variance sources which function as the basis for the investigation of instructional sensitivity (see left-hand side of Figure 3). Naumann and colleagues (2016) label these perspectives (a) the Time Points-Perspective, (b) the Groups-Perspective, and (c) the Time Points- and Groups-Perspective. Each perspective relates to a specific research design that targets the same variance source in the evaluation of the intervention effectiveness (right-hand side of Figure 3).

Time Points-Perspective. The Time Points-Perspective refers to the capacity of a test or an item of differentiating students' learning progress at different points in time. For example, scores of instructionally sensitive tests are expected to increase over time (Baker, 1994). Also, items are expected to get easier over time (Cox & Vargas, 1966). More precisely: An item is considered to be instructionally sensitive, if there is a change in item difficulty between a pretest and a posttest. To investigate instructional sensitivity following a Time Points-Perspective, the Pretest-Posttest-Difference Index (PPDI; Cox & Vargas, 1966) is the most widespread approach. PPDI quantifies instructional sensitivity simply as the difference in item difficulty between posttest and the pretest. With regard to educational intervention studies, the research design underlying this perspective is a one group pre-posttest design.

Groups-Perspective. A second perspective is the groups-perspective (Naumann et al., 2016). This perspective refers to the sensitivity aspect that students should show different performances due to their learning group allocation. To investigate

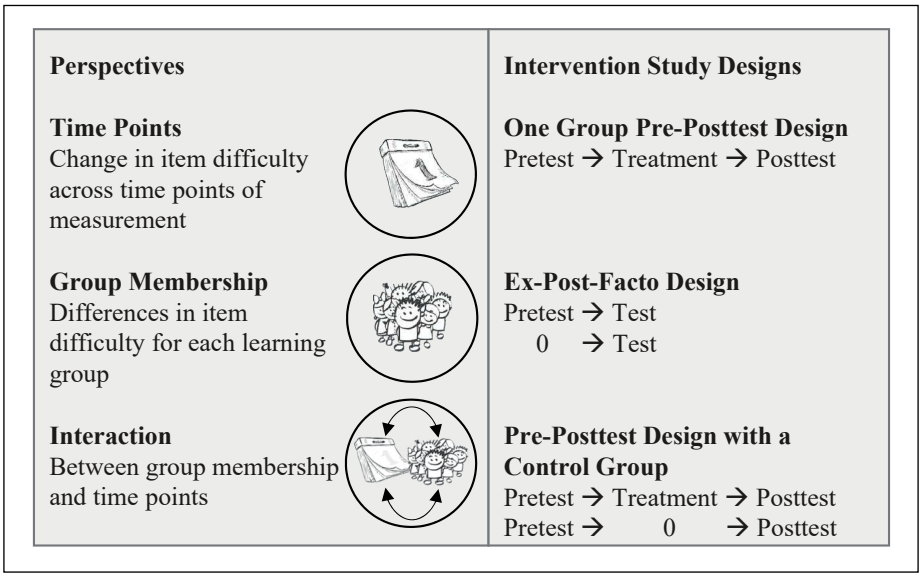


Fig. 3: Three perspectives to evaluate instructional sensitivity and different intervention study designs (Naumann et al., 2016, 2017).

the instructional sensitivity from a Groups-Perspective, analyses of the variation of test scores or item difficulty across groups can be carried out (e.g., Naumann et al., 2017). In educational intervention studies, the research design corresponding to this perspective is a cross-sectional Ex-Post-Facto design. Therefore, assessments in intervention studies are instructionally sensitive, if students' performances in the experimental group significantly differ from the students' performances in the control group.

Time Points and Groups-Perspective. The third perspective refers to the interaction of measurement occasions and learning group membership. Thus, it is called the Time Points and Groups-Perspective (Naumann et al., 2016). The Time Points and Groups-Perspective utilizes information about the group-specific change in test scores or item difficulty respectively. When taking on this perspective, it is optional whether change is modelled as change scores or via covariance-analytic approaches (Naumann et al., 2019b). Accordingly, instructional sensitivity can, for example, be assessed using a longitudinal multilevel Item Response Theory model (LMLIRT; Naumann et al., 2017) or regression of test scores on teaching characteristics while adjusting for prior learning prerequisites (e.g., Polikoff, 2016). In the context of educational intervention studies, the underlying research design corresponds to a pretest-posttest design with a control group. Based on the Time Points- and Groups-Perspective, instruments can be considered as instructionally sensitive in two ways: A test's or an item's (a) global sensitivity describes the average change in item difficulty or test scores across learning groups between measurement occasions, while (b) differential sensitivity indicates the variance of group-specific change in item difficulty or test scores, respectively (Naumann et al., 2016). The

two facets play specific roles in educational intervention studies. The higher the global sensitivity, the higher effect sizes from pretest to posttest can be found in the evaluation of the intervention. If tests or items are capable of differentiating learning progress between the treatment groups, we find indication of differential sensitivity. The higher the instrument's differential sensitivity, the better its capacity to detect specific intervention effects in comparison to the control condition.

Overall, we would like to emphasize that the perspective on instructional sensitivity should fit the research design that is used to evaluate the effectiveness of the intervention. That is, if a test or an item is sensitive from one perspective, it is not necessarily sensitive from another perspective (Naumann et al., 2014). For this reason, it is important to carefully decide whether the analyses are based on Time Points, Groups, or Time Points and Groups-Perspective, respectively. With regard to educational intervention studies, (quasi-)experimental pre-posttest designs with a control group oftentimes are regarded as the gold standard. Therefore, in the following, we will focus on a Time Points and Groups-Perspective to point out ways of ensuring instructional sensitivity in educational intervention studies.

Implementation in Educational Intervention Studies

There are various ways of implementing and ensuring instructional sensitivity in educational intervention studies. The easiest way is to use instruments whose instructional sensitivity has been proven in previous studies (Naumann et al., 2019a). However, a major requirement is that the current study needs to be similar to the research question(s) and the research design of the previous studies. This implies that the new study needs to be comparable with respect to (a) the target population, (b) the intended curriculum, and (c) the assessment design. Otherwise, if previous studies targeted students with learning opportunities differing from those in the current study or provided evidence for instructional sensitivity from a different perspective, it may be hard to assume that the instruments will as well be sensitive in the current study.

A second way of ensuring instructional sensitivity is resorting to expert ratings (Popham, 2007; Popham & Ryan, 2012). Expert ratings on instructional sensitivity are beneficial as they are comparably easy to implement. In principle, expert ratings do not need any empirical student test data or item responses. For this reason, expert ratings appear as a very economical method. However, there are currently only few empirical studies on expert ratings concerning instructional sensitivity (e.g., Chen, 2012; Musow, Naumann, Ruiz-Primo, Hartig, & Hochweber, 2019b). On the one hand, studies found that experts tend to classify more items as sensitive than statistical approaches (Chen, 2012; Musow, Naumann, Hochweber, & Hartig, 2019a; Musow et al., 2019b); on the other hand, recent work by Musow and colleagues (2019a; 2019b) indicates that raters and statistics may coincide depending on the kind of rating and the group of experts recruited. In summary, further research is needed before making a final recommendation.

Lastly, educational interventions may conduct pilot studies for ensuring instructional sensitivity of instruments. Ensuring instructional sensitivity in the context of pilot studies appears beneficial when no instruments whose instructional sensitivity has already been proven are available, and when the aim is to validate instructional sensitivity empirically. Ideally, pilot studies are conducted in samples with similar or at least comparable learning opportunities as the sample of the main study. Then, the aforementioned statistical methods can be applied to the item response data for determining instruments' instructional sensitivity.

In many scenarios, however, sample sizes and/or expertise in elaborate statistical methods may not be sufficient for implementing sophisticated approaches like the aforementioned LMLIRT model. While the LMLIRT model is methodologically sound, its implementation lacks user-friendliness as it requires advanced knowledge in Bayesian estimation and corresponding software packages. Thus, in the following, we will provide a screening procedure that is easy to implement and still allows for the evaluation of instructional sensitivity from a Time Points and Groups-Perspective.

A Screening Procedure for Instructional Sensitivity

Our screening procedure follows comparable methodological principles as the two versions of the LMLIRT model. While the LMLIRT model either utilizes (a) estimates of group-specific change in IRT item difficulty parameters or (b) baseline-adjusted posttest IRT item parameters as a basis for measuring global and differential sensitivity (Naumann et al., 2019b), the proposed screening procedure resorts to Classical Test Theory (CTT) item difficulties. Compared to the latent variable models, the main drawbacks are that the proposed sensitivity measures are purely descriptive, covariance structures between measurement occasions are neglected, and that CTT item difficulties are prone to measurement error. In other words, the observed CTT item difficulties capture an item's true difficulty plus – to some degree – measurement error (e.g., Rost, 2004). The practical advantage is that CTT item difficulties are easy to compute using standard software and applicable in many scenarios that are common to educational intervention studies. Analogous to the LMLIRT model, we will provide two versions of our screening procedure, one suitable for the change-score approach and the other one appropriate for the covariance-analytical approach. The choice of the approach depends on how change will be modelled in the main study. Both versions essentially require three steps for evaluating instructional sensitivity from a Time Points and Groups-Perspective.

Change-Score Approach. The change-score approach requires repeated measurements of the same item. When following the change-score approach, the first step is to calculate CTT item difficulties separately for the treatment and the control condition at pretest and posttest, respectively. If there is a hierarchical data structure with multiple learning groups (e.g., classes) within each condition, we calculate item difficulties for each learning group. Second, we compute the difference in

item difficulty between pretest and posttest for each learning group. Conceptually, this corresponds to group-specific PPDI values. Finally, mean and variation of the group-specific change in item difficulty serve as indicators for absolute global and differential sensitivity. Mean values may range from –1 to 1, with zero indicating that an item is not globally sensitive. Similarly, the higher the variation in group-specific change in item difficulty, the higher the item's differential sensitivity. In the simplest case, when there is only one treatment and one control group, we use the difference in PPDI between the two groups as a measure of differential sensitivity.

Covariance-Analytical Approach. The covariance-analytical approach does not require repeated measurements of the same item, yet, it also does not preclude them. When following the covariance-analytical approach, the first step is to calculate CTT item difficulties at posttest separately for the treatment and the control condition. If there is a hierarchical data structure with multiple learning groups (e.g., classes) within each condition, we calculate item difficulties for each learning group. Second, we regress the group-specific posttest item difficulties on covariates that account for prior learning prerequisites, for example, prior achievement. Then, the residual variance in group-specific item difficulty serves as an indicator for an item's differential sensitivity. If residual variance is near zero, the item under investigation can be considered as not differentially sensitive. In the simplest case, when there is only one treatment and one control group, we use the difference in item difficulty between the two groups as a measure of differential sensitivity. In contrast to the change-score approach, there is no measure of global sensitivity (cf. Naumann et al., 2019b).

Illustrative Data Example. For illustration of the proposed methods, we use data from the study "Individual support and adaptive learning environments in primary school" (IGEL; Decristan et al., 2015b). IGEL was a quasi-experimental intervention study in grade-level three of German primary school science education. More specifically, IGEL was a cluster-randomized controlled trial using a pretest-posttest-follow-up assessment design. Participation was voluntary. First, all participating teachers were trained in the content area of floating and sinking. Then, teachers were assigned to the treatment conditions or the control condition, respectively. Randomization was carried out at the school level. Teachers within the treatment conditions received training in one of three adaptive teaching methods, that is, formative assessment, peer-learning, or scaffolding. Teachers within the control condition received training in parental counseling, which was not expected to show effects on students in the course of the IGEL intervention. After training, teachers implemented the teaching methods in a pre-structured curriculum on floating and sinking in class. The curriculum was adapted from an empirically evaluated primary school inquiry-based science education unit (Hardy, Jonen, Möller, & Stern, 2006; Möller, Jonen, Hardy, & Stern, 2002). It consisted of two consecutive teaching units with five lessons each. The first teaching unit was devoted to the concept of density, while the second one was devoted to the concepts of buoyancy force and displacement. All classes were checked for adherence to the intended curriculum (Adl-Amini, Decristan, Hondrich, & Hardy, 2014). For detailed results regard-

ing the IGEL intervention, see Decristan and colleagues (2015a, 2015b) as well as Hondrich, Hertel, Adl-Amini and Klieme (2016). Our exemplary analyses focus on data from the first teaching unit, as the assessments framing that teaching unit have been extensively investigated for their global and differential instructional sensitivity before, using both the change-score and the covariance-analytic versions of the LMLIRT model (see Naumann et al., 2019b).

The data used for analyses comprises about 1045 students in 54 classes (M_{age} = 8.8 years, SD_{age} = 0.5, 50% female) who participated in the pre- and posttests of students' conceptual understanding of floating and sinking. Students' conceptual understanding served as the main outcome for judging the interventions' effectiveness in fostering students' learning. Corresponding assessments took place with an average time lag of three weeks between pretest and posttest. The tests were administered in classroom-wide assessments by trained personnel. To ensure students' understanding of the tasks, each task was read aloud and visualized using projectors. Then, students had the opportunity to respond to the task. The pretest comprised sixteen items while the posttest consisted of thirteen items, with seven items in common to both measurement occasions. The items were either adapted from previous work done by Hardy and colleagues (2006, 2010), the German TIMSS 2007 science assessment (Bos et al., 2008), or self-constructed. All items were (re)worded to be appropriate for grade level three. Response formats comprised multiple-choice and open-ended tasks. Scoring followed previous research on students' conceptual understanding of floating and sinking (Hardy et al., 2006; Kleickmann et al., 2010). All items fit the partial-credit model (PCM; Masters, 1982).

For our analyses, we split polytomous items into separate dichotomous step indicators. Analyses were carried out using R 3.6.1 (R Core Team, 2019). Markov-Chain-Monte Carlo sampling for the LMLIRT models was conducted via RStan (Stan Development Team, 2019) in a Bayesian framework using vague priors. For details on the technical implementation of the LMLIRT models, see Naumann and colleagues (2017, 2019b). In the covariance-analytic approaches, we adjusted sensitivity measures for students' prior achievement and students' cognitive abilities.

In the change-score approach, calculated group-specific change in CTT item difficulty is highly correlated with latent LMLIRT change estimates, with Pearson correlation ranging from –.95 to –.65 across items (*Mean r* = –.88). Table 1 shows results for items' global and differential sensitivity obtained from the change-score CTT procedure and the change-score LMLIRT model. While the LMLIRT model identifies all repeatedly-administered items as globally sensitive with Bayesian Credible Intervals not comprising zero, the CTT approach seems to identify the items as less globally sensitive. For example, the second step indicator within item 13 appears comparably insensitive. One reason for this finding is that item 13 was very difficult at both measurement occasions, which cannot be captured by the CTT approach in an adequate way. With respect to differential sensitivity, the CTT change-score approach indicates at least some variation in change across groups, while the LMLIRT model identifies more items as differentially sensitive.

Results for the covariance-analytic approach are shown in Table 2. Baseline-adjusted measures of group-specific posttest CTT item difficulty are highly correlated with latent LMLIRT baseline-adjusted estimates, with Pearson correlation ranging from .97 to .79 across items (*Mean r* = .93). Similar to the scenario using change-scores, the CTT approach suggests fewer items to be differentially sensitive compared to the LMLIRT model.

In summary, results support the use of both CTT screening procedures for approximating items' global and differential sensitivity. When comparing LMLIRT and CTT measures of global and differential sensitivity, one has to keep in mind that the measures obtained from the different approaches have different metrics. In the CTT approaches, global sensitivity is expressed in terms of average change between pretest and posttest across groups in the proportion of students who get an item correct, while differential sensitivity describes the degree these proportions vary across groups, expressed in standard deviations. In both cases, the underlying metric is percent correct. In contrast, the LMLIRT models provide measures on a logit scale, with variation across groups expressed as variance. Accordingly, the values from the CTT approaches may appear lower than or even different from those from the LMLIRT models, especially for very easy and very difficult items as they are usually more prone to measurement error.

When screening for instructional sensitivity, we generally recommend excluding such items for which CTT sensitivity measures take on the value zero. However, we do not recommend only selecting items with high global and differential sensitivity values. Depending on the item content, we also recommend considering items with lower sensitivity indices if these items capture hard-to-learn facets of the achievement construct. Nevertheless, we would like to emphasize that the screening procedures may help avoiding the selection of insensitive items, yet they are not ideal for a deeper analysis of the extent of an item's sensitivity.

The Role of Time Spans and how Change is Modelled

Usually when planning an intervention, the question arises whether and to what extent effects are to be expected during a specific period of time. When evaluating instructional sensitivity, a similar question arises: How sensitive are the items in a specific time span? In addition, if data from more than two measurement points are available, there is more than one option to conceptualize change values. To date, there is only little knowledge on the role of time and the ways of modelling it when measuring instructional sensitivity. Yet, when planning an intervention, time plays an important role with regard to the expectation on its effectiveness (Kauffeld, 2010). That is, researchers usually have hypotheses on what intervention effects are expected in which period of time. Consequently, instruments' sensitivity must fit the time span that is covered by the intervention programme.

In addition, pretest-posttest-follow-up design utilize multiple measurement occasions, each associated with specific expectations on effect sizes. When dealing

Tab. 1: Change-score Approach: Item sensitivity results for repeatedly-administered IGEL-items

		LMLIRT model				CTT Screening Procedure	
		Global Sensitivity			Differential Sensitivity	Global Sensitivity	Differential Sensitivity
Item	Cat	M	(SD)	95% BCI	M (SD)		
2	2.1	–4.01	(.16)	[–4.33, –3.70]	.11 (.12)	0.64	0.13
3	3.1	–3.86	(.15)	[–4.16, –3.57]	.06 (.08)	0.65	0.13
4	4.1	–1.83	(.16)	[–2.15, –1.51]	.51 (.21)	0.30	0.20
	4.2	–1.62	(.27)	[–2.13, –1.07]	.52 (.41)	0.13	0.13
5	5.1	–0.99	(.16)	[–1.30, –0.67]	.55 (.26)	0.17	0.19
6	6.1	–0.98	(.14)	[–1.25, –0.70]	.37 (.20)	0.18	0.17
9	9.1	–2.80	(.14)	[–3.08, –2.53]	.15 (.15)	0.50	0.15
13	13.1	–1.94	(.25)	[–2.42, –1.45]	.55 (.33)	0.17	0.15
	13.2	–1.82	(.42)	[–2.65, –1.01]	.62 (.50)	0.06	0.08

Note. M = posterior mean; SD = standard deviation of the posterior mean; BCI = Bayesian credible interval; Cat = score category.

with multiple measurement occasions, there are multiple ways of conceptualizing change values. For example, Steyer, Eid, and Schwenkmezger (1997) distinguish two types of change values in latent variable models: The change values type I imply that values are calculated with reference to the initial value, that is, in a pretest-posttest-follow-up design, the pretest serves as reference. In contrast, change values type II quantify change relative to the nearest measurement points. While Naumann and colleagues' (2017) LMLIRT model is capable of providing both types of measures for group-specific change, the practical implications of the choice of type I and/or type II change values and their appropriateness in different contexts of evaluating instructional sensitivity have not been discussed yet.

To illustrate possible practical implications when modelling change values in educational intervention studies, three prototypical examples with information on the students' performance development of the experimental and control groups are depicted in Figure 4 (adapted from Kauffeld, 2010). In Figure 4, diagram a shows the condition "sensitive items capture effects of the ideal type of intervention", diagram b represents the condition "sensitive items capture effects of a successful intervention", and diagram c shows the condition "sensitive items capture effects of a successful intervention with later development". These three diagrams in Figure 4 underline that depending on the type of change values (type I or type II) chosen, the change values vary differently. Accordingly, it is important to specify whether we are interested in the sensitivity of test items for short-time effects (pretest – posttest) or for long-time effects (pretest – follow-up test, posttest – follow-up test). We thus recommend checking sensitivity of the pretest, posttest and/or follow-up accordingly.

Tab. 2: Covariance-Analytical Approach: Item sensitivity results for all IGEL-items

		LMLIRT Model Differential Sensitivity	CTT Screening Procedure Differential Sensitivity
Item	**Cat**	**M (SD)**	
1	1.1	0.43 (0.14)	0.18
	1.2	0.62 (0.20)	0.16
2	2.1	0.21 (0.13)	0.13
3	3.1	0.07 (0.07)	0.10
4	4.1	0.34 (0.12)	0.18
	4.2	1.19 (0.43)	0.14
5	5.1	0.62 (0.21)	0.18
6	6.1	0.54 (0.19)	0.18
	6.2	1.37 (0.56)	0.11
7	7.1	0.28 (0.11)	0.16
	7.2	0.28 (0.16)	0.12
8	8.1	0.40 (0.17)	0.16
9	9.1	0.24 (0.14)	0.11
	9.2	0.74 (0.29)	0.12
10	10.1	0.26 (0.13)	0.13
11	11.1	0.20 (0.10)	0.13
12	12.1	1.24 (0.40)	0.14
	12.2	1.95 (0.73)	0.12
13	13.1	0.75 (0.28)	0.14
	13.2	0.93 (0.43)	0.08

Note. M = posterior mean; SD = standard deviation of the posterior mean; BCI = Bayesian credible interval; Cat = score category.

Concluding Remarks

In the present chapter, we first provided a brief overview on the concept of instructional sensitivity and then pointed out differences and communalities in its role in educational effectiveness research and educational intervention studies. After presenting common ways of measuring instructional sensitivity, we proposed a screening procedure based on CTT that allows for approximating the absolute instructional sensitivity of single items in situations where more complex approaches are not feasible, for example, when sample sizes are small. Finally, we discussed the role of time lapses in the context of instructional sensitivity. We are confident that the ideas presented in this book chapter help fostering the valid use and interpretation of test scores in the context of educational intervention studies. Again, we would like to point out that the screening procedures presented in this chapter can

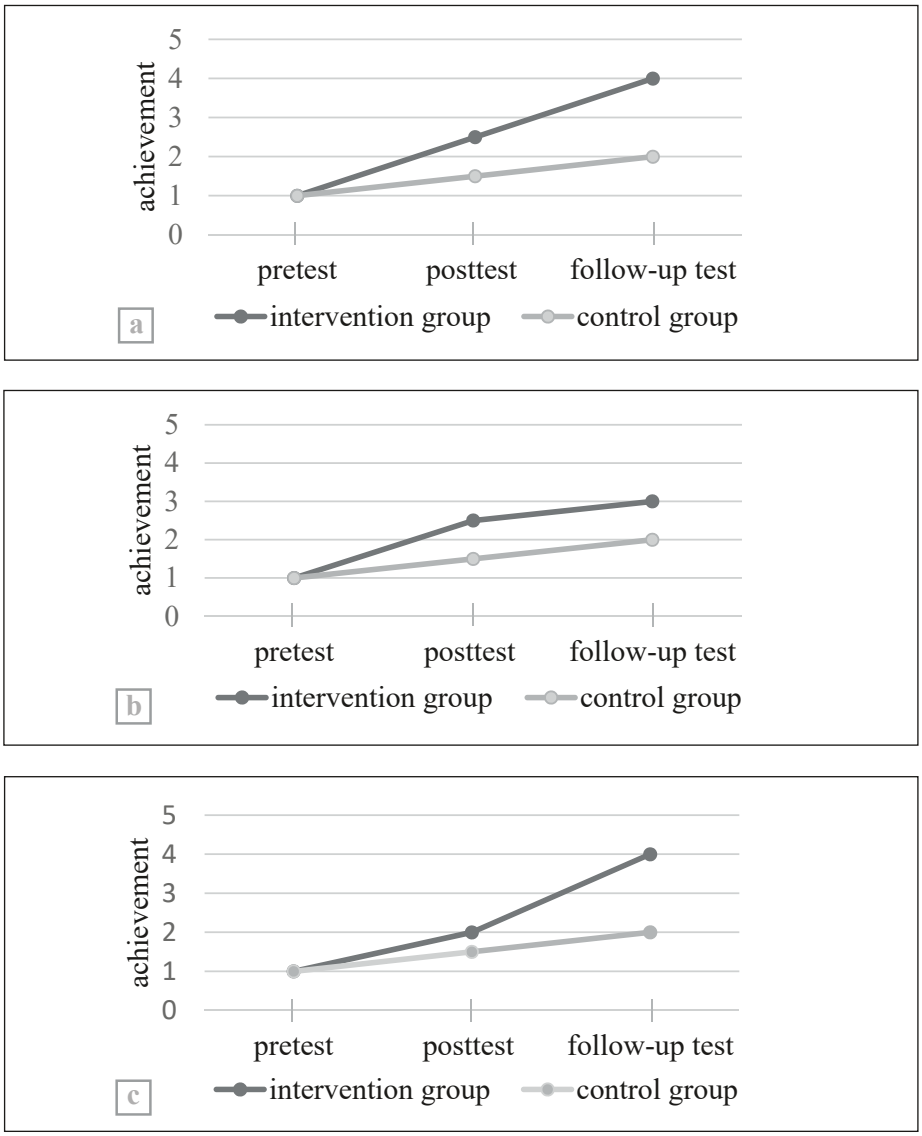


Fig. 4: Ways how sensitive items capture effects of a successful intervention: (a) ideal type of intervention, (b) successful intervention, and (c) with delayed effect (adapted from Kauffeld, 2010).

only help preventing insensitivity. They cannot fully replace a deeper analysis of instructional sensitivity.

Sensitivity of instruments to change due to treatments is regularly discussed in various domains like psychology or medicine (e.g., Benoy et al., 2019; Hays & Hadorn, 1992). Similar to these domains, sensitivity to teaching is oftentimes neglected in educational intervention studies, compromising the validity of inferences drawn from test scores (cf. Burstein, 1989; D'Agostino et al., 2007). In education-

al intervention studies, researchers need to make sure that instruments used for measuring the outcome criteria are capable of detecting potential intervention effects. Hence, measures of instructional sensitivity establish links between student responses and the inferential target and thus serve as validity evidence (cf. AERA et al., 2014; Naumann et al., 2019b). More specifically, information on instructional sensitivity supports (a) the evidence model in evidence-centered design (e.g., Mislevy & Haertel, 2006) or (b) the instructional and inferential facets within Pellegrino, DiBello and Goldman's (2016) validity framework, respectively. Accordingly, without sufficient information on instructional sensitivity, there is no argument supporting a specific instrument's use for measuring the intervention's outcome criteria. Consequently, instructional sensitivity is a necessary prerequisite for the valid use and interpretation of test scores in educational effectiveness research, as well as in educational intervention studies.

Following the psychometric framework by Naumann and colleagues (2017), absolute measures of instructional sensitivity essentially address the reliability of item responses or test scores on the level of learning groups (e.g., classes or schools) with respect to differences between (a) the learning environments students are exposed to (i.e. their learning groups or intervention conditions) or (b) the different stages of learning (i.e. time points of measurement), respectively. However, researchers should not be guided solely by the degree of sensitivity when designing a test, as otherwise effect sizes may increase as a function of item sensitivity (see Naumann et al., 2019b). As a result, inferences on teaching or intervention effectiveness may become invalid if the resulting test is not representative for the underlying task universe. That is, researchers need to clarify which test (Grossman et al., 2014) or which configuration(s) of items (Naumann et al., 2019b) is representative for the desired construct and provides the desired level of instructional sensitivity.

The previous considerations notwithstanding, item selection is not trivial even when information on instructional sensitivity is available. Despite van der Linden's (1981) request for validating instructional sensitivity measures, valid use and interpretation of measures with respect to teaching is still unclear for most of the item sensitivity statistics presented by Polikoff (2010). At best, statistics try approximating influences of learning environments students are exposed to on item responses by using classroom-membership as grouping variable when estimating item parameters (e.g., Robitzsch, 2009). At least partly, the LMLIRT model overcomes this issue as Naumann and colleagues (2019b) were able to provide empirical evidence supporting LMLIRT differential sensitivity measures validity.

Still, it appears hard to define upfront which specific teaching aspect(s) a single item can detect and which not (see also Ing, 2018). Ideally, items represent learnable leaps from one level of sophistication to the next level of sophistication within a domain. As such, they should be sensitive to adequate teaching of content and skills. Yet, while there are strong requests on what tests and items should not be sensitive to (e.g., inherited ability or SES; Popham, 2007), there is no consensus on which specific teaching aspects instruments should be able to capture (Polikoff, 2010). In

our view, the answer to this question largely depends on the purpose(s) of the assessment and the desired test score interpretation. For example, a test that serves as a criterion for judging whether or not a specific facet of teaching quality is effective should be sensitive to the quality of teaching. In educational effectiveness studies that resort to natural variation within a population, tests oftentimes serve for multiple purposes at the same time. Then, operationalizing instructional sensitivity may become all the more complex the more purposes have to be fulfilled, as each intended test score interpretation requires fitting validity evidence in the form of a proof of sensitivity (cf. Kane, 2013). Nevertheless, in the case of educational intervention studies, the purpose of the assessment can usually be expected to be clearly defined. Accordingly, tests should at least be sensitive to those teaching/intervention characteristics whose effectiveness intervention studies are about to judge.

References

Adl-Amini, K., Decristan, J., Hondrich, A.L., & Hardy, I. (2014). Umsetzung von peer-gestütztem Lernen im naturwissenschaftlichen Sachunterricht der Grundschule [Implementation of peer-supported learning in scientific science teaching]. *Zeitschrift für Grundschulforschung, 7*, 74–87.

AERA, APA, & NCME (2014). *Standards for educational and psychological testing*. Washington: AERA.

Airasian, P.W., & Madaus, G.F. (1983). Linking testing and instruction: Policy issues. *Journal of Educational Measurement, 20*, 103–118. doi: https://doi.org/10.1111/j.1745-3984.1983.tb00193.x

Anderson, L.W. (2002). Curricular alignment: A re-examination. *Theory Into Practice, 41*, 255–260. doi: https://doi.org/10.1207/s15430421tip4104_9

Baker, E. L. (1994). Making performance assessment work: The road ahead. *Educational Leadership, 51,* 58–62. Retrieved from http://www.ascd.org/publications/educational-leadership/mar94/vol51/num06/Making-Performance-Assessment-Work@-The-Road-Ahead.aspx

Benoy, C., Knitter, B., Schumann, I., Bader, K., Walter, M., & Gloster A. (2019). Treatment sensitivity: Its importance in the measurement of psychological flexibility. *Journal of Contextual Behavioral Science, 13*, 121–125. doi: https://doi.org/10.1016/j.jcbs.2019.07.005

Bos, W., Valtin, R., Hornberg, S., Arnold, K.-H., Faust, G., Fried, L., & Schwippert, K. (2008). Zusammenfassung und Schlussfolgerungen [Summary and conclusions]. In W. Bos, R. Valentin, S. Hornberg, K.-H. Arnold, G. Faust, L. Fried, L. Lankes, Schwippert, K. & Valtin, R. (Eds.), *IGLU-E 2006. Die Länder der Bundesrepublik Deutschland im nationalen und internationalen Vergleich* [IGLU-E 2006. A National and International Comparison of the Federal States of Germany.] (pp. 143–156). Münster: Waxmann.

Burstein, L. (1989, March). *Conceptual considerations in instructionally sensitive assessment*. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Chen, J. (2012). *Impact of instructional sensitivity on high-stakes achievement test items: A comparison of methods* (Unpublished doctoral dissertation). University of Kansas, Lawrence.

Cox, R.C., & Vargas, J.S. (1966, February). *A comparison of item-selection techniques for norm referenced and criterion referenced tests*. Paper presented at the Annual Conference of the National Council on Measurement in Education, Chicago, IL.

D'Agostino, J.V., Welsh, M.E., & Corson, N.M. (2007). Instructional sensitivity of a state standards-based assessment. *Educational Assessment, 12*, 1–22. doi: https://doi.org/10.1080/10627190709336945

Decristan, J., Hondrich, A. L., Büttner, G., Hertel, S., Klieme, E., Kunter, M., et al. (2015a). Impact of additional guidance in science education on primary students' conceptual understanding. *The Journal of Educational Research, 108*, 358–370. doi: https://doi.org/10.1080/00220671.2014.899957

Decristan, J., Klieme, E., Kunter, M., Hochweber, J., Büttner, G., Fauth, B., et al. (2015b). Embedded formative assessment and classroom process quality: How do they interact in promoting science understanding? *American Educational Research Journal, 52*, 1133–1159. doi: https://doi.org/10.3102/0002831215596412

Deutscher, V., & Winther, E. (2018). Instructional sensitivity in vocational education. *Learning and Instruction, 53*, 21–33. doi: https://doi.org/10.1016/j.learninstruc.2017.07.004

Grossman, P., Cohen, J., Ronfeldt, M., & Brown, L. (2014). The test matters: The relationship between classroom observation scores and teacher value added on multiple types of assessment. *Educational Researcher, 43*, 293–303. doi: https://doi.org/10.3102/0013189X14544542

Hardy, I., Jonen, A., Möller, K., & Stern, E. (2006). Effects of instructional support within constructivist learning environments for elementary school students' understanding of "floating and sinking". *Journal of Educational Psychology, 98*, 307–326. doi: https://doi.org/10.1037/0022-0663.98.2.307

Hardy, I., Kleickmann, T., Koerber, S., Mayer, D., Möller, K., Pollmeier, J., Schwippert, K., & Sodian, B. (2010). Die Modellierung naturwissenschaftlicher Kompetenz im Grundschulalter [Modeling scientific competence in primary-school age]. *Zeitschrift für Pädagogik, 56*, 115–125.

Hartig, J., Frey, A., & Jude, N. (2012). Validität [Validity]. In H. Moosbrugger & A. Kelava (Eds.), *Testtheorie und Fragebogenkonstruktion* [Testing Theory and Design of Questionnaires]. (2nd ed., pp. 143–171). Berlin: Springer. doi: https://doi.org/10.1007/978-3-642-20072-4_7

Hascher, T., & Schmitz, B. (2010). *Pädagogische Interventionsforschung. Theoretische Grundlagen und empirisches Handlungswissen* [Educational intervention research. Theoretical basics and empirical action knowledge]. Weinheim: Juventa.

Hays, R.D., & Hadorn, D. (1992). Responsiveness to change: An aspect of validity, not a separate dimension. *Quality of Life Research, 1*, 1–73. doi: https://doi.org/10.1007/BF00435438

Hondrich, A. L., Hertel, S., Adl-Amini, K., & Klieme, E. (2016). Implementing curriculum-embedded formative assessment in primary school science classrooms. *Assess-*

ment in Education: Principles, Policy & Practice, *23*, 353–376. doi: https://doi.org/10.1080/0969594X.2015.1049113

Ing, M. (2018). What about the "instruction" in instructional sensitivity? Raising a validity issue in research on instructional sensitivity. *Educational and Psychological Measurement, 78*, 635–652. doi: https://doi.org/10.1177/0013164417714846

Kane, M.T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73. doi: https://doi.org/10.1111/jedm.12000

Kauffeld, S. (2010). *Nachhaltige Weiterbildung. Betriebliche Seminare und Trainings entwickeln, Erfolge messen, Transfer sichern* [Sustainable training. Developing operational seminars and trainings, measuring success, ensuring transfer]. Berlin: Springer. doi: https://doi.org/10.1007/978-3-540-95954-0

Kleickmann, T., Hardy, I., Möller, K., Pollmeier, J., Tröbst, S., & Beinbrech, C. (2010). Die Modellierung naturwissenschaftlicher Kompetenz im Grundschulalter: Theoretische Konzeption und Testkonstruktion [Modeling scientific competence in primary-school age. Theoretical conception and test construction]. *Zeitschrift für Didaktik der Naturwissenschaften, 16*, 265–284. Retrieved from http://hdl.handle.net/11858/00-001M-0000-0024-F649-7

Klieme, E. (2019). Unterrichtsqualität [Quality of instruction]. In M. Gläser-Zikuda, M. Harring & C. Rohlfs (Eds.), *Handbuch Schulpädagogik* [Handbook School Pedagogics]. (pp. 393–408). Münster: Waxmann.

Kosecoff, J.B., & Klein, S.P. (1974, April). *Instructional sensitivity statistics appropriate for objectives-based test items*. Paper presented at the Annual Conference of the National Council on Measurement in Education, Chicago, IL.

Linn, R.L. (1983). Testing and instruction: Links and distinctions. *Journal of Educational Measurement, 20*, 179–189. doi: https://doi.org/10.1111/j.1745-3984.1983.tb00198.x

Linn, R.L., & Harnisch, D.L. (1981). Interactions between item content and group membership on achievement test items. *Journal of Educational Measurement, 18*, 109–118. doi: https://doi.org/10.1111/j.1745-3984.1981.tb00846.x

Marsh, H.W., Lüdtke, O., Nagengast, B., Trautwein, U., Morin, A.J.S., Abduljabbar, A. S., & Köller, O. (2012). Classroom climate and contextual effects: Conceptual and methodological issues in the evaluation of group-level effects. *Educational Psychologist, 47*, 106–124. doi: https://doi.org/10.1080/00461520.2012.670488

Masters, G.N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174. doi: https://doi.org/10.1007/BF02296272

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). Phoenix: Oryx Press.

Mislevy, R.J., & Haertel, G. (2006). Implications of evidence-centered design for educational assessment. *Educational Measurement: Issues and Practice, 25*, 6–20. doi: https://doi.org/10.1111/j.1745-3992.2006.00075.x

Möller, K., Jonen, A., Hardy, I., & Stern, E. (2002). Die Förderung von naturwissenschaftlichem Verständnis bei Grundschulkindern durch Strukturierung der Lernumgebung [Fostering scientific understanding of primary school children by structuring their learning environment]. In M. Prenzel & J. Doll (Eds.), *Bildungsqualität von Schule: Schulische und außerschulische Bedingungen mathematischer, naturwis-*

senschaftlicher und überfachlicher Kompetenzen [Quality of Education at School: Academic and Out-of-School Requirements of Mathematical, Scientific and Interdisciplinary Competencies]. (pp. 176–191). Weinheim: Beltz.

Musow, S., Naumann, A., Hochweber, J., & Hartig, J. (2019a, April). *Multilevel IRT as a validation strategy for expert judgements on instructional sensitivity.* Paper presented at the Annual Conference of the National Council on Measurement in Education, Toronto, CAN.

Musow, S., Naumann, A., Ruiz-Primo, M.A., Hartig, J., & Hochweber, J. (2019b). *Expert judgments – Is it an appropriate approach to evaluate instructional sensitivity?* Manuscript in preparation for publication.

Muthén, B.O., Huang, L., Jo, B., Khoo, S.-T., Goff, G.N., Novak, J.R., & Shih, J. C. (1995). Opportunity-to-learn effects on achievement: Analytical aspects. *Educational Evaluation and Policy Analysis, 17*, 371–403. doi: https://doi.org/10.3102/01623737017003371

Naumann, A., Hartig, J., & Hochweber, J. (2017). Absolute and relative measures of instructional sensitivity. *Journal of Educational and Behavioral Statistics, 42*, 678–705. doi: https://doi.org/10.3102/1076998617703649

Naumann, A., Hochweber, J., & Hartig, J. (2014). Modeling instructional sensitivity using a longitudinal multilevel differential item functioning approach. *Journal of Educational Measurement, 51*, 381-399. doi: https://doi.org/10.1111/jedm.12051

Naumann, A., Hochweber, J., & Klieme, E. (2016). A psychometric framework for the evaluation of instructional sensitivity. *Educational Assessment, 21*, 1–13. doi: https://doi.org/10.1080/10627197.2016.1167591

Naumann, A., Musow, S., Aichele, C., Hochweber, J., & Hartig, J. (2019a). Instruktionssensitivität von Tests und Items [Instructional sensitivity of tests and items]. *Zeitschrift für Erziehungswissenschaft, 22*, 181–202. doi: https://doi.org/10.1007/s11618-018-0832-0

Naumann, A., Rieser, S., Musow, S., Hochweber, J., & Hartig, J. (2019b). Sensitivity of test items to teaching quality. *Learning and Instruction, 60*, 41–53. doi: https://doi.org/10.1016/j.learninstruc.2018.11.002

Pellegrino, J.W. (2002). Knowing what students know. *Issues in Science & Technology, 19*, 48–52. doi: https://doi.org/10.17226/10019

Pellegrino, J.W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*, 59–81. doi: https://doi.org/10.1080/00461520.2016.1145550

Polikoff, M.S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issues and Practice, 29*, 3–14. doi: https://doi.org/10.1111/j.1745-3992.2010.00189.x

Polikoff, M.S. (2016). Evaluating the instructional sensitivity of four states' student achievement tests. *Educational Assessment, 21*, 102–119. doi: https://doi.org/10.1080/10627197.2016.1166342

Popham, W.J. (2007). Instructional insensitivity of tests: accountability's dire drawback. *Phi Delta Kappan, 89*, 146–155. doi: https://doi.org/10.1177/003172170708900211

Popham, W.J., & Ryan, J.M. (2012, April). *Determining a high-stakes test's instructional sensitivity*. Paper presented at the Annual Conference of the National Council on Educational Measurement in Education, Vancouver, Canada.

R Core Team (2019). *R: A language and environment for statistical computing.* Vienna, Austria. Retrieved from http://www.R-project.org.

Robitzsch, A. (2009). Methodische Herausforderungen bei der Kalibrierung von Leistungstests [Methodological challenges in the calibration of performance tests]. In D. Granzer, O. Köller & A. Bremerich-Vos (Eds.), *Bildungsstandards Deutsch und Mathematik* [Scholastic Standards German and Mathematics]. (pp. 42–106). Weinheim: Beltz.

Rost, J. (2004). *Lehrbuch Testtheorie – Testkonstruktion [Course book on test theory – test construction]*. Bern: Huber.

Ruiz-Primo, M.A., Li, M., Wills, K., Giamellaro, M., Lan, M.C., Mason, H., & Sands, D. (2012). Developing and evaluating instructionally sensitive assessments in science. *Journal of Research in Science Teaching, 49*, 691–712. doi: https://doi.org/10.1002/tea.21030

Schmidt, W.H., Porter, A.C., Schwille, J.R., Floden, R., & Freeman, D. (1983). Validity as a variable: Can the same certification test be valid for all students. In G.F. Madaus (Ed.), *The courts, validity, and minimum competency testing* (pp. 133–151). Hingham: Kluwer-Nijhoff Publishing. doi: https://doi.org/10.1007/978-94-017-5364-7_6

Stan Development Team (2019). *RStan: The R interface to Stan. R package version 2.19.2*. http://mc-stan.org/

Steyer, R., Eid, M., & Schwenkmezger, P. (1997). Modeling true intraindividual change: True change as a latent variable. *Methods of Psychological Research Online, 2*, 21–33. Retrieved from https://www.dgps.de/fachgruppen/methoden/mpr-online/issue2/art2/steyer.pdf

van der Linden, W. J. (1981). A latent trait look at pretest-posttest validation of criterion-referenced test items. *Review of Educational Research, 51*, 379–402. doi: https://doi.org/10.3102/00346543051003379