Naumann, Alexander; Hochweber, Jan; Klieme, Eckhard

A psychometric framework for the evaluation of instructional sensitivity

formal und inhaltlich überarbeitete Version der Originalveröffentlichung in:
formally and content revised edition of the original source in:
Educational assessment 21 (2016) 2, S. 89-101

A Psychometric Framework for the Evaluation of Instructional Sensitivity

(This is a preprint, the definitive version is available at
http://www.tandfonline.com/doi/full/10.1080/10627197.2016.1167591)

Alexander Naumann

German Institute for International Educational Research (DIPF), Frankfurt, Germany

IDeA Research Center, Frankfurt, Germany

Jan Hochweber

University of Teacher Education St. Gallen, Switzerland

German Institute for International Educational Research (DIPF), Frankfurt, Germany

Eckhard Klieme

German Institute for International Educational Research (DIPF), Frankfurt, Germany

IDeA Research Center, Frankfurt, Germany

Please cite as:

Naumann, A., Hochweber, J., & Klieme, E. (2016). A Psychometric Framework for the Evaluation of Instructional Sensitivity. Educational Assessment, 21(2), 1–13, DOI:10.1080/10627197.2016.1167591

Abstract

Although there is a common understanding of instructional sensitivity, it lacks a common operationalization. Various approaches have been proposed, some focusing on item responses, others on test scores. As approaches often do not produce consistent results, previous research has created the impression that approaches to instructional sensitivity are noticeably fragmented. To counter this impression, we present an IRT-based framework which can help us to understand similarities and differences between existing approaches. Using empirical data for illustration, this paper identifies *three perspectives* on instructional sensitivity: One perspective views instructional sensitivity as the capacity to detect differences in students' stages of learning across points of time. A second perspective treats instructional sensitivity as the capacity to detect differences between groups that have received different instruction; for a third perspective, the previous two are combined to consider differences between both time points and groups. We discuss linking sensitivity indices to measures of instruction.

Keywords: Instructional sensitivity, psychometrics, validity

A Psychometric Framework for the Evaluation of Instructional Sensitivity

Assessments of students' competencies and achievement are widely used in educational research and evidence-based policy making (Pellegrino, 2002; Hartig, Klieme, & Leutner, 2008). In many cases, results of these tests are more or less explicitly related to the effectiveness of the instruction that students have received (Creemers & Kyriakides, 2008). Yet, more often than not there is only little validity evidence to support this way of test score use and interpretation. In response, many researchers have asserted that construction and evaluation of assessments for diverse educational purposes need detailed information about how test scores and item responses are influenced by classroom instruction (e.g., Burstein, 1989), giving rise to the concept of *instructional sensitivity*.

Instructional sensitivity is defined as the psychometric capacity of a test or a single item to capture the effects of instruction (Polikoff, 2010). More specifically, instructional sensitivity refers to the extent to which test scores and item responses are influenced by the content and the quality of instruction that students have received (e.g., D'Agostino, Welsh, & Corson, 2007; Muthén, Kao, & Burstein, 1991). In instructionally sensitive assessments, scores are expected to be positively related to more or better teaching (Baker, 1994). Also, students who have received different kinds of instruction should respond differently to instructionally sensitive items (Ing, 2008). Thus, instructional sensitivity can be seen as a necessary requirement when drawing inferences on instruction based on test scores (Popham, 2007).

Although there is a shared understanding of the concept of instructional sensitivity, it lacks a common approach to operationalization. Various approaches have been proposed (see Polikoff, 2010): some focus on item responses (item level), others on test scores (test level), with most approaches producing inconsistent results (e.g., Li, Ruiz-Primo, & Wills, 2012). As a

result, previous research has created the impression that approaches to instructional sensitivity are noticeably fragmented.

In this paper, we seek to contribute to a more systematic understanding of similarities and differences between common approaches to instructional sensitivity. We assume the reason for commonly used approaches not providing consistent results in their focus on different facets of instructional sensitivity. Thus, we argue that testing the hypothesis of whether or not a test or a single item is instructionally sensitive involves the question "instructionally sensitive in what respect?". To provide guidance for researchers in answering this question, we identify three perspectives within an IRT framework that cover the different facets of instructional sensitivity: (1) the differences-between-groups perspective, related to the extent a single item or test can differentiate between groups of students who receive different instruction; (2) the differences-between-time-points perspective, related to the direction and extent of progress reflected in tests and items; and (3) a combination of (1) and (2) to consider a differences-between-groups-and-time-points perspective. That is, we seek to contribute knowledge about how tests and items reflect variability due to the instructional context (e.g., classroom-membership) and how this variability can be related to measures of instruction to determine the portion of variance relevant to the construct of instructional sensitivity.

Approaches to Instructional Sensitivity

Fundamental to the concept of instructional sensitivity is the expectation that student responses change as a consequence of instruction (Burstein, 1989). In a recent review of strategies to evaluate instructional sensitivity, Polikoff (2010) categorized the diverse approaches according to the kind of evidence used: (1) expert judgment (e.g., Popham, 2007), (2) instruction-focused

(e.g., Niemi, Wang, Steinberg, Baker, & Wang, 2007), or (3) item statistics (e.g., Clauser, Nungester, & Swaminathan, 1996; Cox & Vargas, 1966).

Instruction-focused approaches and item statistics rely on empirically measured student responses, which are not considered in expert judgment approaches. Viewing students' responses as critical to the evaluation of instructional sensitivity, we examine instruction-focused approaches and item statistics. The perspectives we describe, however, at least in principle, might also be applicable to expert judgment approaches.

Evaluation of Instructional Sensitivity Using Instruction-Focused Approaches

Current analyses of instructional sensitivity that incorporate instructional measures, such as measures of teaching content or quality, use multilevel regression models (Raudenbush & Bryk, 2002) with students nested in classes (or higher cluster levels) to relate teaching characteristics to achievement (e.g., D'Agostino et al., 2007; Ing, this issue). If meaningful relationships between students' test scores and measures of amount and/or quality of instruction are found, the test is considered instructionally sensitive to the corresponding facet of instruction. Researchers following this approach have introduced a wide range of instructional measures (e.g., content coverage, emphasis of the content, alignment between instruction and test) into the evaluation of instructional sensitivity. However, in contrast to work by Muthén and colleagues (Muthén, 1989; Muthén et al., 1991), recent studies (e.g., D'Agostino et al., 2007) have investigated instructional sensitivity exclusively at the test level, neglecting the relationship between instructional measures and students' responses on single items.

Evaluation of Instructional Sensitivity Using Item Statistics

Current approaches using item statistics focus on the difficulty or the discrimination of items (Haladyna, 2004). The most prominent approaches are the Pretest-Posttest Difference

Index (PPDI; Cox & Vargas, 1966) and the investigation of differential item functioning (DIF; Holland & Wainer, 1993). According to PPDI, instructional sensitivity is measured as the difference between the proportion of students who correctly answer an item at pretest and those who correctly answer the item at posttest measurement. The more an item's difficulty changes across time, the higher its instructional sensitivity. In contrast, DIF studies usually investigate groups of students who were assumed to have been exposed to different opportunities-to-learn (OTL) or educational experiences (e.g., Linn & Harnisch, 1981; Zwick & Ercikan, 1989), including country-specific teaching cultures (Klieme & Baumert, 2001). Accordingly, the variation of item difficulties across groups is conceived as an indicator of an items' instructional sensitivity (e.g., Robitzsch, 2009). The more an item's difficulty varies across groups, the higher its instructional sensitivity.

Recently, Naumann and colleagues (2014) combined the PPDI and DIF approaches to instructional sensitivity in a longitudinal multilevel DIF (LMLDIF) model. The model allows estimating classroom-specific change in item difficulties across measurement occasions. In contrast to PPDI and DIF, two statistical indicators are proposed to describe an item's instructional sensitivity, that is, the item's average change in difficulty across time points and the variation of change across classes. The authors concluded that if only one item statistic is used, the evaluation of instructional sensitivity might be partially incomplete.

However, recent studies have shown that these items statistics do not lead to consistent results when applied to the same set of data (Li, Ruiz-Primo, & Wills, 2012; Naumann, Hochweber, & Hartig, 2014). Additionally, the connection to instructional sensitivity of the test and to instructional measures has not been widely investigated for the item statistics described above (Polikoff, 2010).

Levels of Analysis in the Evaluation of Instructional Sensitivity

We begin with a more general systematization of the approaches presented above by describing sources of variance relevant to the evaluation of instructional sensitivity. That is, in addition to the evidence that is used, existing approaches to evaluate instructional sensitivity are related to different levels of analysis within a hierarchical framework.

From a multilevel IRT perspective, item responses can be regarded as hierarchically nested, cross-classified within persons and items (Van den Noortgate, De Boeck, & Meulders, 2003). Persons and items again are nested within higher clusters. Higher cluster levels for persons include classes, teachers, schools, districts, or countries (Raudenbush & Bryk, 2002). Similarly, items are nested within testlets and tests (Lee, Brennan, & Frisbie, 2000). The hierarchical level on the item side determines the level at which evaluation of instructional sensitivity is conducted—that is, whether analysis focuses on single items (item level) or on what is common to a set of items (i.e., test level). On the person side, clusters (such as classes) represent potential sources of different instruction.

An additional hierarchy emerges when item responses are nested within time points. That is, each student provides responses on items at multiple time points of measurement, allowing analysis of intra-individual change (e.g., in ability) or intra-item change (e.g., in difficulty) across time. Generally, group clustering and time point clustering may occur simultaneously (Gelman & Hill, 2006). That is, sources of variance are between groups and between time points. According to the clustering considered in the evaluation of instructional sensitivity, we distinguish between the following *three perspectives*: (1) the differences-between-time-points perspective, (2) the differences-between-groups perspective, and (3) a combination of (1) and

(2), that is, a differences-between-groups-and-time-points perspective. Figure 1a presents a framework based on these three perspectives on instructional sensitivity.

- Insert Figure 1a about here –

Within this framework, the perspectives can be distinguished with respect to the variance components under investigation. These variance components are critical to testing the hypothesis of whether or not a test or a single item is sensitive to instruction, as they determine the way one conceives of how instructional sensitivity becomes noticeable in test scores and item responses. In the following, we discuss modeling of instructional sensitivity at item and test levels for each perspective.

Modeling Different Perspectives on Instructional Sensitivity

Based on this framework, we describe the perspectives on instructional sensitivity in more detail. We first discuss a necessary prerequisite for instructional sensitivity, variability in item parameters that are due to the instructional context (e.g., classroom membership), and how this variability can be evaluated in an IRT framework. Instructional sensitivity, then, is defined as the proportion of variance due to content and quality of instruction. The following section describes the empirical data we use for illustration.

Illustrative Data

To support our reasoning, we use empirical data from the study "Individual Support and Adaptive Learning Environments in Primary School" (IGEL; Hardy et al., 2011). IGEL is a cluster-randomized controlled trial on adaptive teaching methods in 3rd grade in German primary schools. Teachers were trained in one of three adaptive teaching methods, which they implemented in a standardized curriculum on "floating and sinking." The content of instruction was intended to be identical in all classrooms, but the teaching methods varied according to the

treatment group to which each teacher was assigned. The data used in this paper are from 991 students (M_{age}= 8.8 years, 49% female) in 54 classes.

Students' content knowledge of floating and sinking was assessed immediately before and after implementation of the teaching unit in classroom-wide assessments. Items were either developed by IGEL staff or adapted from the SCIENCE-P project (Hardy et al., 2010) and from TIMSS 2007 (Martin, Mullis, & Foy, 2008). Scoring followed students' conceptual understanding of floating and sinking (Kleickmann, 2010): (1) naive conceptions, (2) explanations of everyday life, and (3) scientific explanations.

Each of the sections below introduces one of the perspectives together with an illustration of the underlying concepts using students' results on a single trichotomously scored IGEL item. The "stone" item required students to explain why a stone sinks when it is placed in water. Responses were assigned to the intermediate-score category if students used explanations based on everyday life (e.g., reference to material, "a stone sinks because it is made of stone"), and to the highest-score category for scientific explanations (e.g., concept of density, "a stone sinks because it is heavier than the same amount of water").

Perspective 1: Differences-between-Time-Points

The aim underlying the differences-between-time-points perspective on instructional sensitivity is to engage the directionality and extent of learning progress reflected in tests and items in a sample. This is inherently a longitudinal perspective; and between-time-points instructionally sensitive assessments should reliably depict the change in achievement due to instruction across time. Items and tests sensitive to instruction should become easier across measurement occasions—or, more difficult, in certain cases of conceptual change (Vosniadou, 2007) or bad teaching. In contrast, tests and items that are insensitive between time points should

not significantly change in difficulty over time even if adequate or high-quality instruction is provided. Depending on the research question, the scale of time spanned might be at a macro level, e.g., assessments at the beginning and at the end of a school year—or at a micro level, e.g., multiple assessments within a teaching unit. Technically, the number of measurement occasions is unlimited. Nevertheless, the approaches presented thus far define between-time-point sensitivity as the change in item difficulty between a pretest and a posttest.

Table 1 shows the relative frequencies of response categories in the "stone" item at pretest and posttest. Most students (86%) demonstrated naïve conceptions at pretest. At posttest, however, only 55% of students showed naïve conceptions. That is, 27% of the students reached the intermediate score category (pre: 10%), and 18% of the students offered scientific explanations (pre: 4%). Hence, empirical data suggested that the item might effectively detect differences in achievement due to instruction between time points of measurement.

- Insert Table 1 about here –

A strikingly intuitive representation of the differences-between-time-points perspective and a well-known example of measuring instructional sensitivity is the approach taken by the PPDI (Cox & Vargas, 1966) when there is no untreated control group. PPDI essentially measures item-specific learning, with a clearly evident link to learning on the test as a whole. We will demonstrate this relationship based on a two-dimensional one-parameter logistic (1PL) model (Reckase, 2009), with each dimension representing one time-point of measurement (Hartig & Kühnbach, 2006). In this longitudinal 1PL model, the probability that person *v* correctly answers item *i* at time *t* is expressed by:

$$logit[p(X_{tvi} = 1)] = \theta_{tv} - \beta_{ti}, \quad (2)$$

where, for each time point of measurement t, θ_{tv} denotes the individual ability of person v, and β_{ti} is the difficulty of item i. Individual ability and item parameters may be conceived as multivariate normally distributed with time-point specific means and variances (e.g., Fox, 2010):

$$\begin{aligned} \theta_{tv} &\sim MNorm(\mu_t, \sigma^2{}_t), \\ \beta_{ti} &\sim MNorm(\beta_t, v^2{}_t). \end{aligned} \tag{3}$$

Similar to the PPDI, the change in item difficulty between two measurement occasions $\Delta\beta$ is defined as:

$$\Delta\beta_i = \beta_{ti} - \beta_{t-1i} \tag{4}$$

The model is not identified. Identification can be achieved by imposing constraints on either ability or item parameters. When each μ_t is set to equal zero, all effects of learning will be reflected in $\Delta\beta_i$, which is the item-specific learning gain on item i. Because each item is nested within an assessment a, the test-specific learning $\Delta\beta_a$ can be modeled as the expected value for item-specific learning:

$$\Delta\beta_a = \mathbf{E}(\Delta\beta_{ia}), \tag{5}$$

where $\Delta\beta_a$ denotes the average change in item difficulties across all items i between two time-points t. The average change in item difficulty for all items in the test reflects the average learning within the sample, which can be regarded as a measure of learning on the test as a whole—in other words, the test scores' instructional sensitivity according to the differences-between-time-points perspective.

Evaluation of instructional sensitivity within this perspective is not necessarily limited to two time points of measurement. Equation 5 can be readily applied to multiple measurement occasions, defining $\Delta\beta$ values for each segment between two time points of measurement. Sensitivity of a single item or test score between time points is then investigated across multiple

measurement occasions, which, in principle, would allow examination of the shape of the change in item difficulties across time.

Perspective 2: Differences-between-Groups

The differences-between-groups perspective on instructional sensitivity relates to the question about the extent to which a single item or a test can differentiate between two or more groups of students who receive different instruction. The intent is to examine scores and item responses for effects due to students' membership in groups (e.g., courses, classes, schools, educational systems) given varying opportunities to learn. Tests and items that are sensitive according to the differences-between-groups perspective should, of course, depict differences between groups. And insensitive tests and items should not significantly differentiate between groups even if different instruction is actually provided. In contrast to the differences-between-time-points perspective, evaluation of an instrument's instructional sensitivity from a differences-between-groups-perspective generally relies on cross-sectional (usually posttest) data.

Table 2 depicts students' responses to the stone item for two IGEL classes at posttest. More students in Class B than in Class A responded via explanations of everyday life. Scientific explanations are prevalent in Class A. Hence, the relative frequencies of the score categories suggest that the item can detect differences between both groups.

- Insert Table 2 about here –

The differences-between-groups perspective is evident in many studies on instructional sensitivity (e.g., Clauser et al., 1996). In evaluating instructional sensitivity, DIF methods typically have been used to investigate how items distinguish between two groups (a focal and a reference group) of students who received different instruction. Approaches describe between-

group sensitivity as to how item difficulty for the focal group deviates from item difficulty in the reference group:

$$logit[p(X_{vi} = 1)] = \theta_v - \beta_i - \beta_{i,focal.} \quad (6)$$

where θ_v denotes the ability of an individual person v, β_i is the difficulty of item i in the reference group, and $\beta_{i,focal}$ is the deviation in difficulty for the focal group. The greater the difference in difficulty of item i between both groups, the greater the item's ability to differentiate between them, that is, the higher the item's instructional sensitivity.

Analyses of item responses following approaches in the differences-between-groups perspective are not restricted to two groups. Accordingly, the DIF approach can be extended to settings with n ≥ 2 focal groups. Robitzsch (2009) introduced multilevel-DIF (Meulders & Xie, 2004) as an approach to instructional sensitivity, taking into account the hierarchical data structure that is inherent in complex samples. When groups (e.g., classes) are sampled randomly from a population of groups, between-groups sensitivity may be expressed as the variation of item difficulty across the groups:

$$logit[p(X_{kvi} = 1)] = \theta_v + \theta_{kv} - \beta_{ki},$$

with (7)

$$\beta_{ki} \sim Norm(\beta_i, \upsilon_i^2).$$

where θ_k denotes the average ability of group k; θ_{vk} is the individual deviation in ability of person v from the corresponding group-mean; and β_{ik} is the difficulty of item i in group k. In Robitzsch's (2009) approach, the group-specific item difficulties are assumed to be normally distributed and centered around an item-specific mean β_i; and the variance component υ_i^2 is taken to describe the extent to which a single item's difficulty varies across the groups within the sample. The greater the difficulty of a single item varies across groups, the greater the item's

instructional sensitivity. Regardless of whether two or more groups are being investigated, DIF effects should favor the group(s) exposed to teaching that is better aligned with the test (e.g., providing higher quality instruction or more time for learning the required content and skills).

Analogous to analyses on the item level, the (multilevel) regression approach can be applied at the test score level to investigate the between-groups instructional sensitivity of tests. In contrast to evaluations of item responses, most studies using multilevel regression analysis have included empirical measures of instruction in their analyses to investigate the amount of variance between groups explained by the predictor(s). Nevertheless, we start with a model that contains no predictors, and later in this paper will discuss its relationship to instructional measures.

In a two-level intercept-only hierarchical model, the variance in test scores is decomposed into the variance within groups (individual level) and the variance between groups (group level):

$$\begin{aligned} Y_{vk} &\sim Norm(\mu_k, \sigma^2), \\ \mu_k &\sim Norm(\mu, \tau^2), \end{aligned} \tag{8}$$

where Y_{vk} is the test score of person v in group k, μ_k is the average test score in group k centered around the sample mean μ. While σ^2 denotes the individual-level variance, τ^2 is the variation at the group level. The higher the τ^2, the more pronounced are the differences between groups. Accordingly, following a differences-between-groups perspective, overall group-level variation might be regarded as a statistical indicator for the instructional sensitivity of a test. This would be analogous to the item level, where differences in item difficulty between groups have been interpreted as indicators of an item's instructional sensitivity.

This similarity between statistical indicators of instructional sensitivity at the item and test levels can be made even more explicit. The difference between groups on average test scores is equivalent to their difference in average item difficulty. If item difficulties do not balance out across a test, then the greater each item's difficulty within a test varies across groups, the greater the likelihood that variance between groups will appear at the test level. At the same time, variation between groups on the item level or the test level may not be attributable solely to instruction, but may originate in other sources such as group composition (e.g., ability sorting).

Perspective 3: Differences-between-Groups-and-Time-Points

The previous sections were devoted to two essentially different perspectives on instructional sensitivity, the differences-between-groups and the differences-between-time-points perspectives. Both perspectives can be combined to create a single differences-between-groups-and-time-points perspective. To illustrate, Table 3 provides information on the change in relative frequencies for each score category between pretest and posttest on the stone item within two IGEL classes. As is evident, the relative frequencies of the score categories in each class change between pretest and posttest, and this change differs between classes. That is, the item may be regarded as (potentially) sensitive to instruction from both the between-time-points and the between-groups perspectives. However, the most prominent approaches are related to perspectives that do not capture both types of information. And this could be a crucial factor in examinations of instructional sensitivity, as the assessment of the sensitivity of a test or a single item might be incomplete.

- Insert Table 3 about here –

Naumann, Hochweber, and Hartig (2014) combined the longitudinal between-time-points perspective of the PPDI approach with the differences-between-groups perspective of the ML-

DIF approach in a longitudinal multilevel-DIF model. The result was to integrate the indices' perspectives and derive the concepts of global sensitivity and differential sensitivity. In this approach, the probability of a correct response from person *v* in group *k* on item *i* at time *t* is modeled as a function of classroom ability, individual ability, and the difficulty of the item at the respective time point of measurement:

$$logit[p(X_{tkvi} = 1)] = \theta_{tv} + \theta_{tkv} - \beta_{tki}, \quad (9)$$

with classroom-specific pretest-posttest differences for each item defined as:

$$\Delta\beta_{ki} = \beta_{2ki} - \beta_{1ki} \quad (10)$$

The average pretest-posttest difference (PPD)—or average item-specific learning—across all groups in a sample reflects global sensitivity. Global sensitivity therefore refers to the extent to which the difficulty of a single item changes *on average* across groups and time points of measurement, i.e., the average directionality and extent of item-specific learning across time points of measurement for all groups. In contrast to approaches based solely on a differences-between-time-points perspective, the concept of global sensitivity explicitly considers nesting of students in groups. Differential sensitivity is conceptualized as the *variation* of item-specific learning between time points across groups within a sample. As such, the concept of differential sensitivity addresses differences in item-specific learning between groups, i.e., the extent to which an item can differentiate between group learning rates. In contrast to approaches based solely on a differences-between-groups perspective, the concept of differential sensitivity explicitly considers learning progress instead of achievement at a single time point.

A combination of information on global and differential sensitivity leads to a 2 × 2-typology of instructional sensitivity (see Table 4), which allows for a multiperspective judgment of a single item's potential sensitivity to instruction. Naumann and colleagues' (2014) results

suggest that if evaluation of instructional sensitivity is based on indices from only one perspective, the judgment may be somewhat incomplete and potentially misleading as there may be items that show sensitivity across time but not across groups, and vice versa.

- Insert Table 4 about here –

Despite originating in a statistical approach to instructional sensitivity at the item level, the concepts of global and differential instructional sensitivity can be extended to encompass both the item and the test level. Global sensitivity then may be more generally conceived as the extent to which the difficulty of a single item or a test changes, on average, across groups and time points of measurement. Similarly, differential sensitivity refers to the variation in learning on a single item or on the test as a whole, between time points across groups within a sample. The 2 × 2 typology can thus be applied, in principle, to the evaluation of both item responses and test scores.

In contrast to the quite recently published LMLDIF approach to the sensitivity of items, there's already a well-established approach to the instructional sensitivity on the test level taking a between-groups-and-time-points perspective: the multilevel regression analysis of students' posttest scores while controlling for prior achievement (e.g., D'Agostino et al., 2007). From a methodological point of view, the application of such covariate-adjusted models in longitudinal studies with two time points of measurement has been criticized, as, for example, they confound status and growth (Rowan, Correnti, & Miller, 2002).Nevertheless, the model represents a comprehensible example for extending the concepts of global and differential sensitivity to the test level. By considering prior achievement, the perspective on instructional sensitivity shifts from a between-groups perspective to a differences-between-groups-and-time-points perspective.

Then, a simple two-level regression model for achievement *Y* with students *v* nested within groups *k* is:

$$\begin{aligned} Y_{vk} &= \beta_{0k} + \beta_{01}*\text{PRIOR ACHIEVEMENT}_v + e_{vk}, \\ \beta_{0k} &= \gamma_{00} + u_k, \end{aligned} \tag{11}$$

where the group-level intercept γ_{00} may be seen as an indicator of global sensitivity, and the group-level variation (variance of u_k) as an indicator of differential sensitivity. However, the degree of global sensitivity of a test often is not of significant interest or cannot be determined due to the lack of a common metric across time points. Hence, most multilevel analyses for test scores that consider prior achievement focus on the differential aspect of instructional sensitivity (e.g., Ing, 2008).

Relationship to Empirical Measures of Instruction

In the previous discussion, instructional sensitivity in tests and items has been described as statistical variability between time points, between groups, and between groups and time points. However, this variability cannot per se be validly attributed to instruction (van der Linden, 1981). Because statistical variability is a necessary but insufficient requirement for instructional sensitivity, its validity for instruction cannot be considered without taking into account empirical measures of instruction. It has to be shown that instruction and not other sources such as student composition, drives the observed variability.

Educational research uses a wide variety of measures of instruction. Most of these can be classified into three types: (1) content matter (e.g., content coverage and content focus, often called “opportunity-to learn,” see Schmidt & Maier, 2009); (2) instructional approach (e.g., direct instruction or inquiry-based learning, teacher- or student-centered; for an international perspective, see Vieluf, Kaplan, Klieme, & Bayer, 2011); and (3) quality of how content matter,

teaching and learning activities are enacted (Raudenbush, 2008), which can be described, for example, as classroom management, supportive climate, and cognitive activation (Klieme, Pauli, & Reusser, 2009; for a similar approach, see Pianta & Hamre, 2009).

Given the availability of at least some empirical measures of instruction, statistical variation as an indicator of instructional sensitivity can be related to measures of the instruction students receive. For example, in a multilevel IRT framework (Adams, Wilson, & Wu, 1997; Kamata, 2001), using an explanatory IRT approach (De Boeck & Wilson, 2004), between-groups sensitivity indicators can be related to instruction by adding instructional measures as predictors for either group-specific test scores or group-specific item difficulties. Similarly, statistical indices of test or item sensitivity between measurement time points can be regressed on content and quality measures of instruction obtained at the different time points. For example, Klieme, Pauli, and Reusser (2009) showed that TIMSS-like mathematics items repeated with a one-year time lag are sensitive to different facets of instruction. This is also true for items covering specific mathematical content repeated with time lags of only some lessons. Generally, sufficient empirical evidence for instructional sensitivity is available only if item responses and test scores are related to instructional measures, i.e., variation in test scores and item difficulties must be explained by facets of instruction. That is, ideally, statistical modeling rules out other alternative explanations for time-variation or group-variation, such as resource allocations.

Conclusions and Discussion

This paper presented an IRT framework allowing to distinguish three perspectives underlying different approaches to instructional sensitivity: (1) between-time-points perspective related to how a test or an item differentiates between time points of measurement, (2) between-groups perspective related to how a test or an item differentiates between groups within a sample, and

(3) combined perspective related to how a test or an item differentiates between groups and time points, as captured by the concepts of global sensitivity and differential sensitivity. Generally, all perspectives are applicable to the evaluation of single items or item clusters (e.g., test scores). Also, evaluations of instructional sensitivity consistent with any of the perspectives can be implemented in settings with two or more groups and/or time points. In principle, indices rooted in any perspective can be related to instructional measures as evidence for a valid interpretation of instructional sensitivity indices. In summary, we provided a framework based on IRT for modeling instructional sensitivity that aims to integrate these perspectives and improve understanding of similarities and differences between common approaches to instructional sensitivity, even across Polikoff's (2010) categories.

A Multilevel and Multiperspective View of Instructional Sensitivity

Existing approaches thus far have appeared inconsistent in their evaluation of instructional sensitivity. Yet, as illustrated by our framework, each approach examines instructional sensitivity by relying on different variance components related to different levels of analysis (e.g., single items or tests) and different underlying perspectives. That is, approaches target different facets of instructional sensitivity and therefore do not necessarily have to be consistent. Thus, systematizing existing approaches based on their underlying perspectives helps us better understand differences and similarities in the evaluation of instructional sensitivity.

The question in which a test should be instructionally sensitivity is closely linked to the use and interpretation of the final test scores (Kane, 2013). For example, if test results are to be used to judge the effectiveness of various kinds of instruction groups of students within a sample have received, the instrument's capacity to detect differences between groups (i.e., the between-groups facet of the instrument's instructional sensitivity) should be investigated beforehand

(Glaser, 1963). Otherwise, if test results indicated no differences in the effectiveness of teaching, the interpretation "each kind of instruction appeared equally effective" might be flawed as it remains unclear whether the instrument was able to detect differences in instruction at all. Analogously, the between-time-points facet should be examined when studies focus on students' progress, ultimately leading to the differences-between-groups-and-time points perspective if both facets are of interest. With respect to the different sources of variance addressed in the evaluation of instructional sensitivity, a single statistic might not be sufficient to thoroughly describe the sensitivity of a test or a single item, because a single indicator cannot express all of the variability in item responses (Naumann et al., 2014). That is, differences between time points and differences between groups need to be expressed with appropriate items or test statistics. To date, empirical studies rarely have made their perspective on instructional sensitivity explicit. We believe this is a serious oversight, as on the one hand, results may diverge substantially when they are based on different approaches rooted in different perspectives and therefore based on different sources of variance (Li et al., 2012; Naumann et al., 2014), and on the other hand, it remains unclear whether the facet of instructional sensitivity under investigation has fitted the intended use and interpretation of test scores (Kane, 2013).

Technically, the level for modeling instructional sensitivity should not be chosen arbitrarily. By treating items as hierarchically nested within tests, all variance common to items is attributed to the test level, while the item level contains variance unique to single items. Consequently, within a set of homogeneously sensitive items, a single item may not be identified as sensitive (using DIF approaches, for example) unless instructional sensitivity is modeled at the test level. Conversely, potentially beneficial information on a single item's sensitivity might be overlooked if instructional sensitivity is engaged solely at the test level. Accordingly,

simultaneous analyses of item responses on multiple levels may enhance understanding of the instructional sensitivity of assessments and therefore may be especially helpful in test construction.

In summary, a thorough evaluation of an instrument's (instructional) sensitivity requires evaluation of instructional sensitivity on multiple levels of analysis and from multiple perspectives'. If the analysis takes into account only one perspective, judgment of instructional sensitivity may be somewhat incomplete. In contrast, analyses from multiple perspectives are less likely to incorrectly label tests and items as insensitive. For example, if only between-groups sensitivity is investigated, insensitivity would be diagnosed in a sample where all groups received the same amount and quality of instruction. In contrast, approaches taking into account differences between time points could still determine there is sensitivity with regard to students' progress. Additionally, more detailed information on the sensitivity of instruments becomes available when the differences-between-time-points and the differences-between-groups perspectives are combined in a common approach—that is, how the instrument under investigation is capable of distinguishing between groups treated differently and different stages of learning.

Two drawbacks compared to single-perspective investigations are that data requirements may be substantially higher, and statistical modeling becomes more complex. Data requirements for the investigation of instructional sensitivity from a between-groups-and-time-points perspective comprise a) longitudinal data, with a set of items administered within the same classroom at each time point of measurement when item level effects are of interest, and b) a reasonably large sample size on the classroom-level. For example, with small sample sizes, multilevel logistic regression estimates, for example, as in the LMLDIF model, might be biased

due to outliers (see Maas and Hox (2005) for a thorough discussion of sufficient sample sizes for multilevel modeling). Additionally, information on classroom-instruction, for example, obtained via video-observations (e.g., Klieme et al., 2009), is needed to relate test or item sensitivity to instructional sensitivity. In contrast, data requirements for single-perspective investigations are comparably modest since either longitudinal data or large sample sizes on the classroom-level are needed, and hence these approaches might be easier to implement. For example, PPDI may be calculated based on classical item difficulty *p* using standard software (e.g., SPSS), while the LMLDIF model requires Bayesian estimation (see Naumann et al., 2014, for details).

The previous considerations notwithstanding, we encourage the use of experimental designs in research on instructional sensitivity. Most studies on instructional sensitivity are re-analyses of existing instruments (e.g., D'Agostino et al., 2007). At best, studies on instructional sensitivity have been conducted in the context of interventions or other settings where a lot of information on instruction has been available. However, information on the characteristics of the employed instruments and items is often comparably superficial. Experimental variation of item and test characteristics is rare in studies on instructional sensitivity (e.g., see Wills, Li, & Ruiz-Primo, this issue, for an application of experimental design in instructional sensitivity research). Detailed information on item, test, and classroom characteristics is crucial to the ability of researchers to formulate hypotheses and to acquire a deeper understanding of the interplay between instruments, single items, and instruction. Accordingly, for a thorough and valid evaluation of existing instruments, and to build knowledge that facilitates the construction of new instruments, future research will need to rely on sophisticated experimental designs to study the instructional sensitivity of items and tests from multiple perspectives.

References

Adams, R. J., Wilson, M., & Wu, M. (1997). Multilevel item response models: An approach to errors in variables regression. *Journal of Educational and Behavioral Statistics*, *22*, 47–76.

Baker, E. L. (1994). Making performance assessment work: The road ahead. *Educational Leadership*, *51*, 58–62.

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

Clauser, B. E., Nungester, R. J., & Swaminathan, H. (1996). Improving the matching for DIF analysis by conditioning on both test score and an educational background variable. *Journal of Educational Measurement*, *33*, 453–464.

Cox, R. C., & Vargas, J. S. (1966). *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.

Creemers, B. P. M., & Kyriakides, L. (2008). *The dynamics of educational effectiveness: A contribution to policy, practice and theory in contemporary schools*. London and New York: Routledge.

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

De Boeck, P., & Wilson, M. (Eds.). (2004). *Explanatory item response models: A generalized linear and nonlinear approach*. New York: Springer.

Fox, J.-P. (2010). *Bayesian item response modeling: Theory and applications*. New York, NY: Springer.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist, 18*, 519–521.

Gelman, A. B., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Haladyna, T. M. (2004). *Developing and validating multiple-choice test items*. Mahwah, N.J: Lawrence Erlbaum Associates.

Hardy, I., Hertel, S., Kunter, M., Klieme, E., Warwas, J., Büttner, G., & Lühken, A. (2011). Adaptive Lerngelegenheiten in der Grundschule: Merkmale, methodisch-didaktische Schwerpunktsetzungen und erforderliche Lehrerkompetenzen [Adaptive learning environments in primary school]. *Zeitschrift für Pädagogik*, *57*, 819–833.

Hardy, I., Kleickmann, T., Koerber, S., Mayer, D., Möller, K., Pollmeier, J., Sodian, B., & Schwippert, K. (2010). Die Modellierung naturwissenschaftlicher Kompetenz im Grundschulalter [Modeling science competence in primary school]. In E. Klieme, D. Leutner, M. Kenk (Eds.), *Kompetenzmodellierung. Zwischenbilanz des DFG-Schwerpunktprogramms und Perspektiven des Forschungsansatzes*. *56*. Beiheft der Zeitschrift für Pädagogik, Weinheim u.a.: Beltz.

Hartig, J., Klieme, E., & Leutner, D. (Eds.) (2008). *Assessment of competencies in educational contexts*. Göttingen: Hogrefe & Huber.

Hartig, J., & Kühnbach, O. (2006). Schätzung von Veränderung mit Plausible Values in mehrdimensionalen Rasch-Modellen [Estimating change using plausible values in multidimensional Rasch models]. In A. Ittel & H. Merkens (Eds.), *Veränderungsmessung*

und Längsschnittstudien in der Erziehungswissenschaft (pp. 27–44). Wiesbaden: Verlag für Sozialwissenschaften.

Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning: Theory and practice*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Ing, M. (2008). Using instructional sensitivity and instructional opportunities to interpret students' mathematics performance. *Journal of Educational Research & Policy Studies*, *8*, 23–43.

Kamata, A. (2001). Item analysis by the hierarchical generalized linear model. *Journal of Educational Measurement*, *38*, 79–93.

Kane, M. T. (2013). Validating the interpretations and uses of test scores. *Journal of Educational Measurement, 50*, 1–73.

Kleickmann, T., Hardy, I., Möller, K., Pollmeier, J., Tröbst, S., & Beinbrech, C. (2010). Die Modellierung naturwissenschaftlicher Kompetenz im Grundschulalter: Theoretische Konzeption und Testkonstruktion [Modeling science competence of primary school children]. *Zeitschrift für Didaktik der Naturwissenschaften*, *16*, 263–281.

Klieme, E. & Baumert, J. (2001). Identifying national cultures of mathematics education: Analysis of cognitive demands and differential item functioning in TIMSS. European Journal of Psychology of Education, *16*, 383–400.

Klieme, E., Pauli, C., & Reusser, K. (2009). The Pythagoras study: Investigating effects of teaching and learning in Swiss and German mathematics classrooms. In T. Janík & T. Seidel (Eds.), *The power of video studies in investigating teaching and learning in the classroom* (pp. 137–160). Münster: Waxmann.

Li, M., Ruiz-Primo, M. A., & Wills, K. (2012). *Comparing methods to evaluate the instructional sensitivity of items*. Paper presented at the Annual Conference of the National Council on Measurement in Education, Vancouver.

Lee, G., Brennan, R. L., & Frisbie, D. A. (2000). Incorporating the testlet concept in test score analyses. *Educational Measurement: Issues and Practice*, *19*, 9–15.

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.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Maas, C. J. M., & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences, 1* (3), 86–92.

Martin, M. O., Mullis, I. V. S., & Foy, P. (with Olson, J. F., Erberber, E., Preuschoff, C., & Galia, J.). (2008). *TIMSS 2007 International Science Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Meulders, M., & Xie, Y. (2004). Person-by-item predictors. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized linear and nonlinear approach* (pp. 213–240). New York: Springer.

Muthén, B. O. (1989). Using item-specific instructional information in achievement modeling. *Psychometrika, 54*, 385–396.

Muthén, B. O., Kao, C.-F., & Burstein, L. (1991). Instructionally sensitive psychometrics: Application of a new IRT-based detection technique to mathematics achievement test items. *Journal of Educational Measurement*, *28*, 1–22.

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.

Niemi, D., Wang, J., Steinberg, D. H., Baker, E. L., & Wang, H. (2007). Instructional sensitivity of a complex language arts performance assessment. *Educational Assessment*, *12*, 215–237.

Pellegrino, J. W. (2002). Knowing what students know. *Issues in Science & Technology*, *19* (2), 48–52.

Pianta, R. C., & Hamre, B. K. (2009). Conceptualization, measurement, and improvement of classroom processes: Standardized observation can leverage capacity. *Educational Researcher, 38*, 109–119.

Polikoff, M. S. (2010). Instructional sensitivity as a psychometric property of assessments. *Educational Measurement: Issues and Practice*, *29* (4), 3–14.

Popham, J. W. (2007). Instructional insensitivity of tests: Accountability's dire drawback. *Phi Delta Kappan*, *89*, 146–155.

Raudenbush, S. W. (2008). The Brown legacy and the O'Connor challenge: Transforming schools in the images of children's potential. *Educational Researcher, 38*, 169–180.

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Thousand Oaks, CA: Sage Publications.

Reckase, M. (2009). *Multidimensional item response theory*. New York, London: Springer.

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

Rowan, B., Correnti, C., & Miller, R. J. (2002). What large-scale, survey research tells us about teacher effects on student achievement: Insights from the Prospects study of elementary schools. *Teachers College Record, 104* (8), 1525–1567.

Ruiz-Primo, M. A., Shavelson, R. J., Hamilton, L., & Klein, S. (2002). On the evaluation of systemic science education reform: Searching for instructional sensitivity. *Journal of Research in Science Teaching*, *39*, 369–393.

Schmidt, W. H., & Maier, A. (2009). Opportunity to Learn. In G. Sykes, B. Schneider, & D. N. Plank (Eds.), *Handbook of education policy research* (pp. 541–559). New York: Routledge.

van den Noortgate, W., De Boeck, P., & Meulders, M. (2003). Cross-classification multilevel logistic models in psychometrics. *Journal of Educational and Behavioral Statistics, 28*, 369–386.

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.

Vieluf, S., Kaplan, D., Klieme, E., & Bayer, S. (2012). *Teaching practices and pedagogical innovation: Evidence from TALIS*. Paris: OECD Publishing.

Vosniadou, S. (2007). The cognitive-situative divide and the problem of conceptual change. *Educational Psychologist*, 42, 55–66.

Zwick, R., & Ercikan, K. (1989). Analysis of differential item functioning in the NAEP history assessment. *Journal of Educational Measurement*, *26*, 55–66.

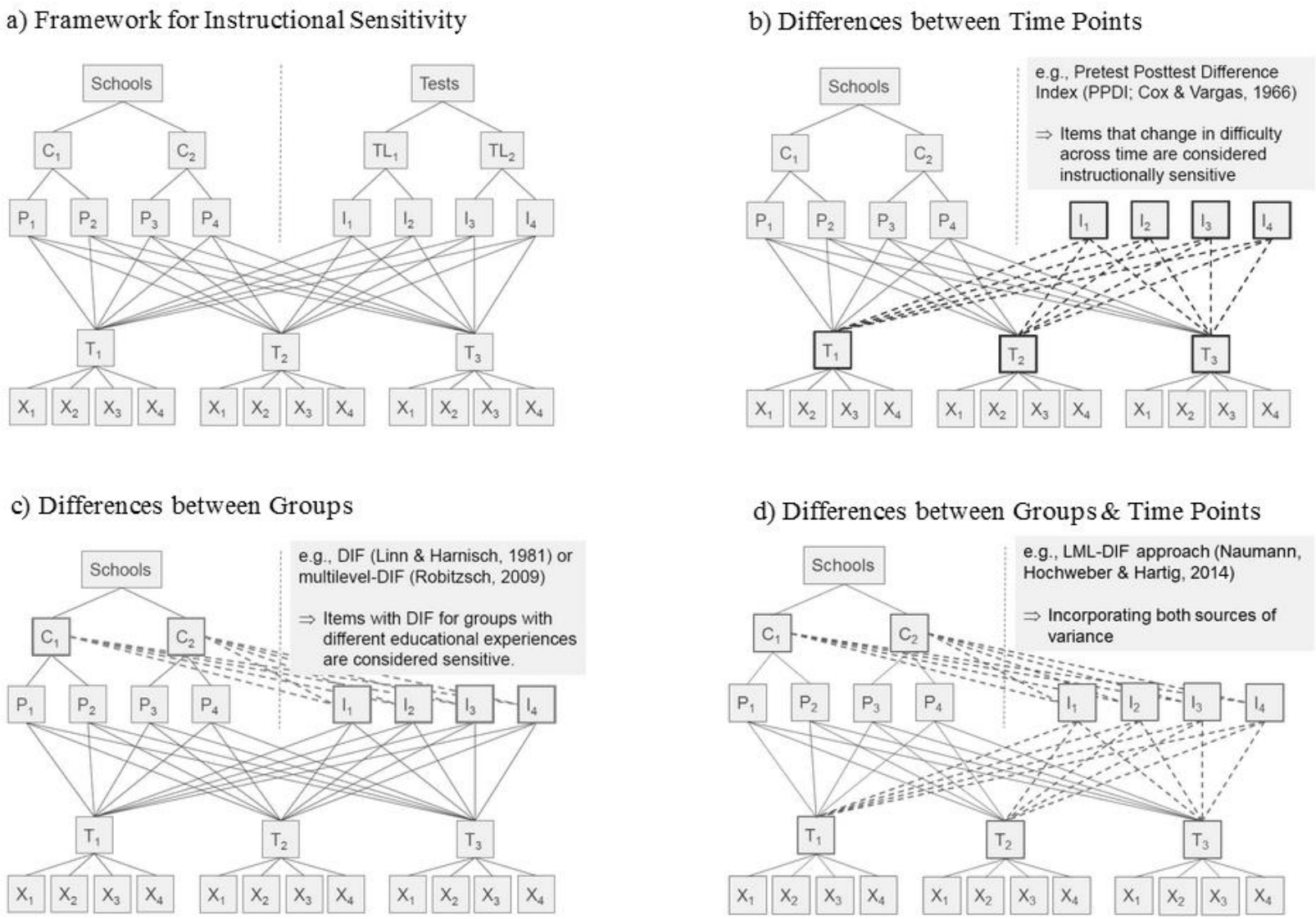

Figure 1. A hierarchical framework for instructional sensitivity with item responses (X) nested in time points (T), persons (P), classes (C) and schools on the person side and nested in items (I), testlets (TL) and tests on the item side.

Table 1

Relative Frequencies of Score Categories in the "Stone" Item at Pretest and Posttest

Score category	Frequencies	
	Pre n=986	Post n=991
Naïve conceptions	86.1 %	55.2 %
Explanations from everyday life	9.7 %	27 %
Scientific explanations	4.2 %	17.8 %

Table 2

Relative Frequencies of Score Categories in the "Stone" Item in Classes A and B at Posttest

	Frequencies	
Score category	Class A	Class B
Naïve conceptions	13.6 %	15 %
Explanations from everyday life	40.9 %	65 %
Scientific explanations	45.5 %	20 %

Note. $N_A = 22$, $N_B = 20$.

Table 3

Change in Score Categories of the "Stone"-Item in Classes A and B from Pre- to Posttest

	Change in relative frequencies (%)	
Score category	Class A	Class B
Naïve conceptions	−81.9	−58.7
Explanations from everyday life	+36.4	+43.9
Scientific explanations	+45.5	+14.7

Note. $N_A = 22$, $N_B = 20$.

Table 4

A 2 × 2 Typology of Instructional Sensitivity

Variance in pretest-posttest difference	Average pretest-posttest difference	
	Low	high
Low	Not sensitive	Global
High	Differential	Global and differential