Naumann, Alexander; Hartig, Johannes; Hochweber, Jan

Absolute and relative measures of instructional sensitivity

Journal of educational and behavioral statistics 42 (2017) 6, S. 678-705

Quellenangabe/ Reference:
Naumann, Alexander; Hartig, Johannes; Hochweber, Jan: Absolute and relative measures of instructional sensitivity - In: Journal of educational and behavioral statistics 42 (2017) 6, S. 678-705 - URN: urn:nbn:de:0111-pedocs-156029 - DOI: 10.25656/01:15602

https://nbn-resolving.org/urn:nbn:de:0111-pedocs-156029
https://doi.org/10.25656/01:15602

Article

Journal of Educational and Behavioral Statistics
2017, Vol. 42, No. 6, pp. 678–705
DOI: 10.3102/1076998617703649

Absolute and Relative Measures of Instructional Sensitivity

Alexander Naumann
Johannes Hartig
German Institute for International Educational Research (DIPF)

Jan Hochweber
University of Teacher Education St. Gallen (PHSG)

Valid inferences on teaching drawn from students' test scores require that tests are sensitive to the instruction students received in class. Accordingly, measures of the test items' instructional sensitivity provide empirical support for validity claims about inferences on instruction. In the present study, we first introduce the concepts of absolute and relative measures of instructional sensitivity. Absolute measures summarize a single item's total capacity of capturing effects of instruction, which is independent of the test's sensitivity. In contrast, relative measures summarize a single item's capacity of capturing effects of instruction relative to test sensitivity. Then, we propose a longitudinal multilevel item response theory model that allows estimating both types of measures depending on the identification constraints.

Keywords: *instructional sensitivity; multilevel IRT; differential item functioning*

Researchers as well as policymakers regularly rely on student performance data to draw inferences on schools, teachers, or teaching (Creemers & Kyriakides, 2008; Pellegrino, 2002). Yet valid inferences drawn from student test scores require that instruments are sensitive to the instruction that students have received in class (Popham, 2007; Popham & Ryan, 2012). Accordingly, measures of test items' instructional sensitivity may provide empirical support for validity claims about the inferences on instruction derived from student test scores.

Instructional sensitivity is defined as the psychometric property of a test or a single item to capture effects of instruction (Polikoff, 2010). Scores of instructionally sensitive tests are expected to increase with more or better teaching (Baker, 1994). Students who received different instruction should produce different responses to highly instructionally sensitive items (Ing, 2008). Fundamentally, instructional sensitivity relates to the observation of change in students' responses on items as a consequence of instruction (Burstein, 1989). If item responses do not change as a consequence of instruction, it may remain unclear

whether teaching was ineffective or the test was insensitive (Naumann, Hochweber, & Hartig, 2014). To test the hypothesis of whether an item is instructionally sensitive, various measures have been proposed (see Haladyna & Roid, 1981; Polikoff, 2010). Most commonly, these item sensitivity measures are based on item parameters, that is, item difficulty or discrimination (Haladyna, 2004).

According to Naumann, Hochweber, and Klieme (2016), each item sensitivity measure refers to one of the three perspectives on how to test the instructional sensitivity of items. From the first perspective, instructional sensitivity is conceived as change in item parameters between two time points of measurement, while from the second perspective instructional sensitivity is conceived as differences in item parameters between at least two groups (e.g., treatment and control groups or classes) within a sample. The third perspective is a combination of the two preceding ones, which allows deriving measures addressing two facets of item sensitivity: global and differential sensitivity. Global sensitivity refers to the extent to which item parameters change on average across time. Differential sensitivity refers to the variation of change in parameters across groups, indicating an item's capacity of detecting differences in group-specific learning. Overall, these perspectives provide an elaborate framework for the measurement of instructional sensitivity based on item statistics by highlighting the relevant sources of variance: variance between (a) time points, (b) groups, and (c) groups and time points. As item sensitivity measures rooted in different perspectives target different sources of variance, they do not necessarily provide consistent results (Naumann et al., 2014).

Yet the three perspectives are not sufficient for describing common characteristics and distinctions of instructional sensitivity measures. Actually, instructional sensitivity measures referring to the same perspective may address two essentially different hypotheses regarding item sensitivity: Some measures relate to the hypothesis of whether an item is sensitive at all, that is, *absolute sensitivity*, while others relate to the hypothesis of whether an item substantially deviates from the test's overall sensitivity, that is, *relative sensitivity*.

This additional distinction has important theoretical and practical implications for the evaluation of instructional sensitivity. For example, studies have shown that the most commonly applied approaches, the Pretest–Posttest Difference Index (PPDI; Cox & Vargas, 1966) and differential item functioning (DIF)-based methods (e.g., Linn & Harnisch, 1981; Robitzsch, 2009), are inconsistent in their judgment of item sensitivity (Li, Ruiz-Primo, & Wills, 2012; Naumann et al., 2014). One reason for this finding lies in the difference of the perspective taken on instructional sensitivity by these approaches (Naumann, Hochweber, & Klieme, 2016): While the PPDI focuses on change in item difficulties across time points, DIF approaches focus on differences in item difficulty between at least two groups of students (e.g., treatment groups or courses or classes) within a sample. Yet another reason is that the approaches differ in the way they measure

instructional sensitivity: While the PPDI is an absolute sensitivity measure, DIF approaches provide relative measures of item sensitivity.

Thus, in the present study, we aim to contribute to the measurement framework of instructional sensitivity by introducing the distinction between absolute and relative measures. Absolute and relative measures may be distinguished within each of the three perspectives on instructional sensitivity and provide unique and valuable information on item functioning in educational assessments when inferences on schools, teachers, or teaching are to be drawn. In the following, we will first elaborate on the distinction of absolute and relative measures. We will point out how absolute and relative measures relate to test sensitivity and current approaches to the instructional sensitivity of items. Second, we will provide a model-based approach that allows testing the hypothesis of whether items are absolutely and/or relatively sensitive within a more general item response theory (IRT) framework. For illustration purposes, we apply our approach to simulated and empirical item response data. Finally, we will discuss implications for the measurement of instructional sensitivity, test development, and test score interpretation.

Extending the Measurement Framework of Instructional Sensitivity

Figure 1 depicts an extended measurement framework. The extended measurement framework comprises the three perspectives as well as the two sensitivity facets—global and differential sensitivity—that can be distinguished within the groups and time points perspective following Naumann and colleagues (2016). In addition, we draw the distinction between absolute and relative item sensitivity measures within each perspective, making explicit that two different hypotheses regarding item sensitivity may be tested via absolute and relative measures.

Absolute measures address the hypothesis of whether a single item is sensitive to instruction. In principle, absolute measures summarize a single item's total capacity of capturing potential effects of instruction in terms of variation in item parameters across time, groups, or both. Hence, absolute measures are expected to approach zero the less sensitive an item is and depart from zero the higher the item's sensitivity to instruction is.

In contrast, relative measures address the hypothesis of whether a single item's sensitivity substantially deviates from test sensitivity. Test sensitivity is a concept that so far has only been implicitly used in the measurement of instructional sensitivity. In consistence with the predominant statistical notion of item sensitivity (see Haladyna & Roid, 1981; Haladyna, 2004; Polikoff, 2010), test sensitivity may be defined as the overall (i.e., unconditional) variation of test scores across either time points, groups, or both (cf. Naumann et al., 2016). Test sensitivity then is a prerequisite for what is commonly conceived as the instructional sensitivity of a test, which typically refers to the proportion of

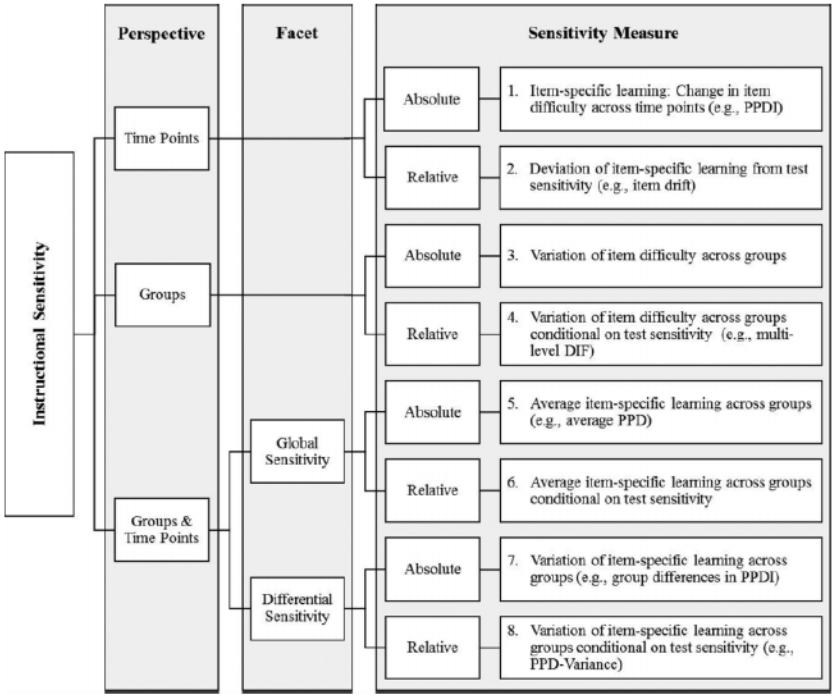

FIGURE 1. *Extended measurement framework of instructional sensitivity comprising the three perspectives, the two facets, and the eight absolute and relative sensitivity measures.*

variance in test scores explained by school, teacher, or teaching characteristics (e.g., D'Agostino, Welsh, & Corson, 2007; Grossman, Cohen, Ronfeldt, & Brown, 2014; Ing, 2008). Generally, test sensitivity captures the degree of item sensitivity that is common to all the items within a test. Technically speaking, the stronger the item sensitivity correlates across all test items, the higher the test sensitivity. Accordingly, relative measures express the degree to which a single item's sensitivity differs from test sensitivity. More precisely, relative measures are expected to approach zero the more an item's sensitivity is in consistence with test sensitivity and to be nonzero if the item's sensitivity deviates from test sensitivity.

In general, whether a specific item sensitivity measure is absolute or relative depends on whether or not the underlying measurement model comprises one or more parameters capturing test sensitivity. Absolute measures of sensitivity are unconditional on test sensitivity while relative measures are conditional on test sensitivity. That is, from each of the three perspectives, measures are obtainable in two ways, either independently of (i.e., absolute) or depending on (i.e., relative) test sensitivity. As a result, there are eight different ways of measuring an

item's sensitivity in total: four ways based on absolute sensitivity measures and four ways based on relative sensitivity measures (see Figure 1, last column). However, not all of these ways have been applied so far in practice.

Absolute and Relative Measures From the Time Points Perspective

In practice, Cox and Vargas's PPDI (1966) is the most prominent approach to measuring item sensitivity when focusing on differences in item parameters between time points of measurement. It is conceptually easy to understand and, provided that longitudinal data are available, technically straightforward to implement. The drawbacks are that PPDI does not account for differences between learning groups, although it is reasonable to assume that content and quality of teaching may vary between classes. Also, separation of instruction effects from maturation is impossible if there is no untreated control group (Polikoff, 2010).

Conceptually, the PPDI is conceived as the difference in difficulty of a single item for instructed and uninstructed students (Polikoff, 2010). Technically, the PPDI is calculated as the change in a single item's difficulty p before and after instruction:

$$\text{PPDI} = p_{\text{post}} - p_{\text{pre}}. \tag{1}$$

The bigger the change in item difficulty p is, the potentially higher the item's instructional sensitivity. For example, an item with difficulties $p_{\text{post}} = 0.6$ and $p_{\text{pre}} = 0.4$ results in a PPDI $= 0.6 - 0.4 = 0.2$. That is, the item became easier over time and accordingly may be considered as sensitive from a between time points perspective following the PPDI. Essentially, this is equivalent to the effect of item-specific learning. As the extent of an item's PPDI solely depends on the proportions of students who get that very item correct at each time point of measurement, the PPDI is an absolute sensitivity measure (Figure 1, Sensitivity Measure 1).

Studies taking on a time points perspective have also regularly applied relative item sensitivity measures. Actually, relative item sensitivity between time points is better known as item parameter drift due to the teaching students were exposed to and is commonly investigated using methods of DIF detection (e.g., DeMars, 2004; French, Finch, Randel, Hand, & Gotch, 2016). Methods of DIF detection calculate item sensitivity measures conditional on test sensitivity. From the time points perspective, test sensitivity may be conceived as the average change in difficulty between time points of measurement across all items, which is commonly conceived as the effect of learning on the entire test. Relative item sensitivity then relates to the extent that item-specific learning deviates from test learning. If a single item changes in difficulty relatively lower or higher than the test, the item is considered as being exposed to parameter drift or as being sensitive, respectively (Figure 1, Sensitivity Measure 2). Still, relative sensitivity

measures differing from 0 indicate a violation of measurement invariance assumptions across time points (e.g., Meade, Lautenschlager, & Hecht, 2005).

To illustrate the principle and the relation of such relative measures to absolute measures, we exemplarily formulate a relative version of the PPDI that accounts for test sensitivity. This relative PPDI is conceptually identical to item difficulty drift in longitudinal settings and is calculated as the change in a single item i's difficulty p before and after instruction centered on the average change in the difficulty of the test:

$$\mathrm{PPDI}_{\mathrm{rel}} = p_{i,\mathrm{post}} - p_{i,\mathrm{pre}} - (\bar{p}_{.,\mathrm{post}} - \bar{p}_{.,\mathrm{pre}}), \tag{2}$$

where $\bar{p}_{.,\mathrm{post}}$ and $\bar{p}_{.,\mathrm{pre}}$ are the average item difficulties at posttest and pretest, respectively. In contrast to Cox and Vargas's original PPDI, the relative version not only utilizes the responses to a single item but the response data of all test items and is nonzero when item sensitivity deviates from test sensitivity and 0 if not.

Remember the PPDI example above. Suppose the average item difficulties in the test are $\bar{p}_{.,\mathrm{post}} = 0.5$ and $\bar{p}_{.,\mathrm{pre}} = 0.3$, respectively. Then, the relative PPDI of the example item is $\mathrm{PPDI}_{\mathrm{rel}} = 0.2 - (0.5 - 0.3) = 0$. That is, although the item has become easier over time, its change in difficulty does not deviate from the learning on the test. Thus, the item is insensitive from a between time points perspective following the relative sensitivity measure.

Absolute and Relative Measures From the Groups Perspective

Absolute sensitivity from the groups perspective refers to the item-wise variation of (unconditional) group-specific item parameter estimates. The higher item parameter variation across learning groups, the higher an item's sensitivity. Yet, as to our knowledge, such a measure has not been applied in practice so far (Figure 1, Sensitivity Measure 3).

In examinations focusing on differences in item parameters between learning groups, item sensitivity has traditionally been measured in terms of uniform DIF (e.g., Clauser, Nungester, & Swaminathan, 1996). As a result, and similar to longitudinal settings, instructional sensitivity has on the one hand been perceived as a violation of measurement invariance assumptions impairing test fairness if not all students within a sample have received comparable instruction (Geisinger & McCormick, 2010). On the other hand, this "instructional bias" (Linn & Harnisch, 1981, p. 117) between groups of students has been regarded as beneficial when drawing inferences on teaching (Linn & Harnisch, 1981; Naumann et al., 2014).

Conceptually, DIF approaches to instructional sensitivity from the groups perspective provide relative measures of item sensitivity due to the conditioning of item parameters on test sensitivity. Test sensitivity from the groups perspective refers to the variation of group-specific ability parameters, which basically

depends on the covariance of group-specific item difficulty estimates unconditional on ability. That is, if group-specific item difficulty correlates across items, then the between group variance in test scores (i.e., test sensitivity) becomes larger.

DIF approaches from the groups perspective focus on cross-sectional data and may become computationally rather demanding when accounting for multilevel structures (multilevel DIF; Meulders & Xie, 2004). Recent multilevel DIF approaches utilize classroom membership as a proxy for the manifold sources of differences in instruction that students may have received (Robitzsch, 2009; see also Naumann et al., 2014). Robitzsch's multilevel DIF approach models the probability of a correct response of person i in class c on item k as follows:

$$\text{logit}(P[X_{cik} = 1]) = \theta_c + \theta_{ci} - \beta_{ck}, \tag{3}$$

where θ_c is the average ability of class c, and θ_{ci} is the individual deviation in ability from the respective class mean. The item parameter β_{ck} is the classroom-specific difficulty of item k. All parameters are assumed to be normally distributed:

$$\begin{aligned} \theta_c &\sim N(0, \lambda^2), \\ \theta_{ci} &\sim N(0, \sigma^2), \\ \beta_{ck} &\sim N(\beta_k, \upsilon_k^2). \end{aligned} \tag{4}$$

Robitzsch suggests using the standard deviation υ_k of the classroom-specific item parameter distribution as an item sensitivity measure: The more an item's difficulty varies across classes, the potentially higher its instructional sensitivity.

As this variation is expressed conditional on the classroom ability parameters θ_c, standard deviation υ_k represents a multilevel DIF effect, that is, the extent of item difficulty variation between groups after variation in overall classroom ability has been taken into account. Consequently, the magnitude of (multilevel) DIF can be regarded as a *relative* measure of item sensitivity, and the variation of the classroom ability parameters θ_c may be conceived as a measure of test sensitivity from the groups perspective.

Absolute and Relative Measures From the Groups and Time Points Perspective

Recently, Naumann et al. (2014) combined the PPDI and Robitzsch's multilevel DIF model in the LMLDIF approach to instructional sensitivity. The advantage of the LMLDIF approach is that it integrates both perspectives and provides two measures for item sensitivity, one dedicated to global sensitivity and one dedicated to differential sensitivity. Information on both global and differential sensitivity allow for a more complete judgment of item sensitivity, which may be partially incomplete or even misleading if one facet of sensitivity is neglected (Naumann et al., 2014).

As a combination of PPDI and multilevel DIF, the LMLDIF approach requires longitudinal data from students within the same set of classes at (at least) two time points of measurement. Similar to the multilevel DIF approach, the LMLDIF approach assumes that meaningful differences in the instruction that students have received are due to their classroom membership. Additionally, the average classroom ability is assumed to be equal, namely 0, across time. This assumption, on the one hand, serves as an identification constraint and, on the other hand, ensures that all growth across time is reflected in the item difficulty parameters. Accordingly, the item difficulties are allowed to vary across classes and time points. Following the LMLDIF approach, the probability of a correct response of person i in class c on item k at time point t is given by

$$\text{logit}(P[X_{tcik} = 1]) = \theta_{tc} + \theta_{tci} - \beta_{tck}, \tag{5}$$

where θ_{tc} is the classroom-level ability component of class c at time point t, θ_{tci} is the time point-specific individual ability component of person i, and β_{tck} is the time point and classroom-specific difficulty of item k. The desired item sensitivity measures are calculated based on the β_{tck} estimates for two time points $t = 1$ and $t = 2$ in terms of classroom-specific pretest–posttest differences (PPD):

$$\Delta\beta_{ck} = \beta_{2ck} - \beta_{1ck}. \tag{6}$$

The classroom-specific PPDs $\Delta\beta_{ck}$ are treated as normally distributed. The mean of $\Delta\beta_{ck}$ across classes, the so-called average PPD, serves as a measure for an item's global sensitivity, and the variance of $\Delta\beta_{ck}$ across classes, the PPD variance, serves as a measure for an item's differential sensitivity. Like the PPDI, the average PPD does not depend on test-level (global) sensitivity. Since the average classroom ability is fixed to 0 at both time points, all learning progress on the items across groups is reflected in the mean of $\Delta\beta_{ck}$. Hence, the average PPD represents an absolute measure of global sensitivity. Nevertheless, the PPD variance captures differential sensitivity conditional on the variation of classroom-level ability. Hence, its magnitude is relative to the test-level (differential) sensitivity.

Despite its conceptual advantages compared to the singular application of PPDI or multilevel DIF, the LMLDIF approach has three major drawbacks. First, the change in item difficulties, and thus the foundation for the item sensitivity measures, is not part of the probability model itself. Instead, the quantities of interest are calculated based on the time point and classroom-specific item parameters. Consequently, whether global or differential sensitivity measures are statistically meaningful can only be evaluated indirectly and not by imposing constraints on the parameters of interest, that is, the mean and variance of $\Delta\beta_{ck}$ (cf. Naumann et al., 2014). Second, correlations of initial status and change in parameter values remain unconsidered. Finally, although the model integrates PPDI and DIF approaches, it does not generalize to a broader view on items' (instructional) sensitivity. For example, there is no explicit consideration and

convenient way of switching between absolute and relative measures for the different facets of instructional sensitivity.

Thus, in the following, we will propose a more general and straightforward model-based approach to measuring item sensitivity within an IRT framework. Our model aims at the evaluation of item sensitivity from a groups and time points perspective, yet reduces to the groups or to the time points perspectives if only one group or one time point is considered. When both longitudinal and multigroup data are available, our model provides absolute and/or relative measures for global and differential sensitivity, respectively. In contrast to the LMLDIF approach, our model allows users to switch between absolute and relative sensitivity measures by simply altering the identification constraints imposed on the model parameters. That way, the model basically allows estimating all types of sensitivity measures depicted in Figure 1.

Modeling Approach

We start by advancing the LMLDIF approach to a more general longitudinal multilevel IRT (LMLIRT) model that is not necessarily restricted to DIF and directly accounts for all parameters of interest. Similar to the multilevel DIF and LMLDIF approaches, we build the LMLIRT model under the assumption that meaningful differences in instruction students have received are tied to their classroom membership. Thus, item parameters are allowed to vary across time points and across classes. In contrast to the LMLDIF approach, we model initial status and classroom-specific change in item difficulty directly in a generalized linear mixed-model framework using person, item, and person-by-item covariates (Rijmen, Tuerlinckx, De Boeck, & Kuppens, 2003; van den Noortgate, De Boeck, & Meulders, 2003), allowing for correlations between initial status and change.

Following our approach, the probability of a correct response of individual i in class c on item k at time point t $(1, \ldots, T)$ is given by

$$\text{logit}\left(P[\mathrm{X}_{tcik} = 1]\right) = \sum_{u=1}^{T} q_{tu}\left(\theta_{tc} + \theta_{tci} - \beta_{tck}\right), \tag{7}$$

where q_{tu} is element of a predefined $T \times T$ lower triangular matrix Q of ones ($q_{tu} = 1$ when $u \leq t$ and 0 otherwise), ensuring that subsequent time points do not contribute to response probabilities at earlier ones:

$$\mathbf{Q} = \begin{bmatrix} 1 & 0 \cdots & 0 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix}. \tag{8}$$

Accordingly, at time point $t = 1$, θ_{tc} denotes the initial average ability of class c and θ_{tci} is the initial individual deviation in ability of person i from the respective

classroom ability component. For each following time point $t > 1$, θ_{tc} and θ_{tci} denote the change in classroom and individual ability components from the preceding time point $t - 1$. Parameters θ_{tc} and θ_{tci} are assumed to be mutually independent and multivariate normally distributed with time point–specific mean θ_t and covariance matrices $\mathbf{\Lambda}$ and $\mathbf{\Sigma}$, respectively:

$$\begin{aligned} \theta_{tc} &\sim MN(\theta_t, \mathbf{\Lambda}), \\ \theta_{tci} &\sim MN(0, \mathbf{\Sigma}). \end{aligned} \tag{9}$$

Analogously, β_{tck} is the initial classroom-specific difficulty of item k at time point $t = 1$ and the item's classroom-specific change in difficulty for each time point $t > 1$. Similar to the classroom and individual ability parameters, classroom-specific item parameters β_{tck} are assumed to be multivariate normally distributed with time point and item-specific mean vector β_{tk} and item-specific covariance matrix $\mathbf{\Phi}_k$:

$$\beta_{tck} \sim MN(\beta_{tk}, \mathbf{\Phi}_k), \tag{10}$$

where β_{tk} denotes the average initial item difficulty across classes of item k at time point $t = 1$ and the item's average change in difficulty parameters for each time point $t > 1$. The diagonal elements of Φ_k, the variance parameters ϕ^2_{tk}, represent a single item k's variation of initial difficulty parameters across classes at time point $t = 1$ and the variation of the item's change in difficulty parameters across classes for each time point $t > 1$. That is, in contrast to the LMLDIF approach, change in and variation of classroom-specific item parameters are not calculated post hoc but estimated directly within the LMLIRT model.

In consequence, the distribution of the classroom-specific item parameters β_{tck} directly provides information about the global and differential sensitivity of each item. While the components of the mean vector β_{tk} relate to global sensitivity, the variance components ϕ^2_{tk} relate to differential sensitivity. If β_{tk} is nonzero for a time point $t > 1$, the average change in item difficulty is either negative or positive across classes, and thus the item can be seen as globally sensitive within this time span. Accordingly, the higher ϕ^2_{tk} is, the higher item k's differential sensitivity within this period.

Combining the information on global and differential sensitivity allows for judging a single item's sensitivity based on the 2×2 typology presented in the LMLDIF approach (see Naumann et al., 2014), yet additionally, the LMLIRT model allows for testing the two hypotheses related to absolute and relative item sensitivity directly. As indicated above, the LMLIRT model in its general form is unidentified. As any other IRT model, the LMLIRT model can be identified by imposing constraints on the item difficulty or ability parameters. Depending on the identification constraints chosen, the LMLIRT model provides absolute or relative measures of item sensitivity. In the following, we will describe how to identify and thus how to obtain absolute and relative measures of global and differential sensitivity in more detail.

TABLE 1.
Constraints on the LMLIRT Model for Obtaining Absolute and Relative Measures of Sensitivity

	Global Sensitivity	
Differential Sensitivity	Absolute	Relative
Absolute	$\theta_t = 0,\ \mathbf{\Lambda} = 0$	$\beta_{tK} = -\sum_{k=1}^{K-1} \beta_{tk}, \mathbf{\Lambda} = 0$
Relative	$\theta_t = 0$	$\beta_{tK} = -\sum_{k=1}^{K-1} \beta_{tk}$

Note. θ_t = mean classroom-level ability at time point t; $\mathbf{\Lambda}$ = person-side classroom-level covariance matrix; β_{tk} = average difficulty or change in difficulty of item k at time point t.

Absolute and Relative Measures of Instructional Sensitivity in the LMLIRT Model

There are several ways to identify the LMLIRT model. An overview of the constraints that are relevant for obtaining absolute and relative measures of global and differential sensitivity is available in Table 1. Some of these constraints might lead to assumptions that are rather unrealistic for models that are commonly used for scaling in education research. However, these constraints are a necessary requirement for determining the instructional sensitivity of items (Naumann et al., 2014).

Absolute sensitivity. To determine a specific item's absolute sensitivity, the item's sensitivity measure may not depend on the global or differential sensitivity of the test. That is, in case of the LMLIRT model, the global and differential sensitivity measures have to be unconditional on classroom ability and its variation, respectively. With respect to global sensitivity, the model is identified and provides absolute measures when fixing all components θ_t of the mean vector of the classroom ability parameters to 0. This constraint is, in principle, equivalent to the procedures in the PPDI or the LMLDIF approach. In consequence, the average ability growth in the sample is entirely reflected in the item parameters. Analogously, obtaining absolute measures of differential sensitivity requires constraining $\mathbf{\Lambda}$, that is, the variance components λ_t^2 of the classroom ability distribution and their covariances, to 0. Taken together, having the LMLIRT model identified and providing solely absolute measures of global and differential sensitivity simply requires fixing all classroom ability parameters θ_{tc} to 0.

Relative sensitivity. To determine a single item's relative sensitivity, the item's sensitivity has to be estimated conditional on the classroom ability estimates. Thus, in contrast to estimating absolute measures, determining an item's relative

sensitivity requires unconstrained classroom ability parameters. Accordingly, instead of fixing classroom ability parameters, item difficulty parameters have to be constrained for identification purposes. With respect to global and differential sensitivity, the LMLIRT model provides relative measures when θ_{tc} is estimated freely and the item difficulty parameter β_{tK} for the last item K in the test is constrained such that the average difficulty of the test equals 0 at each time point:

$$\beta_{tK} = -\sum_{k=1}^{K-1} \beta_{tk}. \tag{11}$$

Alternatively, in the case of random items (De Boeck, 2008), the time point–specific mean of the item difficulty parameter β_{tk} distribution, β_t, can be fixed to 0. In either scenario, the change that is common to all items will be reflected in θ_t for each time point $t > 1$. Then, the parameters β_{tk} are measures of the items' relative global sensitivity. Similarly, the contribution to classroom-level differences that is common to all items will be reflected in λ_t^2. Therefore, the magnitude of the item-side variance components ϕ_{tk}^2, which capture the remaining item-specific variation in item difficulty between classrooms, can be conceived as the items' relative differential sensitivity.

Application to Data

For demonstration purposes, we applied the LMLIRT model to two data sets. First, we used simulated data to illustrate key features of absolute and relative measures. In the literature, the magnitude of sensitivity measures is commonly used as an indicator of an item's degree of instructional sensitivity, regardless of whether the measures are absolute or relative. Yet high absolute sensitivity does not necessarily translate to high relative sensitivity and vice versa. Suppose we assemble a test from items with a varying degree of absolute sensitivity, then the magnitude of relative sensitivity will be influenced by the degree items' absolute sensitivity correlates across the entire test. If absolute sensitivity is highly concordant across test items, the magnitude of relative measures should be low, as none of the items deviates meaningfully from test sensitivity. In contrast, relative sensitivity may be high for items deviating meaningfully from test sensitivity, regardless of their absolute sensitivity. We illustrate this relationship for measures of differential sensitivity using three extreme settings, one setting where item sensitivity is highly concordant, one where item sensitivity is uncorrelated, and one where items are absolute differentially insensitive. For measures of global sensitivity, the relationship is more straightforward as can be seen from the PPDI example provided above.

Afterward, we conduct an exemplary analysis of the absolute and relative global and differential sensitivity of empirical data using the LMLIRT model.

Applying the LMLIRT to empirical data, we try to provide empirical evidence supporting the claim that our analyses are adequate for real applications. The focus of the analyses lies on the intersection of absolute and relative measures with global and differential sensitivity. We expect that absolute and relative measures coincide (or differ) as a function of the concordance of items' absolute sensitivity measures. That is, if an item's absolute sensitivity is high and in accordance with the common change in classroom-specific difficulty estimates, we expect the item's relative sensitivity measures to be low. If the item's change in classroom-specific difficulty estimates is unrelated to the common change, we expect the item's relative sensitivity measures to coincide with the absolute measures.

All analyses were carried out in a Bayesian framework using R (R Development Core Team, 2008), JAGS 4.1.0 (Plummer, 2003), and the runjags package version number 2.0.4-2 (Denwood, in press). The estimation method was Markov chain Monte Carlo (MCMC). To estimate the LMLIRT model, we chose Wishart distributions with $T + 1$ degrees of freedom and scale matrix set to identity as priors for the inverse of the covariance matrices $\boldsymbol{\Sigma}$, $\boldsymbol{\Lambda}$, and $\boldsymbol{\Phi}_t$, resulting in vague priors for the matrices' off-diagonal elements (Gelman et al., 2013). As recommended by Gelman and Hill (2006), we assumed flat normal distributions with mean 0 and variance 10,000 as priors for the means of the classroom-level ability distributions and highest level of the item difficulty distributions. Initial values were randomly drawn from the prior distributions by JAGS.

For each of the analyses, we ran four Markov chains with 30,000 iterations each and discarded the first 5,000 iterations as burn-in. To reduce autocorrelation, we only used every 10th iteration for the analyses. Convergence was checked via visual inspection of trace plots and the Gelman–Rubin $\hat{R}$ statistic. Below, we summarized parameter posterior distributions in terms of point estimates (i.e., maximum a posteriori estimates) followed by the corresponding Bayesian credible intervals (BCI) in square brackets.

Simulation Data Example

Data generation. One simulated data set for two time points of measurement (pretest–posttest) was generated in R based on the LMLIRT model. The simulation was set up as follows: First, individual ability parameters (θ_{tkv}) for 100 classes with 24 students each were randomly drawn from univariate standard normal distributions. That is, correlation of initial status and growth of individual ability was fixed to 0. Each classroom's average ability at the pretest as well as the change in classroom ability (θ_{tk}) across time were constrained to 0. Second, we prepared three sets of items (Sets A–C) with 20 items each. The initial (pretest) item difficulties (β_{1i}) were chosen to be equidistant, ranging from -1.5 to 1.5 logits around the mean of the ability distribution. That is, initially,

there was no variation of item difficulties across classes. We assumed the average change in item difficulty across classes (β_{2i}), that is, the indicator of global sensitivity, to be equal and 0 for all items.

The item Sets A, B, and C varied in their degree of absolute differential sensitivity (ϕ^2_{2i}), that is, in the variation of change in the item difficulties across classrooms, and in the correlation among the classroom-specific change in item difficulty parameters (β_{2ki}). While the items within the Sets A and B were differentially sensitive ($\phi^2_{2i} = 1$), the items within Set C remained equally difficult in each classroom ($\phi^2_{2i} = 0$). Additionally, the items of Set A exhibited high correlations of β_{2ki} with $r = .95$, meaning that classes with high improvement on 1 item also highly improved on the other items, while these correlations were fixed to 0 in Set B. In Set C, classroom-specific change parameters were also uncorrelated due to the lack of variation across classes ($\phi^2_{2i} = 0$). Finally, we generated item responses based on the person and item parameters using the rbinom function in R.

To demonstrate the distinctions of absolute and relative measures of instructional sensitivity, we applied the LMLIRT model to four different sets of items put together from the aforementioned item Sets A through C: (a) the 20 items of Set A, (b) the 20 items of Set B, (c) 15 randomly drawn items from Set A combined with 5 randomly drawn items from Set B, and (d) 15 randomly drawn items from Set A combined with 5 randomly drawn items from Set C. These newly combined item sets were each analyzed twice. First, the covariance matrix $\mathbf{\Lambda}$ was fixed to 0 to estimate measures of absolute differential sensitivity, and second, all elements of $\mathbf{\Lambda}$ were estimated freely to obtain measures of relative differential sensitivity, corresponding with the constraints depicted in Table 1.

Results. Figure 2 displays the simulation results. Applying the LMLIRT model to the highly absolute differentially sensitive items of Set A, relative measures appear low as sensitivity is concordant across items (Figure 2a). In contrast, the uncorrelated items of Set B are highly differentially sensitive according to both sensitivity measures (Figure 2b). Mixing items of Sets A and B yields relative measures of differential sensitivity higher for those items whose sensitivity is unrelated to the rest of the items (i.e., items from Set B; Figure 2c). This is in line with expectations, as these items' differential sensitivity is not captured by the corresponding test sensitivity parameter (λ^2_t), which rather captures the common change in difficulty across the items from Set A. Finally, mixing items of Set A and invariant items from Set C, the invariant items appear relatively differential sensitive despite being absolutely insensitive (Figure 2d). The reason corresponds to the differences in relative sensitivity observed in the aforementioned condition. The invariant items from Set C deviate strongly in their differential sensitivity from the variant items of Set A, which additionally are highly correlated. In consequence, τ^2_t captures the common change in difficulty across the

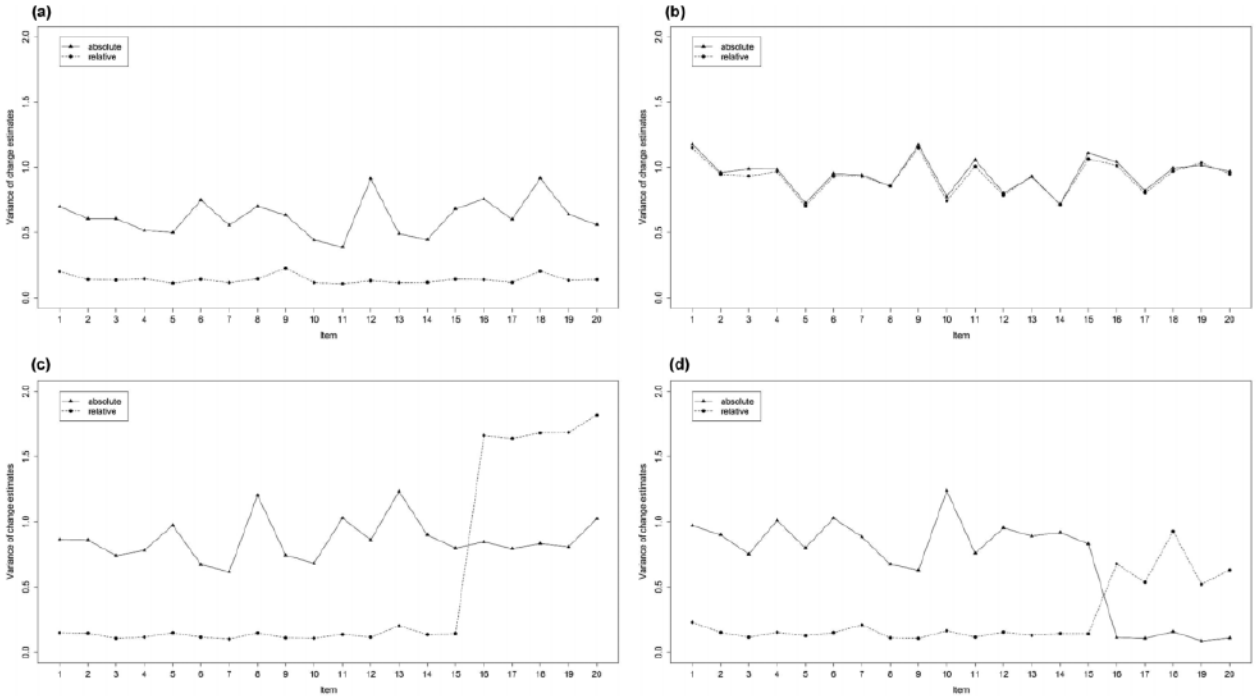

FIGURE 2. *Absolute and relative differential sensitivity of simulated data for (a) 20 items with high correlation between classroom-specific estimates (Set A), (b) 20 items with low correlation between classroom-specific estimates (Set B), (c) a random draw of 15 items from Set A and 5 items from Set B, and (d) the combination of a random draw from Set A and 5 items with low absolute differential sensitivity (Set C).*

items, resulting in a low relative differential sensitivity measure for items from Set A, while the items from Set C deviate very strongly from the common change, that is, test sensitivity, as expressed by high relative differential sensitivity measures.

Real Data Example

For the exemplary analysis of empirical item responses, we used data from the German DESI large-scale assessment study (DESI-Konsortium, 2008). DESI investigated the development of students' language competencies and language instruction in Grade Level 9 of German secondary schools during the school year 2003–2004. The target population was all German ninth graders attending a regular secondary school type (i.e., all school types except special needs schools). To reflect effects of instruction, the tests in DESI were aligned with the curricula of German as native and English as a foreign language in the ninth grade. Data were collected from representative samples from all 16 German federal states. For demonstration purposes, we focused on a subsample from the German lower secondary schools comprising 3,613 students in 135 classes.

Method. We exemplarily applied the LMLIRT model to a language awareness test comprising 34 items, administered at the beginning and the end of the school year (Eichler, 2007). The items were administered in a multimatrix testlet design with anchoring. On average, eight students per class received the same item at one time point. No student received the same item twice. With the exception of 1 item, all items were scored dichotomously as either correct or incorrect. For the analysis, score categories of the polytomous item were recoded into two dichotomous step indicators, defining the respective step functions in a cumulative approach (Agresti, 1990).

Before applying the LMLIRT model, we evaluated item fit to a two-level one-parameter logistic model via infit statistics. Infit statistics were calculated following Wright and Masters's (1990) study. Fit was acceptable for all items including the dichotomous step indicators with weighted mean square values ranging from 0.88 (0.84, 0.93) to 1.11 (1.07, 1.14) at pretest and from 0.87 (0.82, 0.93) to 1.15 (1.03, 1.30) at posttest (cf. Wright & Linacre, 1994). Latent intraclass correlation of ability parameters was .23 (.18, .29) at pretest and .22 (.18, .28) at posttest.

In addition to the LMLIRT analyses, we checked whether the extent of global and differential sensitivity found in DESI data was statistically meaningful. That is, each item's absolute and relative measures were checked for statistical importance. Absolute and relative global sensitivity was judged based on the 95% BCI corresponding to the sensitivity measure β_{2i}, considering an item as insensitive if the interval comprised 0 and sensitive if it did not. We checked items' absolute and relative differential sensitivity, that is, the variance components ϕ^2_{2i},

following a procedure by Verhagen and colleagues (Verhagen & Fox, 2013; Verhagen, Levy, Millsap, & Fox, 2015). Verhagen and colleagues utilize the Savage–Dickey density ratio to compute Bayes's factors for the null hypothesis of invariance. As neither the prior nor the posterior distribution of ϕ^2_{2i} may comprise 0, the null hypothesis is defined by an "about equality" constraint. That is, the procedure provides Bayes's factors based on the cumulative probabilities under the prior and the posterior distributions below an a priori set threshold δ:

$$\mathrm{BF}_{01} = \frac{p(\phi^2_{2i} < \delta | H_1, X)}{p(\phi^2_{2i} < \delta | H_1)}, \tag{12}$$

with observed responses X. Then, a Bayes's factor larger than 3 is considered as substantial support for the null hypothesis, while a Bayes's factor smaller than 0.33 is considered as substantial support for the alternative hypothesis, pointing to a statistically meaningful variance across groups. For our exemplary analyses of the DESI item response data, we chose $\delta = 0.0225$, corresponding to a standard deviation of 0.15 on the latent scale.

Results. An MCMC estimation yielded good convergence with $\hat{R}$ approximately 1.00 for all model parameters. Table 2 provides the estimation results for the absolute global and differential sensitivity measures of DESI items. Absolute global sensitivity ranged from −1.60 to 1.30 and was statistically meaningful in 30 items. That is, 5 items' change in difficulty was nondirectional across classes, as the corresponding BCIs comprised 0. This means that 29 items became easier over time, 5 items remained equally difficult, and 1 item became harder over the school year. Absolute differential sensitivity ranged from 0.13 to 0.59. Bayes's factors supported the hypothesis of variance in 27 items. That is, item-specific learning varied across classes in 27 items, while learning was equal on 8 items. In summary, 4 items were absolutely insensitive, 4 items were absolutely globally sensitive, and 27 items were absolutely globally and differentially sensitive. None of the items was absolutely differentially sensitive only.

Relative sensitivity measures are provided in Table 3. Relative global sensitivity ranged from −1.07 to 1.84. As indicated by BCIs, 15 items' sensitivity did not deviate statistically meaningful from (global) test sensitivity, meaning that 20 items function differently than those 15 that are in accordance with test sensitivity. More specifically, 6 items' change was relatively smaller (i.e., more negative) than test sensitivity, while 11 items' change was relatively higher (i.e., less negative). Relative differential sensitivity ranged from 0.12 to 0.56.

Bayes's factors labeled 31 items' variance as statistically meaningful, indicating that these items' sensitivity differed from test sensitivity. Combining the information on relative global and relative differential sensitivity, 4 items were relatively insensitive, 11 items were solely relatively differentially sensitive, and 20 items were relatively globally and differentially sensitive. That is, 4 items

TABLE 2.
Absolute Measures of Item Sensitivity for DESI Items

	β_{2i}		ϕ^2_{2i}		
Item	MAP (*SD*)	95% BCI	MAP (*SD*)	95% BCI	BF_{01}
1	−0.32 (0.10)	[−0.53, −0.12]	.48 (.18)	[0.20, 0.86]	.28
2	0.00 (0.10)	[−0.19, 0.20]	.34 (.14)	[0.12, 0.65]	.21
3	−0.11 (0.10)	[−0.29, 0.08]	.36 (.15)	[0.13, 0.70]	.19
4	−0.64 (0.10)	[−0.81, −0.45]	.24 (.12)	[0.10, 0.52]	.03
5	−0.44 (0.10)	[−0.65, −0.21]	.59 (.20)	[0.27, 1.05]	.08
6	−0.33 (0.10)	[−0.58, −0.13]	.37 (.21)	[0.10, 0.87]	.08
7	−1.01 (0.10)	[−1.19, −0.81]	.28 (.14)	[0.10, 0.60]	.00
8	−0.55 (0.10)	[−0.73, −0.36]	.12 (.06)	[0.05, 0.28]	.11
9	1.30 (0.20)	[0.99, 1.68]	.30 (.27)	[0.07, 1.00]	.06
10	−0.64 (0.10)	[−0.95, −0.39]	.44 (.23)	[0.15, 0.97]	.02
11	−1.47 (0.10)	[−1.77, −1.24]	.26 (.17)	[0.08, 0.67]	.11
12	−1.23 (0.10)	[−1.45, −1.02]	.45 (.18)	[0.19, 0.88]	.34
13	−1.60 (0.10)	[−1.83, −1.40]	.36 (.18)	[0.12, 0.77]	.04
14	−0.78 (0.10)	[−1.04, −0.51]	.26 (.15)	[0.09, 0.63]	.15
15	−0.20 (0.10)	[−0.38, −0.02]	.25 (.11)	[0.10, 0.51]	.01
16	−0.64 (0.10)	[−0.84, −0.46]	.26 (.12)	[0.09, 0.54]	.11
17	−0.94 (0.10)	[−1.13, −0.75]	.30 (.14)	[0.12, 0.63]	.96
18	−1.30 (0.10)	[−1.49, −1.11]	.18 (.09)	[0.07, 0.40]	.07
19	−0.79 (0.10)	[−1.08, −0.53]	.41 (.27)	[0.12, 1.09]	.31
20	0.06 (0.20)	[−0.39, 0.51]	.49 (.58)	[0.09, 1.99]	.35
21	−0.38 (0.20)	[−0.80, 0.00]	.40 (.48)	[0.08, 1.63]	.04
22	−0.72 (0.10)	[−0.90, −0.53]	.14 (.07)	[0.06, 0.32]	.11
23	−0.48 (0.10)	[−0.68, −0.27]	.20 (.10)	[0.08, 0.45]	.21
24	−0.34 (0.10)	[−0.55, −0.13]	.15 (.08)	[0.06, 0.34]	.04
25	−0.55 (0.20)	[−0.86, −0.26]	.42 (.30)	[0.10, 1.15]	.36
26	−0.39 (0.10)	[−0.54, −0.23]	.16 (.07)	[0.07, 0.33]	.16
27	−0.48 (0.10)	[−0.64, −0.31]	.14 (.07)	[0.06, 0.30]	.39
28	−0.21 (0.10)	[−0.39, −0.02]	.15 (.08)	[0.06, 0.35]	.00
29	−0.47 (0.10)	[−0.64, −0.29]	.16 (.08)	[0.07, 0.35]	.00
30	−0.94 (0.10)	[−1.13, −0.71]	.24 (.14)	[0.08, 0.59]	.18
31	−0.54 (0.10)	[−0.73, −0.36]	.16 (.08)	[0.06, 0.35]	.40
32	−0.96 (0.10)	[−1.17, −0.74]	.24 (.15)	[0.07, 0.61]	.00
33	−0.12 (0.10)	[−0.32, 0.09]	.25 (.13)	[0.08, 0.56]	.55
34	−0.01 (0.10)	[−0.18, 0.15]	.14 (.07)	[0.06, 0.32]	.48
35	−0.32 (0.10)	[−0.51, −0.15]	.29 (.13)	[0.11, 0.57]	.27

Note. MAP = maximum-a-posteriori estimate; *SD* = standard deviation of the posterior mean; BCI = Bayesian credible interval; BF_{01} = Bayes's factor in favor of the null hypothesis of invariance.

TABLE 3.
Relative Measures of Item Sensitivity for DESI Items

	β_{2i}		ϕ^2_{2i}		
Item	MAP (*SD*)	95% BCI	MAP (*SD*)	95% BCI	BF_{01}
1	0.20 (0.10)	[0.00, 0.42]	.48 (.18)	[0.19, 0.88]	.29
2	0.51 (0.09)	[0.34, 0.71]	.31 (.14)	[0.11, 0.62]	.06
3	0.42 (0.1)	[0.24, 0.62]	.39 (.16)	[0.15, 0.73]	.16
4	−0.10 (0.09)	[−0.28, 0.08]	.21 (.10)	[0.09, 0.46]	.65
5	0.09 (0.11)	[−0.12, 0.30]	.56 (.20)	[0.26, 1.02]	.00
6	0.18 (0.12)	[−0.05, 0.40]	.35 (.20)	[0.10, 0.83]	.46
7	−0.47 (0.09)	[−0.67, −0.30]	.24 (.12)	[0.09, 0.53]	.00
8	−0.03 (0.09)	[−0.21, 0.15]	.12 (.06)	[0.06, 0.27]	.41
9	1.84 (0.18)	[1.52, 2.21]	.28 (.25)	[0.07, 0.93]	.22
10	−0.11 (0.14)	[−0.39, 0.16]	.36 (.20)	[0.11, 0.83]	.00
11	−0.96 (0.13)	[−1.22, −0.71]	.24 (.16)	[0.07, 0.62]	.00
12	−0.68 (0.11)	[−0.89, −0.47]	.43 (.18)	[0.15, 0.86]	.10
13	−1.08 (0.11)	[−1.29, −0.87]	.35 (.16)	[0.12, 0.72]	.00
14	−0.23 (0.13)	[−0.49, 0.02]	.26 (.15)	[0.08, 0.60]	.00
15	0.34 (0.09)	[0.15, 0.50]	.22 (.11)	[0.09, 0.48]	.20
16	−0.12 (0.09)	[−0.29, 0.06]	.22 (.11)	[0.08, 0.48]	.00
17	−0.42 (0.09)	[−0.60, −0.24]	.27 (.13)	[0.09, 0.57]	.10
18	−0.77 (0.09)	[−0.95, −0.59]	.17 (.09)	[0.07, 0.38]	.01
19	−0.26 (0.14)	[−0.55, −0.02]	.32 (.25)	[0.09,0 .95]	.00
20	0.63 (0.21)	[0.19, 1.04]	.39 (.54)	[0.08, 1.83]	.11
21	0.16 (0.20)	[−0.22, 0.54]	.36 (.46)	[0.08, 1.60]	.02
22	−0.19 (0.09)	[−0.36, 0.00]	.14 (.07)	[0.06, 0.32]	.00
23	0.06 (0.10)	[−0.14, 0.25]	.19 (.10)	[0.07, 0.41]	.05
24	0.21 (0.10)	[0.01, 0.42]	.15 (.07)	[0.06, 0.32]	.30
25	−0.05 (0.14)	[−0.32, 0.24]	.31 (.23)	[0.09, 0.88]	.19
26	0.14 (0.08)	[−0.02, 0.30]	.16 (.07)	[0.07, 0.34]	.07
27	0.05 (0.08)	[−0.11, 0.22]	.13 (.07)	[0.06, 0.31]	.41
28	0.32 (0.10)	[0.13, 0.51]	.15 (.08)	[0.06, 0.34]	.02
29	0.06 (0.09)	[−0.12, 0.22]	.16 (.08)	[0.06, 0.34]	.17
30	−0.39 (0.11)	[−0.61, −0.20]	.24 (.15)	[0.08, 0.60]	.00
31	−0.03 (0.09)	[−0.22, 0.15]	.15 (.08)	[0.06, 0.35]	.32
32	−0.42 (0.11)	[−0.65, −0.21]	.28 (.18)	[0.08, 0.70]	.11
33	0.41 (0.11)	[0.19, 0.60]	.25 (.15)	[0.09, 0.62]	.18
34	0.52 (0.09)	[0.35, 0.69]	.15 (.07)	[0.06, 0.31]	.03
35	0.20 (0.09)	[0.02, 0.38]	.24 (.12)	[0.09, 0.53]	.00

Note. MAP = maximum-a-posteriori estimate; *SD* = standard deviation of the posterior mean; BCI = Bayesian credible interval; BF_{01} = Bayes's factor in favor of the null hypothesis of invariance.

TABLE 4.
Comparison of Judgment Based on Absolute and Relative Measures

	Absolute Measures	
Relative Measures	Insensitive	Sensitive
Global sensitivity		
Insensitive	1	14
Sensitive	6	14
Differential sensitivity		
Insensitive	1	3
Sensitive	7	24

Note. Judgment of global sensitivity is based on 95% Bayesian credible intervals. Judgment of differential sensitivity is based on Bayes's factors.

were in accordance with test sensitivity, while 30 deviated from test sensitivity in different ways.

Table 4 provides a comparison of the judgment of the DESI items' sensitivity based on the absolute and relative measures. Absolute and relative measures of global sensitivity do differ not only in meaning but also in their judgment of the DESI items' sensitivity. While 28 items are sensitive following the absolute measures, only 20 items are sensitive following the relative ones. Yet, even when the judgment on an item's global sensitivity appears to be in accordance, the sign of the estimate differs. For example, Item 35's absolute global sensitivity measure indicates a negative change in difficulty on average across classes, while the change is still positive relative to the other items (see Tables 2 and 3). This means that the item on the one hand is capable of detecting learning progress yet on the other hand violates measurement invariance assumptions by deviating from test sensitivity. In contrast, while absolute and relative measures of differential sensitivity appear almost equally high and thus seem to coincide, support for the null hypothesis of invariance from Bayes's factors differs. In summary, absolute measures identify 27 items as sensitive while 31 items are sensitive following relative measures.

Discussion

In the present article, we introduced the distinction of absolute and relative sensitivity. The distinction addresses two essentially different yet interrelated hypotheses regarding item sensitivity. Absolute sensitivity relates to the hypothesis of whether or not a single item is sensitive. Accordingly, absolute measures summarize the overall sensitivity of a single item to instruction. In contrast, relative sensitivity relates to the hypothesis of whether item sensitivity deviates

from test sensitivity. Hence, relative measures summarize a single item's degree of deviation from test sensitivity.

Technically, the distinction is based on whether or not item sensitivity measures are estimated conditional on test sensitivity. Within an IRT framework, test sensitivity refers to the (latent) classroom-level ability and variance components. The classroom-level ability and variance components capture what is common to all of the test items, so that mean and variance of classroom- specific and time point–specific item parameters are relative measures of instructional sensitivity, or DIF, if classroom ability is included in the estimation process. If classroom ability is not included in the estimation, the mean and variance of classroom-specific and time point–specific item parameters become absolute measures of instructional sensitivity.

Absolute and relative sensitivity add to the measurement framework of instructional sensitivity in several ways. First, distinguishing absolute and relative measures allows for a better understanding of common characteristics and differences of item sensitivity measures. Existing item sensitivity measures like the PPDI, DIF approaches or the LMLDIF approach directly relate to one of these two categories, which are complementary to the three perspectives on instructional sensitivity defined by Naumann et al. (2016). That is, instructional sensitivity measures are either absolute or relative *and* take on a specific perspective on instructional sensitivity, related to sensitivity (a) between groups, (b) between time points, or (c) both.

Second, distinguishing absolute and relative sensitivity has implications for the ways we conceive instructional sensitivity. Following Polikoff's (2010) definition, instructional sensitivity is the capacity of a test or a single item to capture effects of instruction. Absolute measures obviously comply with this definition. They provide information on whether or not a specific item is capable of capturing effects of instruction at all. If there is no change in or variation of item parameters across time or groups, the very item does not contribute to the measurement of different learning stages or classroom-level ability. In contrast, relative measures do not obviously comply with Polikoff's definition, at least at the first glance. As our results from simulated data demonstrate, relative measures are not necessarily high when item parameters vary across time or groups. In fact, relative measures may be high even if item parameters do not vary at all. That is, relative measures basically identify items functioning differently from the tests as a whole, either across time (relative global sensitivity) and/or groups within the sample (relative differential sensitivity). Hence, a high degree of relative sensitivity is a violation of measurement invariance assumptions in the first place. While some researchers have argued that such violations of measurement invariance assumptions are a necessary prerequisite for instructional sensitivity and therefore might be beneficial in testing of classroom-level characteristics (Linn & Harnisch, 1981; Naumann et al., 2014), we cannot comply with this argument. Given that items may be sensitive in an absolute way

without being sensitive in a differential way, relative sensitivity (i.e., DIF) is not a necessary requirement for instructional sensitivity and accordingly, a high degree of instructional sensitivity does not necessarily impact test fairness negatively (Geisinger & McCormick, 2010).

At the second glance, relative measures in fact do contribute to the concept of instructional sensitivity. While absolute measures target the item level, relative measures relate to the intersection of the item level and the test level. As simulation results (Set B) suggest, having items sensitive in an absolute way does not necessarily result in a sensitive test. In fact, the group-level variance component of a test built from these items is almost 0. That is, information on absolute sensitivity alone is insufficient for building instructionally sensitive assessments.

In practice, test items oftentimes are not equally sensitive for several reasons. For example, measures indicate more or less relative differential sensitivity for almost all DESI items. As previous studies have revealed (e.g., Muthén, Kao, & Burstein, 1991), relative item sensitivity may actually relate to teaching characteristics. That is, finding absolute item sensitivity related to teaching tells us whether a single item alone is instructionally sensitive, while finding relative sensitive related to teaching tells us that the very item captures effects of teaching differently than the test. In the latter case, the test might be either (a) insensitive to one or more facets of instruction the item is sensitive to or (b) sensitive, yet more or less than the specific item under investigation. As both absolute and relative sensitivity may originate in the teaching students have received, we consider both aspects important for the evaluation of instructional sensitivity.

The previous considerations notwithstanding, both absolute and relative measures may carry important information on item sensitivity to instruction. With respect to test construction, absolute measures provide information on overall sensitivity to instruction separately for each item. Relative measures then may be beneficial in examining the consequences of test assembly by highlighting which items deviate from the sensitivity of the assembled test. That is, relative measures allow investigating the extent to which absolute sensitivity is in accordance across multiple items. Ideally, a test serving as a foundation for drawing inferences on schools or teaching should comprise items with high absolute sensitivity in combination with low relative sensitivity. Then, items would be capable of capturing effects of instruction concordantly and were less prone to violations of measurement invariance assumptions, possibly resulting in test unfairness (e.g., Geisinger & McCormick, 2010).

Building on the distinction of absolute and relative item sensitivity, we provided a coherent and straightforward IRT model in consistence with the framework presented in Figure 1, the LMLIRT model. When both longitudinal and clustered data are available, the LMLIRT model allows quantifying item sensitivity in terms of sensitivity between time points, that is, global sensitivity, as well as groups within the sample, that is, differential sensitivity. The model is a generalization of the LMLDIF approach (Naumann et al., 2014) fit to provide

both absolute and relative measures in a convenient way, that is, conditional on the identification constraint chosen (see Table 1). When only one group of students or one time point of measurement is observed, the model reduces either to a longitudinal PPDI-like (i.e., time points perspective) or to cross-sectional multilevel DIF-like (i.e., groups perspective) approach that still allows deriving absolute and relative measures. Given the model's flexibility, we are confident that the LMLIRT model allows for a more complete evaluation of item sensitivity.

The LMLIRT model worked well in an exemplary application to empirical data from the DESI study. All parameters of interest were estimated with reasonable (un-)certainty. Absolute and relative measures were distinguishable empirically. As may be inferred from the empirical data example, item sensitivity measures provide inconsistent results not only due to their perspective on instructional sensitivity (see Naumann et al., 2014, 2016) but also due to their absolute or relative nature. That is, judgment based on absolute and relative sensitivity measures may differ not only between but also within the perspectives on instructional sensitivity. Interestingly, the DESI-Konsortium (2008) has already provided empirical support for the DESI test's instructional sensitivity. Yet the latent intraclass correlations for DESI test scores are comparably low despite many absolutely sensitive items. Thus, we suspect one reason for the high number of relatively differentially sensitive items in a rather low correlation of absolute sensitivity across items. Accordingly, further analyses on the extent to which relative sensitivity found in DESI data relates to teaching characteristics might allow for a deeper understanding of what aspects of teaching do contribute to students' DESI test scores—and what aspects affect certain items only.

Nevertheless, there are two issues that need to be addressed regarding our work and measuring instructional sensitivity in general. First, although we checked whether item sensitivity was statistically meaningful, we did not engage its practical relevance. To date, there are no criteria for judging the practical impact for a given degree of item sensitivity. Yet such criteria might allow for judging item sensitivity beyond simply labeling items as either (absolutely and/or relatively) sensitive or insensitive and ideally refer to the consequences for test construction and test score interpretation associated when using items with a given degree of sensitivity. Second, statistical indicators of item sensitivity need validation. As criticized by van der Linden (1981) before, statistical indicators are not per se valid to instruction. Even today, empirical evidence on valid interpretations of item sensitivity indices is rather scarce (Polikoff, 2010), as only very few studies included actual measures of teaching and classroom characteristics in their analyses of items' (instructional) sensitivity (e.g., Muthén et al., 1991). The last issue is at least partly addressed by the LMLIRT model by using classroom membership as a proxy for the diverse educational settings students experience in school.

Yet, using such proxy variables like classroom, course, or school membership does not ensure that item sensitivity may be solely attributed to learning in class. On the one hand, item sensitivity may be driven by classroom characteristics rather unrelated to teaching, for example, students' individual background or classroom composition. On the other hand, even when item sensitivity is related to effects of teaching, these effects may still to some extent originate in various sources, for example, when some content is taught in different (yet related) school subjects or students have more than one teacher. In practice, the origin of effects is only traceable when sufficient information on such cross-classified structures (e.g., van den Noortgate et al., 2003) is present in the data to appropriately deal with confounders of opportunities-to-learn and students' group membership. In a like manner, researches may want to consider dependencies among items, for example, testlet structures (e.g., Lee, Brennan, & Frisbie, 2000). The LMLIRT model may account for such complexities by choosing appropriate cluster variables on the person- and/or the item-side of the model, for example, by considering (cross-classified) nesting of students within (multiple) teachers or items nesting within testlets.

Nevertheless, the (statistical) variation of item parameters is a necessary prerequisite for instructional sensitivity (Naumann et al., 2016). Researchers may foster valid interpretation of sensitivity measures in three ways: (a) implementing strong experimental designs or (b) making sure capturing the construct-relevant variation within a domain when implementing less standardized experimental designs, for example, by random sampling and using sufficiently large samples sizes, and ideally in either case by (c) incorporating direct measures of instruction that are deeply rooted in learning and instructional theories. Then, guided by strong hypotheses, researchers may be able to validly draw inferences on items' instructional sensitivity in terms of item sensitivity explained by or attributable to teaching characteristics.

With respect to our work, although the LMLIRT model is descriptive in nature, advancing the model to an explanatory IRT model (see, e.g., De Boeck & Wilson, 2004; van den Noortgate & De Boeck, 2005) should, in principle, be straightforward and provide empirical evidence for items' instructional sensitivity (in contrast to statistical item sensitivity). Overall, we emphasize the need for a deeper discussion on to what aspects of instruction items should or should not be sensitive to, and we are confident that the LMLIRT model may serve as one foundation for further analyses on the link between teaching and item sensitivity, fostering valid use and interpretation when inferences on instruction are to be drawn based on test scores.

Nonetheless, researchers as well as policymakers need to be aware that students' achievement only is one of many outcomes of schools and teaching. While tests serving as criteria for judging the effectiveness of teaching need to be instructionally sensitive (cf. Popham, 2007), test scores alone do not tell us much

about the factors and processes that contribute to the success of teaching. Consequently, even highly instructionally sensitive tests and items cannot replace the direct observation of classroom processes.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This research was funded by the German Research Foundation (DFG), Grant HA 5050/5-1, and the Swiss National Science Foundation, Grant 100019E-157261.

References

Agresti, A. (1990). *Categorical data analysis. Wiley series in probability and mathematical statistics*. New York, NY: Wiley.

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

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.

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, 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.

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

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

De Boeck, P. (2008). Random item IRT models. *Psychometrika*, *73*, 533–559.

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

DeMars, C. E. (2004). Detection of item parameter drift over multiple test administrations. *Applied Measurement in Education*, *17*, 265–300.

Denwood, M. J. (in press). runjags: An R package providing interface utilities, parallel computing methods and additional distributions for MCMC models in JAGS. *Journal of Statistical Software*. Retrieved from http://runjags.sourceforge.net

DESI-Konsortium. (Ed.). (2008). *Unterricht und Kompetenzerwerb in Deutsch und Englisch: Ergebnisse der DESI-Studie* [Learning and instruction of German and English. Results from the DESI study]. Weinheim, Germany: Beltz.

Eichler, W. (2007). Sprachbewusstheit [Language awareness]. In B. Beck & E. Klieme (Eds.), *Sprachliche Kompetenzen: Konzepte und Messung* (pp. 147–157). Weinheim, Germany: Beltz.

French, B. F., Finch, W. F., Randel, B., Hand, B., & Gotch, C. M. (2016). Measurement invariance techniques to enhance measurement sensitivity. *International Journal of Quantitative Research in Education*, *3*, 79–93.

Geisinger, K. F., & McCormick, C. M. (2010). Adopting cut scores: Post-standard-setting panel considerations for decision makers. *Educational Measurement: Issues and Practice*, *29*, 38–44.

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

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., & Rubin, D. (2013). *Bayesian data analysis*. New York, NY: CRC/Chapman & Hall.

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.

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

Haladyna, T. M., & Roid, G. (1981). The role of instructional sensitivity in the empirical review of criterion-referenced test items. *Journal of Educational Measurement*, *18*, 39–53.

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

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.

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

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.

Meade, A. W., Lautenschlager, G. J., & Hecht, J. E. (2005). Establishing measurement equivalence and invariance in longitudinal data with item response theory. *International Journal of Testing*, *5*, 279–300.

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, NY: Springer.

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.

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

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

Plummer, M. (2003, March). *JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling*. Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), Vienna, Austria.

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

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

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

R Development Core Team. (2008). *R: A language and environment for statistical computing*. Vienna, Austria. Retrieved from http://www.R-project.org

Rijmen, F., Tuerlinckx, F., De Boeck, P., & Kuppens, P. (2003). A nonlinear mixed model framework for item response theory. *Psychological Methods*, *8*, 185–205.

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, Germany: Beltz.

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.

van den Noortgate, W., & De Boeck, P. (2005). Assessing and explaining differential item functioning using logistic mixed models. *Journal of Educational and Behavioral Statistics*, *30*, 443–464. doi:10.3102/10769986030004443

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.

Verhagen, J., & Fox, J. P. (2013). Bayesian tests of measurement invariance. *The British Journal of Mathematical and Statistical Psychology*, *66*, 383–401.

Verhagen, J., Levy, R., Millsap, R. E., & Fox, J.-P. (2015). Evaluating evidence for invariant items: A Bayes factor applied to testing measurement invariance in IRT models. *Journal of Mathematical Psychology*. doi:10.1016/j.jmp.2015.06.005

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, 370.

Wright, B. D., & Masters, G. N. (1990). Computation of outfit and infit statistics. *Rasch Measurement Transactions*, *3*, 84–85.

Authors

ALEXANDER NAUMANN is a member of the research staff at the Department of Educational Quality and Evaluation, German Institute for International Educational Research (DIPF), D-60486 Frankfurt am Main, Germany; email: naumanna@dipf.de. His primary research interests include explanatory IRT modeling, multilevel modeling, and teaching quality.

JOHANNES HARTIG is a professor for Educational Measurement at the Department of Educational Quality and Evaluation, German Institute for International Educational Research (DIPF), D-60486 Frankfurt am Main, Germany; email: hartig@dipf.de. His primary research interests include explanatory IRT models, multidimensional IRT, multilevel IRT, and modeling achievement in English as a foreign language.

JAN HOCHWEBER is a professor at the University of Teacher Education St. Gallen (PHSG), CH-9000 St. Gallen, Switzerland; email: jan.hochweber@phsg.ch. His primary research interests include classroom assessment, teaching quality, and multilevel modeling.

Manuscript received April 21, 2016
First revision received November 30, 2016
Second revision received March 7, 2017
Accepted March 10, 2017