Naumann, Alexander; Hochweber, Jan; Hartig, Johannes

Modeling instructional sensitivity using a longitudinal multilevel differential item functioning approach

formal und inhaltlich überarbeitete Version der Originalveröffentlichung in:
formally and content revised edition of the original source in:
Journal of educational measurement 51 (2014) 4, S. 381-399, 10.1111/jedm.12051

Modeling Instructional Sensitivity Using A Longitudinal Multilevel Differential Item Functioning Approach

Alexander Naumann

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

IDeA Research Center, Frankfurt, Germany

Jan Hochweber and Johannes Hartig

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

Abstract

Students' performance in assessments is commonly attributed to more or less effective teaching. This implies that students' responses are significantly affected by instruction. However, the assumption that outcome measures indeed are instructionally sensitive is scarcely investigated empirically. In the present study, we propose a longitudinal multilevel-DIF model to combine two existing yet independent approaches to evaluate items' instructional sensitivity. The model permits for a more informative judgment of instructional sensitivity, allowing the distinction of global and differential sensitivity. Exemplarily, the model is applied to two empirical datasets, with classical indices (PPDI and posttest multilevel-DIF) computed for comparison. Results suggest that the approach works well in the application to empirical data, and may provide important information to test developers.

Keywords: Instructional sensitivity, differential item functioning, multilevel IRT

Modeling Instructional Sensitivity Using A Longitudinal Multilevel Differential Item Functioning Approach

Students' performance in assessments is commonly attributed to more or less effective teaching. Research on educational effectiveness as well as policy-makers regularly rely on student performance data to hold schools and teachers accountable for their actions (Creemers & Kyriakides, 2008), hoping to retrieve useful information for scientific, pedagogical or political decisions (Pellegrino, 2002). For instance, with the enactment of the No Child Left Behind Act of 2001 (NCLB) a nation-wide accountability system based on high-stakes tests was introduced in the US, bearing direct impact on school funding (e.g., Taylor, Stecher, O'Day, Naftel, & Le Floch, 2010). This course of action implies that students' responses on tests are affected by instruction to a significant degree. However, the assumption that outcome measures indeed are sensitive to instruction is scarcely empirically engaged. Although many researchers have emphasized the issue of instructional sensitivity (e.g., Baker, 1994; Linn, 1983), the question whether instruments and items are in fact sensitive to instruction remains open more often than not. At least partly this might be due to the lack of a commonly accepted operationalization of instructional sensitivity. Over the years, various indices have been proposed (see Polikoff, 2010), most of them representing unique perspectives on item sensitivity due to the context of instruction and thereby providing valuable information for test development.

In the present study, we want to contribute to the measurement of instructional sensitivity by combining two prominent yet independent statistical approaches, the Pretest-Posttest Difference Index (PPDI; Cox & Vargas, 1966) and differential item functioning (DIF; Holland & Wainer, 1993). We believe that an integration of both approaches allows for a more informative

and comprehensive judgment of items' instructional sensitivity. Conceptually, our approach can be seen as a generalization of existing indices to investigations with $n > 1$ classrooms and $n > 1$ time points of measurement that reduces to the classical approaches when observing only one classroom or one time point, respectively. In the following, we will first give a brief overview of the theoretical framework of instructional sensitivity and its measurement. We will then describe our modeling approach and apply it to data from two different settings, an intervention study in primary schools and a large-scale assessment in secondary schools.

The Issue of Instructional Sensitivity

Instructional sensitivity is defined as the psychometric property of a test or a single item to be sensitive to instruction (Polikoff, 2010). The concept relates to the extent to which tests and items are capable of detecting effects of the implemented curriculum (Travers & Westbury, 1989), in particular, the content and the quality of instruction (D'Agostino, Welsh, & Corson, 2007). Airasian and Madaus (1983) emphasize instructional sensitivity as an important aspect of construct validity (Cronbach & Meehl, 1955). Specifically, instructional sensitivity can be seen as a necessary, though not sufficient, requirement for consequential validity (Messick, 1989) when test scores are used to draw inferences on instruction (Yoon & Resnick, 1998). In terms of current validity theory, the evaluation of instructional sensitivity provides pieces of evidence for valid and invalid test score use and interpretation (Kane, 2001, 2013).

Fundamental to the concept of instructional sensitivity is the expectation that student responses change as a consequence of instruction (Burstein, 1989). In instructionally sensitive tests, scores are expected to be positively related to more or better teaching (Baker, 1994). Students who received different instruction should produce different responses to highly instructionally sensitive items (Ing, 2008). Hence, as a function of an instruments' instructional

sensitivity, the measured construct reflects influences of instruction but also other sources of inter-individual ability differences (Burstein, 1989; Muthén, Kao, & Burstein, 1991). Correspondingly, Geisinger and McCormick (2010) point out consequences for test fairness in diagnosis of individual abilities when highly instructionally sensitive instruments are used while not all students in a sample received comparable instruction.

Operationalizations

Empirical approaches to quantify instructional sensitivity emerged in the context of criterion-referenced testing (e.g., Popham, 1971). Polikoff (2010) recently categorized the numerous approaches according to the evidence used: (1) expert judgment (e.g., Popham, 2007), (2) instructional measures (e.g., D'Agostino et al., 2007), or (3) item statistics (e.g., Cox & Vargas, 1966; Linn & Harnisch, 1981). To date, the validity of ratings on instructional sensitivity seems rather unclear, and it still remains unknown which instructional measures really are relevant for the investigation of instructional sensitivity (Polikoff, 2010). The present study focuses on statistical approaches.

Statistical approaches to measure the instructional sensitivity of items commonly focus on item difficulty or discrimination (Haladyna, 2004). Two prominent approaches are the PPDI and the investigation of DIF. Haladyna and Roid (1981) as well as Polikoff (2010) emphasize the use of PPDI as it is technically easy to implement and conceptually easy to understand. In an experimental study, Ruiz-Primo and colleagues (2012) found items' PPDI to be proportional to the alignment of item characteristics and the implemented curriculum. Also, item selection based on PPDI does not negatively impact reliability (Crehan, 1974). PPDI requires longitudinal data and is the difference between the proportions of students who get an item right at pretest and posttest measurement:

$$\text{PPDI} = \text{Difficulty}_{post} - \text{Difficulty}_{pre} \quad (1)$$

Conceptually, PPDI is conceived as the difference in item difficulty between instructed and uninstructed students. High absolute PPDI values indicate high instructional sensitivity.

The first DIF study on instructional sensitivity was conducted by Linn and Harnisch (1981). They found uniform DIF for ethnic subgroups in a math achievement test and assumed its origin in characteristics of the schooling these subgroups received. Consequently, they discussed the use of DIF methodology to detect item bias due to differences in instruction. Subsequent studies investigated uniform DIF conditional on different opportunities-to-learn or educational experiences (e.g., Clauser, Nungester, & Swaminathan, 1996; Muthén, 1989). While these DIF studies ignored the clustering of students within classes, Robitzsch (2009) suggested applying multilevel DIF (ML-DIF; Meulders & Xie, 2004) models to analyze instructional sensitivity. In contrast to previous DIF approaches that used information on students' background or their learning environment to assign them to a focal or a reference group, this ML-DIF approach accounts for the hierarchical data structure and utilizes the clustering of students in their respective classes without using further information for the assignment.

In the ML-DIF model, performance on all items depends on the same latent ability variable θ, which is decomposed into a classroom-specific and an individual component. In this respect, the model is similar to a unidimensional IRT model. However, the item difficulties are allowed to vary across classes, which would constitute a violation of assumptions for most IRT models commonly used for measurement purposes. Consequently, the ML-DIF model is less a scaling model than a model to analyze response processes to single items, accounting for overall individual and classroom-specific ability levels. In a three-level ML-DIF model with responses nested in students and classes, the probability that person *v* answers item *i* correctly is given by:

$$logit[p(X_{vik} = 1)] = \theta_k + \theta_{vk} - \beta_{ik}, \quad (2a)$$

where θ_k is the average ability of class *k*, that is, the classroom-specific ability component, and θ_{vk} is the individual deviation in ability of person *v* from the corresponding class mean, that is, the individual ability component. Parameter β_{ik} is the difficulty of item *i* in class *k*. Both the classroom-specific and the individual ability component are assumed to be mutually independent and normally distributed, with means μ and μ_k and variances σ^2 and τ^2. Similarly, the classroom-specific item difficulties are normally distributed with mean β_i and variance ${\upsilon_i}^2$:

$$\begin{aligned} \theta_k &\sim Norm(\mu, \tau^2), \\ \theta_{vk} &\sim Norm(\mu_k, \sigma^2), \\ \beta_{ik} &\sim Norm(\beta_i, {\upsilon_i}^2). \end{aligned} \quad (2b)$$

The model is not identified. Similar to common IRT models, the ML-DIF model can be identified by imposing constraints on the ability or the item difficulty parameters, for example, by fixing the means of the latent ability distributions of θ_k and θ_{vk} to zero.

As the ML-DIF approach assumes that meaningful differences in instruction received by students are due to their class membership, the variation of classroom-specific item difficulties, ${\upsilon_i}^2$, has been conceived as indicating an item's instructional sensitivity (Robitzsch, 2009). That is, the more an item's difficulty varies between classes, the higher its instructional sensitivity.

Either approach has major drawbacks. On the one hand, PPDI does not allow modeling differences in item learning between classes, 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). On the other hand, DIF studies on instructional sensitivity have so far focused solely on cross-sectional data,

neglecting variation in item difficulty between classes before instruction. Consequently, it seems plausible that ML-DIF and PPDI will oftentimes not lead to consistent results.

Aims of the Study

Despite their drawbacks, PPDI and the cross-sectional ML-DIF approach provide valuable information on item sensitivity. Hence, we assume that a combination of their unique perspectives allows for a more informative judgment of items' instructional sensitivity in accordance with demands of theory. Technically, we propose a longitudinal ML-DIF (LML-DIF) model to estimate the change in classroom-specific item difficulties between two time points of measurement. In summary, our study aims at investigating three main research questions: (1) Do PPDI and the ML-DIF approach provide consistent results in the judgment of items' instructional sensitivity? (2) Can PPDI and ML-DIF be combined in a common approach to quantify instructional sensitivity? (3) Does a combination of both approaches allow for a more informative judgment? In the following, we first briefly describe the LML-DIF approach. Then, we exemplarily apply the LML-DIF model to empirical data from two studies and compare the results to those obtained from the PPDI and the cross-sectional ML-DIF approach.

Modeling Approach

We combine PPDI and ML-DIF by extending Model (2a) to account for pretest item difficulty. Similar to the ML-DIF approach, we assume that meaningful differences in instruction received by students are due to their class membership. Accordingly, the item difficulties are allowed to vary across classes. Additionally, we assume that the average proficiency is constant across time, so that all growth is reflected in the item difficulties. That is, the item difficulties are also allowed to vary across time points. In educational research, the assumption that the average proficiency is the same before and after instruction is rather unrealistic. Normally, one would

assume that item parameters are invariant and hence the growth in proficiency is reflected in the ability parameters (e.g., Kolen & Brennan, 2004). However, allowing the item difficulties to vary across time is necessary to determine each item's sensitivity. Otherwise, change in proficiency as captured by a single item would not become noticeable in the item's difficulty parameter, but in the ability parameters, rendering the identification of instructionally sensitive items impossible. In the resulting LML-DIF model, the probability that person *v* answers item *i* at time *t* correctly is given by:

$$logit[p(X_{tvik} = 1)] = \theta_{tk} + \theta_{tvk} - \beta_{tik}, \tag{3a}$$

with:

$$\begin{aligned} \theta_{tk} &\sim Norm(\mu_t, {\tau_t}^2), \\ \theta_{tvk} &\sim Norm(\mu_{tk}, {\sigma_t}^2), \\ \beta_{tik} &\sim Norm(\beta_{ti}, {\upsilon_{ti}}^2). \end{aligned} \tag{3b}$$

For each time point of measurement, the mean ability of class *k*, θ_{tk}, the individual deviation of person *v* from the corresponding class mean, θ_{tvk}, and the classroom-specific item difficulty, β_{tik}, are estimated. Given β_{tik} for two time points $t = 1$ and $t = 2$, we define classroom-specific pretest-posttest difference values:

$$\Delta\beta_{ik} = \beta_{2ik} - \beta_{1ik} \tag{3d}$$

As the differences between two normally distributed variables are normally distributed, we assume that:[1]

$$\Delta\beta_{ti} \sim Norm(\mu_i, {\phi_i}^2). \tag{3e}$$

Expression (3e) contains beneficial information for the judgment of an item's instructional sensitivity. Parameter μ_i indicates the average Pretest-Posttest-Difference (PPD) of item *i* across classes, and thus reflects the *global sensitivity* of item *i* to instruction. Its variance component,

ϕ_i^2, describes the variation of the PPD between classes. Therefore, PPD-variance reflects the item's *differential sensitivity* to instruction.

Table 1
2 × 2-Typology of instructional sensitivity

	Average PPD	
PPD-variance	Low	High
Low	not sensitive	global
High	differential	global & differential

A combination of information on global and differential sensitivity allows the distinction of four types of instructional sensitivity (see Table 1). When average PPD and PPD-variance are low, items are considered not sensitive to instruction. Items with high average PPD, but low PPD-variance, are globally instructionally sensitive. Such items strongly change in difficulty across time points, but the change is almost equal for all classes. In contrast, items with low average PPD and high PPD-variance are differentially instructionally sensitive. These items reveal disparities in item-specific learning between classrooms, but as the low average PPD indicates, the global learning effect is rather nondirectional. While item difficulty decreases within some classes, it increases in other classes. Although this pattern is not very likely to be observed in practice, it might occur, for example, if items tap into very specific knowledge that is disremembered in a rather short time span after being learned. Finally, items with both high absolute PPD values and high PPD-variance are considered globally and differentially sensitive. Such items reveal disparities in item-specific learning with a global learning effect that is rather unidirectional across classes. Although the boundaries between these categories are fluent, we will debate the potential use of each pattern of instructional sensitivity in the discussion section.

Application to Empirical Data

We exemplarily applied the LML-DIF approach to empirical data from two studies. First, we used data from the project 'Individual support and adaptive learning environments in primary school' (IGEL; Hardy et al., 2011). IGEL is a quasi-experimental intervention study evaluating adaptive teaching methods in grade level three of German primary schools. Teachers were trained in adaptive teaching methods and implemented these in a prestructured curriculum on floating and sinking. Hence, content of instruction was intended to be identical in all classrooms while the teaching methods varied. Immediately before and after the four and a half lessons teaching unit, students' content knowledge was assessed via multiple-choice and open items. The IGEL data set comprised responses from 916 students in 54 classes that participated both in pre- and posttest. On average, the time lapse between pretest and posttest was three weeks.

The IGEL-tests were closely aligned to the intended curriculum of the teaching unit. Following the item classification system by Ruiz-Primo, Shavelson, Hamilton, and Klein (2002), the items may be considered close to proximal to instruction. Scoring followed students' conceptual understanding of floating and sinking (Kleickmann et al., 2010): naïve conceptions, explanations of everyday life, and scientific explanations. Sixteen items were administered at pretest and twelve at posttest with seven items common to both measurement points. Each student received the same items. Originally, items were selected to fit to the partial credit model (Masters, 1982). Intraclass correlation based on weighted likelihood estimates (WLE; Warm, 1989) for students' content knowledge was .06 at pretest and .16 at posttest.

Our analyses focused on the seven joint items. As four items were trichotomous, their score categories were recoded into separate dichotomous step indicators, defining the respective step functions in a cumulative approach (Agresti, 1990). We dropped the upper score categories

of two items as only two students reached them, resulting in a total of nine items (five items + four step indicators) included in the analyses. The recoded items showed acceptable fit to the one-parameter logistic (1PL) model with weighted mean-square (WMNSQ) values ranging from 0.95 to 1.06 at pretest and from 0.92 to 1.15 at posttest (cf. Wright & Linacre, 1994).

Second, we applied the LML-DIF model to empirical data from the DESI study (DESI-Konsortium, 2008). DESI is a longitudinal German large scale assessment study investigating students' language competencies and language instruction in grade level nine of German secondary schools. The target population was all German ninth-graders attending a regular secondary school type (i.e., all school types except special needs schools). Data were collected from representative samples from all 16 German federal states. A total of 219 schools were sampled with probabilities proportional to size and up to two classes within each school were sampled randomly. The DESI data set comprised responses from 10,965 students in 427 classes.

Our analyses focused on a language-awareness test comprising 34 items, administered at the beginning and the end of the school year 2003/2004 (Eichler, 2007). Following the classification system by Ruiz-Primo et al. (2002), the DESI items may be considered as rather distal to instruction. The DESI items were administered in a multi-matrix testlet design with anchoring. On average, eight students per class received the same item at one time point of measurement. No student received the same item twice. All items were scored dichotomously as either correct (1) or incorrect (0). Item-fit to a 1PL model was acceptable with WMNSQ-values ranging from 0.80 to 1.18 at pretest and from 0.82 to 1.20 at posttest. 45% (pre) and 50% (post) of the variance in WLEs was situated at the school level, while 50% (pre) and 57% (post) was situated at the class and school level. The level-two intraclass correlation was .89 (pre) and .87 (post). That is, the school level contributed more to variability than the class level. Yet in DESI,

at most two classes were assessed per school, and hence, the basis for estimating variation of classes within schools is very limited.

The IGEL and DESI data represent two different settings commonly found in educational research, a quasi-experimental intervention design and a large-scale assessment study. However, it has to be noted that some caution is advisable when interpreting and particularly comparing the results from these data sets. Both studies do not only differ with respect to the proximity of the item content to instruction, but also regarding the test administration mode. It cannot be ruled out that results are affected by the fact that in IGEL all items were answered twice by each student, while in DESI individual students responded to a different subset of items at each time point.

Estimation

For the IGEL and the DESI items, we estimated PPDI, ML-DIF for posttest data and the LML-DIF model. PPDI was computed based on IRT item difficulty estimates obtained from a two-level (students within classes) 1PL model (e.g., Kamata, 2001):

$$logit[p(X_{vik} = 1)] = \theta_k + \theta_{vk} - \beta_i \tag{4}$$

ML-DIF and LML-DIF were estimated based on the three-level item response models described in Equations 2a and 3a, respectively. Estimation of the unknown parameters was carried out using R (R Development Core Team, 2008), WinBUGS (Lunn, Thomas, Best, & Spiegelhalter, 2000) and the R2WinBUGS package (Sturtz, Ligges, & Gelman, 2005) in a Bayesian framework. The estimation method was MCMC. To identify Models (2a) and (3a), we adjusted parameters following Bafumi, Gelman, Park, and Kaplan (2005):

$$\begin{aligned} \theta_{tk}^{adj} &= \theta_{tk} - \overline{\theta}_{tk}, \\ \theta_{tvk}^{adj} &= \theta_{tvk} - \overline{\theta}_{tvk}, \\ \beta_{tik}^{adj} &= \beta_{tik} - \overline{\theta}_{tk} - \overline{\theta}_{tvk}, \end{aligned} \tag{5}$$

dropping index *t* for Model (2a). Similarly, the means of the latent ability distributions in the two-level 1PL models were adjusted to zero.

As recommended by Gelman and Hill (2006), we assumed noninformative normal distributions with mean zero and variance 10,000 as priors for ability and difficulty parameters and uniform (0, 100) distributions as priors for standard deviations. Initial values for standard deviations were randomly drawn from uniform (0, 1) distributions, while difficulty parameters' initial value was zero. For the two-level 1PL models, we ran four Markov-Chains with 5,000 iterations each and discarded the first 1,000 iterations as burn-in. Each ML-DIF and LML-DIF model ran for 25,000 iterations per chain with a burn-in of 10,000. To reduce autocorrelation, we only used every tenth iteration for ML-DIF and LML-DIF analyses. We assessed convergence by visual inspection of trace plots and the Gelman-Rubin $\hat{R}$ statistic. Point estimates were given by the means of the posterior distributions (*expected a posteriori* estimates).

To evaluate the statistical significance of the items' posttest ML-DIF variances, we specified ML-DIF models with one item's random DIF effect fixed to zero. We then compared model fit to the unrestricted ML-DIF model using the deviance information criterion (DIC; Spiegelhalter, Best, Carlin, & van der Linde, 2002). If the unrestricted model's DIC was substantially smaller than the DIC of the restricted model for a specific item, the item's DIF effect was considered as statistically significant. Differences in DIC greater than ten were regarded as substantial (Lunn, Jackson, Best, Thomas, & Spiegelhalter, 2010). Similarly, we determined the statistical significance of the PPD-variance by specifying LML-DIF models with one item's random DIF effect restricted to zero at both time points of measurement. With respect to PPDI, the standard deviation of the posttest expected a posteriori item difficulty estimate was used as a criterion for statistical significance. Items with PPDI values exceeding two times the

standard deviations of the posttest expected a posteriori item difficulty estimates were considered as instructionally sensitive. Directionality of average PPD was judged based on 95% Bayesian Credible Intervals (BCIs), considering an item as not globally instructionally sensitive if the respective 95%-BCI comprised zero and as globally instructionally sensitive if it did not.

Results

All models yielded good convergence with $\hat{R}$ approximately 1.00 for all parameters. The highest upper bound among $\hat{R}$-credible intervals was 1.07. Additionally, the distributions of item difficulties at the two time points as well as the distribution of the pretest-posttest differences were evaluated using the Kolmogorov-Smirnov test, histograms and quantile-quantile plots. For all estimates, no meaningful deviations from a normal distribution could be detected. In the following, we describe the results for IGEL and DESI data.

Table 2
IGEL data: Estimation results for PPDI and posttest ML-DIF variance

	Difficulty					ML-DIF variance	
Item	β_{pre}	(SD)	β_{post}	(SD)	PPDI	*M* (*SD*)	95% BCI
1	2.54	(0.13)	–1.34	(0.12)	–3.89	0.19 (.12)	[0.01, 0.46]
2	2.19	(0.12)	–1.61	(0.12)	–3.82	0.19 (.12)	[0.01, 0.45]
3.1	2.11	(0.12)	0.19	(0.11)	–1.91	0.32 (.12)	[0.11, 0.62]
3.2	3.56	(0.18)	1.85	(0.12)	–1.71	0.72 (.31)	[0.30, 1.48]
4	1.46	(0.10)	0.49	(0.11)	–0.97	0.51 (.19)	[0.23, 0.97]
5	0.92	(0.09)	0.01	(0.11)	–0.92	0.49 (.17)	[0.24, 0.90]
6	0.95	(0.09)	–1.85	(0.12)	–2.80	0.21 (.13)	[0.01, 0.49]
7.1	3.59	(0.19)	1.57	(0.12)	–2.02	0.45 (.19)	[0.16, 0.89]
7.2	4.97	(0.34)	3.01	(0.16)	–1.96	0.46 (.27)	[0.09, 1.15]

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

Results IGEL. Table 2 provides pretest and posttest IRT item difficulty estimates, the PPDI computed from these estimates, and the posttest item difficulty variation between classes from the ML-DIF model. According to PPDI, all IGEL items and step indicators were instructionally sensitive with values ranging from −3.89 to −0.92 (*M* = −2.22, *SD* = 1.08).

Posttest ML-DIF-variances ranged from 0.19 to 0.72 (M = 0.39, SD = 0.18). Three items showed statistically negligible ML-DIF-variances. Hence, in contrast to PPDI, only six items (including the four step indicators) were instructionally sensitive following the posttest ML-DIF approach. Interestingly, items 1, 2 and 6 – with highest PPDI – appeared insensitive. Contrariwise, the step indicator 3.2 showed a comparatively moderate PPDI but the highest posttest ML-DIF-variance.

Estimation results for the LML-DIF model are shown in Table 3. All items revealed an average PPD ranging from −3.98 to −0.95 (M = −2.22, SD = 1.15). As none of the BCIs comprised zero, all items performed globally sensitive to instruction. IGEL items' PPD-variances, indicating variation in change of item difficulties between classes, ranged from 0.22 to 0.86 (M = 0.52, SD = 0.25). Three items showed no statistically meaningful PPD-variances. Hence, six items (including the four step indicators) were differentially sensitive to instruction. Combining these information, three IGEL items performed solely globally sensitive, no item was differentially sensitive only, two items and the four step indicators appeared globally and differentially sensitive, and none of the items was insensitive.

Table 3
IGEL data: Estimation results for the LML-DIF model

	Average PPD			PPD-variance	
Item	*M*	(SD)	95% BCI	*M* (*SD*)	95% BCI
1	–3.98	(0.18)	[–4.35, –3.62]	0.22 (.12)	[0.02, 0.49]
2	–3.95	(0.18)	[–4.31, –3.60]	0.36 (.21)	[0.05, 0.85]
3.1	–1.90	(0.16)	[–2.21, –1.58]	0.34 (.12)	[0.12, 0.62]
3.2	–1.51	(0.25)	[–2.01, –1.04]	0.77 (.29)	[0.32, 1.46]
4	–1.03	(0.16)	[–1.34, –0.73]	0.80 (.22)	[0.42, 1.31]
5	–0.95	(0.14)	[–1.23, –0.67]	0.50 (.14)	[0.26, 0.81]
6	–2.89	(0.16)	[–3.20, –3.00]	0.22 (.12)	[0.03, 0.50]
7.1	–2.03	(0.29)	[–2.67, –1.53]	0.86 (.55)	[0.24, 2.27]
7.2	–1.68	(0.38)	[–2.54, –1.00]	0.60 (.41)	[0.14, 1.66]

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

Results DESI. Results from PPDI and ML-DIF approaches for the DESI data are presented in Table 4. PPDI values ranged from −1.52 to 1.02 (M = −0.49, SD = 0.47). Given that six items had low PPDI values below two times the standard deviations of the posttest 1PL item difficulty estimates, these items were considered as insensitive. Accordingly, twenty-eight items were regarded instructionally sensitive following the PPDI approach, with one item showing an increase in item difficulty. Evaluating DESI items' posttest ML-DIF, variances ranged from 0.03

Table 4
DESI data: Estimation results for PPDI and posttest ML-DIF variance

	Difficulty					ML-DIF variance	
Item	β_{pre}	(SD)	β_{post}	(SD)	PPDI	*M* (*SD*)	95% BCI
1	–0.91	(0.06)	–1.10	(0.07)	–0.19	0.45 (0.07)	[0.31, 0.61]
2	–0.58	(0.06)	–0.63	(0.07)	–0.06	0.19 (0.05)	[0.09, 0.31]
3	–0.48	(0.06)	–0.45	(0.07)	0.03	0.29 (0.06)	[0.18, 0.42]
4	–0.74	(0.06)	–1.47	(0.07)	–0.73	0.06 (0.04)	[0.01, 0.15]
5	–1.39	(0.07)	–1.72	(0.07)	–0.33	0.12 (0.06)	[0.03, 0.26]
6	–1.62	(0.07)	–1.80	(0.07)	–0.18	0.13 (0.06)	[0.03, 0.27]
7	–2.29	(0.07)	–2.83	(0.08)	–0.54	0.06 (0.04)	[0.01, 0.16]
8	–0.34	(0.06)	–0.70	(0.07)	–0.36	0.04 (0.02)	[0.01, 0.09]
9	–0.07	(0.06)	–0.45	(0.07)	–0.38	0.03 (0.02)	[0.00, 0.08]
10	–1.46	(0.07)	–1.50	(0.07)	–0.04	0.32 (0.07)	[0.19, 0.48]
11	0.05	(0.06)	–0.41	(0.07)	–0.46	0.15 (0.05)	[0.07, 0.26]
12	0.14	(0.07)	–0.40	(0.07)	–0.54	0.17 (0.05)	[0.07, 0.28]
13	–0.98	(0.06)	–1.44	(0.07)	–0.46	0.08 (0.04)	[0.01, 0.17]
14	0.50	(0.07)	–0.24	(0.07)	–0.74	0.05 (0.03)	[0.01, 0.12]
15	–1.63	(0.07)	–1.72	(0.07)	–0.08	0.11 (0.05)	[0.03, 0.23]
16	0.04	(0.06)	0.00	(0.07)	–0.04	0.24 (0.05)	[0.14, 0.35]
17	0.44	(0.06)	0.07	(0.07)	–0.37	0.08 (0.04)	[0.02, 0.17]
18	0.54	(0.06)	–0.14	(0.07)	–0.68	0.15 (0.05)	[0.06, 0.25]
19	1.10	(0.06)	0.69	(0.07)	–0.41	0.37 (0.07)	[0.25, 0.52]
20	0.89	(0.07)	0.44	(0.07)	–0.45	0.13 (0.04)	[0.05, 0.22]
21	–0.19	(0.06)	–1.12	(0.07)	–0.93	0.07 (0.04)	[0.01, 0.15]
22	1.16	(0.06)	0.43	(0.07)	–0.73	0.10 (0.04)	[0.03, 0.20]
23	2.63	(0.07)	3.65	(0.09)	1.02	0.75 (0.19)	[0.42, 1.18]
24	–1.95	(0.07)	–2.40	(0.08)	–0.45	0.20 (0.09)	[0.05, 0.39]
25	–1.45	(0.06)	–2.72	(0.08)	–1.27	0.11 (0.06)	[0.02, 0.26]
26	0.73	(0.06)	–0.13	(0.07)	–0.86	0.82 (0.12)	[0.60, 1.08]
27	–0.27	(0.06)	–1.79	(0.07)	–1.52	0.20 (0.07)	[0.07, 0.36]
28	–0.23	(0.06)	–0.79	(0.07)	–0.56	0.48 (0.08)	[0.34, 0.65]
29	0.49	(0.06)	–0.29	(0.07)	–0.78	0.10 (0.04)	[0.03, 0.19]
30	–0.07	(0.06)	–1.08	(0.07)	–1.01	0.20 (0.05)	[0.10, 0.32]

31	0.87	(0.07)	–0.32	(0.07)	–1.09	0.25 (0.06)	[0.14, 0.39]
32	–0.66	(0.07)	–1.72	(0.09)	–1.06	0.40 (0.14)	[0.16, 0.69]
33	2.37	(0.08)	2.16	(0.09)	–0.21	0.95 (0.22)	[0.55, 1.42]
34	2.28	(0.08)	2.15	(0.08)	–0.13	1.00 (0.20)	[0.62, 1.42]

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

to 1.00 (*M* = 0.26, *SD* = 0.26). Twenty-five items were instructionally sensitive, while nine items' ML-DIF-variances were statistically negligible. In summary, more items were instructionally sensitive following PPDI compared to the ML-DIF approach.

Table 5

DESI data: Estimation results for the LML-DIF model

	Average PPD			PPD-variance	
Item	*M*	(SD)	95% BCI	M (SD)	95% BCI
1	–0.19	(0.10)	[–0.39, 0.00]	0.72 (0.09)	[0.55, 0.91]
2	–0.05	(0.10)	[–0.25, 0.14]	0.41 (0.08)	[0.27, 0.57]
3	0.03	(0.10)	[–0.16, 0.23]	0.53 (0.08)	[0.39, 0.69]
4	–0.74	(0.10)	[–0.94,–0.54]	0.20 (0.06)	[0.10, 0.32]
5	–0.35	(0.11)	[–0.55,–0.14]	0.25 (0.07)	[0.12, 0.39]
6	–0.17	(0.11)	[–0.38, 0.04]	0.30 (0.07)	[0.17, 0.46]
7	–0.53	(0.12)	[–0.76,–0.30]	0.20 (0.07)	[0.09, 0.36]
8	–0.37	(0.10)	[–0.56,–0.17]	0.16 (0.04)	[0.08, 0.26]
9	–0.39	(0.10)	[–0.48,–0.19]	0.20 (0.05)	[0.11, 0.30]
10	–0.02	(0.10)	[–0.22, 0.18]	0.67 (0.10)	[0.48, 0.87]
11	–0.50	(0.10)	[–0.69,–0.30]	0.31 (0.06)	[0.20, 0.44]
12	–0.59	(0.10)	[–0.79,–0.38]	0.34 (0.08)	[0.20, 0.50]
13	–0.47	(0.10)	[–0.67,–0.25]	0.17 (0.05)	[0.08, 0.30]
14	–0.76	(0.10)	[–0.97,–0.56]	0.18 (0.06)	[0.07, 0.31]
15	–0.08	(0.11)	[–0.29, 0.12]	0.28 (0.07)	[0.15, 0.43]
16	–0.05	(0.10)	[–0.24, 0.15]	0.41 (0.07)	[0.29, 0.55]
17	–0.38	(0.10)	[–0.57, 0.18]	0.36 (0.06)	[0.25, 0.49]
18	–0.70	(0.10)	[–0.89,–0.50]	0.29 (0.06)	[0.17, 0.42]
19	–0.44	(0.10)	[–0.64,–0.24]	0.56 (0.08)	[0.42, 0.72]
20	–0.48	(0.11)	[–0.69,–0.28]	0.33 (0.09)	[0.18, 0.52]
21	–0.95	(0.10)	[–1.16,–0.76]	0.15 (0.05)	[0.07, 0.27]
22	–0.76	(0.10)	[–0.96,–0.55]	0.24 (0.06)	[0.13, 0.36]
23	1.13	(0.13)	[0.87, 1.39]	1.24 (0.22)	[0.85, 1.69]
24	–0.49	(0.11)	[–0.71,–0.27]	0.29 (0.09)	[0.13, 0.50]
25	–1.31	(0.11)	[–1.54,–1.09]	0.21 (0.07)	[0.09, 0.38]
26	–0.97	(0.10)	[–1.17,–0.77]	1.13 (0.12)	[0.91, 1.37]
27	–1.59	(0.10)	[–1.80,–1.39]	0.36 (0.08)	[0.22, 0.53]
28	–0.61	(0.10)	[–0.81,–0.41]	0.81 (0.09)	[0.64, 1.00]
29	–0.81	(0.10)	[–1.01,–0.62]	0.27 (0.06)	[0.16, 0.40]
30	–1.02	(0.10)	[–1.22,–0.82]	0.45 (0.07)	[0.31, 0.61]

31	–1.23	(0.10)	[–1.43,–1.03]	0.39 (0.07)	[0.25, 0.54]
32	–1.11	(0.12)	[–1.35,–0.87]	0.70 (0.16)	[0.41, 1.05]
33	–0.12	(0.14)	[–0.38, 0.15]	1.24 (0.24)	[0.81, 1.75]
34	–0.12	(0.13)	[–0.37, 0.15]	1.39 (0.24)	[0.95, 1.89]

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

Estimation results for the LML-DIF model are shown in Table 5. In the longitudinal approach, average PPD values ranged from −1.59 to 1.13 (*M* = −0.51, *SD* = 0.50). As nine of the BCIs comprised zero, the respective items were considered not globally sensitive to instruction. Hence, twenty-five items were globally sensitive according to their average PPD. PPD-variances ranged from 0.15 to 1.39 (*M* = 0.46, *SD* = 0.34) with thirty items showing non-negligible PPD-variances. Combining the information, four items were solely globally sensitive, nine items were differentially sensitive only, twenty-one items performed globally and differentially sensitive, and none of the items was insensitive.

Relationships between indices

Table 6 summarizes frequentist intercorrelations and percent agreement of indices for IGEL and DESI data. Values of the conceptually connatural indices – PPDI and average PPD on the one hand, and posttest ML-DIF and PPD-variance on the other hand – were highly positively associated. PPDI-based measures and dispersion-based approaches were also positively related, suggesting that high values on both dispersion-based indices seemed to relate to positive global item learning. However, while PPDI-based indices were significantly linked to PPD-variance in DESI, statistical relationship to posttest ML-DIF did not reach the .05-significance level.

Of more importance than the correlations are the judgments on items' sensitivity based on the indices. Within both the IGEL and the DESI data sets, PPDI and ML-DIF do not provide consistent results. PPDI and average PPD highly agree in both data sets, with disagreement for three DESI items only. Whereas ML-DIF and PPD-variance provide the same results for all the

IGEL items, agreement is lower for the DESI items. Finally, average PPD and PPD-variance do not coincide for about one third of the items in both data sets.

Table 6
Intercorrelations and percent agreement of instructional sensitivity indices

	Intercorrelations				Percent agreement			
	1	2	3	4	1	2	3	4
1. PPDI	–	.78	.99	.65	–	67	100	67
2. ML-DIF	.32	–	.82	.82	62	–	67	100
3. Average PPD	.99***	.32	–	.67	91	53	–	67
4. PPD-variance	.39*	.98***	.39*	–	71	85	62	–

*Note. *p < .05, **p < .01, ***p < .001; Correlations and percent agreement for DESI data (n = 34) are provided below the diagonals, for IGEL data (n = 9) above. For IGEL, tests of significance are not reported due to small sample size.*

Conclusion & Discussion

We presented a LML-DIF model to estimate change in classroom-specific item difficulties between two time points of measurement, denoting the mean item-specific learning across classes as average PPD and the variation as PPD-variance. Directionality of average PPD was judged based on 95%-BCIs, considering an item as globally instructionally sensitive if the respective 95%-BCI did not comprise zero. We evaluated the statistical, though not the practical, relevance of PPD-variance via comparison to models where the respective variance components were restricted to zero, considering an item as differentially instructionally sensitive if the unrestricted model had a substantially lower DIC. Information on global and differential sensitivity was combined to judge an item's instructional sensitivity. Conceptually, our approach can be conceived as integrating two existing approaches, PPDI and ML-DIF, which were applied for comparison. As we used IRT item difficulties to calculate PPDI, values reported are relative measures of instructional sensitivity for items within a test and are not directly comparable across tests.

Technically, we found the LML-DIF model to work well in its application to empirical data. All parameters of interest were estimated with reasonable certainty. As indicated by trace plots and $\hat{R}$, convergence was satisfying after 10,000 iterations. Admittedly, mixing was slow for variance parameters ${\upsilon_i}^2$and ${\phi_i}^2$ Markov-Chains, which tended to be highly autocorrelated when ${\upsilon_i}^2$and ${\phi_i}^2$ values were close to zero.

Results indicated that judgment of items' instructional sensitivity based on PPDI and the ML-DIF approach was not consistent. In contrast, a combination of both indices yielded a more detailed judgment of instructional sensitivity. PPDI and average PPD mostly agreed whether an item was sensitive or not, whereas PPD-variance identified more items instructionally sensitive than the ML-DIF approach. That is, while average PPD and PPDI generally capture the same information, ML-DIF and PPD-variance do not. Technically, the ML-DIF approach focuses on differences in item difficulty between classes at a given time point. In contrast, PPD-variance relates to differences in item-specific learning across time points between classes. Accordingly, both indices relate to different sources of variance. Given its developmental perspective, PPD-variance comprises more construct-relevant variance compared to the ML-DIF approach.

Whether judgment based on the ML-DIF approach and PPD-variance is consistent depends on specific constraints. As PPD-variance takes the classroom-specific base level in item difficulty into account, the ML-DIF and PPD-variance should agree when there is no variation in item difficulty before instruction. Then, the extent of variation in item difficulties across classes at posttest should be equivalent to the extent of variation in item-specific learning across classes. In fact, this is the case for the IGEL items, but not for the DESI data. Although PPD-variance identified more DESI items as sensitive than the posttest ML-DIF approach, we assume that this result need not always occur in empirical data. Suppose item-specific learning is equal across all

classes in a sample and at the same time there is significant ML-DIF variance at pretest and posttest. That is, item difficulties vary between classes at both time points of measurement while the trajectories of item difficulties across time points form parallel lines. Then, the posttest ML-DIF approach will identify such items as sensitive, whereas PPD-variance does not. Consequently, the combined interpretation of PPD-variance and average PPD in the LML-DIF approach may dissent from results of analyses applying both PPDI and ML-DIF concurrently.

For three reasons, we believe our approach to be beneficial in the evaluation of instructional sensitivity compared to existing PPDI and DIF approaches: (1) By including baseline data, changes in item difficulties can be more unequivocally attributed to instruction, (2) by adding classrooms as units of observation, the model accounts for differences at the level where instruction actually happens, and (3) with respect to the different sources of variance that are relevant in empirical research, that is, variance between time points and variance between groups within a sample, a single statistical indicator is not able to describe an item's instructional sensitivity sufficiently. While PPDI and DIF approaches focus either on variation between time points (PPDI) or variation between groups (DIF), the LML-DIF approach accounts for both sources of variance as it combines existing indices' perspectives in a longitudinal model-based approach. Within the LML-DIF approach, average PPD and PPD-variance are inextricably linked to describe an item's global and differential sensitivity. Nevertheless, effects of maturation can still only be eliminated by including an untreated control group, and classroom-specific PPD values might be affected by student, teacher or classroom characteristics, for instance classroom composition regarding students' socio-economic status or migrational background.

Regarding instructional sensitivity, average PPD and PPD-variance provide comprehensive statistical information on test items' capability to capture effects of instruction,

allowing the distinction of global and differential sensitivity. As Glaser (1963) stated, test construction needs information on how well items differentiate between groups treated differently. Our approach is in accordance with this demand by expressing both how an item differentiates between a) instructed and uninstructed students, and b) instructional settings within a sample. For existing instruments, empirical evidence to support test score interpretation is gathered by determining the extent to which item responses are – or are not – influenced by instruction. For the construction of achievement tests assessing the effectiveness of instructional settings, the combination of information on global and differential item sensitivity allows test developers to select items fitting to the purposes of their assessments. As empirical data indicate, when using only either the PPDI or posttest ML-DIF approach, results on instructional sensitivity may be partially incomplete and potentially misleading.

In practice, items that are globally, but not differentially, sensitive might be preferred in testing of individual abilities, competencies, or maturation. While instruction may contribute to answering items correctly, item responses are less dependent on whether a student received specific instruction or not. In contrast, responses on differentially sensitive items are significantly affected by class membership, potentially leading to a misattribution of results to inter-individual differences. Accordingly, the concept of differential instructional sensitivity highlights Geisinger and McCormick's (2010) concerns on test fairness when high-stakes tests are administered to students with heterogeneous educational backgrounds. Hence, in diagnosis of individual abilities, differentially sensitive items should presumably be avoided, unless including them on purpose.

If one is less interested in individual ability but rather classroom-level characteristics, items that are differentially sensitive appear beneficial. For instance, let us assume we wanted to check the implementation of a certain curriculum in class. In this case, we might choose items

that can only be solved via knowledge or abilities gained through the exact implementation of the curriculum, consciously penalizing students who did not receive instruction according to the intended curriculum. Consequently, the measured construct would comprise both individual ability and classroom characteristics, and scores would reflect the interplay between students' ability and the degree to which the curriculum was implemented. Yet, relating the results to the effectiveness of the curriculum would be an ecological fallacy, creating artificial effect sizes when comparing classes that implemented the curriculum to those who did not.

Selecting items that are both globally and differentially sensitive may appear to be a natural choice in testing effects of classroom-level characteristics. Due to instruction, learning is typically expected to be rather unidirectional across classes. Nevertheless, some items might become more difficult for students as a consequence of instruction. Apparently, item difficulty may increase when instruction in classrooms is detrimental to learning, for example, badly structured or error-prone. But difficulty may also increase when students' gain in knowledge leads to more intra-individual cognitive conflicts, inducing greater uncertainty in students' answering process than before instruction (see Vosniadou, 2007).

Items that appear insensitive need special attention. If an item performs neither globally nor differentially sensitive, its instructional insensitivity might be due to three reasons (cf. Haladyna, 2004): a) the item has nothing to do with the instruction students in a sample received, b) the item is so difficult or easy in both pre- and posttest that either everybody or nobody can solve it – despite instruction, or c) there actually is no effect of instruction. Option b can be verified by inspecting the pre- and posttest data. To decide whether options a or c apply, further information on the item in question and the instruction are needed. For example, when all classes in a sample did not implement the part of the curriculum the item aims at, information on

whether the item has proven to be sensitive in previous studies are beneficial. If so, the item might indeed be instructionally sensitive, but the content needed had not been learned yet. Hence, we advise careful examination of reasons why items are flagged insensitive.

Notwithstanding the previous considerations, we would like to emphasize that in many situations, different types of instructionally sensitive items will be required. As the same assessment commonly addresses various recipients, its developers might want to consider their respective needs and purposes for test score interpretation, and hence inclusion of different types of instructionally sensitive items might be necessary (Linn, 1983; Ruiz-Primo et al., 2002).

Finally, the proposed statistical indicators themselves need validation. The LML-DIF approach provides information on the amount of variability in learning between classrooms, but the source of this variation remains unknown. Classroom membership can be confounded with other variables affecting student learning, for example, socio-economic classroom composition. Thus, the variation may also originate in student or classroom characteristics unrelated to instruction. Nevertheless, statistical variation is a necessary, though not sufficient, requirement for instructional sensitivity. Instructional sensitivity may then be conceived as the proportion of variance in item parameters explained by content and quality of teaching. Accordingly, average PPD and PPD-variance may mainly function as minimum criteria for the evaluation of instructional sensitivity. They support – but cannot replace – the analysis of the relationships between item responses and empirical measures of instruction. Consequently, further analyses will extend the LML-DIF model to incorporate teaching characteristics.

Note

[1]As β_{1ik} and β_{2ik} each have their own normal distributions, the distribution in expression (3e) is theoretical in nature and therefore not part of the model itself.

References

Agresti, A. (1990). *Categorical data analysis*. New York: Wiley.

Airasian, P. W., & Madaus, G. F. (1983). Linking testing and instruction: Policy issues. *Journal of Educational Measurement, 20*, 103-118.

Bafumi, J., Gelman, A., Park, D. K., & Kaplan, N. (2005). Practical issues in implementing and understanding Bayesian ideal point estimation. *Political Analysis*, *13*, 171–187.

Baker, E. L. (1994). Making performance assessment work: The road ahead. *Educational Leadership, 51*(6), 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, April). *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.

Crehan, K. D. (1974). Item analysis for teacher-made mastery tests. *Journal of Educational Measurement, 11*, 255–261.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281–302.

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

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 and Basel: Beltz.

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

Geisinger, K. F., & McCormick, C. M. (2010). Adopting cut scores: Post-standard-setting panel considerations for decision makers. *Educational Measurement: Issues and Practice, 29(1)*, 38–44.

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

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

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

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

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.

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*(1), 23–43.

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

Kane, M. T. (2001). Current concerns in validity theory. *Journal of Educational Measurement, 38*, 319–342.

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 naturwissenschaftliche Kompetenz im Grundschulalter: Theoretische Konzeption und Testkonstruktion [Modeling scientific competence of primary school children]. *Zeitschrift für Didaktik der Naturwissenschaften, 16*, 263–281.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. New York: Springer Science.

Linn, R. L. (1983). Testing and instruction: Links and distinctions. *Journal of Educational Measurement, 20*, 179–189.

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.

Lunn, D., Jackson, C., Best, N., Thomas, A., & Spiegelhalter, D. (2010). *The BUGS Book: A practical introduction to Bayesian analysis*. London: Chapman & Hall.

Lunn, D., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing, 10*, 325–337.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149–174.

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

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.

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

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

Popham, J. W. (1971). Indices of adequacy for criterion-reference test items. In J. W. Popham (Ed.), *Criterion-referenced measurement (an introduction)* (pp. 79–98). Englewood Cliffs, NJ: Educational Technology Publications.

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

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

Robitzsch, A. (2009). Methodische Herausforderungen bei der Kalibrierung von Leistungstests [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 and Basel: Beltz.

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

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

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., & van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology), 64*, 583–639.

Sturtz, S., Ligges, U., & Gelman, A. (2005). R2WinBUGS: A package for running WinBUGS from R. *Journal of Statistical Software, 12*(3), 1–16.

Taylor, J., Stecher, B., O'Day, J., Naftel, S., & Le Floch, K. C. (2010). *State and local implementation of the No Child Left Behind Act: Volume IX-Accountability under NCLB.* Washington, DC: U.S. Department of Education.

Travers, K. J., & Westbury, I. (1989). *The IEA study of mathematics I: Analysis of mathematics curricula (Vol. 1)*. Elmsford, NY: Pergamon Press.

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

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427–450.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions,* 8:3, 370.

Yoon, B., & Resnick, L. B. (1998). *Instructional validity, opportunity to learn, and equity: New standards examinations for the California Mathematics Renaissance*. Los Angeles, CA: Center for the Study of Evaluation.