A Psychometric Framework for the Evaluation of Instructional Sensitivity

Alexander Naumann\textsuperscript{1,2}, Jan Hochweber\textsuperscript{1}, Johannes Hartig\textsuperscript{1}, & Eckhard Klieme\textsuperscript{1,2}

1) German Institute for International Educational Research (DIPF), Frankfurt, Germany
2) IDeA Research Center Frankfurt, Germany

First International Conference on Instructional Sensitivity, Lawrence, KS, 2013
Instructional Sensitivity

• Students’ performance in assessments is regularly used as the main outcome measures in research on educational effectiveness (Creemers & Kyriakides, 2008).

• This implies outcomes being affected by instruction to a significant degree.

⇒ To what extent are tests and items capable to detect effects of instruction?
Approaches to Instructional Sensitivity

• Various approaches to instructional sensitivity (Polikoff, 2010):

  a) Prediction of test scores using empirical measures of instruction, e.g., content coverage or teaching styles (e.g., Ing, 2008)
  b) Expert judgment (e.g., Popham, 2007)
  c) Item statistics (e.g., Haladyna & Roid, 1981)

• Approaches oftentimes do not provide consistent results (e.g., Li, Ruiz-Primo, & Wills, 2012)
Approaches to Instructional Sensitivity

• Various approaches to instructional sensitivity (Polikoff, 2010):

  a) Prediction of test scores using empirical measures of instruction, e.g., content coverage or teaching styles (e.g., Ing, 2008)
  b) Expert judgment (e.g., Popham, 2007)
  c) Item statistics (e.g., Haladyna & Roid, 1981)

⇒ Lack of knowledge on differences /similarities within and across the categories
Aims of the Study

• Building up a psychometric framework for the evaluation of instructional sensitivity based on how tests and items reflect variability due to the instructional context.

• Providing a model-based approach to quantify items’ variability due to the instructional context.

• Relating the estimated variability to measures of instruction.
A Psychometric Framework for IS

Responses
A Psychometric Framework for IS
A Psychometric Framework for IS

Persons

Time points

Responses
A Psychometric Framework for IS

Classes
Persons
Time points
Responses
A Psychometric Framework for IS

Schools

C₁

P₁

T₁

X₁, X₂, X₃, X₄

Schools

C₂

P₁

T₁

X₁, X₂, X₃, X₄

Persons

P₂

T₂

X₁, X₂, X₃, X₄

Persons

P₃

T₃

X₁, X₂, X₃, X₄

Time points

Responses
A Psychometric Framework for IS

Schools

C₁

C₂

P₁

P₂

P₃

P₄

T₁

X₁

X₂

X₃

X₄

T₂

X₁

X₂

X₃

X₄

T₃

X₁

X₂

X₃

X₄

Schools

Classes

Persons

Time points

Responses
A Psychometric Framework for IS
A Psychometric Framework for IS

Diagram showing the relationships between Schools, Classes | Testlets, Persons | Items, Time points, and Responses.
A Psychometric Framework for IS
A Psychometric Framework for IS

Person side clusters that proxy sources of different instruction (e.g. classes, courses, teachers, schools, and so on)
A Psychometric Framework for IS

**Item side** determines the level at which the evaluation of instructional sensitivity takes place (e.g., whole test or single items)
Three Perspectives on Instructional Sensitivity

- **Schools**
  - C₁
  - C₂
  - P₁, P₂, P₃, P₄

- **Tests**
  - TL₁
  - TL₂
  - I₁, I₂, I₃, I₄

- **Time points**
  - T₁
  - T₂
  - T₃

- **Responses**
  - X₁, X₂, X₃, X₄

- **Classes | Testlets**

- **Schools | Tests**

- **Persons | Items**

- **Time points**

- **Responses**
(a) Differences between time points

- Schools
  - C₁
    - P₁
      - T₁
        - X₁
      - X₂
    - P₂
      - T₁
      - X₃
    - P₃
      - T₁
      - X₄
- C₂
  - P₄
    - T₂
      - X₁
      - X₂
      - X₃
      - X₄
    - T₃
      - X₁
      - X₂
      - X₃
      - X₄

E.g., Pretest Posttest Difference Index (PPDI; Cox & Vargas, 1966)

⇒ Items that change in difficulty across time are considered instructionally sensitive

- Schools | Tests
- Classes | Testlets
- Persons | Items
- Time points
- Responses
(b) Differences between groups

- e.g., DIF (Linn & Harnisch, 1981) or multilevel-DIF (Robitzsch, 2009)

⇒ Items with DIF for groups with different educational experiences are considered sensitive.
(c) Differences between groups and time points

- Schools
- Classes | Testlets
- Persons | Items
- Time points
- Responses

- e.g., LML-DIF (Naumann, Hartig & Hochweber, 2013)
  - Incorporating both sources of variance
Modeling approach

Longitudinal multilevel IRT model to estimate change in item difficulties with interaction of items, time points, and classrooms:

\[
\text{logit}[p(X_{kvt} = 1)] = \theta_{kt} + \theta_{kvt} - \sum_{q=1}^{T} Z_{qkvit} \beta_{qki}
\]

- $\theta_{kt}$: Average ability of class $k$ at time $t$
- $\theta_{kvt}$: Deviation of person $v$ from its class $k$ average ability at time $t$
- $\beta_{qik}$: Difficulty of item $i$ in class $k$ at time $t = 1$ and classroom-specific change in item difficulty from the preceding time point for $t > 1$
- $Z_{qkvit}$: Dummy variable with values 1 for $q \leq t$ and 0 otherwise

Lawrence, 14.11.2013 | A psychometric framework | Naumann | First International Conference on Instructional Sensitivity
Modeling approach

Longitudinal multilevel IRT model to estimate change in item difficulties with interaction of items, time points, and classrooms:

\[
\text{logit}[p(X_{kvit} = 1)] = \theta_{kt} + \theta_{kt} - \sum_{q=1}^{T} Z_{qkvit} \beta_{qki}
\]

The model provides...

(1) an estimate of baseline classroom-specific item difficulty (t = 1)
(2) an estimate of classroom-specific change in item difficulty across time
Modeling approach

Longitudinal multilevel IRT-model to estimate change in item difficulties

with:

\[ \beta_{q_i} \sim \text{Norm}(\beta_{q_i}, \phi_{q_i}^2), \]

For \( t = 1 \)

\( \beta_{q_i} \) Average baseline item difficulty
\( \phi_{q_i}^2 \) Variation of baseline item difficulty across classes (multilevel DIF)
Modeling approach

Longitudinal multilevel IRT-model to estimate change in item difficulties

with:

\[ \beta_{qi} \sim \text{Norm}(\beta_{qi}, \phi_{qi}^2), \]

For \( t > 1 \)

\[ \beta_{qi} \] Average Pretest-Posttest-Difference in item difficulty (average PPD)

\[ \phi_{qi}^2 \] Variation of PPDs across classes (PPD-variance)
Benefits to Instructional Sensitivity

• Two kinds of instructional sensitivity:

⇒ **Global Sensitivity** refers to the average extent and direction of change in classroom-specific item difficulty across time points

⇒ **Differential Sensitivity** refers to the variation of change in item difficulty across classes
Benefits to Instructional Sensitivity

- Combination of information leads to a 2 x 2 typology of instructional sensitivity:

<table>
<thead>
<tr>
<th>Differential Sensitivity</th>
<th>Global Sensitivity</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>insensitive</td>
</tr>
<tr>
<td>high</td>
<td>differential</td>
</tr>
<tr>
<td>high</td>
<td>global</td>
</tr>
<tr>
<td>low</td>
<td>global &amp; differential</td>
</tr>
</tbody>
</table>
Benefits to Instructional Sensitivity

The modeling approach can easily be extended
Adding a latent regression term to the model yields an explanatory IRT model (e.g., van den Noortgate & De Boeck, 2005) to explain multilevel-DIF and PPD-variance:

\[
\beta_{qki} \sim \text{Norm}(\gamma_{0qi} + \sum_{j=1}^{J} \gamma_{jqi} X_{jk}, \phi_{qi}^2),
\]

\[
\gamma_{0qi} \quad \text{Regression intercept}
\]

\[
\gamma_{jqi} \quad \text{Effect of the } j \text{ th covariate } X \text{ on } \beta_{qki}
\]
Exemplary

APPLICATION TO EMPIRICAL DATA
Empirical data example

Estimate items' variability due to the instructional context (i.e. classes)

Judge items’ global and differential sensitivity based on statistical significance of variance components

Relate the estimated variability in classroom-specific item difficulties to measures of instruction (controlling for SES)
IGEL-Study (Hardy et al., 2011)

“Individual Support and Adaptive Teaching Methods in Elementary School”

• Quasi-experimental intervention study in German elementary school science education (grade level 3)

• Teachers were trained in one of three adaptive teaching methods (or control),

  a) Peer Learning,
  b) Scaffolding,
  c) Formative Assessment
  d) Control group (Parental Counseling)

• and implemented these methods in a prestructured curriculum on “Floating & Sinking”

• N = 901 students in 53 classes
Content Knowledge Test on Floating & Sinking

- Administered immediately before and after instruction (time span: 3 weeks)
- Nine items in common two both time points
- Scoring followed students’ conceptual understanding in “floating & sinking” (Kleickmann et al., 2010) and closely aligned to the intended curriculum:
  - Naïve conceptions (e.g. “a stone sinks because it is so heavy.”)
  - Explanations of everyday life (e.g. “Things made of stone do always sink”)
  - Scientific explanations (e.g. “The stone sinks because it is heavier than the same amount of water.”)
Instructional Measures

Three measures of instruction as predictors in latent regression:

a) A measure of the degree of implementation of the curriculum,

b) A measure of how teaching was enacted, that is, belonging to the formative assessment group, and

c) A measure of the quality of the enactment (cognitive activation; Klieme, Pauli, & Reusser, 2009).
Teaching Quality

Framework for teaching quality based on three dimensions (Klieme et al., 2009):

a) Classroom management focuses on classroom rules and procedures, coping with disruptions, and smooth transitions
b) Supportive climate covers specific aspects of the teacher-student relationship (e.g., positive and constructive teacher feedback)
c) Cognitive activation integrates challenging tasks, the exploration of concepts, ideas, and prior knowledge, and Socratic Dialogue practice as key features
Teaching Quality

Framework for teaching quality based on three dimensions (Klieme et al., 2009):

a) Classroom management focuses on classroom rules and procedures, coping with disruptions, and smooth transitions

b) Supportive climate covers specific aspects of the teacher-student relationship (e.g., positive and constructive teacher feedback)

c) Cognitive activation integrates challenging tasks, the exploration of concepts, ideas, and prior knowledge, and Socratic Dialogue practice as key features

⇒ In accordance with other international theoretical models and empirical findings (Baumert et al., 2010; Pianta & Hamre, 2009)
Teaching Quality

Framework for teaching quality based on three dimensions (Klieme et al., 2009):

a) Classroom management focuses on classroom rules and procedures, coping with disruptions, and smooth transitions

b) Supportive climate covers specific aspects of the teacher-student relationship (e.g., positive and constructive teacher feedback)

c) Cognitive activation integrates challenging tasks, the exploration of concepts, ideas, and prior knowledge, and Socratic Dialogue practice as key features

⇒ Predictive for student outcomes (e.g., Klieme & Rakoczy, 2008; Kunter, Baumert, & Köller, 2007)
Supportive Climate (w)  
Classroom Management (w)  
Cognitive Activation (w)  

Supportive Climate (b)  
Classroom Management (b)  
Cognitive Activation (b)  

(Fauth et al., 2014)
RESULTS
Results – Item Sensitivity

<table>
<thead>
<tr>
<th>Differential sensitivity (PPD-variance)</th>
<th>Global sensitivity (Average PPD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>low 0</td>
</tr>
<tr>
<td>high</td>
<td>high 3</td>
</tr>
</tbody>
</table>

Lawrence, 14.11.2013 | A psychometric framework | Naumann | First International Conference on Instructional Sensitivity
Results – Item Sensitivity

<table>
<thead>
<tr>
<th>Differential sensitivity (PPD-variance)</th>
<th>Global sensitivity (Average PPD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>low 0</td>
</tr>
<tr>
<td>high</td>
<td>high 3 6</td>
</tr>
</tbody>
</table>

⇒ All items became easier over time (on average)
Results – Item Sensitivity

<table>
<thead>
<tr>
<th>Differential sensitivity (PPD-variance)</th>
<th>Global sensitivity (Average PPD)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>low</td>
</tr>
<tr>
<td>low</td>
<td>0</td>
</tr>
<tr>
<td>high</td>
<td>0</td>
</tr>
</tbody>
</table>

⇒ Six items’ change in difficulty varied across classes
Results – Item Sensitivity

<table>
<thead>
<tr>
<th>Differential sensitivity (PPD-variance)</th>
<th>Global sensitivity (Average PPD)</th>
</tr>
</thead>
<tbody>
<tr>
<td>low</td>
<td>low: 0, high: 3</td>
</tr>
<tr>
<td>high</td>
<td>low: 0, high: 6</td>
</tr>
</tbody>
</table>

⇒ Combined information: three items globally sensitive, six items globally and differentially sensitive
Results – Latent Regression

• no relationship between degree of implementation and baseline item difficulties or change.

• a negative impact of formative assessment on two items’ change in difficulty across time points (95% certainty that $\gamma \neq 0$).

• a negative impact of teaching quality on pretest multilevel-DIF and change of trichotomous items’ highest score categories (95% certainty that $\gamma \neq 0$).
Summary

• A psychometric framework for modeling instructional sensitivity based on how tests and items reflect variability due to the instructional context was derived.

• Within this framework, three perspectives on instructional sensitivity can be distinguished, each related to different variance components.

• Based upon the framework, we derived a model-based approach to the instructional sensitivity of items.

• Items’ sensitivity was found to be related to instructional measures.
Limitations & Future Work

• We assumed that meaningful differences in instruction received by students are due to classroom membership, but other grouping variables might also be plausible.

• We focused on item difficulties and two time points only.

• We did not evaluate the practical relevance of (statistical) item sensitivity.

• Although (differential) sensitivity was found to be related to instructional measures, there’s still variance left unexplained.
Thank you!

Corresponding author:
Alexander Naumann
E-Mail: naumanna@dipf.de
## ML-DIF & PPDI for IGEL data

### IGEL data: Estimation results for PPDI and posttest ML-DIF

<table>
<thead>
<tr>
<th>Item</th>
<th>Difficulty Pre</th>
<th>Difficulty Post</th>
<th>PPDI</th>
<th>M (SD)</th>
<th>95% BCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2.54</td>
<td>-1.34</td>
<td>-3.89</td>
<td>0.19 (.12)</td>
<td>[0.01, 0.46]</td>
</tr>
<tr>
<td>2</td>
<td>2.19</td>
<td>-1.61</td>
<td>-3.82</td>
<td>0.19 (.12)</td>
<td>[0.01, 0.45]</td>
</tr>
<tr>
<td>3</td>
<td>2.11</td>
<td>0.19</td>
<td>-1.91</td>
<td>0.32 (.12)</td>
<td>[0.11, 0.62]</td>
</tr>
<tr>
<td>4</td>
<td>3.56</td>
<td>1.85</td>
<td>-1.71</td>
<td>0.72 (.31)</td>
<td>[0.30, 1.48]</td>
</tr>
<tr>
<td>5</td>
<td>1.46</td>
<td>0.49</td>
<td>-0.97</td>
<td>0.51 (.19)</td>
<td>[0.23, 0.97]</td>
</tr>
<tr>
<td>6</td>
<td>0.92</td>
<td>0.01</td>
<td>-0.92</td>
<td>0.49 (.17)</td>
<td>[0.24, 0.90]</td>
</tr>
<tr>
<td>7</td>
<td>0.95</td>
<td>-1.85</td>
<td>-2.80</td>
<td>0.21 (.13)</td>
<td>[0.01, 0.49]</td>
</tr>
<tr>
<td>8</td>
<td>3.59</td>
<td>1.57</td>
<td>-2.02</td>
<td>0.45 (.19)</td>
<td>[0.16, 0.89]</td>
</tr>
<tr>
<td>9</td>
<td>4.97</td>
<td>3.01</td>
<td>-1.96</td>
<td>0.46 (.27)</td>
<td>[0.09, 1.15]</td>
</tr>
</tbody>
</table>

*Note.* M = posterior mean; SD = standard deviation of the posterior mean; BCI = Bayesian credible interval.
LML-DIF model for IGEL data

IGEL data: Estimation results for the LML-DIF model

<table>
<thead>
<tr>
<th>Item</th>
<th>Average PPD</th>
<th>PPD-variance</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M (SD)</td>
<td>95% BCI</td>
</tr>
<tr>
<td>1</td>
<td>-3.98 (0.18)</td>
<td>[-4.35, -3.62]</td>
</tr>
<tr>
<td>2</td>
<td>-3.95 (0.18)</td>
<td>[-4.31, -3.60]</td>
</tr>
<tr>
<td>3</td>
<td>-1.90 (0.16)</td>
<td>[-2.21, -1.58]</td>
</tr>
<tr>
<td>4</td>
<td>-1.51 (0.25)</td>
<td>[-2.01, -1.04]</td>
</tr>
<tr>
<td>5</td>
<td>-1.03 (0.16)</td>
<td>[-1.34, -0.73]</td>
</tr>
<tr>
<td>6</td>
<td>-0.95 (0.14)</td>
<td>[-1.23, -0.67]</td>
</tr>
<tr>
<td>7</td>
<td>-2.89 (0.16)</td>
<td>[-3.20, -3.00]</td>
</tr>
<tr>
<td>8</td>
<td>-2.03 (0.29)</td>
<td>[-2.67, -1.53]</td>
</tr>
<tr>
<td>9</td>
<td>-1.68 (0.38)</td>
<td>[-2.54, -1.00]</td>
</tr>
</tbody>
</table>

Note. $M = \text{posterior mean}; \ SD = \text{standard deviation of the posterior mean};\ BCI = \text{Bayesian credible interval}$
Estimation

• Estimation in WinBUGS (Spiegelhalter et al., 2003) in a Bayesian framework

• Non-informative prior-distributions on parameters (Gelman & Hill, 2007):
  
  ➢ Normal distributions for ability and difficulty parameters
  ➢ Uniform distributions for parameter standard deviations

• Identification of IRT models by adjusting the means of the latent ability distributions to zero (Bafumi et al., 2004)
Item-statistical Approaches

Various statistical indices (see Polikoff, 2010), with unique perspectives on sensitivity:

1) Differences *between time points*
Differences between time points

Pretest-Posttest-Difference Index (PPDI; Cox & Vargas, 1966)

⇒ Change in item difficulty between pretest and posttest:

\[ \text{PPDI} = \text{Difficulty}_{\text{post}} - \text{Difficulty}_{\text{pre}} \]
Item-statistical Approaches

Various statistical indices (see Polikoff, 2010), with unique perspectives on sensitivity:

1) Differences between time points (e.g., PPDI)

2) Differences between groups
Differences between groups

**Differential item functioning** (DIF; Holland & Wainer, 1993) within tests after a treatment (e.g., Linn & Harnisch, 1983)

⇒ Items with uniform DIF for groups with different educational experiences are instructionally sensitive
Differences between groups

Multilevel Differential Item Functioning

Assumption: Differences in instruction are due to classroom-membership

Multilevel-DIF-model (ML-DIF; Meulders & Xie, 2004) with responses nested in items, students and classes (Robitzsch, 2009):

\[
\text{logit}[p(X_{kvi} = 1)] = \theta_k + \theta_{vk} - \beta_{ik}.
\]

\(\theta_k\): Average ability of class \(k\)

\(\theta_{vk}\): Deviation of person \(s\) from its class \(k\) average ability

\(\beta_{ik}\): Difficulty of item \(i\) in class \(k\)
Differences between groups

Multilevel Differential Item Functioning

\[
\text{logit}[p(X_{kvi} = 1)] = \theta_k + \theta_{vk} - \beta_{ik}
\]

\[
\theta_k \sim \text{Norm}(\mu, \tau^2)
\]

\[
\theta_{vk} \sim \text{Norm}(\mu_k, \sigma^2)
\]

\[
\beta_{ik} \sim \text{Norm}(\beta_i, \nu_i^2)
\]
Differences between groups

Multilevel Differential Item Functioning

\[
\text{logit}[p(X_{kvi} = 1)] = \theta_k + \theta_{vk} - \beta_{ik}
\]

\[
\theta_k \sim \text{Norm}(\mu, \tau^2)
\]

\[
\theta_{vk} \sim \text{Norm}(\mu_k, \sigma^2)
\]

\[
\beta_{ik} \sim \text{Norm}(\beta_i, v^2_i)
\]

⇒ Variation of item difficulty across classes as indicator for instructional sensitivity
Item-statistical Approaches

Various statistical indices (see Polikoff, 2010), with unique perspectives on sensitivity:

1) Differences **between time points** (e.g., PPDI)

2) Differences **between groups** (e.g., multilevel-DIF)
Item-statistical Approaches

Various statistical indices (see Polikoff, 2010), with unique perspectives on sensitivity:

1) Differences between time points (e.g., PPDI)

2) Differences between groups (e.g., multilevel-DIF)

3) Differences between groups and time points
Results – Instructional Sensitivity

<table>
<thead>
<tr>
<th>Item</th>
<th>Cognitive Activation</th>
<th>Implementation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$M$ ($SD$)</td>
<td>$95%$ BCI</td>
</tr>
<tr>
<td>1</td>
<td>0.11 (0.14)</td>
<td>$[-0.38, 0.15]$</td>
</tr>
<tr>
<td>2</td>
<td>0.05 (0.14)</td>
<td>$[-0.22, 0.33]$</td>
</tr>
<tr>
<td>3.1</td>
<td>$-0.20$ (0.14)</td>
<td>$[-0.47, 0.07]$</td>
</tr>
<tr>
<td>3.2</td>
<td>$-0.56$ (0.18)</td>
<td>$[-0.90, -0.21]$</td>
</tr>
<tr>
<td>4</td>
<td>0.00 (0.15)</td>
<td>$[-0.27, 0.29]$</td>
</tr>
<tr>
<td>5</td>
<td>$-0.14$ (0.14)</td>
<td>$[-0.42, 0.13]$</td>
</tr>
<tr>
<td>6</td>
<td>$-0.11$ (0.14)</td>
<td>$[-0.39, 0.18]$</td>
</tr>
<tr>
<td>7.1</td>
<td>$-0.28$ (0.15)</td>
<td>$[-0.28, 0.01]$</td>
</tr>
<tr>
<td>7.2</td>
<td>$-0.58$ (0.21)</td>
<td>$[-0.97, -0.18]$</td>
</tr>
</tbody>
</table>

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