Impact of Item Phrasing on Detection of Differential Item Functioning: An Empirical Study of the Chinese General Self-Concept Scale

Rationale

In order to avoid response bias such as agreement bias or acquiescence, it was a common recommendation that both positively- and negatively-worded items should be included in survey instruments (e.g., Anastasi, 1982; Nunnally, 1978;). More recently, many empirical studies found that the practice of using positively- and negatively-worded items may introduce systematic measurement errors that disrupt analyses and interpretations of the results (DiStefano & Motl, 2006; Quilty, Oakman, & Risko, 2006). These phenomena are referred to as method effects. So far, few related research has been conducted to examine if item phrasing affects the detection of differential item functioning (e.g., Fletcher & Hattie, 2005). This study will focus on the impact of positively- and negatively-worded items on detection of differential item functioning (DIF). Three approaches, including confirmatory factor analysis (CFA), item response theory (IRT), and the logistic regression method, are used to detect DIF items in this study. A CFA-based approach plays a key role for detecting DIF item due to its availability for incorporating method effects into the models.

Research Questions

The major interest in this study is to explore whether DIF items exhibited in the Chinese general self-concept (CGSC) scale are affected by item wording. The more specific question is whether more negatively-worded items are tapped as DIF items. The second research question is to examine whether three different approaches used to detect DIF items produce consistent results. As far as DIF detection, we intend to investigate DIF items on the Chinese general self-
concept (CGSC) scale across male and female students as well as across the third and sixth grades students.

Methods

Participants

Stratified sampling was used to sample students by classroom in 10 schools from four cities in Taiwan. Participants consisted of 752 elementary school students. There were 324 (43%) students in grade 3 and 428 (57%) students in grade 6. The sample was 51.9% male and 48.1% female.

Instrument

There were eight items in a Chinese general self-concept (CGSC) scale (refer to Table 1) for elementary students. Considering Taiwan’s elementary school context, the original general self-concept scale was created in Chinese. The 8-item Chinese general self-concept scale contained four positively worded items and four negatively worded. Students responded to the items on this questionnaire using a 4-point scale to rate the extent of their agreement (1=strongly disagree; 2=disagree; 3=agree; 4=strongly agree). Negatively worded items were reverse-scored.

Statistical Analysis

The strategy, called the free baseline with Bonferroni correction (Stark, Chernyshenko, & Drasgow, 2006), was applied for CFA- and IRT-based DIF detection. In addition, Zumbo’s (1999) ordinal logistic regression (OLR) method with corresponding effect size statistic (Jodoin & Gierl, 2001) was also applied to detect DIF items. When an item was flagged as a DIF item, a graphic analysis was conducted to better understand the DIF item by using Ramsay’s TestGraf (2001).
Finding unbiased items as a referent. To find an unbiased item as a referent is a crucial step for DIF detection. In order to acquire an unbiased item as a referent for a series of DIF comparisons, an approach of the constrained baseline versus augmented model comparisons was used for CFA-based DIF detection. Two rather than one unbiased items with the highest loading on the common factor would be found (Stark, Chernyshenko, & Drasgow, 2006). Each unbiased item was used to make a series of DIF comparisons separately.

A sequences of DIF comparisons. Mean and covariance structure (MACS; So¨rbom, 1974) as a confirmatory factor analysis (CFA) method was applied using Mplus version 4.2 (Muthén & Muthén, 1998-2004) and was a major approach for DIF detection. Based on the previous research using the same data, one general factor with correlated uniquenesses for the negatively worded items was determined to have the best fit. The fit indices were CFI=.969, RMSEA=.044, and SRMS=.026 (Chen, Rendina-Gobioff, & Dedrick, 2007). The MACS approach along with one general factor with correlated uniquenesses for the negatively worded items sets all the items except for a referent item to be freely estimated across groups as a baseline model. The loading (discrimination) of a referent item is constrained to be 1 and the intercept (location) is set to be equal across groups in the baseline model. Then a nested model is set by constraining an item at a time that involves its loading and intercept to be equal simultaneously across groups. A comparison between the baseline and nested models is made to determine if the constrained item exhibits DIF by checking chi-square difference with a Bonferroni critical $p$ value (Maydeu-Olivares & Cai, 2006; Stark, Chernyshenko, & Drasgow, 2006).

The same free-baseline strategy was applied for IRT-based DIF detection. MULTILOG software (Thissen, 1991) with the graded response model (Samejima, 1969) was applied to
perform IRT-based DIF detection in this study. For the IRT approach, the baseline model was set to be freely estimated for all the items except for the referent item with its equal discrimination and location parameters across reference and focal groups. Like the CFA approach, the constrained model was that an item at a time was set to be equal discrimination and location parameters across groups. The likelihood ratios (LR) among a series of comparisons between the baseline and constrained models were checked with the Bonferroni critical \( p \) values to identify DIF items.

In Zumbo’s OLR DIF approach, total scale score for each examinee, a grouping variable, and an interaction between total score and a grouping variable are used as independent variables to linearly predict examinee item response. Chi-square changes with two degree of freedom and effect size measure were checked simultaneously to tap DIF items. Items would be tapped as DIF if \( p \) values for chi-square change are less than .01 and effect size \( (R^2) \) values are greater than .035 (Jodoin & Gierl, 2001; Zumbo, 1999).

**Educational Contribution**

CFA-based method incorporating method effects into the models would be more appropriate way to detect DIF items on the instrument that involves positively- and negatively-worded items. This study is a unique demonstration under this circumstance. The findings from this study will make contributions to the empirical instrument development and application of the DIF detecting methods. The entire analyses and the written paper will be finished by January 2008.
Table 1

*Chinese General Self-Concept Scale*

<table>
<thead>
<tr>
<th>Item</th>
<th>Wording</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC 1</td>
<td>Generally speaking, I like myself.</td>
</tr>
<tr>
<td>SC 2</td>
<td>I feel I am a happy person.</td>
</tr>
<tr>
<td>SC 3</td>
<td>I am satisfied with myself.</td>
</tr>
<tr>
<td>SC 4</td>
<td>I believe I will be very successful.</td>
</tr>
<tr>
<td>SC 5*</td>
<td>I do not do most things well.</td>
</tr>
<tr>
<td>SC 6*</td>
<td>I feel I am not a promising person.</td>
</tr>
<tr>
<td>SC 7*</td>
<td>I feel my academic performances are not as good as others.</td>
</tr>
<tr>
<td>SC 8*</td>
<td>My behavior always doesn’t meet others’ expectation.</td>
</tr>
</tbody>
</table>

*Note.* Negatively worded items are marked with an *.
References


