National and international educational surveys are important empirical tools for monitoring student achievement in particular content domains such as mathematics, science, and reading. For instance, the Program for International Student Assessment (PISA) (http://nces.ed.-gov/assessments/pisa/) and the Trends in International Mathematics and Science Study (TIMSS) (http://nces.ed.gov/timss/) are well-established international surveys while the National Assessment of Educational Progress (NAEP) (http://nces.ed.gov/nationsreportcard/) is a well-established national survey in the United States whose technical developments have driven many of the current standards in like surveys around the world (e.g., Mislevy, 1991; Mislevy, Beaton, Kaplan, & Sheehan, 1992; Mislevy, Johnson, & Muraki, 1992). There also exist numerous independent educational surveys conducted by provinces or states that are implemented at regular intervals.

These educational surveys are designed to support system-wide accountability systems, which requires that any inferences about mean performance differences across groups of students can be made reliably and validly (Rutkowski, Gonzalez, Joncas, & von Davier, 2010; von Davier, Sinharay, Oranje, & Beaton, 2006). Statistically, the examination of differential item functioning (DIF) is regarded as an important component in this overall process (Mapuranga, Dorans, & Middleton, 2008) and is reflected in various quality assurance statements in the Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Associating, & National Council on Measurement in Education, 1999).

Our goal for this paper is to introduce key concepts for DIF detection as well as six model specifications under a unified estimation framework for DIF detection within educational surveys, which is the framework of hierarchical generalized linear models (HGLMs) (e.g., Kamata, 2001; Prowker & Camilli, 2007; Raudenbush & Bryk, 2002). We investigate the practical utility of the six models using a small scale simulation study and demonstrate their use with real data from TIMSS 2007.

We have organized the paper as follows. In the first section, we provide an overview of key concepts for DIF detection, whether in educational surveys or other application contexts. In the second section we discuss important features of data collection procedures in educational surveys and the way in which they impact parameter and standard error / variance estimation in parametric statistical models. We then provide a rationale for a designed-based estimation process of DIF effects as implemented in the HGLM framework. In the third section, we conduct a small simulation study to investigate the DIF detection ability of six different HGLMs. In the fourth section we apply these models to a subset of the 2007 TIMSS data. We close the paper with a summary and brief discussion of key findings.

Basic Concepts in DIF Detection

The literature on the theory and practice of DIF is vast and we recommend the following sources for further reading. For a general overview of DIF methods, we recommend, for example, Ferne and Rupp (2007), Mapuranga, Dorans, and Middleton (2008), Osterlind (2009), and Zumbo (1999). For an overview of key implications of the complex sampling designs for educational surveys for secondary analyses generally, we recommend the articles by Rutkowski et al. (2010), von Davier, Gonzalez, and Mislevy (2010), and von Davier et al. (2006). For examples of how DIF analyses can be conducted within unified estimation frameworks for parametric statistical models, which are most appropriate for educational survey data, we recommend Binici (2008), Kamata (2001), Kamata and Binici (2003), Kim (2003), and Prowker and Camilli (2007).

DIF vs. Impact

DIF occurs when different *item response probabilities* are observed for students with identical levels of proficiency (i.e., equal values on the observed or latent variables in the statistical model) who belong to at least two distinct groups. Importantly, DIF is not the same as true mean differences proficiency, which are known as *impact* in the literature (Camilli & Shepard, 1994; Hauger & Sireci, 2008). Put differently, DIF reflects *conditional performance differences* whereas impact reflects *unconditional / marginal performance differences*. DIF can be viewed as caused by distributional differences on a variable that reflects a secondary construct that an instrument is not intended to measure but that the items that display DIF appear to require during responding. Hence, biased inferences for items displaying DIF will only result if the students in different groups actually differ in their proficiency distribution on the secondary variables (Shealy & Stout, 1993).

Uniform vs. Non-uniform DIF

Researchers distinguish between two general types of DIF, which are known as *uniform DIF* and *non-uniform DIF*. Uniform DIF refers to the condition when one of the groups – typically the one denoted as the *reference group* – is predicted to perform either better or worse than the other group(s) – typically denoted as the *focal group(s)* – throughout the entire proficiency range. In contrast, non-uniform DIF exists when there is a point on the proficiency continuum where the predicted difference in performance across the groups reverts. This is important to remember because certain statistical approaches to DIF detection do not allow for the detection of non-uniform DIF. Importantly, this is equally true of some methods that employ *observed-score matching* (e.g., the Mantel-Haenszel method from the area of multivariate statistics / categorical data analysis) and