Journal of Vocational Behavior 74 (2009) 295–307

Contents lists available at ScienceDirect

Journal of Vocational Behavior

journal homepage: www.elsevier.com/locate/jvb

Gender bias and construct validity in vocational interest measurement: Differential item functioning in the Strong Interest Inventory ☆

Sif Einarsdóttir [a,*], James Rounds [b]

[a] Faculty of Social and Human Sciences, University of Iceland, Gimli Sæmundargötu 2, 101 Reykjavík, Iceland
[b] Department of Psychology, University of Illinois at Urbana-Champaign, 603 East Daniel Street, Champaign, IL 61820, USA

ARTICLE INFO

Article history:
Received 25 September 2008
Available online 22 January 2009

Keywords:
Interest inventories
Vocational interests
Interest measurement
Gender bias
Gender-fairness
Differential item functioning
Construct validity

ABSTRACT

Item response theory was used to address gender bias in interest measurement. Differential item functioning (DIF) technique, SIBTEST and DIMTEST for dimensionality, were applied to the items of the six General Occupational Theme (GOT) and 25 Basic Interest (BI) scales in the Strong Interest Inventory. A sample of 1860 women and 1105 men was used. The scales were not unidimensional and contain both primary and minor dimensions. Gender-related DIF was detected in two-thirds of the items. Item type (i.e., occupations, activities, school subjects, types of people) did not differ in DIF. A sex-type dimension was found to influence the responses of men and women differently. When the biased items were removed from the GOT scales, gender differences favoring men were reduced in the R and I scales but gender differences favoring women remained in the A and S scales. Implications for the development, validation and use of interest measures are discussed.

1. Introduction

Since the pioneering work of Strong (1943), researchers have reported large differences in the vocational interests of men and women. Women tend to express interests that fit their traditional gender role, whereas men express more interests in domains that have been considered masculine (Betz & Fitzgerald, 1987; Hackett & Lonborg, 1994). Research and debate on the issue of gender differences and possible bias in interest measurement reached a peak in the 1970s (then referred to as sex-bias and fairness see Diamond, 1975; Tittle & Zytowski, 1978). Much of the debate centered on the Strong Interest Inventory, one of the oldest and most widely used interest measures. The debate resulted in new perspectives and guidelines to reduce bias in interest inventories based on the psychometric knowledge and techniques of the time. After 1980, the sex bias debate seemed to fade away, but as is evident in the major interest inventories used today, a common agreement on how to best resolve gender bias in interest measurement has not been reached (cf. Donnay, Morris, Schaubhut, & Thompson, 2005; Harmon, Hansen, Borgen, & Hammer, 1994; Holland, Powell, & Fritzsche, 1994; Swaney, 1995).

Recent expansion of sophisticated psychometric modeling grounded in item response theory (IRT) has provide new methods to address the issue of bias and fairness (Bolt & Rounds, 2000). These methods and developments in validity theory may also offer new insights into the nature of the construct of vocational interest, especially those factors that differently affect

☆ We thank David A. Donnay and Consulting Psychologist Press for providing the archived data and Daniel Bolt and Terry A. Ackerman for technical assistance. We also thank Patrick I. Armstrong and especially, Christopher A. Moyer for their close reading of an earlier draft of the manuscript. This article is based on a first authors dissertation submitted to the University of Illinois at Urbana-Champaign. This research was supported in part by a grant from Iceland University of Education.

* Corresponding author. Fax: +354 552 1331.
E-mail address: sif@hi.is (S. Einarsdóttir).

0001-8791/$ - see front matter © 2009 Elsevier Inc. All rights reserved.
doi:10.1016/j.jvb.2009.01.003

the responses of men and women (Smith, 2002). The purpose of this study is to apply differential item functioning (DIF) techniques, to examine gender bias in items, explore its sources and influence on gender differences detected in the General Occupational Theme (GOT) scales and the Basic Interest (BI) scales of the Strong Interest Inventory (SII).

1.1. Gender differences and bias in vocational interests

Gender differences in the responses to interest inventories have been observed both at the scale and item level. Women tend to score higher on Holland's Artistic, Social and Conventional types and men score higher on the Realistic, Investigative and Enterprising types (Betz & Fitzgerald, 1987; Hackett & Lonborg, 1994). One of the main concern in the sex-bias and fairness debate in the seventies was that differences between men and women in vocational interest assessment can have consequences for individuals seeking career counseling and for society as a whole. In particular, scale level differences can lead to sex-restrictive career options being suggested to students (Cole & Hanson, 1975; Prediger & Hanson, 1974). Interest inventories may serve to maintain and perpetuate the limited range of occupations considered appropriate for men and women.

Two main positions were taken on the issue. Prediger and Cole (1975) stated that the primary purpose of using an interest inventory is occupational exploration (also, see Prediger, 1977). Since differences between men and women are extraneous to the goal of occupational exploration, these differences should be removed from interest measures. In contrast, Gottfredson and Holland (1978) argued that because the constructs measured are dependent on differential experiences of men and women, the removal of sex differences from interest scores would decrease the predictive validity of the measure. These positions foreshadowed the wider debate in psychology on construct validity, measurement bias and its social consequences (Cole & Moss, 1989; Linn, 1997; Messick, 1989, 1995; Shepard, 1997).

A consensus on how to define sex-bias or what now would be termed gender-bias in interest measures has yet to be reached. Nevertheless, several strategies have been used to eliminate bias and sex-restrictiveness. In the 1974 revision of the Strong and construction of one form for both women and men, Campbell (1974) changed the wording of items (e.g., policeman to police officer) and used a variety of norms for reporting standard scores (i.e., both same and combined sex-norms). Each revision of the Strong since 1974 has focused on removing sex-role bias in items and norming the scales with both female and male samples (Hansen & Campbell, 1985; Harmon et al., 1994). The most recent revisions uses only combined norms (Donnay et al., 2005). Another strategy has been to remove items showing large gender differences during test development. For the Strong, items that show large gender differences in endorsement have been eliminated during the 1994 revision (Harmon et al., 1994). These strategies, indicative of a classical test theory approach for reducing bias, are necessary but not sufficient to optimally reduce gender bias in interest measures. The removal of items showing gender differences can be confounded by real group differences in the trait being measured.

The lack of consensus about how to deal with gender differences is not surprising because it has not yet been adequately explained why measured interests are different for men and women. It is possible that these differences may be partly explained by item bias in interest inventories and the influence of construct-irrelevant factors (Messick, 1989) on the scales used in counseling. Fouad and Walker (2005) suggested that perceived barriers and opportunities may be such a factor influencing the assessment of interests of ethnically diverse clients. They examined racial/ethnic group differences in the SII using differential bundle functioning (DBF). Large racial/ethnic DBF was detected, implying that the items were influenced by other constructs in addition to the traits the Holland scales were designed to measure. This is also likely to be the case for men and women who work in occupations that are largely sex-segregated. Numerous barriers for entering certain types of jobs have been identified for women (Betz, 1994). It is possible that gendered opportunity structure and stereotyping of the job market differently influences the interest trait being measured for men and women.

Aros, Henly, and Curtis (1998) showed that occupational stereotypes influence the responses to items in interest inventories. They used DIF, specifically Mantel-Haenszel log-odds ratios, to explore gender differences in responses to 28 occupational title items from the GOT scales measuring the six RIASEC interest types in the SII. Gender-related DIF was detected on most of the items. However, they only explore DIF in a few items and focused their investigation on one item type. Occupational titles, for example, may be more susceptible to stereotyping than activities (Crites, 1969; Kuder, 1977; Osipow, 1983). The present study examines the full range of items used in interest inventories that may influence the gender-related differences found at the scale level.

1.2. Multidimensional model of DIF applied to interest measurement

Item response theory differs from classical test theory by modeling the interaction of the person and the individual items to a latent trait. By modeling responses in terms of their relations to a common underling trait, IRT models have an important feature that allows us to determine if people from two groups respond differently to the same item given that they have the same level of a trait (Bolt & Rounds, 2000; Embretson & Reise, 2000). For example, by using DIF techniques it can be determined if women and men who are equally realistic in their interests (a trait being measured) are as likely to endorse a highly sex-stereotyped occupations like "auto racer" or "nurse."

A theoretical framework called multidimensional item response theory has been developed to account for how item bias as defined by DIF relates to item and test validity (Ackerman, 1992; Bolt & Stout, 1996; Kok, 1988; Shealy & Stout, 1993). The underlying mechanism producing the DIF is addressed by making a distinction between the main trait that the researcher intends to measure, alternately called the *target trait* or *primary dimension*, and other factors influencing test performance