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 not intended to be measured by the test, such as *nuisance determinants* and *secondary dimensions* (Roussos & Stout, 1996a; Shealy & Stout, 1993). The construct validity of a test is threatened if it contains items that capture traits or dimensions the test developer does not intend to measure by the test (construct irrelevant). Item bias can arise if two groups differ in their underlying distribution of this extraneous trait or secondary dimension that is not intended to be captured by the scale (Ackerman, 1992; Bolt & Stout, 1996). In interest measurement, items may function differently for men and women when the two groups differ, for example, in their distribution of a sex-type dimension (a secondary dimension) found to underlie responses to the interest items in the Strong (Aros et al., 1998; Einarsdóttir & Rounds, 2000).

When applying an IRT framework and the multidimensional model of DIF to vocational interest measures, questions arise about the primary traits that are being measured and the dimensionality of the scales. Item bias is determined in reference to a criterion internal to the test. A collection of items defining the internal criterion is referred to as a *valid subtest*. Determination of a valid subtest is an empirical decision based on expert opinion or external data (Shealy & Stout, 1993). In the Strong Interest Inventory the GOT scales and the BI scales are established scales that are considered valid and useful for counseling purposes especially occupational exploration (Donnay et al., 2005; Harmon et al., 1994). These two types of scales served therefore as valid subtests in the DIF analysis.

Dimensionality of a scale is a psychometric issue that is also conceptually important but neglected in the domain of vocational interest assessment. The six General Occupational Theme scales measure six broad interest types (Donnay et al., 2005; Holland, 1997). The dimensionality of the GOT and BI scales in the Strong has not been directly evaluated. In the present study Stout's (1987, 1990) conception of *essential unidimensionality* is applied because it is less restrictive and more realistic for most measurement practices than the traditional conception of local independence in IRT based model. Dimensionality and DIF analysis can give valuable insights into the concept being measured and the construct validity of the scales (Smith, 2002; Smith & Reise, 1998). Application of DIF analysis with the multidimensional model provides information on whether scale level differences are the result of *impact* or bias in interest assessment. Impact is defined as a between group difference on a construct valid trait (Ackerman, 1992).

1.3. The present study

In the present study, an IRT-based approach is used evaluate the dimensionality of the SII scales and to test whether the SII items function the same way for men and women. Our study differs from the Aros et al. (1998) study in three important ways. First, SIBTEST, a similar but more recent DIF detection method than the Mantel-Haenszel statistic, was applied. Simulation studies (Roussos & Stout, 1996b) have shown that the SIBTEST procedure detects DIF better than the Mantel-Haenszel statistic. Second, all the items used to define the GOT and the BI scales in the SII were tested, allowing an exploration of how the different item types (e.g., occupational titles, activities) function. Third, the overall influence of DIF on GOT scale scores was estimated and the construct validity of the scales purged of DIF items was evaluated. We expected gender differences to be related to the sex-typing of the occupations, as was the case in Aros et al. study. Because prestige has also been shown to influence responses to interest inventory items (Einarsdóttir & Rounds, 2000; Tracey & Rounds, 1996) both sex-type and prestige were examined as possible secondary dimensions contributing to DIF.

2. Method

2.1. Participants

The responses of 2965 college students to the Strong Interest Inventory (SII) were sampled using simple random sampling strategy from the test publisher's database (Consulting Psychologist Press) in 2000. The SII had been administered to a college student population and sent to the publisher for scoring. The sample consisted of 1860 (62.7%) women and 1105 (37.3%) men. Information about the ethnicity of the participants showed that, 2.3% identified as American Indian or Alaskan Native, 4.5% as Asian or Pacific Islanders, 5.5% as Latino, Latina/Hispanic, 64.9% as Caucasian, and 14.1% identified as multi ethnic in origin or belonging to other groups than specified. There were no statistically significant differences in the ethnic identification of women and men, $\chi^2 = 2.80$ (5, $N = 2965$), $p = .73$. The mean age of the men in the sample was 22.4 years ($SD = 6.2$) and women 24.1 years ($SD = 8.1$), with the differences in age being statistically significant $t(2766, 4) = 6.37$: $p < .001$.

A second sample was used to obtain the sex-type and prestige ratings for the occupational title items. Students taking career exploration classes at a large Midwestern University were administered the rating scales and measures in exchange for research credit. Due to the extensive length of the scales, the material was split in two. Half of the students ($N = 109$), 53 (48.6%) men and 56 (51.4%) women, completed the sex-type ratings of the occupational items. The other half of the sample ($N = 93$), 39 (41.5%) men and 55 (58.5%) women, rated the items according to their perceived prestige. The mean age in these two samples is 19.4 years ($SD = 1.2$).

2.2. Measures

2.2.1. The Strong Interest Inventory (SII)

The SII consists of 317 items pertaining to a wide variety of activities, occupations, and school subjects (Harmon et al., 1994). The inventory measures respondent's interests for each item on a three-point scale of "like," "indifferent," and "dislike." The items for the General Occupational Theme (GOT) and BI scales are weighted (+1 = like, 0 = indifferent, and −1 = dislike). The GOT and BI scales are normed on the General Reference Sample (9467 women and 9484 men) to create standard scores with $M = 50$ and $SD = 10$.

The GOT scales are designed to assess Holland's (1997) RIASEC types. The number of items (in parentheses) per GOT scale is: R (24), I (20), A (33), S (23), E (22), and C (21). The SII has 25 Basic Interest (BI) scales that assess specific content areas (e.g., agriculture, mathematics, and teaching). Each BI scale contains between 5 to 21 items, with 8 of the 25 scales having less than 10 items and 6 scales having more than 15 items. The 1-month test–retest reliabilities of the GOT scales in college student samples range from .84 to .88, and from .78 to .93 for the 25 BI scales. The Strong's GOT and BI scales have been widely studied (see Harmon et al., 1994).

2.2.2. Sex-type ratings

To obtain sex-type ratings for the 135 occupational titles, college students ($N = 109$) were asked to rate the items in terms of perceived proportion of men and women employed in them. A 7-point bipolar scale, similarly to White, Kruczek, and Brown's (1989) scale, was used, ranging from masculine (low scores) to feminine (high score). To evaluate if men and women produced similar occupational ratings, we calculated the mean ratings for the men and women and correlated these mean ratings across occupations. The mean ratings correlated .98, indicating that men and women, on average, rank order the sex-type of occupations almost identically. Studies comparing subjective method of sex-type ratings with labor market data have indicated that the two methods agree to a large extent (Cooper, Doverspike, & Barrett, 1985; White et al., 1989). To evaluate the similarity of the present sex-type ratings with labor market data, we correlated the mean sex-type ratings with data on the proportion of men and women in occupations. Approximately 70% of the occupational titles in the SII could be matched to the most recent labor market data (U.S. Department of Labor, 1998) reflecting the proportion of men and women in occupations. The 1998 labor data was selected because the proportions of men and women closely matched the time that the study sample was collected. The correlation for the 92 SII occupations that could be matched with the U.S. Department of Labor data was .86, indicating a good fit between these methods.

2.2.3. Prestige ratings

In general, studies of occupational prestige have found that prestige is a stable concept over time and groups and that there is a large degree of correspondence between different methods used to obtain prestige ratings for occupations (Chartrand, Dohm, Dawis, & Lofquist, 1987). Objective prestige ratings (Stevens & Cho, 1985) do not exist for all the occupational items in the Strong Interest Inventory. Therefore, college students ($N = 93$) were asked to rate the prestige of 135 occupations on a 7-point scale with a low score indicating low prestige and a high score indicating high prestige. We correlated these subjective ratings with Stevens and Cho's (1985) objectively derived indicator of prestige, a measure based on occupational level, education and earnings for 82 occupations that were matched. The correlation was .72, providing support that the present ratings assess prestige. The present mean ratings were also calculated by rater sex and then correlated, resulting in correlation of .93, indicating that men and women tend to assess the prestige of occupations very similarly. Subjective ratings of prestige and sex-type may be more suitable than objective ratings because they capture the stereotypical ways student think of occupations, which is important in identifying the dimensions underlying differential influences on interest responses.

2.3. Procedure

The SII item responses for the college students were obtained from the publisher's archived data. Rating scales to evaluate the prestige and sex-type of the SII items were assembled. After background information had been obtained, two samples of college students were asked to rate the 135 occupational title items from the SII either in terms of sex-type or prestige.

2.4. Analysis

2.4.1. Dimensionality of the GOT scales

To test the dimensionality of the GOT scales, Stout's (1990) DIMTEST based on the conceptualization of essential unidimensionality was applied. This method was used because it applies less restrictive model of dimensionality than the traditional local independence conception of unidimensionality and is one of few methods that has been developed for polytomous items. The Poly-DIMTEST, a non-parametric procedure for polytomous items (Nandakumar, Yu, Li, & Stout, 1998; Stout, 1987, 1990) splits the scale into two clusters of items that are as dimensionally distinct as possible. (Finch & Habing, 2007; Stout, Nandakumar, & Habing, 1996). Items that measure the same dominant dimension are first selected for the assessment subtest (AT1) either through expert opinion or exploratory factor analysis (Nandakumar & Stout, 1993; Nandakumar et al., 1998). The assessment subtest cannot contain more than one quarter of the total number of items comprising the scale being tested, and it is also recommended that the scale contain no less than 20 items. The program selects a second subset of equal number of items AT2 so that the items match the difficulty distribution of AT1. The purpose of AT2 is to correct for bias in AT1 that can arise when the remaining items comprising the third subset partitioning subscale (PT) is short. The partitioning subtest is used to assign examinees to different subgroups. Conditional covariances for each pair of items in AT1 conditioned on the PT subtest scores are calculated. The Poly-DIMTEST procedure yields Stout's T statistic, to test for unidimensionality by summarizing the conditional covariances. The T statistic approximately follows the standard normal distribution when the unidimensionality assumption holds (for more detailed description see Nandakumar et al., 1998).

2.4.2. Differential item functioning

The POLYSIB procedure is an extension of SIBTEST, a statistical procedure developed for DIF detection (Stout & Roussos, 1996), in polytomous items (Chang, Mazzeo, & Roussos, 1996). POLYSIB was used to detect DIF or differential bundle functioning (DBF) across gender. POLYSIB is a simple and easily understood method that operationalizes item bias. Additionally, the accompanying multidimensional item response theory provides a conceptual framework that ties DIF into questions about the construct validity of the scales. The Mantel-Haenszel statistic and SIBTEST are related methods based on the same general conceptualization of item bias and similar techniques have been successfully used in the two previous studies exploring DIF and DBF in interest measurement (Aros et al., 1998; Fouad & Walker, 2005). Shealy and Stout's (1993) SIBTEST is a non-parametric procedure developed to test DIF, which has the advantages of most IRT models such as statistical sophistication and modeling of item-trait relations, but bypasses the problem of fit to specific IRT functions due to its non-parametric nature.

The procedures estimate the amount of DIF in individual items or a collection of items and whether the resulting DIF/DBF statistic is different from 0. A so called valid subtest is used as an estimate of the target trait being measured and the DIF test evaluates how the items differ in their performance in the two groups that are being compared by conditioning them on the trait level of the examinees. The POLYSIB procedure results in a β statistic for DIF. The β is approximately normally distributed with mean of 0 and variance 1 when the null hypothesis of no DIF holds (for more detailed description, see Bolt & Stout, 1996; Chang et al., 1996; Shealy & Stout, 1993). As the SIBTEST, POLYSIB has a special feature to test for differential bundle functioning (DBF). DBF was designed to test for combined DIF when items are grouped together, and to evaluate if a collection of DIF items amplify or cancel out item DIF at the scale score level.

3. Results

3.1. GOT and BI mean scores

The mean gender differences for the General Occupational Theme (GOT) scales and the Basic Interest (BI) scales were examined to determine if the differences in the present sample are similar to differences detected in previous research. Table 1 shows the mean gender differences on the GOT scales, five of them being statistically significant. The differences are expressed in standard deviation units shown as Cohen's d effect sizes. Women tended to score higher on the Social, Artistic and Conventional scales and men tended to score higher on the Realistic and Investigative scales. The Realistic scale displayed the largest sex differences ($d = .85$) followed by the Artistic and Social scales (both $d = -.45$). These results largely correspond to those found in other studies (e.g., Betz & Fitzgerald, 1987; Fouad, 2002; Hackett & Lonborg, 1994).

Table 2 shows the mean BI scale scores for women and men. Out of 25 BI scales, 20 showed statistically significant differences. The largest scale score differences are for Mechanical Activity ($d = .88$) and Athletics ($d = .78$) where the men have higher scale scores. Largest gender differences favoring women were detected in four scales: Social Services ($d = -.64$), Art ($d = -.49$), Culinary Arts ($d = -.56$), and Office Services ($d = -.52$). These results are very similar to those reported for the standardization sample in the 1994 edition of the Strong (Harmon et al., 1994). In sum, the differences for the GOT and BI scales on the present sample are similar to those reported in previous research.

Table 1
Gender differences for General Occupational Theme scale scores.

GOT	Women		Men		d	t
	M	SD	M	SD		
Realistic	41.93	7.96	50.08	9.74	.85	−24.76*
Investigative	42.91	9.44	45.24	9.97	.30	−6.25*
Artistic	49.53	10.09	44.92	10.06	−.45	12.04*
Social	52.37	10.55	47.55	10.60	−.45	12.00*
Enterprising	51.27	10.11	50.34	10.44	−.09	2.36
Conventional	50.19	10.65	47.66	9.05	−.24	6.87*

Note. Women $N = 1860$, men $N = 1105$.
* $p < .008$, alpha was adjusted using Bonferroni correction to avoid inflating overall type I error rate.

Table 2
Gender differences for the Basic Interest Scales.

BI	Women		Men		*d*	*t*
	M	*SD*	*M*	*SD*		
Agriculture	43.81	8.72	46.48	9.04	.29	–7.96*
Nature	43.96	10.95	44.00	10.42	.00	–.78
Military activity	47.31	8.52	51.73	11.02	–.45	12.22*
Athletics	47.65	9.41	55.72	9.88	–.78	22.16*
Mechanical activity	42.82	7.56	51.11	9.91	–.88	25.66*
Science	43.48	9.14	47.00	9.94	.36	–9.76*
Mathematics	44.22	8.87	47.22	9.51	.32	–8.65*
Medical science	46.49	10.19	46.80	10.14	.03	–.78
Music/dramatics	50.89	10.14	47.17	9.72	–.36	9.79*
Art	50.88	10.00	45.86	10.15	–.49	13.55*
Applied art	46.25	10.45	46.19	10.32	.00	.16
Writing	46.98	10.58	43.95	10.29	–.29	7.62*
Culinary arts	53.31	9.07	47.96	9.43	–.56	15.30*
Teaching	48.87	11.25	47.07	11.55	–.16	4.17*
Social services	54.96	10.00	48.21	10.20	–.64	17.63*
Medical services	51.33	11.74	48.68	10.36	–.23	6.21*
Religious activity	49.66	10.25	48.71	10.38	–.09	2.43
Public speaking	46.56	9.40	49.28	10.28	–.27	7.35*
Law/Politics	45.78	9.81	49.43	10.44	–.36	9.56*
Merchandising	51.29	9.77	49.34	9.80	–.20	5.24*
Sales	51.12	10.08	52.57	11.00	–.14	3.66*
Organizational management	47.83	9.62	48.02	9.82	.02	–.51
Data management	45.84	9.60	47.08	9.23	–.13	3.45*
Computer activities	47.69	11.05	49.65	10.49	–.18	4.74*
Office services	53.00	10.79	47.62	8.30	–.52	14.26*

Note. Women $N = 1860$, men $N = 1105$.

* $p < .002$, alpha was adjusted using Bonferroni correction to avoid inflating overall type I error rate.

3.2. Dimensionality and differential item functioning in the GOT scales

The GOT scales represent six broad interest types. Therefore, it is likely that these scales are not essentially unidimensional. Poly-DIMTEST was used to test Stout's (1990) conception of essential unidimensionality. Dimensionality test of the GOT's aids in determining if there are secondary dimensions or nuisance determinants not intended to be measured in the scales.

To test the dimensionality of the GOT scales, items were split into scales using principal axis factor analysis that was applied to half of the male and female samples for all the items in each of the six scales. Factor analysis was used to identify as dimensionally distinct set of items as possible. This more exploratory approach was chosen because dimensionality of the RIASEC scales has not been tested before. The items loading high on the second factor (but no more than 25% of the total number of items) were chosen for the *assessment subtest* (AT1) for each scale. Poly-DIMTEST was applied to test the essential unidimensionality of the AT1 subtest with regard to the remaining items in the *partitioning subtest* (PT) for each of the GOT scales using the other half of the male and female samples (Nandakumar & Stout, 1993; Nandakumar et al., 1998). Table 3 shows Stout's *T* statistic is significant for all the six GOT scales, indicating that each of them contains dimensionally distinct sets of items. (The "% examinees included" refers to proportion of the respondents from the original sample who where retained in the analysis, an 80% retention is considered sufficient.) These results showed that the scales are not essentially unidimensional, but are instead multidimensional, and consist of two or more factors that influence item responses.

Table 3
Poly-DIMTEST results for the GOT scales by gender.

GOT	Women			Men		
	N	% examinees included	*T*	*N*	% examinees included	*T*
Realistic	764	84	6.75**	528	96	2.73**
Investigative	901	95	7.29**	489	93	6.25**
Artistic	923	99	9.88**	536	98	5.77**
Social	910	97	7.72**	557	98	6.63**
Enterprising	818	88	2.54**	471	88	1.92*
Conventional	858	94	3.79**	519	95	2.73*

* $p < .05$.

** $p < .01$.

Table 4
Number and percentage of items showing DIF for the GOT scales.

	Scale *N*	Differential item functioning (β)					
		N	%	Favor women		Favor men	
				N	*M*	*N*	*M*
Realistic	24	16	67	7	−0.18	9	0.26
Investigative	20	13	65	6	−0.15	7	0.13
Artistic	33	23	70	8	−0.41	15	0.22
Social	23	15	65	8	−0.19	7	0.23
Enterprising	22	16	73	7	−0.24	9	0.18
Conventional	21	18	86	8	−0.22	10	0.19

Note. DIF with $p < .001$.

We then proceeded with the DIF analysis, because multidimensionality indicates that the scales contain factors other than the trait they are intended to measure, which may in turn lead to gender-related DIF. In addition, the DIF method may detect meaningful item bias in the inventory, and thus offer valuable insight into the nature of the constructs being measured in those scales (Smith, 2002). The POLYSIB program (Chang et al., 1996) was applied to the male and female samples to detect DIF. The GOT RIASEC scales were treated as valid subscales with respect to the six theoretically postulated constructs that the SII is designed to measure. A summary of the POLYSIB results is shown in Table 4. Using significance level of $p < .001$ for β, the C scale has the largest proportion of items (86%) showing DIF, followed by the E (73%) and A (70%) scales. Overall, DIF was detected in 70% of all the items used to construct the RIASEC scales. The mean absolute β were .163 ($SD = .131$). Moreover, 33% of the items showed a large amount of DIF (i.e., an absolute β score higher than .200). Interestingly, there is almost equal number of items favoring men and women in each scale. The Artistic scale is an exception where more items favor men but the average DIF score for each item is smaller than for the items favoring women in the scale. In the R scale both more items favor men and the average DIF score is higher for those items. In sum, these results indicate an extensive amount of gender-related item bias in the Strong.

Specific examples of differential item functioning can be seen in Figs. 1–4. These figures show the expected score on an item for each score category on the RIASEC scale the item belongs to. The items three score-categories are coded as 0 = dislike, 1 = indifferent and 2 = like for the POLYSIB analysis. Fig. 1 displays the results for the occupation of auto racer. Participants with a total score of 23 on the R scale had for example very different expected item scores for women (0.9) and men (1.3). Figs. 1–4 show items displaying various amount of DIF. The DIF statistic β captures the vertical difference between the two lines by taking the averaged differences of each score category and summing them. The oscilliation of the curves is due to the non-parametric nature of SIBTEST. The POLYSIB procedure captures only uniform DIF, but not crossing DIF. Crossing DIF occurs when the direction of DIF is dependent on the trait level. For the case of simplification we decided to only test for uniform DIF in this study.

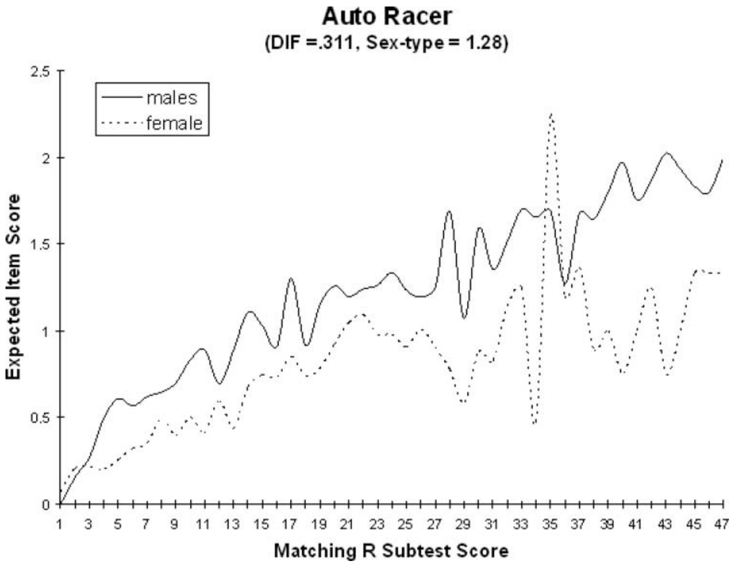

Fig. 1. An example of an item (auto racer) that showed a large amount of DIF favoring men.

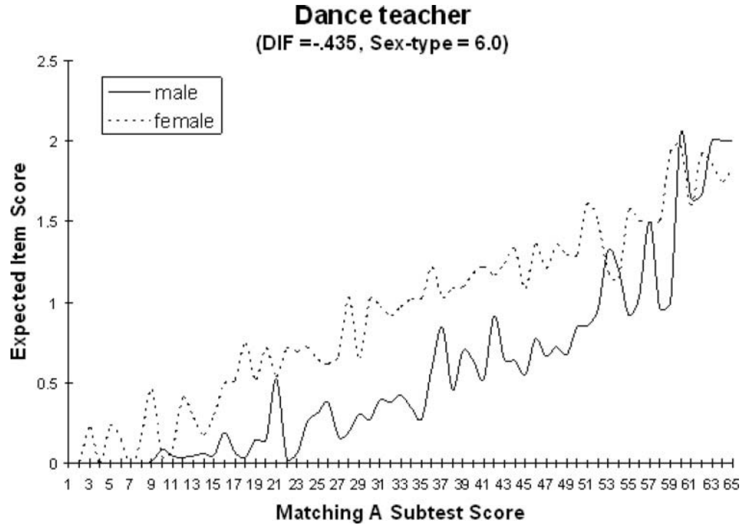

Fig. 2. An example of an item (dance teacher) that showed a large amount of DIF favoring women.

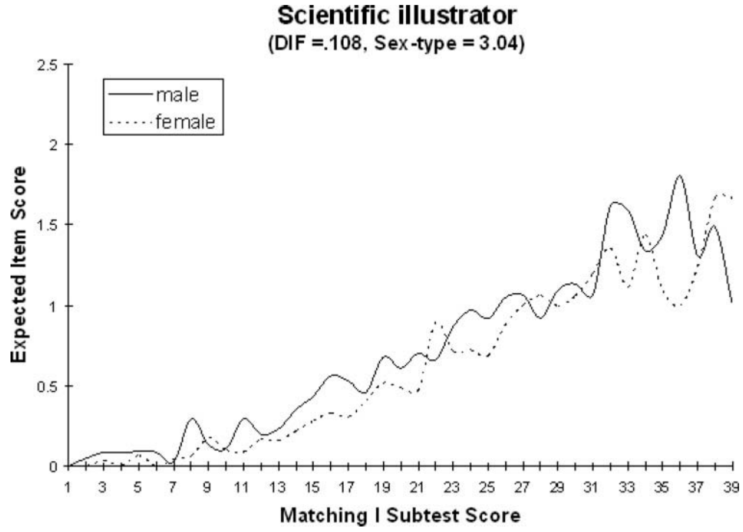

Fig. 3. An example of an item (scientific illustrator) that showed a moderate amount of DIF favoring men.

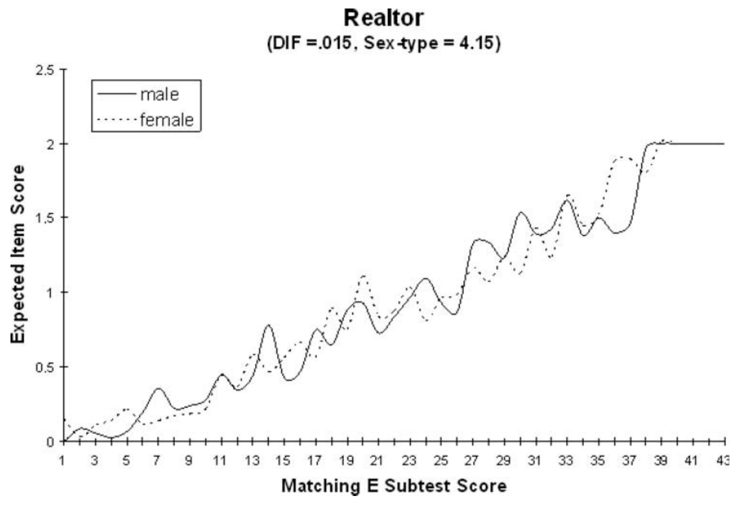

Fig. 4. An example of an item (realtor) that showed no DIF.

To examine if DIF was related to stereotypical characteristics of the items, gender related β scores were correlated with the sex-type and prestige ratings for the occupational titles. The correlation between the DIF scores and the sex-type ratings scores was −.54 ($p < .001$) and the correlation between DIF scores and prestige ratings was .24 ($p < .05$). The relatively higher negative correlation for the DIF scores and sex-type showed that a high DIF score for an item (item endorsed more frequently by men) tended to have a more masculine sex-type score. Because the sex-type and prestige of occupations is correlated the partial correlation was calculated between the DIF scores and sex-type after prestige was partialled out. This resulted in a partial correlation of −.53 almost identical to the original correlation. However, when the sex-type was partialled out of the DIF correlation with prestige the correlation dropped from .24 to .18, indicating that the relationship of prestige with the DIF is, to a small extent, due to the correlation between prestige and sex-type. Overall, the correlation analysis supports Aros et al. (1998) and Einarsdóttir and Rounds (2000) contention that sex-type is an underlying dimension or irrelevant factor that influences responses in the SII occupational title items.

3.3. Dimensionality and differential item functioning in the BI scales

The DIF analysis was also conducted on the Basic Interest scales to determine if a similar amount of sex-related DIF is present in the items contributing to scales that are more narrowly defined, and thus more likely to be essentially unidimensional. However, because the BI scales contain few items, the DIMTEST procedure could not be conducted to test their dimensionality. Therefore, factor analysis with principal axis extraction was conducted separately on data for men and women to examine the dimensionality of the 25 BI scales. Scales can be considered unidimensional if the ratio of the first to second eigenvalue is high (Hattie, 1985), and although no definitive value for this ratio has been specified, Smith (2002) has suggested that 10:1 as a reasonable guideline. Alternately, the existence of a common factor may be indicated when the first factor explains 40% of the total variance (Smith & Reise, 1998). Factor analytic results for the BI scales yielded eigenvalue ratios that ranged from 2.14 to 6.07 for women, and 2.23 to 6.50 for men. Total variance accounted for by the first factor ranged from 32.6% to 66.4% for women, and 32.1% to 70.1% for men. For the majority of the BI scales, the first factor explained 40% or more of the variance, indicating a common factor in those scales. Eigenvalue ratios, however, were not close to 10 suggesting the existence of additional minor factors in the BI scales.

DIF results for the BI scales are shown in Table 5. The percentage of items in each scale that exhibit differential functioning for men and women ranges from 36% to 83%. Scales with the highest proportion of such items were Agriculture (83%), Medical Services (83%), Organizational Management (82%), and Teaching (78%). Overall, 66% (197/298) of the items used to construct the BI scales showed DIF. The correlation between the DIF scores and the sex-type and prestige ratings for the 147 occupational title items in the BI scales was −.49 ($p < .001$) and .18 ($p < .05$), respectively.

Table 5
Number and percentage of items showing DIF for the BI.

BI	Scale *N*	Differential item functioning (β)	
		N	(%)
Agriculture	6	5	83
Nature	8	6	75
Military activity	5	2	40
Athletics	13	9	69
Mechanical activity	21	11	52
Science	18	12	67
Mathematics	11	8	73
Medical science	8	4	50
Music/dramatics	13	10	77
Art	14	10	71
Applied art	12	9	75
Writing	15	10	67
Culinary arts	7	3	43
Teaching	9	7	78
Social services	14	5	36
Medical services	12	10	83
Religious activity	7	4	57
Public speaking	10	7	70
Law/politics	14	7	50
Merchandising	12	9	75
Sales	11	8	73
Organizational management	17	14	82
Data management	17	12	71
Computer activities	5	3	60
Office services	19	11	58

Note. DIF with $p < .001$.

Table 6
The mean absolute DIF (β) by GOT and BI item type.

Item type	GOT			BI		
	N	*M*	*SD*	*N*	*M*	*SD*
Occupations	84	.165	.132	155	.160	.128
Activities	28	.166	.171	73	.137	.129
School subjects	20	.139	.077	44	.123	.088
Types of people	4	.150	.087	13	.127	.071
Total	136	.161	.132	285	.145	.121

To compare DIF by item type, the mean absolute value of DIF was grouped by the type of item used in the inventory (see Table 6). The mean absolute DIF scores for the item type for the GOT scales ($F(3,132) = .23$, $p = .87$) and the BI scales ($F(3,281) = 1.29$, $p = .28$) were not statistically significant. The total percentages of DIF and the amount of DIF detected are comparable for the GOT and BI scales, indicating consistency in the results across scale type. In addition, the correlation between the DIF scores for the 132 items common to the GOT and the BI scales is .79, supporting the robustness of the differential item functioning results.

3.4. Item-bias and GOT scale level gender differences

Two procedures were employed to estimate the influence of DIF on GOT scale scores. First, a feature of the SIBTEST called differential bundle functioning (DBF; Stout & Roussos, 1996), was used to estimate whether the DIF items for the GOT scales amplify or negate each other. If the DIF detected in the items is primarily in the direction of a particular group, the influence of item bias will be amplified at the scale level. Conversely, if the DIF items favoring either group are similar in number as was the case here, the influence of item bias will tend to be cancelled out at the scale level. To conduct a bundle DIF analysis a subtest, called a *suspect subtest*, is created that includes items that show DIF and then, a valid or a matching subtest is created for the analysis that includes items that do not show DIF. Based on the automatic single DIF analysis reported above, the suspect subtest items were those items that were rejected at $p < .001$ and showed a moderate amount of DIF around .10 or more. These criteria were chosen because the GOT scales contain only 20–33 items and the numbers of statistical tests for DIF items are large, inflating the overall error rate. For each of the GOT scales the remaining items that did not show DIF were defined as a purified subscale (few items remained that did not meet both these criterions). The second method consisted of recalculating the GOT gender difference effect sizes with the valid or "purified" scales consisting of items that did not display DIF (Huang, Church, & Katigbak, 1997; Smith & Reise, 1998). Table 7 shows the results for these two types of analysis.

The bundle DIF results indicated that the DIF items tend to cancel each other out at the scale level. For most of the GOT scales, the bundle results were statistically significant but small, because the bundle β value is the sum of the items values (Stout & Roussos, 1996). The S scale, for example, contains the largest bundle DIF favoring women, but items contributing to this bundle displayed only .05 (0.71/15) DIF on average, substantially less than the nominal value of .10 that is considered small (Douglas, Roussous, & Stout, 1996). However, when the effect sizes reflecting differences for the original scales and the scales that have been "purified" were compared, the gender differences in the R and I scales were reduced. The A and S scale differences remained the same and the effect sizes for C and E showed a slight change. These results indicated that due to differential item functioning and sex-type influence on the items, the gender differences in the R and I scales were inflated. Surprisingly, the gender differences in the A and S scale did not seem to be influenced by the sex-type of the items. Furthermore, a randomization test (Hubert & Arabie, 1987) was used to test if the purified RIASEC scales manifest Holland's circular order structure of vocational interests. The Correspondence Index (CI) for the model data fit was .81, indicating a good fit and supporting the construct validity of the scales containing only unbiased items. In comparison, the CI index for the published scales in the sample was .89.

Table 7
Differential bundle functioning and effect sizes in the published and purified GOT scales.

GOT	*N* items in valid scale	*DBF* (β)	*d* published scale	*d* purified scale
Realistic	9	0.40*	.86	.64
Investigative	6	−0.65**	.30	.14
Artistic	10	0.19	−.45	−.44
Social	8	0.71**	−.45	−.45
Enterprising	7	−0.48**	−.09	−.11
Conventional	5	0.37*	−.24	−.29

* $p < .05$.
** $p < .01$.

4. Discussion

The results of this study showed that about two-thirds of the General Occupational Themes and Basic Interest items function differently for women and men, indicating that there is extensive gender-related item bias in the Strong. The six GOT scales were not essentially unidimensional, but contained dimensionally distinct sets of items. A major dimension was detected in the BI scales, but these were also found to contain some minor dimensions. The relationship of DIF with the sex-type ratings indicated that there is a sex-type nuisance dimension underlying the responses with a different distribution for men and women. When the DIF items are removed from the GOT scales, differences favoring men are reduced in the R and I scales, but differences favoring women remained in the S and A scales. Contrary to expectations, the occupational title items did not show more DIF than other item types.

The results indicate that there are different meanings and implications for measured interests between the men and women. Most importantly, women and men with the same level of interest or trait being measured by the GOT's tend to respond differently to sex-stereotyped items. For example, women are not as likely as men to say they like highly male stereotyped items like auto racer even though they have equally realistic interests. The dimensional analysis shows that the scales are not unidimensional and other factors than the primary trait (e.g. RIASEC interest type) being measured influence the responses to interest items. The gender-DIF and sex-type ratings are correlated and along with previous research (e.g. Einarsdóttir & Rounds, 2000) suggests that a sex-type dimension "an irrelevant factor" is affecting the responses of men" and women differently. This supports the contention that gendered barriers and opportunities in the world of work, possibly internalized through socialization as sex-role stereotypes (Betz, 2005), influence responses to interest measures.

Although the majority of the items function differently for women and men in the Strong, at the scale level the influence of DIF items is cancelled out for most of the GOT scales. The present study showed that scale level gender differences in the R and I types are to a certain degree influenced by item bias. However, there are also real gender differences in the scale scores as a result of impact (Ackerman, 1992), especially in the R, S, and A types, probably reflecting the persistent differential socialization of men and women. These gender differences for RIASEC types are consistent with Lippa's (1998) study that showed that gender-related individual differences are linked to the People-Things dimension a construct valid dimension that underlies Holland's RIASEC structure.

The Realistic scale is illustrative of inflated gender differences due to item bias. This is consistent with the contention that the content of items making up the R scales was sampled from domains that are more typically experienced by men (Walsh & Betz, 1995). By comparison, the S and A scales contained more balanced sets of sex-typed items. This contrast highlights the importance of careful item selection, such that items represent the domain that researchers intended to measure. The different preferences of women and men with the same trait level for specific items needs to be taken into account, because detected scale differences are, to some extent, a function of items in relation to their sex-typing. It is important for counselors using interest inventories to be aware not only of the fact that men and women do have different interests but that gender differences in interests may be inflated by the measures used. This underscores the extensive influence of gendered stereotypes on the client's perceptions of themselves in relation to the world of work, an issue that needs to be explicitly addressed at different stages in the counseling process.

Throughout the history of interest measurement there has been a discussion concerning what kind of items are best for measuring interests. Several interest measurement researchers (e.g., Kuder, 1977) have suggested that occupation-related items are more susceptible to stereotyping than activity-related items. Nevertheless, we found that items corresponding to occupational titles were no more susceptible to gender bias than other types of items. This suggests that gender differences cannot be dealt with by using activity items or other types instead of occupational titles.

An attempt has been made to eliminate gender differences from the UNIACT, an inventory that captures Holland's RIASEC types. During construction of its scales, only activity items showing similar distributions for women and men were used, resulting in minimal gender differences in the scales (Swaney, 1995). The main problem with this approach is the fact that it systematically removes real gender differences in the construct being measured because the item distribution is confounded by valid group differences. The use of same-gender, combined-gender and opposited-gender norms for interpretation as is suggested in the 1994 Strong manual (Harmon et al., 1994) is helpful for clients exploring and expanding their options. This was a necessary effort to reduce gender-restrictiveness in the interpretation of the Strong but has been abandoned in the latest version of the SII (Donnay et al., 2005).

DIF methods represent an improvement by allowing us to remove the influence of the sex-typed characteristics, either by using only items showing no DIF, or by balancing the number of male and female sex-typed (DIF) items in each scale. This results in an inventory that accurately reflects real differences in the vocational interests of women and men. An approach that is already standard practice in the construction of ability tests used in academic selection (e.g., Educational Testing Service, 1994; see also Willingham & Cole, 1997).

The finding that the GOT scales are not essentially unidimensional raises further questions about their content, construct validity, and psychometric quality as measures of the RIASEC types. In addition to a sex-type dimension that was detected, it is possible that the Strong scales may also contain other undefined dimensions. The multidimensionality of the scales also suggests that DIF is likely to be detected when other groups besides men and women are compared as was the case in a recent study of ethnic/racial differences in responses (Fouad & Walker, 2005). Theoretical issues must also be kept in mind because new scale construction methods do not solve the problem alone. Dawis (1991) has pointed out that interest measurement is not grounded in theory. It has largely centered on operational definitions of interests that are not connected with other psychological theories, so theoretical considerations do not offer any insight on how the influence of sex-role socialization and gender related barriers in the world of work could be addressed in the conceptualization of vocational interests.

Traveling back in history again to the social context that fostered the birth of current interest theories it is important to point out that Holland's (1997) idea for the RIASEC types emerged from his experience as a vocational counselor in the 1950s. It was also influenced by Guilford, Christiansen, Bond, and Sutton (1954) factor analytic studies of interest measures that were based entirely on male samples. Since that time the world of work has changed especially in regard to women's participation in paid work. New realities call for re-evaluation of RIASEC type definitions. For example, it may be argued that Holland's R type is too narrowly defined, focusing on manual work and the use of tools and machines traditionally operated by men (Walsh & Betz, 1995). The description of the R type and its measurement does not include experiences women have (e.g. gardening, making clothes, driving children to school, cooking).

A limitation of this study is that several issues relevant to the application of IRT models and methods in vocational interest measurement have yet to be resolved. Differential item functioning, for example, is not always stable across samples and DIF detection methods. We applied a non-parametric method because the fit of IRT models and methods have not yet been tested with vocational interests, it is simple to understand and easy to use and they have been used successfully in previous studies of interest measures (Aros et al., 1998; Fouad & Walker, 2005). It cannot be ruled out that the item-bias detected between groups are due to other characteristics (e.g. age) that differ in the male and female samples used. However, the results indicate that the DIF is related to sex-typing of the items supporting the contention that item bias is related to different experiences of women and men. It is important to cross-validate the present results using other samples and inventories and different types of DIF techniques. Parametric methods may also be useful in determining the scale level influence of DIF items (Reise, Smith, & Furr, 2001). Another problem encountered in this study is the limited number of items in the BI scales. Therefore, the issue of dimensionality could only be addressed in a somewhat preliminary fashion. Essential unidimensionality could not be tested directly in the BI scale analysis. The present study does, however, presents valuable information about item level gender bias in the Strong that supports the need for further research on the issues of dimensionality and construct validity of the RIASEC interest types.

The application of IRT and DIF techniques bring interest measurement up to date with current thinking on validity theory, and may fuel a new debate about the inherent problem of gender differences. This modern approach for addressing construct validity emphasizes that the issue of gender bias in interest measures must also be discussed at a theoretical level, and highlights the importance of considering the social consequences of interest theories and the definitions from which they are constructed.

References

Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29*, 67–91.

Aros, J. R., Henly, G. A., & Curtis, N. T. (1998). Occupational sextype and sex differences in vocational preferences-measured interest relationships. *Journal of Vocational Behavior, 53*, 227–242.

Betz, N. E. (1994). Basic issues and concepts in career counseling for women. In W. B. Walsh & S. H. Osipow (Eds.), *Career counseling for women* (pp. 1–41). New Jersey: Lawrence Erlbaum Associates, Inc..

Betz, N. E. (2005). Women's career development. In S. D. Brown & R. W. Lent (Eds.), *Career development and counseling: Putting theory and research to work* (pp. 253–277). New Jersey: John Wiley & Sons.

Betz, N. E., & Fitzgerald, L. (1987). *The career psychology of women*. Orlando, FL: Academic Press.

Bolt, D., & Rounds, J. B. (2000). Advances in psychometric theory and methods. In S. Brown & R. Lent (Eds.), *Handbook of counseling psychology* (3rd ed., pp. 140–198). New York: Wiley.

Bolt, D., & Stout, W. (1996). Differential item functioning: Its multidimensional model and resulting sibtest procedure. *Behaviormetrika, 23*, 67–95.

Campbell, D. P. (1974). *Manual for the Strong–Campbell interest inventory*. Palo Alto, CA: Stanford University Press.

Cole, N. S., & Moss, P. A. (1989). Bias in test use. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 201–219). New York: Macmillan.

Cole, N. S., & Hanson, G. R. (1975). Impact of interest inventories on career choice. In E. E. Diamond (Ed.), *Issues of sex bias and sex fairness in career interest measurement*. Washington, DC: National Institute of Education.

Cooper, E. A., Doverspike, D., & Barrett, G. V. (1985). Comparison of different methods of determining sex type of an occupation. *Psychological Reports, 57*, 747–750.

Chang, H-H., Mazzeo, J., & Roussos, L. (1996). Detecting DIF for polytomously scored items: An adaption of the SIBTEST procedure. *Journal of Educational Measurement, 33*, 333–353.

Chartrand, J. M., Dohm, T. E., Dawis, R. V., & Lofquist, L. H. (1987). Estimating occupational prestige. *Journal of Vocational Behavior, 31*, 14–25.

Crites, J. O. (1969). *Vocational psychology: The study of vocational behavior and development*. New York: McGraw-Hill.

Dawis, R. V. (1991). Vocational interests, values and preferences (2nd ed.. In M. D. Dunnette & L. M. Hough (Eds.). *Handbook of industrial and organizational psychology* (Vol. II, pp. 833–867). Palo Alto, CA: Consulting Psychologist Press.

Diamond, E. E. (Ed.). (1975). *Issues of sex bias and sex fairness in career interest measurement*. Washington, DC: National Institute of Education.

Donnay, D. A. C., Morris, M., Schaubhut, N., & Thompson, R. (2005). *Strong Interest Inventory manual: Research, development, and strategies for interpretation*. Mountain View, CA: CPP.

Douglas, J. A., Roussous, L. A., & Stout, W. (1996). Item-bundle DIF hypothesis testing: Identifying suspect bundles and assessing their differential functioning. *Journal of Educational Measurement, 33*, 465–484.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Einarsdóttir, S., & Rounds, J. (2000). Application of three dimensions of vocational interests to the Strong Interest Inventory. *Journal of Vocational Behavior, 56*, 363–379.

Educational Testing Service. (1994). DIF procedures (1994 supplement). In *Test development manual*. Princeton, NJ: Author.

Finch, H., & Habing, B. (2007). Performance of DIMTEST- and NOHARM-based statistics for testing unidimensionality. *Applied Psychological Measurement, 31*, 292–307.

Fouad, N. A. (2002). Cross-cultural differences in vocational interests: Between-groups difference on the Strong Interest Inventory. *Journal of Counseling Psychology, 49*, 283–289.

Fouad, N. A., & Walker, C. M. (2005). Cultural influences on responses to items on the Strong Interest Inventory. *Journal of Vocational Behavior, 66*, 104–123.

Gottfredson, G. D., & Holland, J. L. (1978). Toward beneficial resolution of the interest inventory controversy. In C. K. Tittle & D. G. Zytowski (Eds.), *Sex-fair interest measurement: Research and implications*. Washington, DC: U.S. Government Printing Office. pp. 43–51 [National Institute of Educational Report].

Guilford, J. P., Christiansen, P. R., Bond, N. A., & Sutton, M. A. (1954). A factor analytic study of human interests. *Psychological Monographs, 68* (4, Whole No. 375).

Hackett, G., & Lonborg, S. D. (1994). Career assessment and counseling for women. In W. B. Walsh & S. H. Osipow (Eds.), *Career counseling for women* (pp. 43–85). Hillsdale, NJ: Erlbaum.

Hansen, J. C., & Campbell, D. P. (1985). *Manual for the SVIB-SCII* (4th ed.). Palo Alto, CA: Consulting Psychologist Press.

Harmon, L. W., Hansen, J. C., Borgen, F. H., & Hammer, A. L. (1994). *Strong Interest Inventory: Applications and technical guide*. Palo Alto, CA: Consulting Psychologists Press, Inc..

Hattie, J. (1985). Methodological review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*, 139–164.

Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Holland, J. L., Powell, A. B., & Fritzsche, B. A. (1994). *The self-directed search (SDS): Professional user's guide – 1994 edition*. Odessa, FL: Psychological Assessment Resources.

Huang, C. D., Church, A. T., & Katigbak, M. S. (1997). Identifying cultural differences in items and traits: Differential item functioning in the NEO personality inventory. *Journal of Cross-Cultural Psychology, 28*, 192–218.

Hubert, L., & Arabie, P. (1987). Evaluating order hypotheses within proximity matrices. *Psychological Bulletin, 102*, 172–178.

Kok, F. (1988). Item bias and test multidimensionality. In R. Langeheine & J. Rost (Eds.), *Latent trait and latent class models* (pp. 263–275). New York: Plenum Press.

Kuder, F. G. (1977). *Activity interests and occupational choice*. Chicago: Science Research Associates.

Linn, R. L. (1997). Evaluating the validity of assessments: The consequences of use. *Educational Measurement: Issues and Practice, 16*(2), 14–16.

Lippa, R. (1998). Gender-related individual differences and the structure of vocational interests. The importance of the people-things dimension. *Journal of Personality and Social Psychology, 74*, 996–1009.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd. ed., pp. 201–219). New York: Macmillan.

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*, 741–749.

Nandakumar, R., & Stout, W. (1993). Refinements of Stout's procedure for assessing latent trait unidimensionality. *Journal of Educational Statistics, 18*, 41–68.

Nandakumar, R., Yu, F., Li, H. H., & Stout, W. (1998). Assessing unidimensionality of polytomous data. *Applied Psychological Measurement, 22*, 99–115.

Osipow, S. H. (1983). *Theories of career development* (3rd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Prediger, D. J. (1977). Alternatives for validating interest inventories against group membership criteria. *Applied Psychological Measurement, 1*, 275–280.

Prediger, D. J., & Cole, N. S. (1975). Sex-role socialization and employment realities: Implications for vocational interest measures. *Journal of Vocational Behavior, 7*, 239–251.

Prediger, D. J., & Hanson, G. R. (1974). The distinction between sex restrictiveness and sex bias in interest inventories. *Measurement and Evaluation in Guidance, 7*, 96–104.

Reise, S. P., Smith, L. L., & Furr, R. M. (2001). Measurement invariance on the NEO-PI-R Neuroticsim scale. *Multivariate Behavioral Research, 36*, 83–110.

Roussos, L., & Stout, W. (1996a). A multidimensionality-based DIF analysis paradigm. *Applied Psychological Measurement, 20*, 355–371.

Roussos, L., & Stout, W. (1996b). Simulation studies of the effects of small sample size and studied item parameters on SIBTEST and Mantel-Haenszel type I error performance. *Journal of Educational Measurement, 33*, 215–230.

Shealy, R., & Stout, W. (1993). A model based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DTF as well as item bias/DIF. *Psychometrika, 58*, 159–194.

Shepard, L. A. (1997). The centrality of test use and consequences for test validity. *Educational Measurement: Issues and Practice, 16*(2), 5–8.

Smith, L. L. (2002). On the usefulness of item bias analysis to personality psychology. *Society for Personality and Social Psychology, 28*, 754–763.

Smith, L. L., & Reise, S. P. (1998). Gender differences on negative affectivity: An IRT study of differential item functioning on the multidimensional personality questionnaire stress reaction scale. *Journal of Personality and Social Psychology, 75*(5), 1350–1362.

Stevens, G., & Cho, J. H. (1985). Socioeconomic indexes and the new 1980 occupational classification scheme. *Social Science Research, 14*, 142–168.

Stout, W. F. (1987). A nonparametric approach for assessing latent trait unidimensionality. *Psychometrika, 52*, 589–617.

Stout, W. F. (1990). A new item response theory modeling approach with application to unidimensionality assessment and ability estimation. *Psychometrika, 55*, 293–325.

Stout, W., Nandakumar, R., & Habing (1996). Analysis of latent dimensionality of dichotomously and polytomously scored test data. *Behaviormetrika, 23*, 37–65.

Stout, W., & Roussos, L. (1996). *SIBTEST manual*. University of Illinois: Authors.

Strong, A. K. Jr., (1943). *Vocational interests of men and women*. Stanford, CA: Stanford University Press.

Swaney, K. B. (1995). *Technical manual: Revised Unisex edition of the ACT interest inventory (UNIACT)*. Iowa City, IA: American College Testing.

Tittle, C. K., & Zytowski, D. G. (Eds.). (1978). *Sex-fair interest measurement: Research and implications*. Washington, DC: National Institute of Education.

Tracey, T. J. G., & Rounds, J. B. (1996). The spherical representation of vocational interests. *Journal of Vocational Behavior, 48*, 1–41.

U.S. Department of Labor. (1998). In FERRET: Federal, electronic research and review extraction tool (On-line). *Current Population Survey*. Available from: http://ferret.bls.census.gov/items/value/valu_20635.htm.

Walsh, W. B., & Betz, N. E. (1995). *Tests and assessment* (3rd ed.). New Jersey: Prentice-Hall.

White, M. J., Kruczek, T. A., & Brown, M. T. (1989). Occupational sex stereotypes among college students. *Journal of Vocational Behavior, 34*, 289–298.

Willingham, W. W., & Cole, N. S. (1997). *Gender and fair assessment*. Mahwah, New Jersey: Lawrence Erlbaum Associates.