Measurement and Evaluation in Counseling and Development

http://mec.sagepub.com/

Psychometric Examination of an Inventory of Self-Efficacy for the Holland Vocational Themes Using Item Response Theory

Brandon M. Turner, Nancy E. Betz, Michael C. Edwards and Fred H. Borgen

Measurement and Evaluation in Counseling and Development 2010 43: 188
DOI: 10.1177/0748175610384810

Article

Measurement and Evaluation in
Counseling and Development
43(3) 188–198

DOI: 10.1177/0748175610384810
http://mec.sagepub.com
SAGE

Psychometric Examination of an Inventory of Self-Efficacy for the Holland Vocational Themes Using Item Response Theory

Brandon M. Turner[1], Nancy E. Betz[1], Michael C. Edwards[1], and Fred H. Borgen[2]

Abstract

The psychometric properties of measures of self-efficacy for the six themes of Holland's theory were examined using item response theory. Item and scale quality were compared across levels of the trait continuum; all the scales were highly reliable but differentiated better at some levels of the continuum than others. Applications for adaptive testing methods in educational and career assessment are discussed.

Keywords

assessment of self-efficacy, Holland's vocational theory, item response theory, adaptive testing

Introduction

In 2004, a special issue of *Measurement and Evaluation in Counseling and Development* was focused on the use of technology in assessment (Wall, Baker, & Sampson, 2004). Among the contributions to that issue was an article by Weiss (2004) on the usefulness of computerized adaptive testing for effective measurement in our field. In that article, Weiss contrasts conventional testing (where each examinee receives the same set of items and scores are determined by cumulating item scores, be those right/wrong or Likert-type) with adaptive testing, where different items and different numbers of items are administered depending on the examinee's trait level as indicated by item responses as testing continues. Because each examinee receives items "targeted" toward his/her trait level, fewer items per examinee are usually required, and testing can be terminated at a specified standard error criterion, leading to what Weiss (2004) calls "equiprecise measurements" across trait levels (p. 75). Adaptive testing is possible because of the theoretical innovations of item response theory (IRT; e.g., Wainer, 2000) and the computing powers of modern-day systems.

In spite of the publication of this important special issue, there has not been extensive follow-up research or application. Some authors have compared computer-administered (personal computer or web-based) with paper-and-pencil administration (e.g., Lumsden, Sampson,

[1]The Ohio State University, Columbus, OH, USA
[2]Iowa State University, Ames, IA, USA

Corresponding Author:
Nancy E. Betz, Department of Psychology, The Ohio State University, 1835 Neil Avenue, Columbus, OH 43210, USA
Email: betz.3@osu.edu

Reardon, Lenz, & Peterson, 2004). These authors found that the three methods yielded equivalent scores on Holland's self-directed search and high levels of agreement among three letter codes derived from the scores. However, the computer-based versions took less time in administration and were preferred by college students. Seol (2007) used an IRT model—the Rasch model—to evaluate the Marlowe Crown Social Desirability Scale. Other uses of IRT and adaptive testing, however, have not yet appeared herein.

Since many assessment packages involve several inventories, the efficiency in item use provided by adaptive testing may prove especially useful. The assessment objective of interest herein is the joint use of measures of self-efficacy and the measures of vocational interests to provide suggestions for the best fitting college majors and career choices (Betz & Borgen, 2000). Studies of the incremental validity of measures of career self-efficacy—beyond measures of vocational interests—have now been shown to predict choices, performance (e.g., grades), and persistence (e.g., degree completion) above the predictive power of previously used measures (Donnay & Borgen, 1999; Tracey & Hopkins, 2001).

Bandura's (1977) original theory of self-efficacy expectations postulated the existence of behavior domain–specific expectations of self-efficacy. "Perceived self-efficacy is a judgment of one's capability to accomplish a certain level of performance" (Bandura, 1986, p. 391). According to Bandura's theory, the critical importance of self-efficacy expectations is that they are postulated to influence three major classes of subsequent behavior: approach versus avoidance (where "approach" is often conceptualized as "choice" vs. avoidance), level of performance, and persistence. Thus, self-efficacy for a domain such as mathematics would be postulated to lead to choice (vs. avoidance) of math coursework and college majors, to facilitate performance on math assignments and exams, and to lead to persistence in the face of obstacles or failure experiences.

Self-efficacy theory has now been widely applied in the fields of education and counseling. In education, self-efficacy has been a critical concept in the study of academic functioning (Bandura, Barbaranelli, Caprara, & Pastorelli, 1996), middle school science (Britner & Pajares, 2006), and writing (Zimmerman & Bandura, 1994). Results uniformly support the specific effects of self-efficacy on achievement and performance (e.g., Bandura et al., 1996). The self-efficacy of teachers (McCoach & Colbert, 2010), school counselors (Bodenhorn & Skaggs, 2005), and coaches (e.g., Myers, Wolfe, & Feltz, 2005) have also been studied. College self-efficacy was measured by Barry and Finney (2009), and college students' relationship self-efficacy was studied by Lopez, Morua, and Rice (2007). Hansen and Bubany (2008) compared college students' ability self-estimates with their self-efficacy perceptions and found that each measured distinct facets of self-evaluation.

Almost all the measures of self-efficacy developed to date have been developed and examined using classical test theory (CTT). The applicability of IRT methods to item and scale evaluation and to the development of adaptive testing strategies with inventories of self-efficacy has only rarely been examined. An exception to this is Scherbaum, Cohen-Charash, and Kern's (2006) IRT comparison of three measures of generalized self-efficacy, but IRT has not yet been used with domain-specific self-efficacy scales.

IRT has particular advantages over CTT in the evaluation of scales and scale items. First, CTT yields an overall item discrimination index for each item, but IRT provides an index of the item's discriminatory power in the item information function (IIF) for each value of the underlying trait (θ). This specificity is not possible with CTT. The sum of the IIFs is the total test information function (TIF), which yields an index of the precision of the measurement at each point on the trait continuum. The analysis also provides a standard error of measurement (SEM) at each point on the continuum. Thus, both item and test quality can be described for individuals with different levels of the underlying trait.

A comprehensive inventory of career self-efficacy, the Career Confidence Inventory (CCI;

Betz & Borgen, 2006), is a 190-item inventory measuring self-efficacy or confidence with respect to the six Holland themes and 27 basic confidence dimensions. It is used in conjunction with an interest inventory such as the Strong Interest Inventory to provide major and career options for which both interest and confidence are present. If confidence is generally lacking, interventions based on self-efficacy theory may be used to help increase career options. The entire inventory takes 20 to 25 minutes to administer, in addition to the time taken by the interest inventory. Since the CCI is designed to be administered by computer; it is an ideal candidate for testing the use of IRT. Successful application of IRT analysis and adaptive testing could reduce the time and costs of administration by as much as 50% (see Mills & Steffen, 2000). Accordingly, a long-term objective of the current research is the use of IRT as the basis for the development of adaptive approaches to the measurement of self-efficacy expectations.

The purpose of the present study was to conduct an IRT analysis of the six Holland confidence themes measured by the CCI. Prior to the IRT analysis, exploratory factor analysis (EFA) and confirmatory factor analysis (CFA) were used to examine the dimensionality of the combined item pools.

Method

Participants

The participants were 2,406 freshmen enrolled in a university program for students who have not declared a college major. These students are required to take a career exploration course in their first quarter. As part of the course, an online career assessment system (CAPA), which included both an interest inventory and a confidence inventory, was administered; the assessment included a total of about 480 items. The inventory results are entered into previously determined regression equations based on the actual content of majors to generate the best majors from the standpoint of joint consideration of interest and confidence. The system provides immediate online feedback, including profiles for both the interest inventory and confidence inventory and the top major clusters for that individual. Major suggestions can also be provided separately for interest patterns alone and for confidence patterns alone. In this study, we were not concerned with the interest inventory and it will not be mentioned further.

Measures

Career Confidence Inventory. As mentioned in the Introduction, the CCI (Betz & Borgen, 2006) measures self-efficacy or confidence with respect to the six Holland themes: Realistic (R), Investigative (I), Artistic (A), Social (S), Enterprising (E), and Conventional (C). The inventory was originally developed using CTT analyses of a 240-item pool of activities and school subjects for which confidence ratings in 1,103 adults and 1,271 college students had been obtained (Betz & Borgen, 2006; Betz et al., 2003). Additional normative data of 580 college students were obtained later.

Items are prefaced with the phrase "Indicate your confidence in your ability to . . ." Sample activities following are, for example, "Identify the chambers of the heart" (Investigative) or "Write a book report" (Artistic). Responses are obtained on a 5-point scale ranging from *no confidence at all* (1) to *complete confidence* (5). Values of coefficient alphas for the six scales in a sample of 160 college students ranged from .91 to .94 (Betz & Borgen, 2006).

Analyses

Factor analyses. Theoretical considerations suggested that the data would represent six factors, but we wished to begin with EFA and CFA because there is some evidence for high correlations between two of the factors, Enterprising and Social (Harmon, Hansen, Borgen, & Hammer, 1994). The EFA was performed using comprehensive exploratory factor analysis (Browne, Cudeck, Tateneni, & Mels, 2006), and the CFA was performed using Lisrel 8.80

(Joreskog & Sorbom, 2006) using a polychoric correlation matrix obtained from the raw scores. We adhered to common rules for the root mean squared error of approximation for the CFA (Browne & Cudeck, 1992).

Graded response model. The original IRT models were used with dichotomous item responses (usually the right or wrong answers yielded by ability and achievement tests). They required three parameters: item difficulty, discrimination, and in some cases a correction for guessing (Weiss, 2004). However, with the Likert-type response continua used often in personality and attitude measurement, a more common model is Samejima's (1969) graded response model. This model is specifically designed for use for polychotomous data where the assumption of ordinal levels of response options is satisfied. For this model, the probability that person i with some level of the trait will choose a response option with a "score" on item j at or above category k is

$$P_{i,j,k} = \frac{\exp[a_j\,(\theta_i - b_{jk})]}{1 + \exp[a_j\,(\theta_i - b_{jk})]}\,,\ k = 2,3,\ldots,m_j\,,$$

where a_j is the discriminability or slope parameter for item j. The higher the value of a_j, the higher the value of discriminability between persons. Furthermore, there are m_j categories, and b_{jk} is the difficulty parameter for item j on category k. The b_{jk} parameter is the ability level where the probability of endorsing the kth, $(k-1)$th, . . ., or first response option is equal to the probability of endorsing any of the $(k+1)$th, $(k+2)$th, . . ., or m_j categories.

Typical results of applying this equation are illustrated in Figure 1. In this model, each response of a 5-point Likert-type scale is represented by one distribution. That is, the response corresponding to 1 (e.g., *not at all confident*) is the leftmost curve, whereas the response corresponding to 5 (*completely confident*) is the rightmost curve. The horizontal axis is the level of the trait (which has a standard normal distribution by construction), and the vertical axis is the probability of that response for individuals at that trait level. So for an individual with trait level at the mean ($\theta = 0$), the most likely response is 3 (e.g., *moderately confident*).

The R package ltm (Rizopoulos, 2006) was used to fit Samejima's (1969) GRM to the data. We then used the parameter estimates to obtain the IIF, TIF, and SEM.

Item and scale parameters. Item parameters were obtained using both CTT and IRT. For the CTT parameters, we obtained item mean (difficulty), corrected item–total correlations (CITC; indices of discrimination), scale means and standard deviations, and values of Cronbach's alpha for the scales. For the GRM model, we obtained item discrimination, item difficulty, the IIF, and the TIF. The latter two functions supply the amount of information provided at each point on the trait continuum.

Results

Factor Analyses

An EFA was performed on the first data set consisting of 786 participants on 190 items. A Crawford–Ferguson Varimax oblique rotation for polychoric data was performed using comprehensive exploratory factor analysis (Browne et al., 2006) for both the five- and six-factor models. Following close inspection of the factor loadings and the scree plot, we chose the six-factor model because of its interpretability of the items. That is, identifying these factors was clear by the content of each factor's corresponding set of items—the six factors corresponded to the set of six Holland general themes. Furthermore, the root mean squared error of approximation was slightly smaller for the six-factor model ($\hat{\varepsilon}$ = .070, 90% confidence interval [CI] = .070, .071) than for the five-factor model ($\hat{\varepsilon}$ = .076, 90% CI = .076, .077). We then used a subset of 100 items that loaded highly onto only one of the six factors to preserve the unidimensionality assumption necessitated by IRT for each subscale. Another Crawford–Ferguson Varimax oblique rotation was performed on each of these six factors to ensure that each

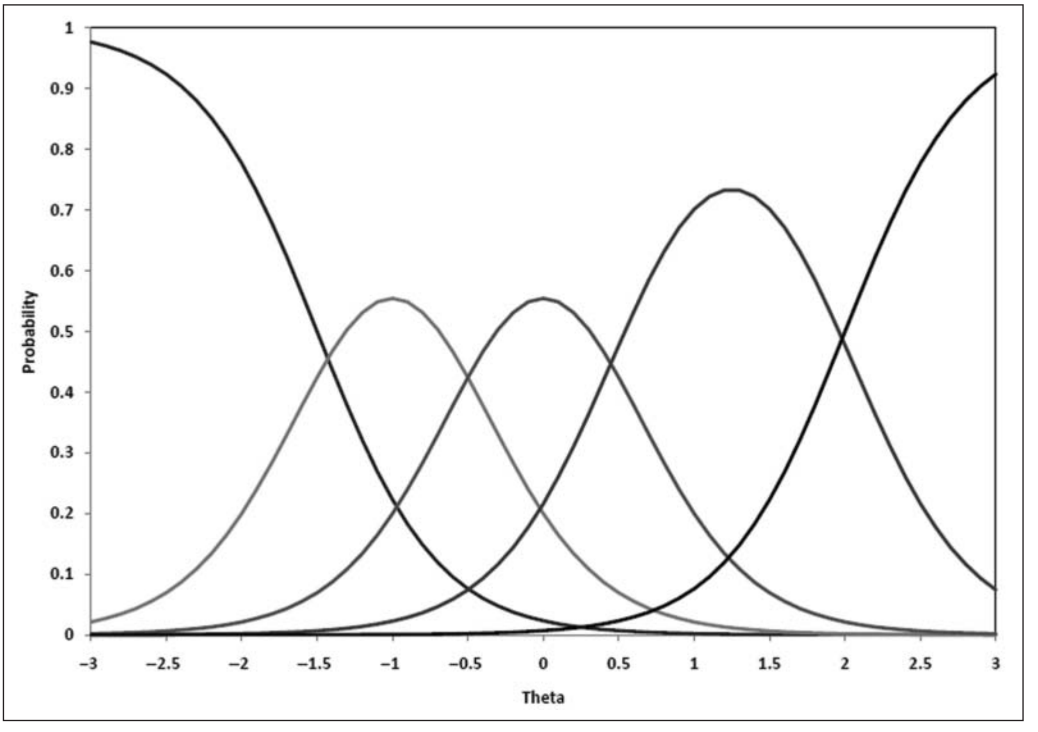

Figure 1. Depiction of a representative item response probability distribution for a graded (e.g., Likert-type) response continuum

subscale upheld the unidimensionality assumption.

Based on the EFA, a model for the subset of items was constructed; numbers of items of each Holland scale were 14 (R), 13 (I), 20 (A), 13 (E), 25 (S), and 15 (C). Using a different sample of college students (N = 1,620), we performed a CFA using Lisrel 8.80 (Joreskog & Sorbom, 2006). The CFA produced an adequate fit to the data ($\hat{\varepsilon}$ = .073, 90% CI = .072, .074; standardized root mean square residual = .076; Browne & Cudeck, 1992; Hu & Bentler, 1999; Steiger & Lind, 1980). This, then, was the item set for which IRT and CTT analyses were done. We have condensed the model to four items per subscale for illustrative purposes here, but a full scale may be obtained by correspondence with the second author. For clarity, Table 1 shows the estimated factor loadings and unique variances for four items per subscale.

Item Analyses

The parameter estimates, both CTT and IRT, for four items per subscale are shown in Table 1. The four items were selected from the total to represent a range of information and difficulty. Shown on the table for each item are the mean, SD, CITC from CTT, and their factor loadings. The IRT analyses yielded the item difficulty or b parameters and the item discrimination or a parameter. Although no simple quality cutoff criterion exists for the a parameter, Zickar, Russel, Smith, Bohle, and Tilley (2002) suggested that all a parameters greater than 1 indicated acceptable discriminability between persons. Hafsteinsson, Donovan, and Breland (2007) suggest that when there are fewer items in a scale (they used three scales of 8, 8, and 5 items), a higher standard of item quality may be needed, perhaps 2.0 or better, to yield sufficiently high quality

Table 1. Items, Item Statistics, Factor Analysis Results, and Item Parameters

Scale	Item	Mean	*SD*	CITC	FL	b_1	b_2	b_3	b_4	a
Artistic	A7. Identify famous works of art	2.33	1.12	0.63	0.71	−0.88	0.32	1.37	2.43	1.79
	A15. Envision an artistic creation	2.58	1.29	0.71	0.82	−0.86	0.08	0.79	1.50	2.58
	A17. Create a work of art	2.46	1.35	0.70	0.82	−0.62	0.24	0.84	1.39	2.68
	A20. Design novel sets for a play	2.28	1.17	0.71	0.80	−0.66	0.35	1.20	1.97	2.35
Conventional	C5. Evaluate applicants for bank loans	2.58	1.09	0.68	0.76	−1.26	−0.04	1.04	2.32	2.05
	C7. Create a budget for a company's fiscal year	2.54	1.08	0.78	0.87	−1.06	0.00	0.98	2.00	3.21
	C8. Record and analyze financial data	2.71	1.13	0.77	0.84	−1.23	−0.17	0.80	1.78	2.87
	C11. Handle money for a bank	3.18	1.14	0.70	0.76	−1.83	−0.78	0.33	1.45	2.09
Enterprising	E3. Prosecute people accused of crimes	2.94	1.09	0.59	0.64	−1.93	−0.54	0.75	2.10	1.55
	E5. Persuade others to support a political candidate	2.74	1.17	0.71	0.79	−1.25	−0.12	0.73	1.71	2.39
	E6. Speak at your class reunion	3.01	1.24	0.51	0.58	−1.99	−0.54	0.66	1.78	1.21
	E11. Assist a legislator	2.55	1.12	0.74	0.82	−1.02	0.05	0.99	1.90	2.73
Investigative	I5. Investigate the cause of a disease	2.85	1.17	0.78	0.85	−1.33	−0.22	0.66	1.69	2.79
	I7. Keep up with new scientific discoveries	2.88	1.14	0.74	0.81	−1.49	−0.27	0.70	1.81	2.35
	I8. Understand the scientific basis of a medical breakthrough	2.79	1.20	0.84	0.91	−1.04	−0.17	0.66	1.62	3.73
	I13. Assist in a medical laboratory	2.77	1.23	0.76	0.82	−1.11	−0.12	0.70	1.73	2.50
Realistic	R1. Construct a patio deck	2.60	1.23	0.65	0.71	−1.01	0.01	1.01	1.96	1.78
	R5. Fight fires	2.26	1.15	0.71	0.79	−0.57	0.41	1.27	2.14	2.34
	R7. Build a cradle	2.54	1.19	0.71	0.78	−0.92	0.11	1.01	1.90	2.21
	R12. Repair mechanical equipment	2.25	1.17	0.62	0.70	−0.65	0.56	1.43	2.33	1.70
Social	S10. Reduce or solve conflict among group members	3.66	0.95	0.67	0.76	−2.80	−1.64	−0.34	1.18	2.00
	S15. Help a group of people to cooperate better	3.47	0.96	0.72	0.79	−2.51	−1.35	−0.01	1.41	2.26
	S22. Convince busy people to take on important volunteer tasks	3.05	1.01	0.65	0.71	−2.33	−0.77	0.64	2.02	1.75
	S25. Work with troubled teens	3.20	1.15	0.66	0.70	−1.97	−0.83	0.28	1.51	1.76

Note: CITC = corrected item–total correlation; FL = factor loading; b_i = difficulty parameter; a = discrimination parameter.

overall measurement. Although we had between 13 and 25 items per scale, we will use their implicit guidelines of *a* values from 1.0 to 2.0 to indicate moderate quality and values more than 2.0 to indicate high quality.

Values of the *a* parameter over the 100 item set ranged from 0.77 to 3.73, with only seven values below 1. For each Holland confidence scale, values fell into the following ranges: Artistic 0.85 to 2.59, Realistic 1.02 to 2.34, Conventional 0.77 to 3.2, Social 1.05 to 2.26, Investigative 0.86 to 3.73, and Enterprising 0.83 to 2.72. Of the 100 items 22 had *a* values more than two, which can be characterized as "high quality," but the number of high-quality items differed across confidence themes: 8 of 13 Investigative items, or 62%, had *a* values more than 2.0, whereas none of the other five scales had more than three items (from 8% to 23%) with *a* values more than 2.0.

Table 1 shows the IRT parameter values and the CITC for four illustrative items from each scale. Overall, there is a relationship between CITC and *a*, with higher values of *a* generally corresponding to higher CITCs. For example, Item E6, "Speak at your class reunion," has an *a* parameter value of 1.21 and a CITC of .51, both are the lowest respective values shown in Table 1. Similarly, the item "Understand the scientific basis of a medical breakthrough" has an *a* of 3.73 and a CITC of .84, both are the highest respective values in the data set.

On the other hand, the values of *a* show much greater differentiation in the indication of item quality than do the CITCs (this is often found in such analyses; e.g., Scherbaum et al., 2006). For example, Items A17 and C11 both have ITCs of .70, yet the *a* for A17 is 2.68 whereas that for C11 is 2.09. Items C7 and I5 both have ITCs of .78 yet have *a*s of 3.21 and 2.79.

Further information is provided by the IIFs, shown in Figure 2. The figure shows that most of the items provide rich information for most values of θ; however, relative to other items, A7, R1, R12, S22, S25, E3, and E6 provide the least amount of information around the most common values of θ but provide more information for the less common values (scores at the lower and higher ends of the distribution). For example, in the Artistic scale, Item 7, "Identify famous works of art," provides the least overall information of those shown but the most information for individuals with high levels of the trait. Item 17, "Create a work of art," provides very high levels of information for θ values near 1.0 but lower levels at the extremes. On the Enterprising scale, Items 3 and 6 provide the lowest overall information but the highest levels at extreme levels of the trait.

Figure 3 shows the TIFs and Figure 4 shows the SEM curves for each subscale. Taken together, these figures show that each subscale provides rich information near the middle and upper ends of the trait distribution. Furthermore, the SEM is quite low for values of θ in the range of −3 to +3, which under IRT should contain approximately 99% of the population. However, it may be noted that the Social scale comes closest to approximating a flat and high distribution of information, which many authors argue is the best result (see Hafsteinsson et al., 2007). The Social information function provides more information, and less error, at the low end of the trait continuum than do the other scales. Although other scales have generally unimodal distributions centered on the mean, the SEM plot suggests that of these subscales Artistic provides the best measurement at the high end of the distribution.

It is useful to compare these information functions and standard errors to the index of reliability usually provided based on CTT assumptions, that is, coefficient alpha. The alphas for the six confidence scales ranged from .90 to .94, with the median .92. The highest alpha, .94, was for the Social scale, which also had the most items (25). The 20-item Artistic scale had an alpha of .92. The homogeneity of these values may be contrasted with the relatively distinct information functions, especially that for Social. The somewhat poorer measurement provided by Social and Investigative at the lower ends of the distribution would not be shown from overall alphas of .94 (S) and .93 (I).

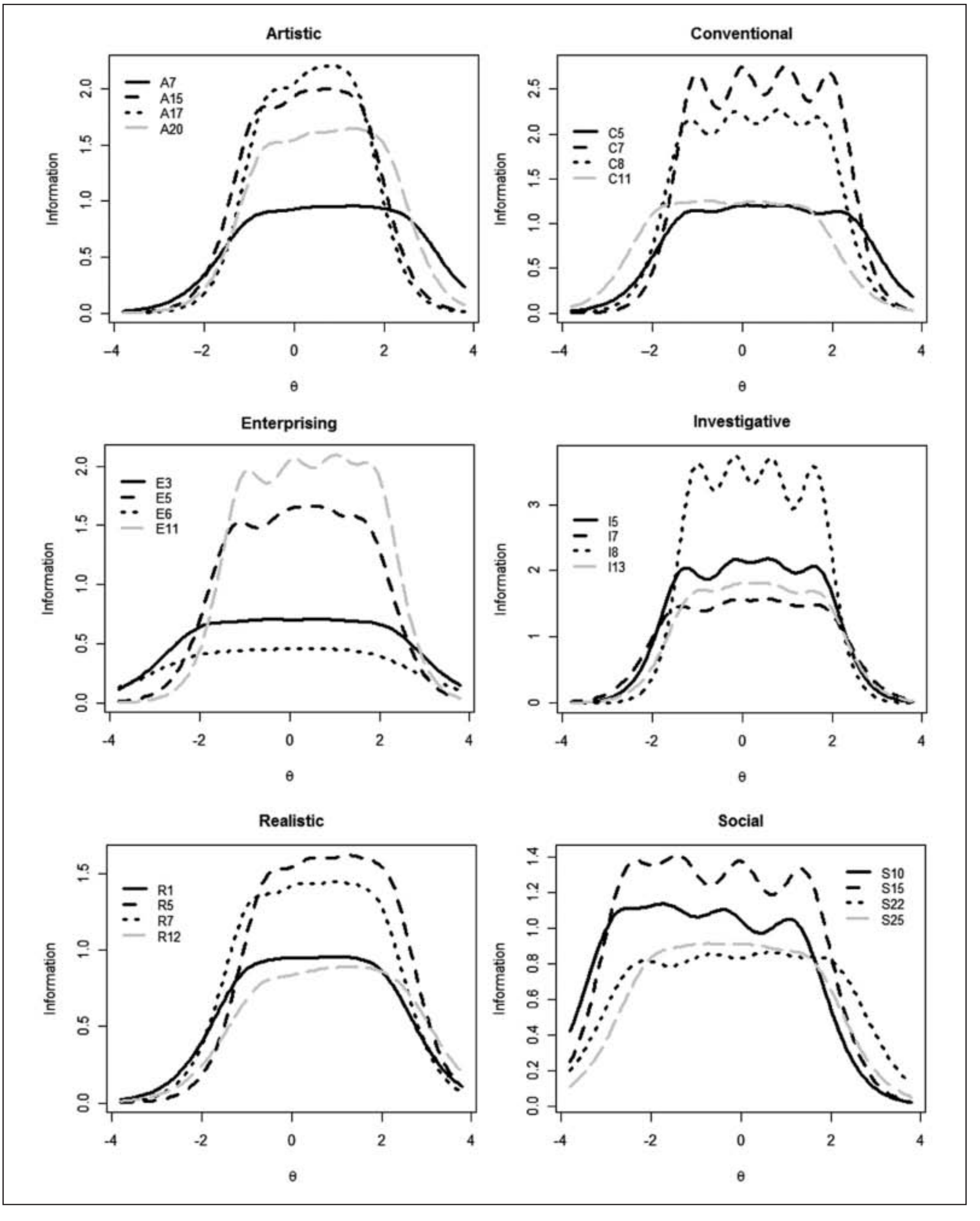

Figure 2. Item information functions (IIF) for the four discussed items for each of the six subscales: Artistic (top left), Conventional (top middle), Enterprising (top right), Investigative (bottom left), Realistic (bottom middle), and Social (bottom right)

Discussion

The current study was designed to evaluate the item and scale properties of an inventory of measures of self-efficacy or confidence for the six themes of Holland's (1997) vocational theory. Following EFA and CFA supporting the six unidimensional scales, both CTT and

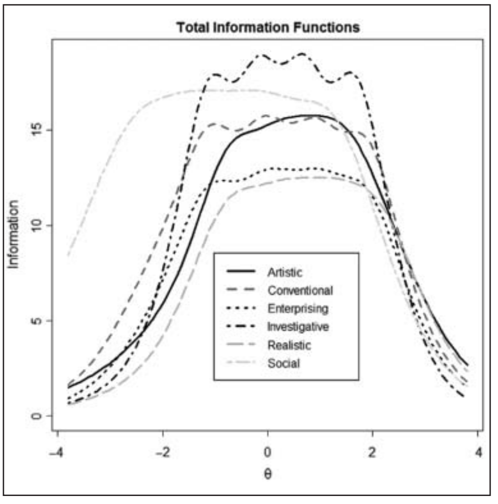

Figure 3. The total or scale information function (TIF) for each of the six subscales

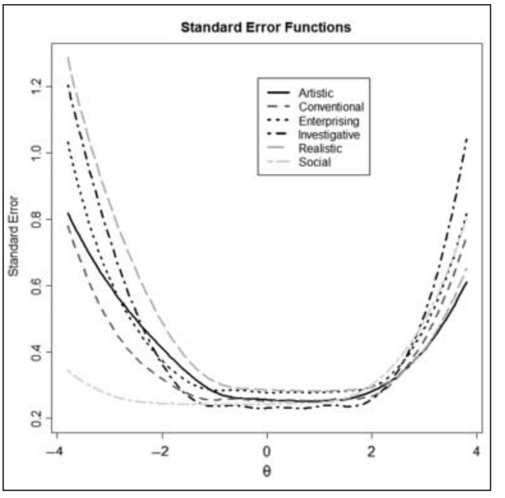

Figure 4. The standard error function for each of the six subscales

IRT analyses were used to evaluate the psychometric quality of the items and scales.

Overall, the expectation that IRT would provide a better differentiated set of indices of item and scale quality was met. First, regarding item discrimination, the *a* parameter provides the index of item quality. Zickar et al. (2002) suggested that all *a* parameters greater than 1.0 indicated acceptable discriminability, whereas Hafsteinsson et al. (2007) suggested that values of more than 2.0 provide a higher standard of quality. In the present data set of 100 total items (between 13 and 25 items per scale), values of the *a* parameter over the 100 item set ranged from .77 to 3.73, with only seven values below 1.0. Of the 100 items 22, or 22%, had *a* values more than 2.0, which can be characterized as "high quality." Of the "high-quality" items the majority were in the 13-item Investigative scale, where 8, or 62%, had values of more than 2.0. The other five scales had between 8% and 23% of items of that quality.

As would be expected, there was generally a relationship between CITC and *a*, with higher values of *a* generally corresponding to higher CITCs. For example, the highest quality item, "Understand the scientific basis of a medical breakthrough," had an *a* of 3.73 and a CITC of .84, both the highest of these values in the data set. On the other hand, the values of *a* showed much greater differentiation in the indication of item quality than do the CITCs, as is often shown in such analyses (see Scherbaum et al., 2006). Furthermore, the IIFs provided in Figure 2 show differences in quality across the trait continuum even when both CITCs and values of *a* are comparable.

The alphas for the six confidence scales ranged from .90 to .94, with the median .92. Even though these values show a very high quality scale, the scale information functions and SEMs across the θ continuum provide much more information than does an overall scale alpha. The figures show that each subscale provides rich information near the middle and upper ends of the trait distribution; values of the SEM are quite low for values of θ in the range of −3 to +3, which under IRT should contain approximately 99% of the population. The Social information function provides more information, and less error, at the low end of the trait continuum than do the other scales, whereas Artistic provides the best measurement at the high end of the distribution. These differ by only .02 in alpha (.94 vs. .92), yet the measurement characteristics are quite different.

For use in adaptive testing, the fact that items differentiate best at different levels of the trait continuum is desirable, because it enables test constructors to tailor the item administration at each level of θ. In addition, with *a* values as high as some of these herein (more than 2.0 or even 3.0), items can be administered that quickly converge on an estimate of the trait level. Thus, a great deal of information about individual differences can be yielded quickly and efficiently. Especially with assessment procedures involving more than one inventory, as is common in educational and career assessment, this efficiency may prove very valuable. We hope to see more research applying IRT and adaptive testing methods in the educational and counseling literature.

Declaration of Conflicting Interests

The author(s) declared no conflicts of interest with respect to the authorship and/or publication of this article.

Funding

The author(s) received no financial support for the research and/or authorship of this article.

References

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*, 191–215.

Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory.* Englewood Cliffs, NJ: Prentice Hall.

Bandura, A., Barbaranelli, C., Caprara, G. V., & Pastorelli, C. (1996). Multifaceted impact of self-efficacy beliefs on academic functioning. *Child Development*, *67*, 1206–1222.

Barry, C., & Finney, S. (2009). Can we feel confident in how we measure college confidence? *Measurement and Evaluation in Counseling and Development*, *42*, 197–222.

Betz, N. E., & Borgen, F. (2000). The future of career assessment: Integrating vocational interests with self-efficacy and personal styles. *Journal of Career Assessment*, *8*, 205–222.

Betz, N. E., & Borgen, F. H. (2006). *Manual for the CAPA Confidence Inventory*. Columbus, OH: Career and Personality Assessments.

Betz, N. E., Borgen, F. H., Rottinghaus, P. R., Paulsen, A., Halper, C. R., & Harmon, L. W. (2003). The Expanded Skills Confidence Inventory: Measuring basic dimensions of vocational activity. *Journal of Vocational Behavior*, *62*, 76–100.

Bodenhorn, N., & Skaggs, G. (2005). Development of the School Counselor Self-efficacy scale. *Measurement and Evaluation in Counseling and Development*, *38*, 14–28.

Britner, S. L., & Pajares, F. (2006). Sources of science self-efficacy beliefs of middle school students. *Journal of Research in Science Teaching*, *43*, 485–499.

Browne, M. W., & Cudeck, R. (1992). Alternative ways of assessing model fit. *Sociological Methods and Research*, *21*, 230–258.

Browne, M. W., Cudeck, R., Tateneni, K., & Mels, G. (2006). CEFA: Comprehensive Exploratory Factor Analysis [computer program]. Retrieved from http://quantrm2.psy.ohio-state.edu/browne/

Donnay, D. A. C., & Borgen, F. H. (1999). The incremental validity of vocational self-efficacy: An examination of interest, self-efficacy, and occupation. *Journal of Counseling Psychology*, *46*, 432–447.

Hafsteinsson, L. G., Donovan, J. J., & Breland, B. T. (2007). An item response theory examination of two popular goal orientation measures. *Educational and Psychological Measurement*, *67*, 719–739.

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.

Hansen, J. C., & Bubany, S. (2008). Do self-efficacy and ability self-estimate scores reflect distinct facets of ability judgements? *Measurement and Evaluation in Counseling and Development*, *41*, 66–88.

Holland, J. L. (1997). *Making vocational choices: A theory of vocational personalities and work environments* (3rd ed.). Odessa, FL: Psychological Assessment.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling*, *6*, 1–55.

Joreskog, K., & Sorbom, D. (2006). *LISREL 8.8 user's reference guide*. Chicago, IL: Scientific Software.

Lopez, F., Morua, W., & Rice, K. (2007). Factor structure, stability, and predictive validity of college students' relationship self-efficacy beliefs. *Measurement and Evaluation in Counseling and Development*, *40*, 80–96.

Lumsden, J., Sampson, J., Reardon, R., Lenz, J., & Peterson, G. (2004). A comparison study of the paper-and-pencil, personal computer, and Internet versions of Holland's Self-directed Search. *Measurement and Evaluation in Counseling and Development*, *37*, 85–94.

McCoach, D., & Colbert, R. (2010). Factors underlying the collective teacher self efficacy scale and their mediating role in the effect of socioeconomic status on academic achievement at the school level. *Measurement and Evaluation in Counseling and Development*, *43*, 31–47.

Mills, C., & Steffen, M. (2000). The GRE Computer Adaptive Test: Operational issues. In W. van der Linden & C. Glas (Ed.), *Computerized adaptive testing: Theory and practice* (pp. 75–99). Boston, MA: Kluwer.

Myers, N. D., Wolfe, E., & Feltz, D. L. (2005). An evaluation of the psychometric properties of the Coaching Efficacy Scale for American coaches. *Measurement in Physical Education and Exercise Science*, *9*, 135–160.

Rizopoulos, D. (2006). ltm: An R package for latent variable modeling and item response theory analyses. *Journal of Statistical Software*, *17*, 1–25.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrica Monograph*, *17*, 1–100.

Scherbaum, C. A., Cohen-Charash, Y., & Kern, M. J. (2006). Measuring general self-efficacy: A comparison of three measures using item response theory. *Educational and Psychological Measurement*, *66*, 1047–1063.

Seol, H. (2007). A psychometric investigation of the Marlowe Crowne Social Desirability Scale using Rasch measurement. *Measurement and Evaluation in Counseling and Development*, *40*, 155–168.

Steiger, J., & Lind, J. (1980). *Statistically based tests for the number of common factors.* Paper presented at the annual meeting of the Psychometric Society, Iowa City, IA.

Tracey, T. J. G., & Hopkins, N. (2001). Correspondence of interests and abilities with occupational choice. *Journal of Counseling Psychology*, *48*, 178–190.

Wainer, H. (2000). *Computerized adaptive testing: A primer.* Mahwah, NJ: Lawrence Erlbaum.

Wall, J., Baker, H., & Sampson, J. (2004). Editorial comments for the special issue on the use of technology in assessment. *Measurement and Evaluation in Counseling and Development*, *37*, 66–69.

Weiss, D. (2004). Computerized adaptive testing for effective and efficient measurement in counseling and education. *Measurement and Evaluation in Counseling and Development*, *37*, 70–84.

Zickar, M. J., Russel, S. S., Smith, C. S., Bohle, P., & Tilley, A. J. (2002). Evaluating two morningness scales with item response theory. *Personality and Individual Differences*, *33*, 11–24.

Zimmerman, B. J., & Bandura, A. (1994). Impact of self-regulatory influences on writing course attainment. *AERA Journal*, *31*, 845–862.

Bios

Brandon M. Turner is a graduate student in Quantitative Psychology at the Ohio State University. His research interests include item response theory, factor analysis, scale construction, mathematical modeling, response time distributions, and Bayesian statistics.

Nancy E. Betz is professor of psychology at the Ohio State University. Her research interests are scale development and evaluation in the areas of vocational interests and confidence, career decision making, and the healthy personality.

Michael C. Edwards is an assistant professor in the quantitative psychology program at the Ohio State University. His research focuses on measurement issues in the social sciences and includes topics such as item response theory, factor analysis, computerized adaptive testing, and computer-intensive estimation methods.

Fred H. Borgen is Emeritus Professor of Psychology at Iowa State University. He has been building career scales since 1967, when he worked with David Campbell to create the Basic Interest Scales for the Strong Interest Inventory. This was the first set of content scales for the Strong. Later, John Holland and Campbell added the six, content-based Holland scales to the Strong. Borgen is currently involved in developing new measures of vocational confidence, interests, and personality.