Using Item Response Theory and Adaptive Testing in Online Career Assessment

Nancy E. Betz and Brandon M. Turner

Journal of Career Assessment 2011 19: 274 originally published online 11 May 2011
DOI: 10.1177/1069072710395534

Journal of Career Assessment
19(3) 274-286

DOI: 10.1177/1069072710395534
http://jca.sagepub.com

Using Item Response Theory and Adaptive Testing in Online Career Assessment

Nancy E. Betz[1] and Brandon M. Turner[1]

Abstract

The present article describes the potential utility of item response theory (IRT) and adaptive testing for scale evaluation and for web-based career assessment. The article describes the principles of both IRT and adaptive testing and then illustrates these with reference to data analyses and simulation studies of the Career Confidence Inventory (CCI). The kinds of information provided by IRT are shown to give a more precise look at scale quality across the trait continuum and also to permit the use of adaptive testing, where the items administered are tailored to the individual being tested. Such tailoring can significantly reduce testing time while maintaining high quality of measurement. This efficiency is especially useful when multiscale inventories and/or a large number of scales are to be administered. Readers are encouraged to consider using these advances in career assessment.

Keywords

item response theory, adaptive testing, web-based career assessment

Introduction

It is an understatement of massive proportions to say that computer technology has transformed the fields of career assessment and counseling. Probably the earliest and most transformative change came with the advent of computer-assisted career guidance systems over 35 years ago, with SIGI (Katz, 1973) and DISCOVER (Rayman & Harris-Bowlsby, 1977) being the first to be introduced. More recently there has been an explosion of both smaller online (web-based) assessment systems (see Borgen & Betz, 2010; Harris-Bowlsby & Sampson, 2001; Reile & Harris-Bowlsby, 2000) and of uses of the Internet in all aspects of career exploration, information gathering, and job searching (see Gore, Bobek, Robbins, & Shayne, 2006; Gore & Leuwerke, 2000). Special issues of the *Journal of Career Assessment* were devoted to uses of the Internet in career assessment in 2000 (Oliver & Chartrand, 2000) and again, recently, in 2010 (Career Assessment and New Technology on the Internet; Walsh, 2010).

[1]The Ohio State University, Columbus, OH, 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

Computer technology and the Internet provide many advantages for the process of career assessment. These advantages include online administration of measures, immediate scoring and profiling, and immediate access to relevant occupational information and to placement aids such as resume writing tools and tips for the job search and interviewing. Research on the effectiveness of these systems has focused primarily on the large-scale career exploration systems—DISCOVER and SIGI or SIGI PLUS. Findings generally indicate that such systems are effective in increasing decidedness and career decision self-efficacy (e.g., Betz & Borgen, 2009; Fukuyama, Probert, Neimeyer, Nevill, & Metzler, 1988; Garis & Niles, 1990; Luzzo & Pierce, 1996) and are well received by student users (Fowkes & McWhirter, 2007; Kapes,Borman, & Frazier, 1989; Peterson, Ryan-Jones, Sampson, Reardon & Shahnasarian, 1994).

Another advantage of computer technology beginning to be exploited is the use of more sophisticated scoring algorithms and methods of combining several types of scores to examine the joint predictive or explanatory power of two or more types of individual differences variables. This can be done and provided to the examinee virtually instantaneously, not possible before the days of the Internet. For example, Borgen and Betz (2008) developed an algorithm, based on regression techniques, to mathematically combine score patterns from interest and confidence inventories to yield ''scores'' for college major clusters. The highest ''scoring'' majors are those with best joint fit to the individual's pattern of interests and confidence and can be provided instantly with completion of the measures administered.

In addition to the administration and scoring of career assessment systems, technological advances now make it possible to develop programs that continually adapt the administration of scale items to the individual examinee, through scoring algorithms that calculate his or her trait level as testing proceeds. This application is known as adaptive testing (Wainer, 2000; Weiss, 2004) and the theory of psychological measurement on which it is based is item response theory (IRT; Bock, 1997). This technology has been used very little in career assessment, but we argue that its potential utility should be more widely exploited in our research and assessment systems.

In the following sections, we will describe IRT and adaptive testing. Next, we will demonstrate their use by means of a data set that had already been analyzed using classical test theory and will show the additional information provided when we analyze the items and scales using IRT. Following this, we will provide the results of a simulated adaptive administration of that inventory to show the dramatic improvement in the efficiency and precision of measurement, which adaptive testing makes possible. We do this hoping that researchers and others involved in career assessment may be encouraged to begin using these exciting methodological advances.

IRT

In its simplest form, IRT (Bock, 1997) is a method of evaluating scale items wherein the items are described using the same metric as the underlying trait being measured. In traditional test or scale construction, usually done using the concepts of classical test theory (CTT; Wainer & Thissen, 2001) we might have 50 correct out of 100 or a score of 60 on a 20-item Likert-type scale. However, what that score means is dependent on the existence of either norms (e.g., percentiles or standard scores) or some type of criterion referencing (e.g., 70% correct), but neither of these metrics yields nor implies a score on an underlying trait. Further we must administer all of the items to get a score which can be subject to normative comparisons. We can describe item statistics (parameters) such as item difficulty and item discrimination, but an individual item cannot yield a test score (unless it is a one-item scale, frowned upon in classical test theory) nor can it yield coefficient alpha, since alpha is based on item inter-correlations.

In contrast, IRT is based on formulas that assign parameters of ''difficulty'' and discrimination'' to each item based on administration in large-scale development samples. Each item can be used as

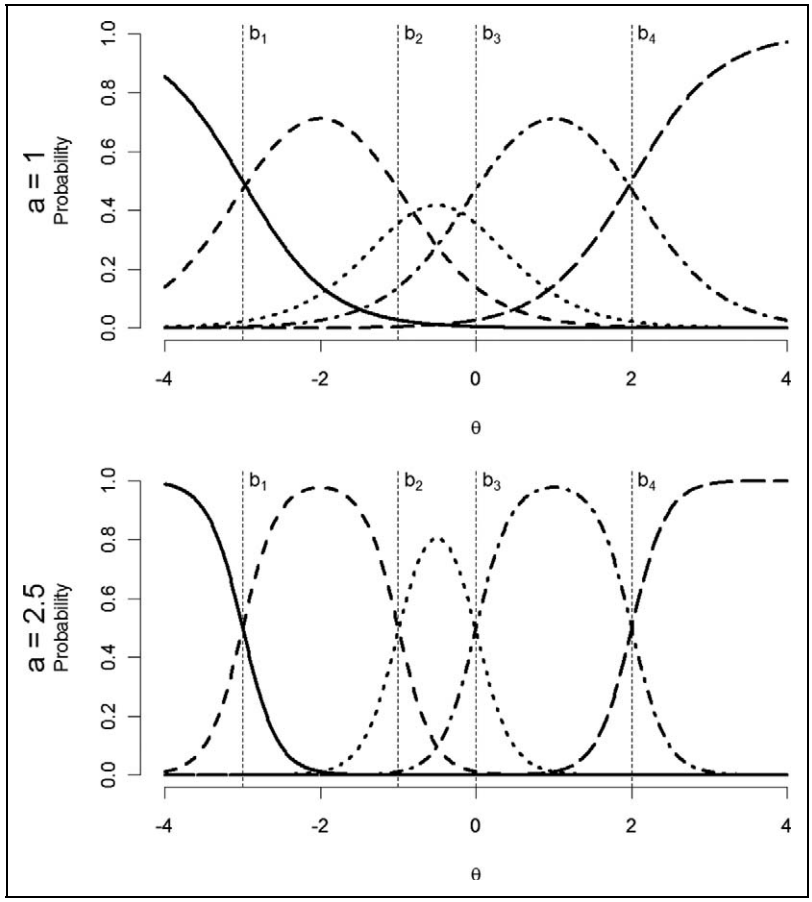

Figure 1. Depiction of representative item response probability distributions for graded (e.g., Likert) response continua.

its own trait estimate because the *b* parameter (item difficulty) is itself an estimate of the level of the latent trait theta (θ); thus, the item's *b* value and the trait are described using the same metric (usually the standard normal distribution with mean 0 and standard deviation of 1, that is, the metric we use with *z* scores). In adaptive testing, the difficulty (trait estimate) of the next item administered can be "adapted" to the individual's response pattern—the next item administered is that which has difficulty level closest to the individual's estimated theta.

IRT also provides an index of the item's discriminatory power (*a*) in the item information function (IIF) for each value of the underlying trait (θ). The sum of the IIFs of the items administered is the total test information function (TIF), which yields an index of the precision of the measurement at each point on the trait continuum. Information is 1 divided by the square root of the standard error of measurement (SEM); the SEM is inversely related to traditional internal consistency reliability coefficients, such as coefficient alpha. However, alpha is an omnibus index, describing the whole test, while information (the inverse of SEM) can be obtained for each point on the trait continuum as well as for the scale as a whole. Accordingly, both item and test quality can be precisely described for individuals with different levels of the underlying trait and using different numbers of items.

Thus, IRT can yield evaluative information that CTT is incapable of providing. For example, a scale may possess an adequate level of internal consistency reliability (usually coefficient alpha) but

be shown by IRT to poorly measure people in one or more areas of the trait distribution (e.g., Hafsteinsson, Donovan, & Breland, 2007). In addition, IRT can be used to provide informative comparisons of two or more measures of the same construct, which may have similar levels of alpha but measure differentially well at different levels of the trait continuum (e.g., Hafsteinsson et al., 2007).

The most frequent uses of IRT have been in educational testing, with aptitude and achievement tests such as the SAT, DAT, and the GRE (Mills & Steffen, 2000; Wainer, 2000). The original IRT models were therefore developed for use with dichotomous item responses (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 (Thissen & Orlano, 2001; Weiss, 2004).

However, with response continua requiring more than two categories, such as the Like/Indifferent/Dislike of some interest measures or the Likert-type response continua used often in personality, attitude, and confidence or self-efficacy measurement, a more common model is Samejima's (1969, 1997) graded response model (GRM). This model is specifically designed for use for polychotomous data where the assumption of ordinality is met and is operationalized using the formula shown in Appendix A. The formula yields indices of item discrimination (a), item difficulty (b), the IIF, and the TIF.

Typical results of applying the equation are illustrated in Figure 1A and 1B. Figure 1A shows an adequately discriminating item while Figure 1B shows a highly discriminating item. In the figures, each response of a 5-point Likert-type scale is represented by one distribution. That is, the response corresponding to 1 (for example, ''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. In Figure 1A, a response of 1 (''not at all confident'') is probably at trait levels below 3 *SD*s below the mean—in normal distribution parlance at or below the 1st percentile. Individuals likely to respond with 5 (completely confident) are likely to have trait levels at or above 2 *SD*s or at least the 99th percentile. The peak of each distribution is the item's difficulty, that is, the trait level corresponding to the greatest frequency of response. The degree to which the distributions do not overlap is the discriminating power of the item. It is clear that the overlap is greater for Figure 1A than that of Figure 1B and the item parameter values (*a*) confirm that, with 1.0 for the item shown in Figure 1A and 2.5 for that shown in Figure 1B. Each item in a scale generates a distribution similar in form to these examples. So for individuals with trait levels at the mean ($\theta = 0$), the most likely responses to Item 1a is 4 (Confident), while for the second item, the most likely response for that individual is 3 (moderately Confident) or 4 (Confident).The second item is somewhat more ''difficult'' than the first.

Adaptive Testing

Adaptive testing (also known a computer-adaptive testing or CAT) has been used since the early 1900s with the Stanford Binet (Weiss, 2004). The metric used in assigning difficulty levels to items was the chronological age (CA) of the average child successfully responding to the item, and the first item administered was usually based on the child's chronological age. Failures led to the administration of easier items (items at lower CAs), and successes led to the administration of more difficult items (higher CAs). Testing was terminated after a specific number of ''failures.'' Not only was a succession of failures providing no new information but it was discouraging to the child. Thus, the difficulty of the items administered was geared to the child's estimated ability level. To do this required the very time-consuming and expensive process of individual test administration. This time and expense were not possible once large-scale group administration of ability tests became the

norm and the necessity, for example, for the placement of vast numbers of military recruits and the testing of millions of students aspiring to go to college and graduate and professional schools.

Now, however, computer technology can instantly "adapt" administration to the individual's level of the trait (whether ability, interests, confidence, etc.) easily. In a special issue of *Measurement and Evaluation in Counseling and Development* on the use of technology in assessment (Wall, Baker, & Sampson, 2004), Weiss (2004) described the usefulness of computerized adaptive testing for effective measurement in counseling and education. Weiss contrasted 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) 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 or 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 has the advantage of minimizing repeated failure experiences when too many too difficult items are administered or leading to boredom for an examinee for whom most or all of the items are too easy. Betz (1977) showed that adaptive ability tests led to higher scores than conventional ability tests for lower ability examinees but that there was no difference for higher ability examinees. A possible interpretation of the findings was that the morale of low ability examinees, and thus possibly the motivation, was better maintained when adaptive testing was used.

Adaptive testing methods may be especially useful in the administration and scoring of multiscale inventories and test batteries, where each scale contains multiple items. Among the most widespread uses of such inventories is in the realm of vocational/career exploration, where multiscale measures of interests, confidence patterns, and sometimes also work values and work-related personality traits are used to guide career exploration and decision making. A combination of such inventories, if well constructed and validated, could result in the necessity of administering several hundred items. Now that much if not most career assessment utilizes online administration and scoring, the possibility of adaptive administration and scoring using IRT offers new possibilities for both streamlined and more precise measurement.

Another advantage of IRT and adaptive testing is that an acceptable degree of error (or its inverse, "information") can be specified and items administered until that level of error is reached. In traditional (CTT-based) scale construction, test developers specified some desired level of alpha for the whole scale. Assume we have a 20-item vocational interest scale (e.g., Science) with an alpha of .80. This seems like efficient measurement until you add 30–40 more interest scales and a like number of confidence scales. Very quickly the number of items needing to be administered can approach several hundred. Some online career assessment systems deal with this problem using short "questionnaires" with no attempt to control psychometric quality, but this seems a move backward rather than forward in terms of testing practice. With adaptive testing, testing ceases for each examinee as soon as the error criterion has been reached, so many fewer items are needed.

Further, although an alpha of .80 characterizes the test as a whole it does NOT describe all test scores in the possible distribution; error (or information) differs along the length of the trait continuum. Thus, one individual may be measured more precisely than another individual who took the same test. This lack of equal precision of measurement could be legally challenged if there were adverse impact for an individual or group of individuals.

In the next section, we present some illustrative findings from a large-scale IRT analysis of the 190-item Career Confidence Inventory (CCI; Betz & Borgen, 2006). The inventory measures self-efficacy or confidence with respect to the six Holland themes and 27 basic dimensions of vocational behavior (e.g., mathematics, science, etc.). The entire inventory takes 20–25min to administer

online. A successful application of computer adaptive testing could reduce this time and the associated costs by as much as 50% for the CCI (see Mills & Steffen, 2000).

IRT Analyses of CCI

Turner, Betz, Edwards, and Borgen (2010) used both classical test theory (CTT) and IRT to evaluate the properties of the CCI, an inventory of measures of self-efficacy for the six Holland (1997) themes and 27 basic dimensions of vocational behavior. The inventory was originally developed using classical test theory analyses of a 240-item pool of activities and school subjects for which confidence ratings in 1,100 adults and 1,800 college students had been obtained (Betz & Borgen, 2006; Betz et al., 2003).

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 ranged from .91 to .94 (Betz & Borgen, 2006). The participants were 2,406 freshmen enrolled in a university career exploration course.

Prior to the IRT analyses we performed an EFA and CFA to derive six unidimensional Holland confidence scales from the original 190-item pool. We did this to facilitate later use of adaptive testing, since adaptive testing is most effective with larger item pools that allow selection of the maximally efficient item at each point in testing a given individual. Following this, the R package ltm (Rizopoulos, 2006) was used to fit Samejima's (1969, 1997) GRM to the data. We then used the IRT parameter estimates (item difficulty and item discrimination, b and a, respectively), to obtain the IIFs, TIFs, and SEMs. For the CTT parameters, we obtained item mean (difficulty), corrected item–total correlations (CITCs; indices of item discrimination), scale means and standard deviations, and values of Cronbach's alpha for each scale as a whole.

Based on the EFA, a model for the subset of items was constructed; numbers of items each Holland scale were 14 (R), 13 (I), 20 (A), 13 (E), 25 (S), and 15 (C), a total of 100 items. A CFA using a different sample of college students (N = 1,620) produced an adequate fit to the data (see Turner et al., 2010). This, then, was the item set for which IRT and CTT analyses were done. We show 4 items per subscale for illustrative purposes here.

Item analyses. The parameter estimates, both CTT and IRT, for four 4 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*, and CITC from classical test theory and the item difficulty or *b* parameters and the item discrimination or *a* parameter from the IRT analysis. 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 et al. (2007) suggest that when there are fewer items in a scale (they used three scales of 8, 8, and 5 items), that a higher standard of item quality may be needed, perhaps 2.0 or better, in order to yield sufficient high-quality overall measurement. Overall, it is the quality of the items that will lead to higher test information and lower standard errors of measurement.

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 E12, "Keep making sales calls even after being rejected",has an *a* parameter value of .83 and a CITC of .39; 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.

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

Scale	Item	*M*	*SD*	CITC	b_1	b_2	b_3	b_4	*a*
Artistic	A4. Create a new logo for a company	3.16	1.10	0.45	−2.85	−0.98	0.47	2.18	1.10
	A5. Play in an orchestra	1.91	1.20	0.45	0.10	1.24	2.00	2.96	1.12
	A11. Sculpt a clay figure	2.38	1.24	0.63	−0.68	0.31	1.10	1.85	2.06
	A17. Create a work of art	2.46	1.35	0.70	−0.62	0.24	0.84	1.39	2.68
Conventional	C3. Invest money in a business opportunity	3.08	1.06	0.61	−2.20	−0.71	0.52	1.84	1.72
	C7. Create a budget for a company's fiscal year	2.54	1.08	0.78	−1.06	0.00	0.98	2.00	3.21
	C8. Record and analyze financial data	2.71	1.13	0.77	−1.23	−0.17	0.80	1.78	2.87
	C13. Audit a company's books	2.20	1.06	0.60	−0.76	0.51	1.66	2.68	1.70
Enterprising	E4. Influence political changes in your community	2.59	1.12	0.71	−1.13	0.01	0.91	1.93	2.45
	E5. Persuade others to support a political candidate	2.74	1.17	0.71	−1.25	−0.12	0.73	1.71	2.39
	E8. Defend people accused of crimes	2.72	1.15	0.62	−1.42	−0.22	0.93	2.03	1.70
	E12. Keep making sales calls in the face of many rejections	2.40	1.19	0.39	−1.38	0.46	1.88	3.52	0.83
Investigative	I4. Successfully complete a demanding course	3.71	0.86	0.42	−6.05	−3.22	−0.52	1.91	0.90
	I6. Conduct a study on the effects of new medications	2.74	1.16	0.77	−1.15	−0.20	0.71	1.68	2.90
	I8. Understand the scientific basis of a medical breakthrough	2.79	1.20	0.84	−1.04	−0.17	0.66	1.62	3.73
	I11. Pass a course in Plant Biology	3.14	1.20	0.63	−1.87	−0.74	0.39	1.51	1.60
Realistic	R2. Hike on a mountain trail	3.67	1.17	0.51	−2.71	−1.65	−0.42	0.91	1.25
	R5. Fight fires	2.26	1.15	0.71	−0.57	0.41	1.27	2.14	2.34
	R8. Work as a police officer	2.42	1.16	0.59	−0.94	0.23	1.25	2.24	1.71
	R11. Be a restaurant chef	2.44	1.19	0.44	−1.21	0.32	1.57	3.04	1.04
Social	S5. Counsel a distressed person	3.46	1.08	0.67	−2.31	−1.18	−0.01	1.21	1.88
	S15. Help a group of people to cooperate better	3.47	0.96	0.72	−2.51	−1.35	−0.01	1.41	2.26
	S20. Collaborate with others to get a job done	4.03	0.84	0.56	−3.81	−2.72	−1.01	0.71	1.54
	S24. Contribute ideas to your work team	3.80	0.88	0.61	−3.59	−2.08	−0.55	1.07	1.72

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

Table 2. Average Numbers of Items Needed to Reach Specified Levels of SEM

	SEM [95% Confidence Interval]			
	0.60	0.50	0.40	0.30
Artistic	1.62 [1.46, 1.78]	2.63 [2.39, 2.87]	4.51 [3.88, 5.13]	10.02 [9.20,10.83]
Conventional	1.23 [1.14, 1.31]	2.20 [2.09, 2.32]	2.92 [2.63, 3.21]	7.02 [6.63, 7.41]
Enterprising	1.26 [1.17, 1.35]	2.36 [2.22, 2.51]	3.99 [3.73, 4.27]	9.23 [8.89, 9.56]
Investigative	1.00 [1.00, 1.00]	1.40 [1.26, 1.54]	2.62 [2.32, 2.91]	5.09 [4.67, 5.51]
Realistic	2.18 [2.10, 2.26]	3.01 [2.74, 3.28]	5.01 [4.60, 5.42]	10.54 [10.08, 10.99]
Social	2.14 [2.07, 2.21]	3.27 [3.15, 3.39]	5.53 [5.17, 5.89]	11.58 [11.11, 12.06]

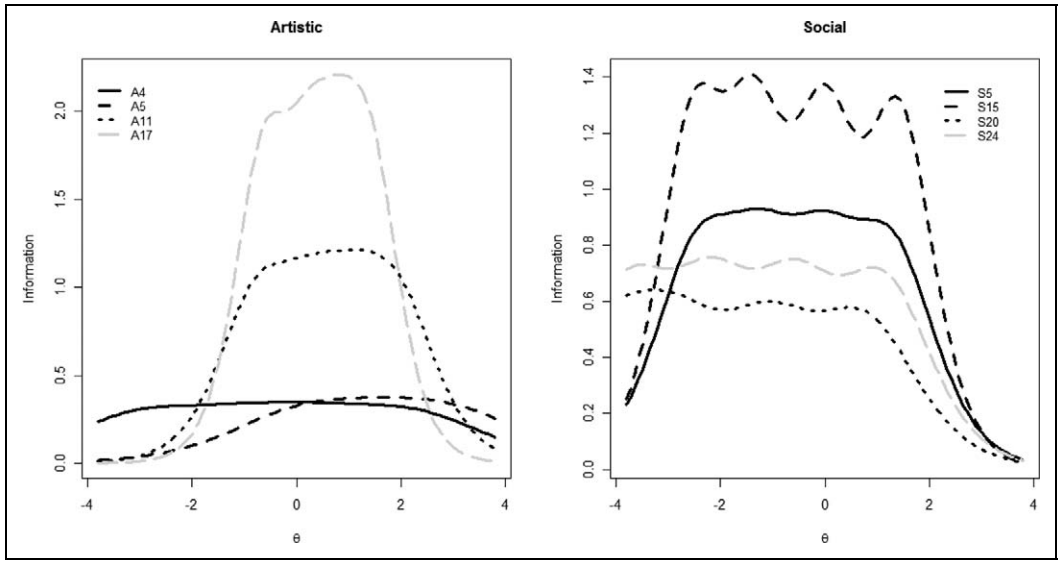

Figure 2. Item information functions (IIF) for four illustrative items for two of the six subscales: Artistic and Social confidence.

On the other hand, the values of *a* show somewhat greater differentiation in the indication of item quality than do the CITCs (this is often found in such analyses; e.g., Scherbaum, Cohen-Charash, & Kern, 2006). For example, Items A17, E4, and E5 have ITCs of .70 or .71, yet the *a* for A17 is 2.68 while that for E4 is 2.45 and that for E5 is 2.39.

Further demonstration of what the analysis yields is provided by the IIFs for the items from the Artistic and Social themes, as shown in Figure 2. The figure shows clear differences in how well the items measure at different points on the trait continuum. For example, A11 and A17 provide rich information for the middle of the θ continuum while A4 and A5 provide poorer measurement around the middle but slightly better measurement at the extremes (the lower end for A4 and the upper end for A5). The four items shown for Social generally provide higher levels of information and, in comparison to the Artistic items, much better measurement at the low end of the trait distribution. None of the four social items provide good measurement at the high end of the trait distribution.

Figure 3 shows a TIF and SEM curves for the Artistic and Social scales. These include all items in each scale (vs. just the four shown for illustrative purposes in the previous figure), but the relatively better measurement shown for the Social items in Figure 2 is paralleled by higher levels of information (solid line) and lower SEM (dotted line) in Figure 3. As shown, the Social scale provides rich information near the middle and lower ends of the trait distribution. The SEM is quite low for values of θ in the range of -3 to $+2$, which under IRT, should contain approximately 99% of the population. The information function for 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).

It is useful to compare these information functions and standard errors to the index of reliability usually provided based on classical test theory 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 of these two scales. The other four Holland confidence scales had equally distinct information functions that would have not appeared different under the metric of coefficient alpha.

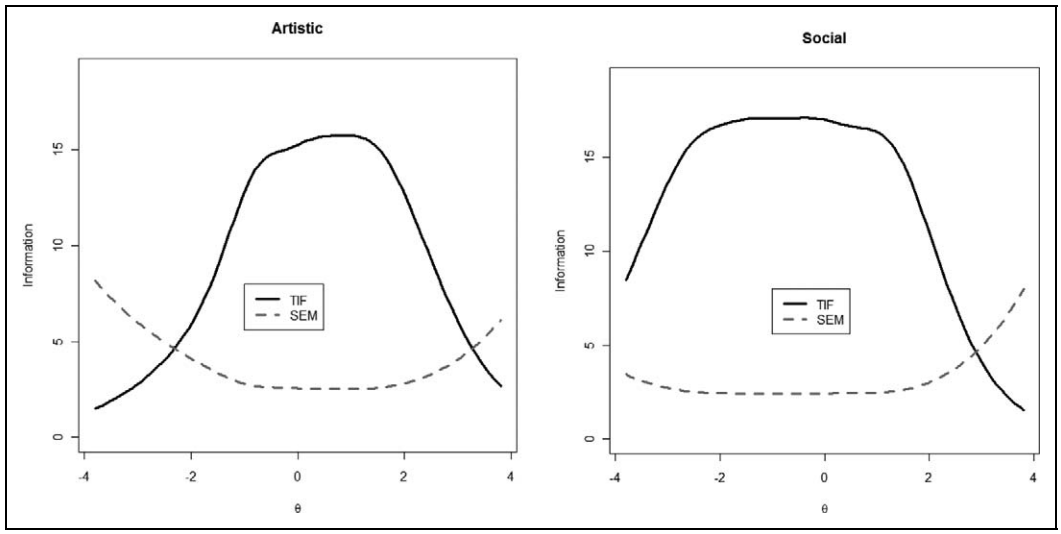

Figure 3. The total or scale information function (TIF) and standard error of measurement (SEM) for Artistic and Social confidence across the trait continuum.

Simulation Study of Adaptive Testing

Following the IRT analysis of the CCI (Turner et al., 2010), we used the *a* and *b* parameters generated for each item in a simulated adaptive administration of each of the six Holland confidence scales. We were interested in the numbers of items necessary for each theme to achieve a given level of precision of measurement (i.e., akin to internal consistency reliability). Note that with adaptive testing, it is possible to state a given level of desired precision for all examines, and the number of items needed depends on the examinee's responses, presumably reflecting his or her level of the trait, and the specified criterion level of precision.

The simulations were done using the software program FIRESTAR Version 1.2.2 (Choi, 2009). To complete each of the simulations below, we used the program to randomly generate values of the trait, assuming the standard normal distribution $\theta \sim N(0,1)$. These values represent examinees with some "known" latent trait ability (θ) and will be referred to as simulated examinees. The program then administers the scale to the simulated examinees adaptively by sequentially selecting items one-at-a-time with the intentions of minimizing the standard error (in a Bayesian framework, the standard error is the *SD*) of the posterior distribution of θ for a particular simulated examinee. This posterior distribution is estimated after each item administration and after each item the error band should be narrower. The test developer should decide on a value for the SEM so that once the CAT has achieved a value of the SEM that is equal to or lower than the specified value, testing is terminated for that individual.

The first simulation allowed us to determine how many items were needed, for each subscale, to reach specified criteria of precision. Precision is inversely related to the standard error of measurement—greater precision is shown by less error. Table 2 shows the average numbers of items needed to achieve given levels of precision, as defined by standard errors of measurement from .60 on the left to .30 on the right, for each of the six Holland confidence themes. For example, to achieve a SEM of .40, equivalent to an alpha of .84, we must use on average 4.5 items (A), 2.92 (C), 3.99 (E), 2.62 (I), 4.6 (R), and 5.5 (Social) or a total 25 of the 100 items, illustrating maximally efficient yet high-quality measurement. Note that as the SEM decreases, the number of items necessary to reach that criterion increases and vice versa.

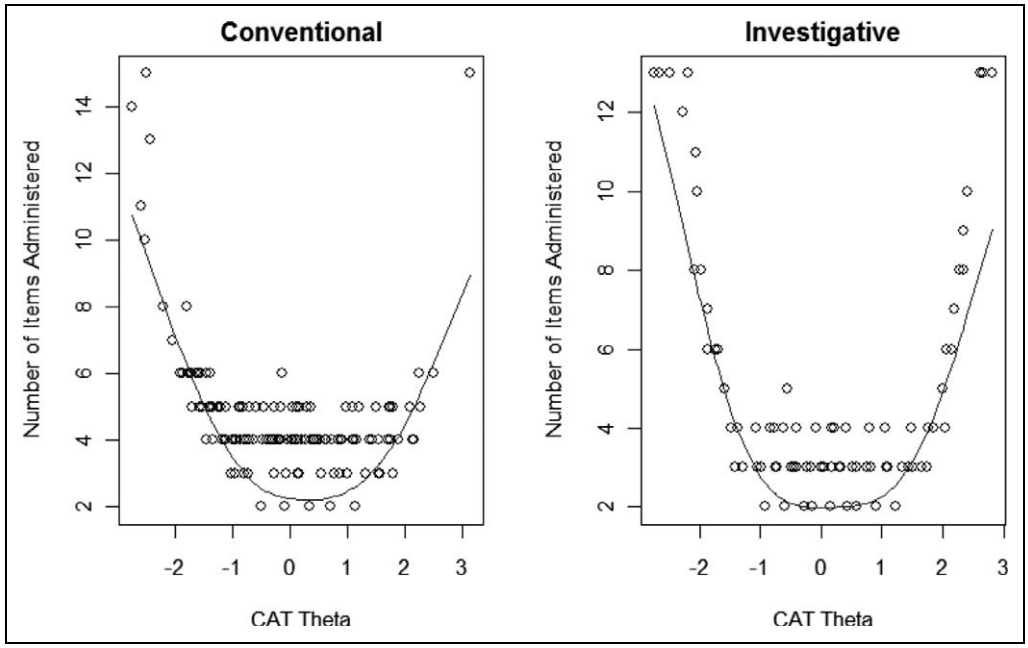

Figure 4. The numbers of items required in adaptive testing to achieve a standard error of .40 ($\alpha = .84$) across the trait continuum for Conventional and Investigative confidence scales.

Further evidence of how well these shortened scales were working was the correlations between the a priori (simulated) trait estimates and the trait score assigned after adaptive testing. Using fewer than 6 items per each scale, Turner et al. (2010) found correlations between the a priori assigned value and the test estimates to be at worst .93.

However, what Table 2 does NOT show is that the numbers of item necessary to reach the error criterion varies for individuals with different levels of the trait. In other words, some values of θ are more easily estimated than others. To examine this, we conducted a simulation wherein we computed the average number of items administered as a function of θ with intentions of identifying values of θ that were not estimated well by the subscales. Two examples are shown in Figure 4—those for the Conventional and Investigative scales, since they provide some contrasts. The figure shows, on the vertical axis, the number of items needed to achieve a SEM of .40 across levels of the trait (θ; the horizontal axis). It is obvious that fewest items are required at the center of the distribution, that is, people with ''average'' levels of the trait are easiest to estimate precisely with fewest items. For people with more extreme levels of the trait, it is necessary to administer more items to achieve this level of error of measurement (.40). This necessity is especially apparent for lower levels of Conventional Confidence, where it takes between 4 and 12 items to estimate precisely below an average level of θ. For the Investigative scale, more items are required for adequate estimation at both high and low levels of the trait.

Discussion

This article has been designed to introduce readers to the ideas of IRT and adaptive testing in the hopes that more use will be made of them in career assessment. The advantages of IRT are a more precise evaluation of item and scale qualities, allowing us to evaluate item quality at different levels of the trait using the criteria of information and its inverse—the standard error of measurement.

Traditional classical test theory indices, most usually internal consistent reliability (coefficient alpha, Kuder-Richardson-20, and split half coefficients) describe the scale as a whole, for all levels of the trait measured.

Further, IRT allows us to get ability estimates for every item administered and to derive valid total scores even when the number of items administered and the items themselves are not the same across individuals. This feature of IRT allows the development of adaptive measures, where the items are adapted or geared to the individual examinee's level of the trait. This can have psychological and motivational advantages for the examinee, reducing testing time and minimizing successive experiences of failure for lower ability examinees (Betz, 1977), but it can also significantly reduce testing time when multiscale inventories or batteries of tests are to be administered. These simulations of adaptive testing provide a good illustration of the latter point. Assuming a standard error of .40, examinees with a trait distribution similar to the standard normal and a value for coefficient alpha of .84, we were able to cease testing with far fewer than the 100 items comprising six Holland confidence scales—specifically a total of about 25 items. Further evidence of how well these shortened scales were working was the correlations between the a priori trait estimate and the trait score assigned after adaptive testing. Using less than 6 items per each scale, we found correlations between θ and $\hat{\theta}$ to be at worst .93.

In summary, IRT and adaptive testing have exciting implications for the future of career assessment. Powerful computers and sophisticated software can provide increasingly efficient and high-quality measurement even when a large number of variables are included in the assessment system.

Appendix A

Samejima's GRM Model

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

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

In this equation, 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.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interests with respect to the authorship and/or publication of this article.

Funding

The authors received no financial support for the research and/or authorship of this article.

References

Betz, N. (1977). Effects of immediate knowledge of results and adaptive testing on ability test performance. *Applied Psychological Measurement*, *1*, 259–266. doi:10.1177/014662167700100212.

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

Betz, N. E., & Borgen, F. H. (2009). Comparative effectiveness of CAPA and FOCUS online: Career assessment systems with undecided college students. *Journal of Career Assessment*, *17*, 351–366. . doi:10.1177/1069072709334229.

Betz, N. E., Borgen, F. H., Rottinghaus, P., Paulsen, A., Halper, C., & Harmon, L. W. (2003). The Expanded Skills Confidence Inventory: Measuring basic dimensions of vocational activity. *Journal of Vocational Behavior*, *62*, 76–100. doi:10.1016/S0001-8791(02)00034-9.

Bock, R. D. (1997). A brief history of item response theory. *Educational Measurement: Issues and practice*, *16*, 21–23. doi:10.1111/j.1745-3992.1997.tb00605.x.

Borgen, F., & Betz, N. (2008). Career self-efficacy and personality: Linking the Career Confidence Inventory and the Healthy Personality Inventory. *Journal of Career Assessment*, *16*, 22–43. doi:10.1177/1069072707305770.

Fowkes, K. M., & McWhirter, E. H. (2007). Evaluation of computer-assisted career guidance in middle and secondary education settings: Status, obstacles, and suggestions. *Journal of Career Assessment*, *15*, 388–400. doi:10.1177/1069072707301234.

Fukuyama, M., Probert, B., Neimeyer, G., Nevill, D., & Metzler, A. (1988). Effects of DISCOVER on career self-efficacy and decision making of undergraduates. *Career Development Quarterly*, *37*, 56–62.

Garis, J., & Niles, S. (1990). The separate and combined effects of SIGI and DISCOVER and a career planning course on undecided university students. *Career Development Quarterly*, *38*, 261–274.

Gore, P. A., Bobek, B., Robbins, S., & Shayne, L. (2006). Computer-based career exploration: Usage patterns and a typology of users. *Journal of Career Assessment*, *14*, 421–436.

Gore, P. A., & Leuwerke, W. C. (2000). Information technology for career assessment on the internet. *Journal of Career Assessment*, *8*, 3–20. doi:10.1177/106907270000800102.

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.

Harris-Bowlsby, J., & Sampson, J. P. Jr. (2001). Computer-assisted career guidance planning systems: Dreams and realties. *Career Development Quarterly*, *49*, 250–260.

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

Katz, M. R. (1973). Career decision making: A computer-based system of interactive guidance and information (SIGI). In: *Proceedings of the Invitational Conference on Testing Problems* (pp. 43–69). Princeton, NJ: Educational Testing Service.

Luzzo, D., & Pierce, G. (1996). Effects of DISCOVER on the career maturity of middle school students. *Career Development Quarterly*, *45*, 170–172.

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. 7599). Dordrecht, Netherlands: Kluwer Academic Publishers.

Oliver, L., & Chartrand, J. (Ed.). (2000).Career assessment and the internet: Special issue. *Journal of Career Assessment, 8*, 1–104. doi:10.1177/106907270000800101.

Peterson, G. W., Ryan-Jones, R. E., Sampson, J. P. Jr., Reardon, R. C., & Shahnasarian, M. (1994). A comparison of the effectiveness of three computer-assisted career guidance systems: DISCOVER, SIGI, and SIGI-PLUS. *Computers in Human Behavior*, *10*, 189–198.

Rayman, J. R., & Harris-Bowslby, J. (1977). Discover: A model for a systematic career guidance program. *Vocational Guidance Quarterly*, *26*, 3–12.

Reile, D. M., & Harris-Bowlsby, J. (2000). Using the internet in career planning and placement. *Journal of Career Assessment*, *8*, 69–84. . doi:10.1177/106907270000800106.

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.

Samejima, F. (1997). Graded response model. In: W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 85–100). New York, NY: Springer.

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. doi:10.1177/0013164406288171.

Thissen, D., & Orlando, M. (2001). Item response theory for items scored in two categories. In: D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73–140). Mahwah, NJ: Lawrence Erlbaum.

Turner, B., Betz, N., Edwards, M., & Borgen, F. (2010). Psychometric examination of an inventory of self-efficacy for the Holland themes using Item Response Theory. *Measurement and Evaluation in Counseling and Development*, *43*, 188–198.

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

Wainer, H., & Thissen, D. (2001). True score theory: The traditional method. In: D. Thissen & H. Wainer (Eds.), *Test scoring* (pp. 73–140). 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.

Career assessment and new technology on the internet. *Journal of career assessment*, *18*, 315–361.

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.