Examining Psychometric and Measurement Properties of the Career Thoughts Inventory: Demonstration and Use of the Rasch Measurement Model in Career Assessment Research

Technical Report No. 51

Christian E. Mueller[1]
University of Memphis

Emily E. Bullock[2]
University of Southern Mississippi

Stephen J. Leierer[3]
University of Memphis

April 2, 2010

[1] Correspondence should be addressed to Christian Mueller at cemuellr@memphis.edu; or, the University of Memphis; Department of Counseling, Educational Psychology & Research; 100 Ball Hall; Memphis, TN 38152-3570; 901-678-4392.

[2] Emily Bullock can be contacted at the Univ. of Southern Mississippi; Dept. of Psychology; 118 College Dr. #5025; Hattiesburg, MS 39406-0001. (601) 266-6603. Emily.Bullock@usm.edu

[3] Stephen Leierer can be contacted at the University of Memphis; Department of Counseling, Educational Psychology & Research; 100 Ball Hall; Memphis, TN 38152-3570. (901) 678-3411. sleierer@memphis.edu

Abstract

The Rasch measurement model for developing and revising career assessment tools has many advantages over traditional test development methods. To better understand this method, the current study met its two purposes (a) to illustrate how the Rasch measurement model can aid vocational psychology researchers in increased precision and accuracy in assessment; and (b) to examine the psychometric and measurement properties the Career Thoughts Inventory (CTI), using the Rasch measurement model with a sample of 232 college students. Results from the Rasch analysis confirmed that most of the CTI items on the decision-making confusion, commitment anxiety, and external conflict subscales were functioning as useful measurement items. Yet, these subscale items were somewhat restricted in the range with which they were measuring their respective type of dysfunctional career thinking.

Examining Psychometric and Measurement Properties of the Career Thoughts Inventory: Demonstration and Use of the Rasch Measurement Model in Career Assessment Research

Technical Report No. 51

Career counselors often rely on vital information gained from personality and psychological inventories to help guide their overall decision-making and in working with the complex issues presented by clients in clinical settings. To be successful in the design and implementation of effective treatment strategies, career counselors and vocational psychologists must use the most appropriate inventories which will provide precise and accurate measurements of the clients' complex career issues. Common assessment types include personality inventories (e.g., NEO-PI-R; Costa & McCrae, 1992), career interest inventories (e.g., Strong Interest Inventory; Harmon, Hansen, Borgen, & Hammer, 1994; Self-Directed Search; Holland, Fritzsche, & Powell, 1994), and career development inventories, such as the Career Thoughts Inventory (CTI; Sampson, Peterson, Lenz, Reardon, & Saunders, 1996), which is typically used to assess dysfunctional thinking, distress associated with career indecision, or readiness for decision making. Historically, quantitative researchers have utilized three methods to develop and validate inventories of this type: (a) psychometrics or classical test theory (CTT), (b) item response theory (e.g., Rasch modeling), and (c) generalizability theory (Fox, 1999); although Pomeranz, Byers, Moorhouse, Velozo, & Spitznagel (2008) note that Rasch modeling is quickly gaining popularity among career and rehabilitation researchers. By far, the most common method used by social science researchers is that of CTT, with Allen and Yen (1979) and Dubois (1970) attributing this popularity to the earlier influences of Karl Pearson and Sir Francis Galton on quantitative research methodology (see Allen & Yen or DuBois for further discussion on this topic). Wright and Stone (1979) and others (e.g., Bond & Fox, 2001) note, however, that there are inherent flaws in exclusively using CTT to develop and validate inventories, because of the increased measurement error introduced by CTT methods. Further, Bond and Fox note, "[researchers within the social sciences] are too narrowly focused on statistical analysis, and not concerned nearly enough about the quality of the measures on which they use these statistics" (p. 1). The purpose of the present study is twofold: first, to illustrate how the Rasch measurement model can aid vocational psychology researchers in their pursuit of increased precision and accuracy in personality and psychological assessment; and second, to examine the psychometric and measurement properties of an existing career assessment inventory, the Career Thoughts Inventory (CTI), using the Rasch measurement model.

Measurement is Fundamental in the Social Sciences

Precision, accuracy, and consistency in measurement are fundamental goals for any scientific researcher, whether in the “hard sciences” or in the social sciences. Bond and Fox (2001) note that these goals have until recently, within the past century or so, mostly been attained in the hard sciences, although this was not always the case historically. Early efforts to develop consistent measurement practices with Celsius and Fahrenheit scales, for example, often resulted in heated debate among researchers in earlier times in those areas (see Bond & Fox for further discussion). Many of these debates now exist in the social sciences, as researchers attempt to replicate similar levels of precision and accuracy in the measurement of human abilities and latent constructs (e.g. personality and psychological traits). In a practical sense, the extent to which researchers can reduce measurement error through the use of precise measurement instruments that are both valid (accurate) and reliable (consistent), is the extent to which research in the social sciences will attain the respect it has achieved in the hard sciences (Fisher, 2008).

Human beings are complex; therefore, the measurement of human traits and abilities is also complex. As a result, prediction of future behavior based upon the measurement of human traits and abilities is also complex. Before highlighting some of the problems that can occur in the attempt to measure human traits and abilities, it is perhaps useful to provide an ideal picture of what fundamental measurement should look like and how the Rasch model can move researchers closer to that ideal. A primary assumption of latent-ability or latent-trait measurement is that as an individual’s ability increases, the probability of answering correctly on any select item also increases (Bond & Fox, 2001). When measuring latent-traits, this translates to increases in any given trait will lead to agreement or endorsement on more extreme items.

According to the Rasch model, a well-constructed psychological assessment will be valid and reliable in that it accurately and consistently measures the latent construct of interest, regardless of time or place measured; will possess sufficient variability and range of items to measure the entire continuum on the latent construct of interest (i.e. lowest to highest); and lastly, can safely be assumed that all items will be measuring the same latent construct of interest (Bond & Fox). Through use of several statistics and procedures, the Rasch model is able to ensure that items and assessments are able to accomplish these goals.

Measurement Problems in the Social Sciences

As most career and vocational researchers are aware, measurement error is one of the most challenging issues facing quantitative researchers in the social sciences. Whereas measurement in the “hard sciences” (e.g., physics) is generally considered to be precise and accurate at all times, where an inch is always an inch no matter the time or place measured, this view has not often been held in respect to social sciences research (Bond & Fox, 2001). There are numerous factors that influence the measurement of personality and psychological constructs, including individual factors, such as being tired, guessing on a difficult item when one does not know an answer, or in wanting to be perceived more favorably by the tester. In addition, numerous factors related to the inventory or to individual measurement items may also introduce measurement error, including poorly worded items, directions that are confusing, inventories that are too short to gain meaningful information or inventories that are too long and may cause respondents to lose interest. In most social science research, measurement has traditionally been defined using Stevens’ (1946) standard of assigning numbers to objects or events using certain rules [and meeting certain assumptions] (Bond & Fox). Further, quantitative researchers have typically assumed that measurement of latent constructs occurs simply as a function of collecting data through use of existing inventories or in the development of new inventories and assessments. As noted above, when quantitative researchers are not fully aware of or are not concerned about fundamental problems in measurement, issues with overall validity and reliability may limit the value and utility of data collected through these procedures, as well as the conclusions that can be drawn from those data.

Bond and Fox (2001) and Wright and Stone (1979), among others, have highlighted three common measurement issues that threaten validity and reliability in traditional quantitative research methods. First, researchers often mistakenly assume that Likert-scaled items are at the interval level of measurement and that item endorsements mean the same thing for all respondents. Second, that all of the items on the inventory are tapping into or measuring all levels on the latent construct of interest. And lastly, that multiple factors influence item response patterns, including those of the individual (e.g., being too tired), as well as those associated with a single item or group of items (e.g., poor wording or confusing directions). Historically, there has been no systematic way to explore these issues using traditional psychometric techniques (Bond & Fox). For reasons outlined later, the Rasch model is not vulnerable to biases and characteristics unique to entire samples or groupings of items, and as such, allows researchers to address each of these issues at both the person and item level. This last point is further explored in a later section of the article (see discussion on calculating of standard errors).

Faulty Assumptions in Likert-Scaled Item Measurement

Two of the most common mistakes that quantitative researchers make in both developing and using Likert-scaled items are assuming that items are always on an interval scale (i.e., equidistant measurement) and that response options are always on an equivalent scale (i.e., participant responses indicate similar levels). For example, on a typical five-point Likert scale, researchers will often treat the difference between a response of “1” or “2” as being equivalent to the difference between “4” and “5” and carry this assumption across all items. In theory, this assumption may or may not hold true; more importantly, however, there has historically been no systematic way of ascertaining whether this is true in practice. Another mistake that quantitative researchers often make in the development and use of Likert scaled items is in assuming that item endorsements are always on an equivalent scale, where a response of “5” means the same thing for all respondents. Again, this may not always be the case in practice. For example, when asking respondents to answer a question related to *How sad are you?*, clients may be indicating different levels of sadness, even though they respond on the highest possible category of “5.” Response patterns to an item assessing sadness are influenced by multiple factors, including the particulars of a given sample (e.g., assessing levels of sadness in college students versus that of psychiatric patients), how items are worded, and whether respondents are interpreting items in the manner they were designed. Thus, while the distressed college student may indicate a “5” because it is the highest level of sadness for them, the score of “5” by the psychiatric patient will almost always reflect a higher level of distress on the part of that individual. While the differences in response patterns caused by these confounding factors may be unintended, the results may be problematic for both researchers and clinicians.

Two related problems associated with the measurement of latent constructs are those of *item redundancy* and ensuring a logical *item hierarchy* in both range and order among the measurement items. Item redundancy is thought to occur when specific items do not provide useful information, from a measurement perspective, about the latent construct of interest. This usually occurs when the items are either highly correlated or when there is dependency among the items (Tang et al., 2005). Ensuring a logical item hierarchy encompasses several factors, including variability, distribution and range (see Methods for further details). In clinical settings, where precise and accurate information must be gained in a timely manner, and with the fewest measurement items possible, these issues have added importance. Other issues that inadvertently increase measurement error and negatively affect latent construct measurement include respondents incorrectly interpreting directions, inconsistent wording among items, and items not measuring the same unidimensional construct (Bond & Fox, 2001; Sampson & Bradley, 2003). While Rasch measurement is not offered as a panacea to all measurement problems in the social sciences, it does offer unique characteristics that can help career and vocational researchers in their pursuit of precision and consistency in latent construct measurement.

Overview of the Rasch Measurement Model

Ideally, the Rasch measurement model is best utilized during the initial development and validation of any new scale or inventory (Sampson & Bradley, 2003; Wright & Stone, 1979), however, there are times when Rasch can be used to improve the reliability and validity of existing measures (e.g., Pomeranz et al., 2008) and, when appropriate, in developing a shorter form of an existing measure (e.g., Cole, Rabin, Smith, & Kaufman, 2004). As Bradley and Sampson (2005) write “Rasch analysis begins at the level of measurement, providing diagnostic information on the quality of the measurement tool, in addition to yielding a more comprehensive and informative picture of the construct under measurement as well as the respondents on that measure” (p. 12). Originally introduced by George Rasch (1960, 1980), the Rasch model assumes that the relationship between an individual’s ability level and an item’s designed difficulty can be mathematically modeled as a probability. As ability increases, so does the probability of answering items correctly. Conversely, as ability level decreases, so does the probability of answering items correctly. This relationship is the fundamental principal that underlies the Rasch measurement model. When measuring personality, attitudinal or psychological constructs, however, the concept of "ability" can be viewed as "endorsability" (see Fox & Jones, 1998 for further discussion on this issue). On the Career Thoughts Inventory, for example, respondents are not likely to rank highly (or endorse) extreme statements such as *No field of study or occupation interests me* unless they are experiencing high levels of decision-making confusion. When items are functioning properly; that is, meeting certain assumptions of the Rasch measurement model, then with increasing levels of an attribute in an individual, the probability of endorsing any item also increases. Because Rasch begins at the level of measurement, with both items and persons, it does not rely on aggregated sample and item characteristics to calculate values for statistics; rather, these are calculated independently for each person and item based upon how well expected and observed response patterns match (Smith, 2004b).

Researchers using only traditional quantitative methods are limited in that they are asking a basic question of "How does a statistical model fit my data?", where characteristics of a given sample or group of items may limit the utility of the statistics calculated from those data. In an applied setting, this question might take the form of "How well does the model describe or diagnose problematic issues in my clients?" In contrast, when researchers use Rasch methodology, they are in essence asking, "How does my data fit the Rasch measurement model?" Again, when certain assumptions are met, items are assumed to be useful indicators of the latent construct of interest, and in applied settings the practitioner is able to assume that information gained from the assessment is precise and accurate. Bradley and Sampson (2006) and others (e.g. Smith, 2004a) note that the essential difference between Rasch measurement and traditional quantitative methodology is in how the standard error is calculated differently using both techniques.

Bradley and Sampson note,

> Rasch measurement places person ability and item difficulty along a linear scale, so these estimates may be used for calculating means and variances. In contrast to CTT, which reports only the error variance for an average person sampled, Rasch measurement produces a standard error (SE) for *each* person and item, specifying the range within which each person's 'true' ability and each item's 'true' difficulty fall. The individual errors can then be used to produce a more accurate average error variance for the sample (p.24).

Because Rasch begins at the level of measurement, with both items and persons, it does not rely on aggregated sample and item characteristics to calculate values for statistics; rather, these are calculated independently for each person and item based upon how well expected and observed response patterns match (Smith, 2004). Thus, Rasch statistics are free from inherent bias generated by aggregating sample data. In this way, Rasch is assumed to test the assumption of how well the data fit the assumptions of the model, rather than how well the statistical model fit the sample data. For readers who may be more familiar with statistical terminology (e.g., factor analysis), Bond and Fox (2001) responded to a Rasch critic (Goldstein) in these terms:

> Goldstein's comment presupposes that the sole objective of data analysis is to manipulate the data analytical procedures until the amount of variance that cannot be explained is reduced to a minimum…From this perspective, the primacy of the empirical data is paramount. The task of data analysis is to account for the idiosyncrasies of the data (p. 191).

In addition, Bond and Fox wrote:

> From the fundamental measurement perspective, the requirements of the measurement model are paramount. The idiosyncrasies of the empirical data are of secondary importance. The measurement ideal, encapsulated in the Rasch model, has primacy. The researcher's task is to work toward a better fit of the data to the model's requirements until the match is sufficient for practical measurement purposes in that field (p. 191).

Therefore, when using the Rasch model the researcher becomes more concerned with how items are measuring a construct of interest rather than dealing after the fact with the idiosyncrasies of the data collected (see Bond & Fox, 2001, pp. 190-191 for further discussion on this issue).

Evaluating Measurement Assumptions through use of the Rasch Model

Rasch analysis, by way of measurement software (e.g. Winsteps), provides several statistics and tools that aid researchers in testing for the assumptions of the Rasch model. When these assumptions are met, researchers can assume that assessments and items on those assessments are functioning as effective indicators of a particular latent construct (i.e. ability or trait). In the present study, we were specifically interested in testing for the assumptions of *unidimensionality; logical item and person hierarchy,* where there is a logically ordered match between the items on an inventory and a given sample of respondents; and lastly, high probability of *replication of results*.

In order to test for the assumption of unidimensionality, the Rasch model utilizes fit statistics (*infit* and *outfit* statistics). Both infit and outfit statistics provide an indication of how well expected and observed response patterns match. To the degree that individual items fall within pre-determined ranges, researchers can confidently conclude that the items are useful indicators of the latent construct of interest, or, are confirming the assumption of unidimensionality. Fit statistics also provide an indication of how consistently individuals respond to particular items. For example, if a few people with low levels of decision-making confusion endorse an item that is representative of high confusion, this response pattern would result in a high outfit statistic. Individuals must answer items somewhat predictably in order to obtain good fit statistics. Once items are found to have good fit, *variable maps* (also referred to as *person/item maps*) can be used by researchers to graphically depict and explore how well a given bank of items is measuring the latent construct for a given group of respondents. Lastly, researchers may also utilize both *person* and *item* reliability estimates to provide further support for the measurement value of a given bank of items. When person and item reliability estimates are high, this indicates high replicability of results across both persons and items where. Items deemed to be problematic in any of these areas may subsequently be dropped or flagged for further inspection by the researcher.

Fit Statistics (Infit and Outfit). Fit statistics, which are provided in Winsteps as both infit and outfit, and in unstandardized mean-square (MNSQ) and standardized (ZSTD) forms, are the main indicator of how well individual items are meeting the assumption of *unidimensionality* in the Rasch model (Linacre, 2002a). In latent construct measurement, it is assumed that only items designed to measure the latent construct of interest will be included on an inventory. Many times, this process has been limited, because of reliance on principal components or factor analysis, which is sample dependent because of how the statistics are calculated; i.e. each statistic is influenced by the responses of everyone in the sample, even when responses may be influenced by extraneous factors such as guessing or fatigue (Bond & Fox, 2001; Bradley & Sampson, 2003). In contrast, Rasch calculates these statistics at both the item and person level based upon expected and observed response patterns and how these work together according to probabilistic expectations.

Infit and outfit statistics are derived from and are thus indicators of how well observed response patterns fit with the expected response patterns. A person possessing more of an attribute should have a higher probability of endorsing any item than a person with less of that attribute (regardless of the endorsability of the item), and an item that reflects less of an attribute should have a higher probability of being endorsed than an item that reflects more of an attribute (regardless of the attitude of the person responding to the item). In the present study, this would take the form of individuals experiencing higher levels of *decision-making confusion*, *commitment anxiety*, and *external conflict* would have a higher probability of endorsing a "4" (using CTI response options) on the items designed to measure the higher extremes on these constructs. Infit statistics are more sensitive to unexpected response patterns located near the person or item measure, such as with the person whose ability or level of attribute is close to the designed sensitivity of a particular item. When observed responses do not match expected responses (according to the Rasch model), these items would be flagged as problematic. Outfit statistics are more sensitive to unexpected response patterns that are outliers, such as when individuals with extreme levels of an attribute or high levels of ability are consistently scoring lower on items designed to be less sensitive or score incorrectly on items designed to be easier. In most cases, unacceptable fit statistics indicate either guessing, confusion on the part of the respondent, or poor wording on the part of the item (Linacre, 2002a). When items fall outside of the acceptable threshold, these items should be further examined to determine the cause of the poor fit statistics (Bond & Fox, 2001; Smith, 2004b).

According to the Rasch model, the expected value for unstandardized infit and outfit statistics is 1.0. When values are substantially less than 1.0, this indicates that observed responses are too predictable, or are redundant. In other words, these items are not adding to the measurement of the latent construct. When values are substantially greater than the expected 1.0, this typically indicates response patterns that are unpredictable and occurs when items do not seem to be measuring the same construct. This happens when items are either poorly worded or are ambiguous in their wording to the point that they are confusing to respondents (Linacre, 2002b). In order to determine whether an item is misfitting, the researcher should review fit statistics using the following criteria: outfit before infit, mean-square (MNSQ) in conjunction with standardized (ZSTD), and addressing extremely high values before extremely low values. Linacre (2005) notes that, with small numbers of responses on a particular item (i.e., less than 30), standardized fit statistics alone may be too insensitive (i.e., "everything fits") and with large numbers of responses to an item (i.e., greater than 300), the standardized fit statistics will be too sensitive, in which "everything misfits"; therefore, standardized fit statistics were used in conjunction with the mean-squares to identify misfitting items (for further discussion on this issue in career and rehabilitation counseling, see Pomeranz et al., 2008, pg. 253).

Pre-determined thresholds or limits with MNSQ infit and outfit statistics are typically used to evaluate how poorly items are fitting the expectations of the Rasch model (see Bond & Fox, 2001, pgs. 178-179 for specific details). Linacre and Wright (1994) suggest that because the Rasch model is stochastic, or viewed as a probabilistic model, the choice of thresholds is never certain or absolute. Instead, they suggest that the choice of cutoff scores reflects the type of analysis to be conducted. Linacre (1990) compares this method to the use of cutoff values in traditional statistics, where the decision to reject the null hypothesis is never absolute and only reflects a pre-accepted level of error (e.g. p<.05). For example, in the current study, the MS cutoff values of .6 to 1.4 were used to identify misfitting items, because Bond and Fox note that these thresholds are appropriate for use with Likert-scaled items. Further, Linacre (2002a) considers items misfitting when they exceed a standardized (ZSTD) value of +/- 2.00; and Pomeranz et al. (2008) note, “For a researcher to consider an item ‘misfitting,’ the item must exceed *both* [italics added] the MS and ZSTD criteria” (p. 253).

Variable maps. When data fit the Rasch model sufficiently, variable maps are perhaps the most practical and useful diagnostic tool available to researchers, because of the amount of measurement information provided. From these variable maps (see Figure 1 for an example), researchers are able to map the entire sample of respondents alongside the entire bank of items in a meaningful way. From this tool, researchers can gain information about the distribution of items and persons, whether there are sufficient items measuring at all levels on the construct, and whether there are gaps where no measurement information is being gained. Information on the left side of the map pertains to all respondents in the sample, and information on the right side of the map pertains to all items on a given measure. Bradley and Sampson (2006) liken the variable map to a ruler where,

> In a well-targeted assessment, mean item and person measures should be approximately equivalent, the difficulty measures of the items should span at least the width of the ability measures of the persons taking the assessment and the items should be distributed such that they accurately measure all persons taking the test. When a group of persons fall in a space between item placements on the ruler, it is comparable to measuring the length of an object with a ruler where the units of measure have been rubbed out at the very length of the object; one could report that the length of the object is between certain units, but could not report the *precise* length. (p. 33)

Person and item reliability. Both person and item reliability provide indications about the replicability of results given how participants in a sample might score if they were to be given a similar bank of items measuring the same construct, or the replicability of results given how a different sample of similar-ability individuals would respond using the same bank of items. Like their use in traditional statistics, reliability estimates in Rasch analysis range from zero to one and reflect the inter-item correlation among all of the responses for a particular item, and thus are used as a measure of internal consistency for scores on that item. Person reliability is equivalent to traditional test reliability (i.e., Cronbach's alpha), whereas item reliability has no statistical equivalent (Linacre, 2005). When below the acceptable level of .70, person reliability estimates can be improved simply by using larger sample sizes with more than 30 observations and ensuring that there are sufficient numbers of items targeting the full range of abilities in a given sample (Linacre, 1994). Item reliability estimates are improved when there is a full range of abilities or attitudes being measured, as well as including sufficient items to measure all levels along the continuum.

Purpose of the Present Study

By examining the psychometric and measurement properties of the Career Thoughts Inventory (CTI) from a Rasch measurement perspective, the authors in the present study hope to accomplish two objectives: first, to illustrate how the Rasch measurement model can assist career assessment researchers in their pursuit of increased precision and accuracy in personality and psychological measurement; and secondly, to provide additional empirical support to the measurement and psychometric value of an established career assessment inventory. Through use of *fit statistics*, *variable maps*, and *person* and *item* reliability estimates, the authors of the present study hope to provide empirical support, from a measurement perspective, for unidimensionality, hierarchical ordering and range, and overall item and person reliability for the decision-making confusion, commitment anxiety, and external conflict sub-domain items of the CTI.

Method

Instrument

The Career Thoughts Inventory (CTI; Sampson et al., 1998) is a 48-item self-report assessment that measures negative thoughts that impede career decision making using a fourpoint Likert-type scale. The instrument includes items such as “I’ll never find a field of study or occupation I really like”, “My interests are always changing”, and “I need to choose a field of study or occupation that will please the important people in my life.” Higher scores on the respective scales indicate more dysfunctional career thinking.

The CTI yields three subscale scores: Decision Making Confusion (DMC; 14 items), Commitment Anxiety (CA; 10 items), and External Conflict (EC; 5 items). The DMC scale reflects an “inability to initiate or sustain the decision making process as the result of disabling emotions and/or a lack of understanding about the decision making process itself” (Sampson, Peterson, Lenz, Reardon, & Saunders, 1996, p. 28). The CA scale reflects an “inability to make a commitment to a specific career choice, accompanied by generalized anxiety about the outcome of the decision-making process,” with the anxiety perpetuating the indecision (Sampson et al., 1996, p. 28). The EC scale reflects an “inability to balance the importance of one’s own self-perceptions with the importance of input from significant others, resulting in a reluctance to assume responsibility for decision making” (Sampson et al., 1996, p. 29).

The Career Thoughts Inventory was designed to be used with multiple populations, including high school students, college students, and adults (Sampson, Peterson, Lenz, Reardon, & Saunders, 1996). The CTI has been used as an instrument to identify individuals within these groups who have a high probability of experiencing problems related to career decision making (Sampson et al., 1996). Moreover, the CTI has been used with adults, college students, and high school students to identity specific types of the career dysfunction thinking, which would impede effective decision making (Sampson et al., 1996). Finally, service providers have used the CTI to develop counseling interventions for individuals with specific dysfunctional career thoughts (Sampson et al., 1996).

Participants

The sample in the present study consisted of 232 college students (51.3% female and 48.7% male, age range 18-39) enrolled in 10 sections of an introductory career development course during successive academic terms at a large southeastern university. The common reason for enrolling in this elective course was to receive assistance in making educational and career decisions. Ethnicity and classification demographics of the sample include African American 15.5%, Asian American 2.2%, Caucasian 64.2%, Hispanic American 12.1%, other 4.7%, prefer not to respond 1.3%; freshmen 9.1%, sophomores 42.2%, juniors 22.8%, seniors 25%, graduate students 0.4%, and other 0.4%.

Procedures

During a regularly scheduled class period at the beginning of the semester, participants were read consent information about the study by research assistants. Participants were then given an informed consent document, demographics questionnaire, and the Career Thoughts Inventory. Surveys were completed and returned during the class period.

Analysis Procedure

Rasch analyses were conducted separately on the 14 items contained on the Decision Making Confusion (DMC) subscale, the 10 items contained on the Commitment Anxiety (CA) subscale, and the five items on contained on the External Conflict (EC) subscale of the Career Thoughts Inventory (CTI). Specifically, through the use of fit statistics, variable maps, and person and item reliability estimates, the authors of the present study sought to explore how items on the three above-mentioned subscales conformed to the assumptions of the Rasch model in regards to *unidimensionality*, adequate *item hierarchy* (variability, distribution and range), and overall item and person reliabilities. All Rasch analyses were conducted using WINSTEPS software (version 3.65; Linacre, 2008).

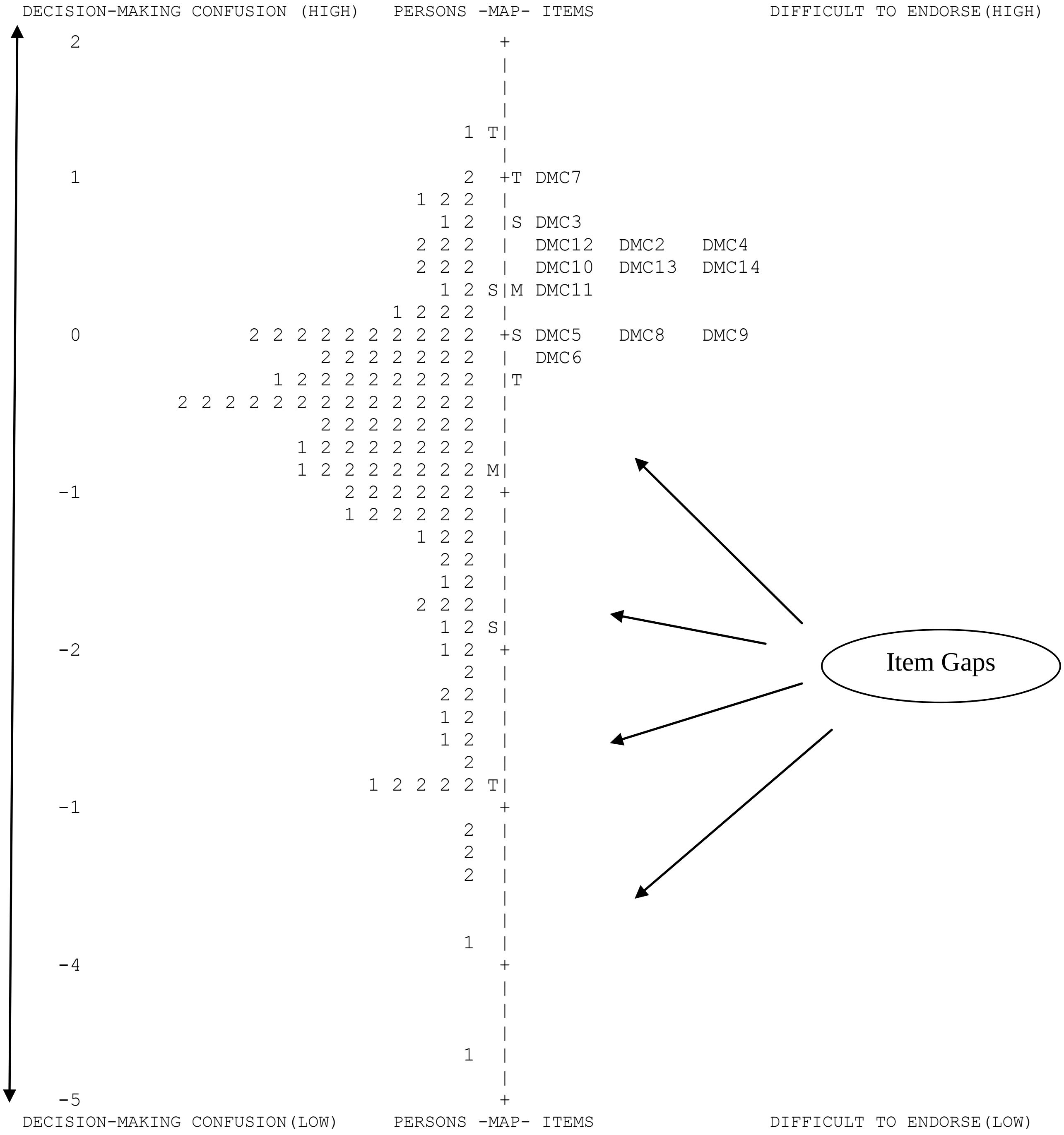
Results and Discussion

In summary, according to the Rasch model a well-constructed assessment will meet the following criteria: (a) contain only items that measure the unidimensional construct of interest; (b) possess a logical item hierarchy in terms of variability, distribution and range; and c) possess high person and item reliability estimates (.70<). These criteria were used to evaluate the measurement properties of the *decision-making confusion*, *commitment anxiety*, and *external conflict* subscales of the Career Thoughts Inventory (CTI). Findings from the Rasch analysis are organized according to each of the subscale findings.

Decision-Making Confusion (DMC) Subscale

Diagnosing misfit. Winsteps software generates infit and outfit statistics, and both are reported in mean-square (unstandardized) and standardized form. Items on the DMC subscale were analyzed according to the following criteria: reviewing outfit statistics, then infit statistics; with mean-square (MNSQ) and standardized (ZSTD) fit statistics being reviewed in combination; and reviewing high values before low values (Linacre, 2008). Items were flagged as misfitting and thus problematic if they were outside of the acceptable range for rating-scale, Likert-type items of .6 to 1.4 (Bond & Fox, 2001) *and* ZSTD values of +/- 2.00. Linacre (2008) suggests always dealing with larger misfit statistics first, because they are a greater threat to overall measurement integrity. Lower misfit statistics are only problematic when the goal is to shorten a test, because these statistics indicate that the item may be overly predictable and will mislead the researcher to think items are measuring better than they are. When removing items to shorten a test, the researchers must be cognizant not to remove the wrong item, and in those cases lower fit statistics are useful. Given that the goal was not to shorten the CTI, lower misfit statistics were not addressed in the present study.

Table 1 reports both infit and outfit mean-square (MNSQ) and standardized (ZSTD) values for all 14 items contained on the DMC subscale (misfitting items in bold). It is evident from these results that only one item, DMC1, is exceeding the pre-established threshold to indicate a misfit (i.e., MS <.6 or 1.4 <MS *and* 2.00 < ZSTD) and would therefore be flagged as problematic. The additional item of DMC12 (outfit MNSQ = .56/ZSTD = -5.2; infit MS = .59/ZSTD = -5.2) falls below the standard, however, this item would only be addressed if the goal were to shorten the test (Linacre, 2002a). As a reminder, when infit or outfit statistics are high, this typically indicates confusion on the part of respondents. Potential causes of confusion would include poor wording, misleading directions, or items that have multiple parts. In other cases, high outfit statistics may indicate that the item is not functioning as a measure of the construct being assessed, i.e., may not function as a measure of decision-making confusion for the present sample. The individual misfitting item in this case is worded: *No field of study or occupation interests me* (DMC1). Normally, the researcher would address this problematic item before moving onto other parts of the analysis, however, the researchers ignored this in the current study. Therefore, this item was simply removed from any subsequent analyses (i.e., not included in variable maps or reliability analyses).



```
DECISION-MAKING CONFUSION (HIGH)   PERSONS -MAP- ITEMS           DIFFICULT TO ENDORSE(HIGH)
    2                                   +
                                        |
                                        |
                                        |
                                     1 T|
                                        |
    1                                2  +T DMC7
                                 1 2 2  |
                                   1 2  |S DMC3
                                 2 2 2  |  DMC12  DMC2   DMC4
                                 2 2 2  |  DMC10  DMC13  DMC14
                                   1 2 S|M DMC11
                               1 2 2 2  |
    0              2 2 2 2 2 2 2 2 2 2  +S DMC5   DMC8   DMC9
                         2 2 2 2 2 2 2  |  DMC6
                     1 2 2 2 2 2 2 2 2  |T
             2 2 2 2 2 2 2 2 2 2 2 2 2  |
                         2 2 2 2 2 2 2  |
                       1 2 2 2 2 2 2 2  |
                       1 2 2 2 2 2 2 2 M|
   -1                      2 2 2 2 2 2  +
                           1 2 2 2 2 2  |
                                 1 2 2  |
                                   2 2  |
                                   1 2  |
                                 2 2 2  |
                                   1 2 S|
   -2                              1 2  +
                                     2  |
                                   2 2  |
                                   1 2  |
                                   1 2  |
                                     2  |
                             1 2 2 2 2 T|
   -1                                   +
                                     2  |
                                     2  |
                                     2  |
                                        |
                                        |
                                     1  |
   -4                                   +
                                        |
                                        |
                                        |
                                     1  |
                                        |
   -5                                   +
DECISION-MAKING CONFUSION(LOW)     PERSONS -MAP- ITEMS           DIFFICULT TO ENDORSE(LOW)
```

Figure 1. Variable (person/item) map for decision-making confusion.[4]

[4] *Note: Right side = Person Map ("2" = 2 persons/"1" = 1 person)*

Item hierarchy (variability, distribution, range). The main function of the variable map is to plot the entire bank of items alongside the entire sample of respondents in order to gain a graphic representation of how well the items measure the sample on the construct of interest. From this, information can be gained about item variability, distribution and range. *Variability* refers to how spread out or how closely grouped items are in comparison to the sample. When items are grouped too closely or are sparsely spread out along the continuum, not enough useful information is gained about the respondents on the construct being measured. *Distribution* of items refers to how spaced out the items are along the continuum, i.e., are items assessing low and moderate levels of decision making confusion just as well as a high level of DMC, how well the items map alongside the person continuum through comparing means, and whether there is redundancy or overlap in the items. Ideally, there will be no significant gaps in the items, the mean of the respondents (bolded "M" on left side of map) will match relatively evenly alongside the mean of the item continuum (bolded "M" on the right side of the map), and there will be few or no items lining up alongside each other on the continuum. Lastly, *range* refers to sufficiently targeting at the extremes or the outliers in any sample. Again, when the goal is to measure the entire distribution in a population, a well-designed measure will contain sufficient items to measure even at the extremes on a given distribution (i.e., at the highest and lowest points).

Several conclusions can be drawn from the information presented in Figure 1. Interpretations for this information will depend upon the goal of the researcher. First, it is evident that there is not much variability to the 13 decision-making confusion subscale items; that is, all 13 items are grouped together toward the top of the continuum for this sample. This is problematic only if the goal is to assess the full range of distress associated with decision-making confusion. In that case, additional items should be designed and included if the goal is to measure lower extremes of decision-making confusion. If the goal, however, is to distinguish between distressed/not distressed students (i.e., use of a cutoff score), then the clumping of items in this manner is quite appropriate. Thus, interpretation of these results is up to the discretion of the researcher. Second, there is inadequate distribution of items along the entire measurement continuum, which limits the amount of information being gained about this sample in regard to their "true" level of decision-making confusion. This fact is evidenced by the significant item gaps along the bottom of the distribution (see ellipses), the mean of the items is significantly higher (one standard deviation) above the mean level for the sample, and there is considerable redundancy or overlap in the items. For example, items DMC12, DMC2, and DMC4; items DMC10, DMC13, and DMC14; and items DMC5, DMC8, and DMC9 are all at equivalent levels in terms of measurement. While there may be several reasons to account for this overlap (i.e. redundancy or dependency), the practical implication for this fact may be twofold. First, this would alert the researcher to possible problems with these items. In a clinical setting, it may very well be that the same amount of useful information can be gained from inclusion of only one or two items. Ultimately, decisions to remove items must always be balanced with theoretical considerations, where items may be measuring a slightly different aspect of the construct (Bond & Fox, 2001). Lastly, the range of items contained on the DMC subscale is not adequately measuring the lower extremes for this sample, but again, this is only problematic when the goal is to measure decision-making confusion at all levels of the continuum. If this were the goal, the researcher could design and include additional less extreme items in order to target the lower levels of distress.

Reliability estimates (person and item). Both person (.94) and item (.91) reliability estimates are quite high, thus indicating that these results would be fairly stable over any given number of administrations. Specifically, the high person reliability estimate indicate that if this sample were given a similar grouping of items, the resultant findings should be consistent. Additionally, the high item reliability estimate indicates that these items should be considered highly reliable indicators of decision-making confusion, and would be expected to produce similar results if given to a sample with similar levels of decision-making confusion. Again, it should be pointed out that the restricted range of items should be included in the interpretation of these estimates as the restricted item range would be problematic if the goal were to measure the full continuum of decision-making confusion. In that case, the high reliability estimates could be misleading.

Table 1
Fit Statistics for Decision-Making Confusion Items (DMC)

DMC Item	CTI Item	Outfit MS	Outfit ZSTD	Infit MS	Infit ZSTD
DMC 1	**CTI 1**	**1.83**	**5.3**	**1.52**	**4.5**
DMC 4	CTI 5	1.02	0.2	1.11	1.2
DMC 2	CTI 3	.98	-0.2	1.02	0.2
DMC 5	CTI 11	.94	-0.6	.96	-0.4
DMC 6	CTI 12	.86	-1.6	.90	-1.2
DMC 8	CTI 16	.86	-1.6	.82	-2.2
DMC 7	CTI 13	.85	-1.3	.93	-0.7
DMC 13	CTI 43	.75	-2.6	.82	-2.0
DMC 3	CTI 4	.74	-2.6	.83	-1.9
DMC 9	CTI 20	.71	-3.5	.71	-3.6
DMC 10	CTI 27	.69	-3.5	.74	-3.1
DMC 14	CTI 44	.67	-3.7	.62	-4.6
DMC 11	CTI 28	.61	-4.7	.63	-4.6
DMC 12	CTI 36	.56	-5.2	.59	-5.2
Mean	---------	.86	-1.8	.87	-1.7
S. D.	---------	.30	2.5	.23	2.5

Summary. From a Rasch measurement perspective, and in combination, the results of the Rasch analysis on the 13 decision-making confusion subscale items convey several pieces of useful information. First, a review of the fit statistics reveals that only one item, DMC1 (*No field of study or occupation interests me*), is misfitting according to the guidelines of the Rasch model. The high fit statistics indicate that wording on the item was possibly confusing for some respondents (Bond & Fox, 2001). Several possible factors may be contributing to the high misfit statistics, including the extreme nature of the wording. It is possible that the absoluteness of the terminology makes it too extreme for those who are not experiencing extreme levels of distress. Another possible factor may be that there are a few students who are not at all confused but do not necessarily have a predetermined career path in mind. Lastly, it may be that the item is “double-barreled” and is assessing both *field of study* and *occupation*, which may be causing confusion for some students about which aspect they should respond. In essence, the double barreled item turns out to be misfitting because individuals respond to different parts of the item.

A review of the person/item map for decision-making confusion shows that, while these items are clustered together at the top of the distribution, this finding is only a problem if the goal is to assess the entire continuum of decision-making confusion. If the goal is to distinguish between high/low levels of decision-making confusion, or to set thresholds, this group of items would meet that goal effectively. Here, the clinical or research goal is an important factor in interpreting these results. Despite this fact, it is still difficult, from a measurement perspective, to assess the lower levels of decision-making confusion for this sample of college students, because of the absence of items toward the lower end of the distribution. From a clinical standpoint for the present sample, another interpretation of this variable map would be that those students with high levels of DMC would be experiencing difficulty with their career development and could be diagnosed with the CTI-DMC.

Lastly, and perhaps most encouraging from a measurement perspective, the high person and item reliability estimates indicate a high level of consistency would be maintained in follow-up testing of the sample (using similar items) or a re-administration of the same items to a different sample of college students. Caution should be used in interpreting these results, however, given the previous findings regarding the misfitting items, as well as the person/item map. Combined, this information does give researchers a good indication of how the items are functioning as indicators of decision-making confusion.

Commitment Anxiety Subscale

Diagnosing misfit. Table 2 reports both infit and outfit MNSQ and ZSTD values for the 10 items contained on the commitment anxiety subscale. A review of the CA items shows that no items are misfitting according to the pre-set criteria, thus indicating that all items are measuring the unidimensional construct of commitment anxiety. Although several of the ZSTD values are outside of the +/-2.0 range, these are not enough to create undue concern. From the Winsteps help section, Linacre (2002a) states, “Ben Wright advises ‘ZSTD is only useful to salvage non-significant MNSQ [mean-square] > 1.5, when sample size is small or test length is short.” Neither of these conditions exists in the present case.

```
COMMITMENT ANXIETY(HIGH)     PERSONS -MAP- ITEMS          DIFFICULT TO ENDORSE(HIGH)
                                  +
                               1  |
                                 T|
                                  |
    1                          2  +
                             1 2  |
                             1 2  |
                             2 2  |
                           1 2 2  |
                           1 2 2  |
                             1 2 S|  CA 5
                     2 2 2 2 2 2  |T
    0            1 2 2 2 2 2 2 2  +
                       2 2 2 2 2  |
               2 2 2 2 2 2 2 2 2  |  CA 10
               2 2 2 2 2 2 2 2 2  |S
                 1 2 2 2 2 2 2 2  |
                   2 2 2 2 2 2 2  |  CA 6
                     1 2 2 2 2 2 M|M CA 4
                 1 2 2 2 2 2 2 2  |  CA 1 CA 8 CA 9
   -1                2 2 2 2 2 2  +
                       1 2 2 2 2  |S CA 7
                           1 2 2  |  CA 2 CA 3
                             2 2  |
                             1 2  |
                             1 2  |T
                           1 2 2  |
                             2 2 S|
   -2                          2  +
                               2  |
                               1  |
                             2 2  |
                           1 2 2  |
                                  |
                           1 2 2  |
                           2 2 2 T|
   -3                             +
                               1  |
                             1 2  |
                                  |
                               2  |
                                  |
                                  |
                               1  |
   -4                             +
                                  |
                                  |
                                  |
                               1  |
                                  +
COMMITMENT ANXIETY(LOW)      PERSONS -MAP- ITEMS                DIFFICULT TO ENDORSE(LOW)
```

Figure 2. Variable (person/item) map for commitment anxiety.[5]

[5] Note: Right side = Person Map (“2” = 2 persons/”1” = 1 person)

Item hierarchy (variability, distribution, range). A review of Figure 2 shows that, similar to the decision-making subscale items, the commitment anxiety subscale items demonstrate similarities in terms of variability, distribution, and subsequent range. Again, interpretation will depend upon the overall goal of the researcher in their intended use of the information. In this case, the variability of the items shows that the items are grouped toward the middle of the distribution, although they are spread a bit further apart when compared with the DMC items. This presents problems when the goal is to measure the full continuum of commitment anxiety.

Table 2
Fit Statistics for Commitment Anxiety Items (CA)

CA Item	CTI Item	Outfit MS	Outfit ZSTD	Infit MS	Infit ZSTD
CA 7	CTI 32	1.37	4.1	1.26	3.0
CA 3	CTI 22	1.25	2.8	1.21	2.4
CA 10	CTI 47	1.13	1.4	1.13	1.5
CA 1	CTI 17	1.12	1.4	1.07	0.9
CA 2	CTI 21	1.09	1.1	1.06	0.8
CA 6	CTI 30	1.07	0.8	1.05	0.6
CA 9	CTI 38	1.05	0.6	1.01	0.2
CA 8	CTI 35	.89	-1.3	.88	-1.4
CA 4	CTI 26	.72	-3.7	.71	-3.9
CA 5	CTI 29	.65	-4.1	.68	-3.9
Mean	---------	1.03	0.3	1.01	0.0
S. D.	---------	.21	2.5	.18	2.3

Distribution of the CA items is improved over the DMC item distribution, however, there are still significant gaps at both the high and low ends of the continuum, thus leaving these individuals inadequately measured on their CA. Further, there is some redundancy in the items (e.g., CA1, CA8, CA), but not as much when compared to the DMC subscale. The range of items is also incomplete in that there are insufficient items at both ends of the continuum to measure or capture the true nature of commitment anxiety in this sample.

Reliability estimates (person and item). The person (.94) and item (.95) reliability estimates for the commitment anxiety subscale were also quite high. These estimates indicate high replicability of results across both persons and items.

Summary. Taken together, all of the information pertaining to the commitment anxiety subscale are quite encouraging from a Rasch measurement perspective. According to the fit statistics, all items are functioning as unidimensional indicators of commitment anxiety. The only area where possible improvement is possible is in the inclusion of additional items that may help measure a broader range of commitment anxiety in college students. The limited variability, narrow distribution, and restricted range simply limits the amount of “measurement” information that can be gained regarding commitment anxiety if the goal is to measure the larger CA continuum. Both high person and item reliability estimates are encouraging and lend empirical support for use of these items in future.

External Conflict Subscale

Before reviewing the measurement information pertaining to the 5 external conflict subscale items, it is important to mention that there are limitations to having such a small number of items assessing any given construct. Wright (1992) states,

> A useful test gives examinees repeated opportunity to demonstrate proficiency. An examinee may guess, make a careless error, or have unusual knowledge. One, two or even three items provide too little evidence. We need enough replications along our one dimension to resolve any doubts about examinee performances. As doubts are resolved, the relevance of each response to our understanding of each examinee's performance becomes clear. We can focus attention on the responses that contribute to examinee measurement, reserving irrelevant responses (guesses, scanning errors, etc) for qualitative investigation (pg. 205).

It should be noted that there are some quantitative researchers who suggest that effective measurement can occur with as few as one item. There is precedent in the literature for using single items in the measurement of health and psychosocial-related outcomes (e.g., DeSalvo, Fisher, Tran, Bloser, Merrill, & Peabody, 2006; Zimmerman, Ruggero, Chelminski, Young, Posternak, Friedman, et al., 2006; see Bergkvist & Rossiter, 2007 for further discussion on the use of single-items for scales of measurement). Unfortunately, there are no hard and fast rules offered about how many items are needed to allow respondents to be measured accurately from a Rasch perspective. In cases such as this, however, Wright does suggest that when there are small numbers of items, reliability estimates are particularly important and should be above .90. It should be noted that on the EC subscale, the person (.94) and item (.88) reliabilities were close or above this threshold.

Diagnosing misfit. A review of Table 3 shows that 3 of the 5 items on the external conflict scale are within the acceptable limits, thus indicating that EC2, EC3, and EC4 are measuring the same unidimensional construct of external conflict. The high outfit and infit statistics for EC1 (outfit MNSQ 1.47/ZSTD 4.6; infit MNSQ 1.40/ZSTD 4.0) and EC5 (outfit MNSQ 1.44/ZSTD 4.0; infit MNSQ 1.28/ZSTD 2.8) are not too far out of the acceptable thresholds, but may indicate some disturbance in the expected response patterns. Normally these items would be flagged as problematic, if only to examine them for possible improvement. Due to the previous noted limitations in word length, however, these items were simply removed and not included in further analysis.

Table 3
Fit Statistics for External Conflict Items (EC)

EC Item	CTI Item	Outfit MS	Outfit ZSTD	Infit MS	Infit ZSTD
EC1	**CTI6**	**1.47**	**4.6**	**1.40**	**4.0**
EC5	**CTI46**	**1.44**	**4.0**	**1.28**	**2.8**
EC2	CTI9	1.24	2.1	1.27	2.7
EC4	CTI23	1.18	1.8	1.09	1.0
EC3	CTI14	.88	-1.3	.93	-0.7
Mean	---------	1.24	2.2	1.19	1.9
S. D.	---------	.21	2.1	.16	1.6

Item hierarchy (variability, distribution, range). The item hierarchy for the three external conflict items is much the same as for the two other subscale items. First, the variability of the items is restricted and clustered together toward the top of the person distribution. Second, the distribution of items is limited in that there are significant gaps where many individuals are not being adequately measured. Lastly, the range of items is also limited in that even though there are no significant gaps between the items, they are not measuring the level of external conflict for about two-thirds of the sample distribution. Figure 3 contains all item hierarchy information pertaining to the three external conflict items.

```
EXTERNAL CONFLICT(HIGH)     PERSONS -MAP- ITEMS         DIFFICULT TO ENDORSE(HIGH)
                                                  +
                                             |
                                          1 T|
                                             |
    1                                     2  +T
                                      1 2 2  |  EC 2
                                        1 2  |S
                                      2 2 2  |
                                      2 2 2  |M EC 4
                                        1 2 S|
                                    1 2 2 2  |S EC 3
    0                   2 2 2 2 2 2 2 2 2 2  +
                              2 2 2 2 2 2 2  |T
                          1 2 2 2 2 2 2 2 2  |
                  2 2 2 2 2 2 2 2 2 2 2 2 2  |
                              2 2 2 2 2 2 2  |
                            1 2 2 2 2 2 2 2  |
                            1 2 2 2 2 2 2 2 M|
   -1                           2 2 2 2 2 2  +
                                1 2 2 2 2 2  |
                                      1 2 2  |
                                        2 2  |
                                        1 2  |
                                      2 2 2  |
                                        1 2 S|
   -2                                   1 2  +
                                          2  |
                                        2 2  |
                                        1 2  |
                                        1 2  |
                                          2  |
                                  1 2 2 2 2 T|
   -3                                        +
                                          2  |
                                          2  |
                                          2  |
                                             |
                                             |
                                          1  |
   -4                                        +
                                             |
                                             |
                                             |
                                          1  |
                                             |
                                             |
   -5                                        +
EXTERNAL CONFLICT(LOW)          PERSONS -MAP- ITEMS           DIFFICULT TO ENDORSE(LOW)
```

Figure 3. Variable (person/item) map for external conflict.[6]

[6] Note: Right side = Person Map (“2” = 2 persons/”1” = 1 person)

Reliability estimates (person and item). The person (.94) and item (.88) reliability estimates for the external conflict subscale were acceptable. These estimates indicate high replicability of results across both persons and items.

Summary. Taken together, all of the information pertaining to the external conflict subscale are quite encouraging from a Rasch measurement perspective, except for a few minor concerns. According to the fit statistics, three of the five EC items are functioning as unidimensional indicators of external conflict. The other two items would need further evaluation. The person/item map does indicate that the five EC items are constricted in both range and distribution and they are clustered near the top of the distribution. Both high person and item reliability estimates are encouraging, which is particularly important when fewer items are analyzed.

Conclusion

Two objectives were established by the authors of the present study: 1) to illustrate how the Rasch measurement model can assist career assessment researchers in their pursuit of increased precision and accuracy in personality and psychological measurement; and 2) to provide additional empirical support to the measurement and psychometric value of an established career assessment inventory, the Career Thoughts Inventory (CTI; Sampson et al., 1998). To meet the first objective, limitations of the Classical Test Theory approach to understanding psychometric properties was discussed while advantages of the Rasch method were highlighted. One advantage discussed was the ability to see a fuller picture of assessment from the individual perspective (e.g., a tired test taker) and item perspective (e.g., a confusing item), which allows a practitioner the confidence that the assessment is measuring the latent construct of interest (e.g., dysfunctional career thinking). Also mentioned were the ability of Rasch to inform a test developer about the item and test taker reliability and validity as well as developing a shorter version of an existing measure, when appropriate. The authors also explained how to understand and interpret the key outcomes of Rasch, explaining the thresholds by which an item should be considered inappropriate to be included in the final measure (e.g., infit statistics; variable maps) or when an item can be eliminated because it accounting for the same variance in the latent construct as another item (i.e., outfit statistics).

To meet the second objective, the authors utilized the CTI results of 232 college students to better understand the psychometric properties of the CTI which originally used Classical Test Theory in its original, rigorous test development process. When analyzing the three CTI subscales (i.e., Decision-Making Confusion, Commitment Anxiety, External Conflict) separately, the authors employed the criteria outlined by the Rasch model for a well constructed assessment. Again, these criteria are: a) contain only items that measure the unidimensional construct of interest; b) possess a logical item hierarchy in terms of variability, distribution and range; c) and possess high person and item reliability estimates. For the Decision-Making Confusion (DMC) subscale only one item was found to misfit with the scale, indicating possible confusing wording that could lead to the item assessing something outside of the unidimensional construct of decision-making confusion. With regards to item hierarchy, the DMC items were found to cluster at the top of the continuum for this sample indicating the scale is measuring a more extreme form of decision-making confusion rather than the full continuum of the construct. Implications for the distribution of DMC items were discussed in the Results and Discussion section. The item and person reliability estimates for the DMC scale were high, indicating a good stability for this scale. There were no misfitting Commitment Anxiety (CA) items and the item and person reliability estimates were similarly high. Again, the hierarchy of items were clustered, but this time in the middle of the distribution indicating most are measuring a moderate level of commitment anxiety. For the 5-item External Conflict (EC) scale, three of the items were measuring the unidimensional construct of external conflict but two items were problematic. The variability of these three items was restricted to a cluster around the top of the hierarchy and item and person reliability were acceptable.

With regards to the findings specific to the CTI, they must be interpreted with the purpose of the measure in mind as discussed in the Results and Discussion section. Ultimately, it appears the CTI is sound psychometrically but Rasch highlighted some ways in which the instrument and individual items could be improved or modified to assess a fuller range of dysfunctional career thinking. Additionally, the authors demonstrated how vocational psychologists can utilize the Rasch to enhance the development, revision, or new version creation of a variety of career-related assessments.

References

Allen, M. J., & Yen, W. M. (1979). *Introduction to measurement theory*. Monterrey, CA: Brooks/Cole Publishing Co.

Bergkvist, L., & Rossiter, J. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research*, *44*(2), 175-184.

Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Bradley, K. D., & Sampson, S. O. (Summer, 2005). A case for using a Rasch model to assess the quality of measurement in survey research. *The Respondent*, 12-13.

Bradley, K. D., & Sampson, S. O. (2006). Constructing a quality assessment through Rasch techniques: The process of measurement, feedback, reflection and change. In X. Liu & W. Boone (Eds.), *Applications of Rasch measurement in science education* (pp. 23-44). Maple Grove, MN: JAM Press.

Cole, J. C., Rabin, A. S., Smith, T. L., & Kaufman A. S. (2004). Development and validation of a Rasch-derived CES-D short form. *Psychological Assessment, 16*(4), 360-372.

Costa, P. T., Jr., & McCrae, R. R. (1992). *The NEO Personality Inventory (NEO PI-R) and NEO Five-Factor Inventory(NEO-FFI) professional manual.* Odessa, FL: Psychological Assessment Resources.

DeSalvo, K., Fisher, W., Tran, K., Bloser, N., Merrill, W., & Peabody, J. (2006). Assessing measurement properties of two single-item general health measures. *Quality of Life Research: An International Journal of Quality of Life Aspects of Treatment, Care & Rehabilitation*, *15*(2), 191-201.

DuBois, P. (1970). Varieties of psychological test homogeneity. *American Psychologist*, *25*(6), 532-536.

Fisher, W. P. (2008). Other historical and philosophical perspectives on invariance in measurement. *Measurement: Interdisciplinary Research and Perspectives, 6,* 190-194.

Fox, C. (1999). An introduction to the partial credit model for developing nursing assessments. *Journal of Nursing Education, 38*(8), 340-6.

Fox, C. M., & Jones, J. A. (1998). Uses of Rasch modeling in counseling psychology research. *Journal of Counseling Psychology, 45*(1), 30-45.

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

Holland, J. L., Fritzsche, B. A., & Powell, A. B. (1994). *The Self-Directed Search technical manual.* Odessa, FL: Psychological Assessment Resources.

Linacre, J. M. (1990). Where does misfit begin? *Rasch Measurement Transactions*, *3*, 80.

Linacre, J. M. (1994). Sample size and item calibration stability. *Rasch Measurement* Transactions, *7*, 328.

Linacre, J. M. (2002a). What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions, 16,* 878.

Linacre, J. M. (2002b). Understanding Rasch measurement: Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*(1), 85-106.

Linacre, J. M. (2008). WINSTEPS Rasch measurement computer program. Chicago: Winsteps.com.

Linacre J.M., & Wright B.D. (1994) Chi-square fit statistics. *Rasch Measurement Transactions, 8,* 350.

Pomeranz, J. L., Byers, K. L., Moorhouse, M. D., Velozo, C. A., & Spitznagel R. J. (2008). Rasch analysis as a technique to examine the psychometric properties of a Career Ability Placement Survey subtest. *Rehabilitation Counseling Bulletin, 51*(4), 251-259.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests.* Chicago: The University of Chicago Press (original work published in 1960).

Sampson, J. P., Peterson, G., Lenz, J., Reardon, R., & Saunders, D. (1996). *Career Thoughts Inventory: Professional manual.* Odessa, FL: PAR, Inc.

Sampson, S. and Bradley, K. D. (2003). Rasch analysis of educator supply and demand rating scale data: An alternative to the true score model. *Research Methods; The Forum*. http://aom.pace.edu/rmd/2003forum.html

Smith, E. V. (2004a). Evidence for the reliability of measures and validity of measure interpretation: A Rasch measurement perspective. In E. V. Smith & R. M. Smith (Eds.). *Introduction to Rasch measurement* (pp. 93-122). Maple Grove, MN: JAM Press.

Smith, E. V. (2004b). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. In E. V. Smith & R. M. Smith (Eds.). *Introduction to Rasch measurement* (pp. 575-600). Maple Grove, MN: JAM Press.

Stevens, S. S. (1946). On the theory of scales in measurement. *Science, 103,* 677-680.

Tang, W. K., Wong, E., Chiu, H. F. K., Lum, C. M., & Ungvari, G. S. (2005). The geriatric depression scale should be shortened: Results of Rasch analysis. *International Journal of Geriatric Psychiatry, 20*, 783-789.

Wright, B. D. (1992). What is the "right" test length? *Rasch Measurement Transactions*, 6, 205.

Zimmerman, M., Ruggero, C., Chelminski, I., Young, D., Posternak, M., Friedman, M., et al. (2006). Developing brief scales for use in clinical practice: The reliability and validity of single-item self-report measures of depression symptom severity, psychosocial impairment due to depression, and quality of life. *Journal of Clinical Psychiatry*, *67*(10), 1536-1541.