

Title: Analysis of the Construct Validity and Measurement Invariance of the Career Decision Self-Efficacy Scale: A Rasch Model Approach

Order of Authors: Guido Makransky, Mary E. Rogers, Peter A. Creed

Corresponding Author: Guido Makransky

Department of Psychology, University of Southern Denmark

Campusvej 55, DK-5230 Odense M, Denmark

gmakransky@health.sdu.dk

Additional Authors:

Mary E. Rogers

Griffith Business School, Griffith University

Building (G01), Room 3.23, Gold Coast campus, Griffith University

Parklands Drive, Southport, QLD 4222, Australia

Peter A. Creed

School of Applied Psychology and Griffith Health Institute, Griffith University

Gold Coast campus, Griffith University

Parklands Drive, Southport, QLD 4222, Australia

Analysis of the Construct Validity and Measurement Invariance of the Career Decision Self-Efficacy Scale: A Rasch Model Approach

Abstract

The Career Decision-Making Self-Efficacy Scale – Short Form (CDMSE-SF) is one of the most frequently used instruments to assess individual levels of career-related self-efficacy. The present study used the partial credit model within the framework of item response theory to examine the content, structural, substantive, and generalizability aspects of validity for the CDMSE-SF in a sample of 534 Australian high school students aged between 14 and 19 years. The results showed clear evidence of multidimensionality for the CDMSE-SF. Furthermore, there was strong support for the content, structural, and substantive aspects of validity when using the five subscales individually. Evidence of measurement invariance was found across grade levels; however, there were individual items that exhibited differential item functioning across gender, achievement level, and age groups. The implications for career counseling and research are discussed.

Keywords: CDMSE-SF, career decision self-efficacy, Rasch model, validation, invariance, differential item functioning

Introduction

The 25-item Career Decision Self-Efficacy Scale – Short Form¹ (CDSE-SF; Betz & Taylor, 2012) is one of the most frequently used instruments to assess individual levels of career-related self-efficacy (Chaney, Hammond, Betz, & Multon, 2007). The CDSE-SF is a shortened version of the 50-item Career Decision Self-Efficacy Scale (Betz & Taylor, 2012; Taylor & Betz, 1983), which was devised to assess self-efficacy regarding the five career choice competencies. Taylor and Betz (1983) based both their theoretical and methodological scale development on Crites (1978) five career competencies from his Career Maturity Inventory. The competencies relate to developing an accurate self-appraisal, gathering occupational information, goal selection, making plans for the future, and problem solving. Self-efficacy generally, and career decision self-efficacy specifically, reflect self-appraisals of competency, which is an important construct, as how individuals behave can often be better predicted by their beliefs about their capabilities rather than by their actual capabilities (Bandura, 1986; Lent, Brown, & Hackett, 1994). Both the CDSE and CDSE-SF tap individuals' beliefs about their capacity to undertake the necessary tasks associated with making career decisions (Betz, Hammond, & Multon, 2005; Taylor & Betz, 1983).

The career decision self-efficacy construct was first introduced to the career area by Hackett and Betz (1981), who demonstrated that college students' beliefs about their educational and career capabilities were related significantly to the variety of career options they considered. Since then, the CDSE and CDSE-SF have been the main instruments used to assess this context-specific self-efficacy. However, due to concerns about the length of the CDSE, and the unwieldy response continuum (10-point Likert-type scale), researchers and practitioners in the career field are

¹ The CDSE-SF was previously called the Career Decision-Making Self-Efficacy Scale – Short Form (CDMSE-SF; Betz, Hammond, & Multon, 2005).

now increasingly using the short form of the scale, which includes 25 items and a 5-point response continuum (Betz et al., 2005; Chaney et al., 2007).

While there is consensus that the career decision self-efficacy construct is useful, there is still much debate over how the construct should be measured, and whether the CDSE-SF is the appropriate instrument to do this (Creed, Patton, & Watson, 2002). Previous research on the CDSE-SF indicates that the data obtained from the instrument are reliable across different language versions and within different populations (Betz, Klein, & Taylor, 1996; Betz & Klein, 1996; Chung, 2002; Creed et al., 2002; Nilsson et al., 2002; Osipow & Gatti, 1998). There is also evidence for criterion-related validity (e.g., Betz et al., 1996; Chung, 2002). However, very different results have been found when assessing the construct validity of the CDSE-SF across different samples. Specifically, different factor structures have been found across different studies and in different populations. Furthermore, little evidence is available about the measurement invariance (MI) of the individual items across different groups. MI is a statistical property of measurement that indicates that the same construct is being measured across several specified groups (Makransky & Glas, 2013). For example, MI can be used to study whether a given measure is interpreted in a conceptually similar manner by respondents representing different genders or age groups. In the present study, we use a detailed framework to investigate different aspects of construct validity for the CDSE-SF, including content, structural, substantive, and generalizability, based on Messick's (1995) framework, using item response theory (IRT; Embertson & Riese, 2000).

The dimensionality of the CDSE and CDSE-SF has previously been investigated in a number of different studies using classical test theory approaches. These studies have resulted in very different conclusions. For instance, two factors were found in a US sample (Peterson & del Mas, 1998). Three factors were found with Australian and South African samples (Creed et al., 2002), as well as with a Chinese sample (Hampton, 2005). Chaney and his colleagues (Chaney,

Hammond, Betz, & Multon, 2007) found four factors, and Betz and her colleagues found five factors (Betz et al., 1996). Furthermore, others (Robbins, 1985; Taylor & Betz, 1983; Taylor & Popma, 1990) reported that the CDSE and CDSE-SF measured a general factor (one-factor model).

Recently, confirmatory factor analysis (CFA) methods have been used to examine numerous alternate models with the objective of identifying the reasons for the differences in construct structure of the CDSE and CDSE-SF. Miller, Roy, Brown, Thomas, and McDaniel (2009) found empirical support for a number of different structural models, but concluded by recommending the use of the original five-factor model, because that model was specified a priori based on theory (i.e., Crites, 1961; Taylor & Betz 1983), as opposed to being identified in an empirical manner. A similar study using data from Italian high school students showed a better fit with the five-factor model than a one-factor model and demonstrated factorial invariance across boys and girls. Other studies using CFA with French and Turkish samples found that a four-factor model was necessary because the five-factor model did not fit (Buyukgoze-Kavas, 2014; Gaudron, 2011). It can be concluded from the existing research that there is still considerable disagreement about the factorial structure of the CDSE and CDSE-SF.

Despite this lack of consensus, the CDSE-SF is still widely used in career assessment research. The scale is typically used according to the original, theoretically based, five-factor model, by either adding up the individual item scores to obtain a total score for each of the five subscales, or as a general measure of career decision self-efficacy by summing the subscales to give a total, general factor score for the entire scale, which provides a general measure of career decision self-efficacy. However, instead of relying on investigations of the factor structure of the measure, testing whether or not the total scores obtained for the five subscales, individually and in combination, are accurate measures of individuals' standing on the latent traits is likely to shed light on how the test might be best applied. In this process, comparisons can also be made across

demographic groups such as gender, grade, age, and ability levels, which are important, as little is known about how the items in the CDSE-SF function across groups. This can be achieved by testing the MI of the CDSE-SF across demographic groups.

Several researchers have suggested that methods such as item response theory (IRT) be used to investigate the validity of scales (Hong, Kim, & Wolfe, 2005; Kim & Hong, 2004; Nam et al., 2010). These methods provide a more thorough assessment of the measurement properties of a scale and test for specific properties such as MI and unidimensionality (Emberson & Riese, 2000). In this study, we use the partial credit Rasch model (PCM; Masters, 1982), within the framework of IRT, to investigate the validity of the CDSE-SF. The PCM describes the association between a person's level of an underlying trait and the probability of a specific item response on a measure. This association places the individual's level of the underlying trait and the item difficulty on the same metric. Observed data are tested against the assumptions of the model, and, if met, the raw score of a scale is statistically sufficient and can be said to reflect the level of the underlying trait on an interval scale of measurement (Tennant & Connaghan, 2007). In other words, when able to fit a Rasch model, the person's total score contains all information available within the specified context about the individual.

The application of the PCM is attractive because it can be used to assess the assumptions of MI at the item level. Item invariance requires that item estimation be independent of the subgroups of individuals completing the measure. In other words, item parameters have to be invariant across populations (Bond & Fox, 2001). Items that do not demonstrate invariance are commonly referred to as exhibiting differential item functioning (DIF). The assessment of MI is important in investigating the validity of the CDSE-SF because conclusions about the differences between different groups, including gender, age, and grade are often reported in the literature (e.g., Arnold & Bye, 1989; Chung, 2002; Creed, Patton, & Watson, 2002; Luzzo, 1995;). Without

evidence of MI, it is impossible to determine whether these differences are due to properties of the CDSE-SF items or if these are actual differences on the latent traits. Although MI has been investigated previously by testing the factor structure across different groups for the CDSE (e.g., Miller et al., 2009; Presti et al., 2012), there is little evidence for the MI of the individual items across groups.

Nam et al. (2010) is the only study we could find that used the Rasch model to investigate the validity of the CDSE-SF. These researchers tested its validity in a sample of South Korean students when used as a global scale (one-factor model). Overall, they found that the 5-point Likert-type format response categories were suitable for the scale in that culture. In addition, they reported that most items, with the exception of three (items 13, 17, and 25), formed a unidimensional trait based on the results from item-fit statistics and principal components analysis with standardized residuals, but conclude that more research is needed.

In the current study, we investigate three research questions. First, we explore if it is valid to use the CDSE-SF as a global scale of career decision self-efficacy. Next, we examine if it is valid to use the theoretical derived five-factor model suggested by Taylor and Betz (1983). These two research questions are examined by using the PCM (Masters, 1982) to investigate: (a) if each scale or subscale is unidimensional, (b) if the individual items fit the intended scale or subscale according to the expectations of the model, (c) if redundancy exists in the items in the form of local dependence, and (d) whether the 5-point Likert-type format response categories are suitable. Finally, we investigate if there is MI across the demographic variables of grade, age, gender, and achievement level. Therefore, the purpose of the current study is to investigate the validity the CDSE-SF from an IRT perspective using the PCM. The results are expected to provide valuable information concerning the dimensionality of the CDSE-SF, as well as evidence regarding how the

instrument functions across different demographic groups. This information will help career counselors and researchers determine how to optimally use, and further develop, the scale.

Method

Participants

Participants were 534 high school students recruited from one co-educational, metropolitan Australian, public high school. There were 263 girls (49.3%) and 271 boys with a mean age of 16.0 years ($SD = .97$), who were attending grades 10 (33.7%), 11 (35.6%), and 12 (30.7%). Response rate was 70%. Two hundred and thirty-five (44%) of the students indicated they had current part-time jobs while attending school. We calculated a proxy for socio-economic status (SES) by using parent education and occupation. Father's level of education was used first, and if that was not provided, mother's level was used. If parent education level was unavailable, we used father's or mother's occupation level (Anderson & Vervoorn, 1983). Based on these calculations, 33.7% of students were rated as low SES, 35.7% as medium, and 26% as high (we were not able to calculate an SES level for 4.6% of students).

Instrument

The 25-item CDSE-SF (Betz & Taylor, 2012) is a shortened version of the 50-item CDSE (Betz & Taylor, 2012), which is typically used as a unidimensional measure (e.g., Duffy, Diemer, & Jadidian, 2012; Page, Bruch, & Haase, 2008). It taps the five factor model which was proposed by Taylor and Betz (1983) who based both their theoretical and methodological scale development on Crites (1978) five career competencies from his Career Maturity Inventory. The five factors include skills in developing an accurate self-appraisal, gathering occupational information, goal selection, making plans for the future, and problem solving. Respondents indicate their level of confidence in carrying out specific behaviors associated with the five competency areas using a 5-point Likert-type response format with endpoints of 1 = *no confidence at all* to 5 =

complete confidence. This gives a possible range for the 25 items of 25-125, with higher scores indicating higher levels of confidence. The scores were recoded to endpoints of 0 = *no confidence at all* to 4 = *complete confidence* in the PCM analyses conducted in this study. Sample items are “How much confidence do you have that you could: accurately assess your abilities?” (accurate self-appraisal domain) / “find out the employment trends for an occupation over the next ten years?” (gathering occupational information domain). Following information from the school, we made minor changes to some items to suit the Australian situation. For example, we amended the question “Choose a major or career that will fit your interests?” to “Choose a career that will fit your interests?”, as the term “major” is not in common use in Australia. Betz et al. (1996) supported construct validity of the CDSE-SF by demonstrating negative associations with career indecision.

Procedure

The study was advertised to parents via the school newsletter, and information sheets and consent forms were provided to parents and students regarding the research. The survey was administered in-school by school personnel, with completed questionnaires returned to the researchers in sealed envelopes. The study was conducted under the auspices of the university’s human ethics committee. Students who did not complete the survey were either absent from school, on a class excursion, or did not return a parental consent form to their class teacher and therefore were not allowed to complete the survey without the consent form. The CDSE-SF was administered to the sample at two time points; as a pre- and post-test. The pre-test results are used in this study to examine the dimensionality and the MI of the instrument. The post-test results are used as a cross-validation of these results.

Data analysis. The validity of the CDSE-SF is evaluated on content, structural, substantive, and generalizability aspects of validity as described in Messick (1995) and Wolf and Smith (2007). These aspects of validity, which are described in more detail below, can be tested by

assessing if the data fit the partial credit model (PCM; Masters, 1982). The RUMM 2030 program (Andrich, Sharidan, & Luo, 2010) was used for this purpose in this study.

Content validity. The content validity of an instrument addresses evidence of content relevance, representativeness, and technical quality (Messick, 1995). In this study, content validity is evaluated by assessing the quality of each item based on item fit indices, and by investigating if there is redundancy between items within the same scale or subscale. Item fit is achieved when the individual items measure the latent trait in a consistent way. Item fit can be viewed graphically by comparing the observed values from the data for groups of responders across the trait (class intervals) plotted against the expected model curve (item characteristic curve; ICC). Items with good fit display each of the group plots lying on the curve. Residuals between -2.5 and +2.5 indicate adequate fit to the model (Pallant & Tennant, 2006). Item redundancy can be tested by assessing the local independence of the items (Write, 1996). Local independence refers to the assumption that once the latent trait has been accounted for, the residual correlations between items should be close to zero. Items with residuals over 0.2 are typically labeled as being locally dependent (LD), which is an indication that the item responses are correlated even after the latent trait has been accounted for.

Structural validity. The structural aspect of validity evaluates the fidelity of the scoring format to the structure of the construct domain at issue (Loevinger, 1957). In unidimensional IRT and Rasch models, the fundamental assumption is that one trait is assessed at a time, this is known as unidimensionality. A scale is unidimensional when all items in a scale measure only one underlying trait. Unidimensionality can be assessed with a formal test proposed by Smith (2002). This test uses the first residual factor in a principal components analysis (of residuals) to determine two groups of items: those with positive- and those with negative residuals. Each set of items is then used to calculate an independent trait estimate for each person in the

sample. When items form a unidimensional scale, it is expected that the person estimates from the two item subsets should be similar. An independent samples t-test is used to determine whether there is a significant difference between the two person estimates. This is repeated for each person with the expectation that the percentage of tests lying outside the range of -1.96 to 1.96 should not exceed 5%.

Substantive validity. The substantive aspect of validity refers to the degree to which the responses provided by examinees are functioning as intended by the item developer (Nam et al., 2010). Substantive validity is assessed in this study by examining whether the category structure of the 5-point Likert-type format response categories is suitable for the scale in this sample. An ordered set of response thresholds for each item is expected when responses to the items are consistent with the metric estimate of the underlying construct (Tennant & Conaghan, 2007).

Generalizability. The generalizability aspect examines the extent to which score properties and interpretations generalize to and across population groups, settings, and tasks (Cook & Campbell, 1979; Shulman, 1970). This aspect of validity can be investigated in terms of MI across populations or settings. Item invariance requires that the probability of a given response on an item is independent of the subgroup an individual belongs to (Bond & Fox, 2001). Items not demonstrating invariance exhibit differential item functioning (DIF; Makransky & Glas, 2013; Makransky, Schnohr, Torsheim, & Currie, in press). For example, DIF occurs when different subgroups within the sample (e.g., males vs. females) have different scores on specific items, despite equal levels of latent trait (e.g., planning). In this study, we investigate the invariance of the CDME-SF across gender (two groups: male/female), age (three groups: 14, 15-16, 17-19), grade (three levels: 10th, 11th, 12th grades), and self-reported achievement level (three groups: high, middle, and low achievers). Presence of DIF is detected through analysis of variance of item scores

across each person factor (e.g., gender) and indicated by a significant main effect of the person factor (in this case at the $\alpha = 0.5$ level with a Bonferroni correction).

In addition to the aspects of validity described above, a general test of fit to the model is also reported. The chi-squared statistic is typically used to assess the general fit of the data to the PCM, including the property of invariance across the trait. A significant chi-squared indicates that the hierarchical ordering of the items varies across the trait, which violates the requirements of the PCM (Pennant & Tennant, 2006). A significance value of 0.5 with a Bonferroni adjustment to account for the number of hypothesis tested was used in this study.

Finally, reliability in the form of Cronbach's alpha and the person separation index (PSI) is also reported in this study. Similar interpretations can be used when using the two measures (Tennant & Conaghan, 2007). However, the PSI relies on an IRT perspective where standard error may vary at different points of the latent construct depending on the item information that is available at that point. Therefore, the PSI is directly related to the targeting of the items, and is important in applied settings because poorly targeted measures often result in floor or ceiling effects.

Results

Using the CDSE-SF as a single, global scale

The results did not support the use of the CDSE-SF as a single, global scale of career decision self-efficacy. Although the scale had a high level of reliability ($\alpha = 0.91$) and separated well between individuals scoring low and high on the scale ($\text{PSI} = 0.91$). Row 3 in Table 1 shows that the scale did not fit the PCM, $\chi^2(200) = 264.94, p = 0.001$. The fundamental issue with the misfit was related to the content and structural aspects of validity. The unidimensionality test indicated that 13.53% of the sample, which is considerably higher than the nominal level of 5%, scored significantly different when dividing the scale into two parts based on a principal component

analysis. This provides strong evidence for multidimensionality in the scale. Furthermore, there were three items (9, 13, and 16) that did not fit the scale, and there were several examples of local dependence between the items. The results support the substantive aspect of validity, evidenced by the finding that the 5-point Likert-type scale functioned appropriately, as there were no reverse thresholds found in the scale.

-----Insert Table 1 here-----

Validity of the five subscales in the CDSE-SF

Occupational information (OI). The five occupational information (OI) items, when treated as a subscale, fit the PCM, $\chi^2(40) = 37.32, p = 0.59$. Furthermore, the unidimensionality of the subscale was supported by the unidimensionality test (1.37% significant tests; Smith, 2002), and there were no positive residuals over 0.2 indicating no redundancy or local dependence (LD) in the items within the subscale. There were also no items with significant chi-squared statistics, indicating that all of the items were good measures of the latent OI construct. Cronbach's alpha for the subscale was 0.68 and the PSI was 0.69, indicating acceptable reliability and person separation for a 5-item scale. The similar level of the Cronbach's alpha and PSI coefficients is an indication that the items had appropriate targeting for the sample. This can be seen in the item-person distribution map in Figure 1. The person distribution is shown on the left side of each frame in the figure, and the item parameter thresholds are shown on the right side of each frame. Appropriate targeting is obtained when the item parameter thresholds correspond to the distribution of respondents. Furthermore, the response categories functioned as expected, which supports the 5-point Likert-type response format. These results support the content, structural, and substantive aspects of validity for the OI subscale. The cross-validation with the post-test data also supports these results.

-----Insert Figure 1 here-----

Self-appraisal (SA). The self-appraisal (SA) subscale did not fit the PCM, $\chi^2(40) = 66.46, p = 0.005$. The unidimensionality of the model was supported (4.71% significant tests), and there were no positive residuals over 0.2 indicating no redundancy or LD in the items. Furthermore, all items fit the PCM. The primary source of misfit was that the response categories did not function as intended in Item 22 “Define the type of lifestyle you would like to live”. Figure 2 illustrates the probability of each response option along the latent trait continuum for Item 22, with the top frame representing the pre-test, and the bottom frame representing the cross-validation. It is clear from the top frame in the figure that there is no point on the latent trait continuum where response option 1 (*Very little confidence*) has the highest probability of being selected. This is an indication that the response option is unnecessary. The cross-validation analysis indicated that the subscale did fit the PCM, $\chi^2(40) = 40.94, p = 0.42$, and the response categories functioned well for all items including Item 22 (see bottom frame in Figure 2). Finally, Cronbach’s alpha was 0.66 and the PSI was 0.65 indicating acceptable reliability and person separation. The scale was also appropriately targeted for the sample (see Figure 1). These results support the content and structural aspects of validity of the SA subscale, but mixed results were found for the substantive aspect of validity. The cross-validation with the post-test data supported all three aspects of validity.

-----Insert Figure 2 here-----

Goal selection (GS). The goal selection (GS) subscale had good general fit, $\chi^2(40) = 44.70, p = 0.28$, and there were no items with significant chi squared statistics, indicating that all of the items were good measures of the latent GS construct. There were also no positive residuals over 0.2 indicating no redundancy or LD in the items. The response categories functioned as expected, which supports the 5-point Likert-type response format. Nonetheless, the unidimensionality of the model was not supported (6.86% significant tests). Specifically, the calculation of a total score made up of items 2 and 6 resulted in too many significant differences from the sum of items 11, 16,

and 20. The lack of unidimensionality was not replicated in the cross-validation (3.87%). Finally, Cronbach's alpha was 0.73 and the PSI was 0.72 indicating adequate reliability and person separation. Figure 1 also shows that the subscale was also appropriately targeted for the sample. These results support the content and substantive aspects of validity of the GS subscale, but mixed results were obtained for the structural aspect of validity. The cross-validation with the post-test data supported all three aspects of validity.

Planning (P). The planning (P) subscale had good general fit to the PCM, $\chi^2(35) = 46.10, p = 0.10$. Unidimensionality was supported (2.94% significant tests), and there were no positive residuals over 0.2 indicating no redundancy or LD. All items were good measures of the latent P construct, and the response categories functioned as expected, which supports the 5-point Likert-type response format. The Cronbach's alpha of the scale was 0.72 and the PSI was 0.69 indicating satisfactory reliability and acceptable person separation. The scale was also appropriately targeted for the sample (see Figure 1). The results support the content, structural, and substantive aspects of validity for the P subscale. The cross-validation with the post-test data also supports these results.

Problem solving (PS). The problem solving (PS) scale had good general fit, $\chi^2(35) = 28.14, p = 0.89$, and unidimensionality was supported (4.71% significant tests). There were no positive residuals over 0.2 indicating no redundancy or LD in the items. Moreover, all of the items were good measures of the latent GS construct, and the response categories functioned as expected, which supports the response format. Cronbach's alpha was 0.68 and the PSI was 0.67, indicating acceptable reliability and person separation. Figure 1 also shows that the subscale was appropriately targeted for the sample. The results support the content, structural, and substantive aspects of validity for the PS scale; however, the cross-validation identified a possible issue with the unidimensionality of the scale as 5.98% of the t-tests were significant.

Measurement invariance in the form of differential item functioning in the CDSE-SF

-----Insert Table 2 here-----

An analysis of MI across different demographic groups showed that there were no items that displayed DIF by grade (see Table 2). Item 21 (“Identify employers, firms, and institutions relevant to your career possibilities”) displayed DIF by age. The group with 15 and 16 year old students scored significantly higher on the item compared to their older (17-19 years of age) and younger colleagues (14 years of age) despite having the same level of the latent trait.

DIF by gender was identified in the OI scale. Specifically, girls scored higher on Item 1 (“Find information at the library about occupations you are interested in”) and boys scored higher on Item 15 (“Find out about the average yearly earnings of people in an occupation”) despite having the same level of the latent trait. Finally, Item 5 (“Accurately assess your abilities”) displayed DIF by gender for the SA scale. Boys scored higher than girls on this item despite having the same level of the SA trait.

DIF by achievement level was identified in the OI scale for Item 23 (“Find information about colleges or universities”) and in the SA scale for Item 5 (“Accurately assess your abilities”). The group of high achievers scored highest on the two items, followed by the middle, and low groups despite having the same level of the latent traits.

Discussion

The construct of career decision self-efficacy is valuable because individuals’ career decisions and behavior are often better predicted by their beliefs about capabilities rather than actual capabilities (Bandura, 1986; Lent, Brown, & Hackett, 1994). Furthermore, in career assessment research, self-report measures of self-efficacy are easier to develop and use than actual assessments of capabilities. In this study, we used the partial credit model (PCM; Masters, 1982) within the framework of item response theory (IRT; Embertson & Riese, 2000), to investigate the content,

structural, substantive, and generalizability aspects of construct validity (Messick, 1995) of the CDSE-SF. We specifically investigated three central research questions related to the general use of the CDSE-SF in a sample of 534 high school students between the age of 14 and 19. First, we examined the validity of using the CDSE-SF as a global scale of career decision self-efficacy. Next, we tested the validity of the theoretically derived five-factor model suggested by Taylor and Betz (1983). Finally, we investigated the MI of the scale across the demographic variables of grade, age, gender, and achievement level.

In relation to the first research question, our findings show clear evidence of multidimensionality for the CDMESE SF and consequently, we recommend that a single unidimensional measure of career decision self-efficacy without consideration for subscales should not be used. This finding supports Taylor and Betz' (1983) theoretical model that describes five career choice competencies including: developing an accurate self-appraisal, gathering occupational information, goal selection, making plans for the future, and problem solving as relevant for the career decision-making process. The results contradict the findings of Nam et al. (2010) and others (e.g., Robbins, 1985; Taylor & Betz, 1983; Taylor & Popma, 1990), who provide general support for the use of the global scale. The contrasting findings between the Nam et al. study and ours deserve closer examination as this is, to our knowledge, the only previous study which has used the Rasch model to investigate the validity of the CDSE-SF.

The main method used to assess unidimensionality was similar in the Nam et al. (2010) study and our study. Both studies derived two person estimates for the items with positive- and negative loadings for each respondent. However, in our study, we used a formal test (cf. Smith, 2002) to determine the percentage of tests that fall outside the recommended range of -1.96 to 1.96, with the expectation that fewer than 5% should do this in order to conclude that the scale is unidimensional. In the Nam et al. study, a graphical depiction of the test was used to determine the

unidimensionality of the scale, so the percentage of tests that exceed the range is not reported.

However, from inspecting their graph it seems unlikely that the number of tests outside this range is as high as in our study (13.53%). Furthermore, Nam et al. supplemented the unidimensionality test from the Rasch model with the use of PCA, which also resulted in the conclusion that the scale was unidimensional. Therefore, it is likely that the causes of the divergent results are a consequence of differences in the studies and not a result of methodological choices.

There are at least three explanations for differences between the two studies. The first is that the Korean version of the CDSE-SF was used in the Nam et al. (2010) study, and an adapted Australian version was used in this study. A related issue is the difference in cultures between the two samples. The construct of career decision self-efficacy may have a different structure in Korea compared to Australia, as large differences have been identified in the structure of the CDSE-SF across cultures in previous studies (e.g., Chaney, Hammond, Betz, & Multon, 2007; Creed et al., 2002; Hampton, 2005; Peterson & del Mas, 1998). A third explanation is that different samples were used in the two studies. College students were used in the Nam et al. study and high school students were used in our study. Therefore, it is also possible that the dimensionality of the CDSE-SF is not the same in high school and college samples.

Strong evidence was found for the second research question, which supports the validity of using the five independent subscales of the CDSE-SF as suggested by Betz et al. (1996). The results indicated that four (OI, GS, P, and PS) of the five subscales, with the exception of the SA subscale, had good fit to the PCM, and the results of the cross-validation indicated that all five subscales fit the model. The primary source of misfit for the SA subscale was that the response categories did not function as intended in Item 22 (top frame of Figure 2). Response option 1 (*Very little confidence*) was not consistently selected, meaning that respondents went from selecting 0 (*No confidence at all*) to 2 (*Moderate confidence*). A rescoring of this item resulted in acceptable fit to

the model. Inappropriate ordering of response thresholds is evidence against the substantive aspect of validity, which refers to the degree to which the responses provided by examinees are functioning as intended by the test developer. In this case, it is unlikely that there is a fundamental problem with the response categories for the item, because the cross-validation analysis indicated that the subscale did fit the PCM, and the response categories functioned well for all of the items. Therefore, the result is most likely a consequence of a Type 1 statistical error due to the large number of tests that were conducted in the study.

There was also evidence of structural validity as four of the five subscales, with the exception of the GS subscale, were unidimensional in the original analysis, and four of the five subscales, with the exception of the PS subscale, were unidimensional in the cross-validation. Again, the inconsistent results here are likely due to statistical artifacts as they were not consistent across the original and the cross-validation analysis. There was also support for content validity as there were no items that were redundant or locally dependent (LD), and all items fit the model within each of the subscales. Finally, there was support for the suitability of the 5-point Likert-type format response categories, which is evidence of substantive validity. Only one item was found where the response format did not function as intended, and the problem did not exist in the cross-validation.

The reliability and person separation of the subscales were also acceptable, with values ranging from 0.66 to 0.73 for Cronbach's alpha and 0.65 to 0.72 for person separation. While these Cronbach's alpha values are similar to those reported by Albion (2000), who found subscale alphas ranging between .64 and .74 in a sample of Australian high school students, they are lower than the subscale alpha values reported in older samples. Betz, Hammond, and Multon (2005) found subscale alpha values ranging from 0.78 to 0.85 in one sample of 400 undergraduate students and from 0.80 to 0.84 in a sample of over 600 undergraduates. These students were predominantly

white. Furthermore, Chaney, Hammond, Betz and Multon (2007) found subscale alpha values ranging from 0.78 to 0.85 in a sample of 220 African American college students. However, the finding that alpha and person separation were similar is an indication that the subscales were well targeted for the sample of Australian high school students used in this study. It is important, particularly in applied settings, that measures are appropriately targeted to assessment populations, as poorly targeted measures often result in floor or ceiling effects.

There were mixed findings related to the third research question, which investigated the MI of the subscales in the CDSE-SF across the demographic variables of grade, age, gender, and achievement level. In general, few items exhibited significant differential item functioning (DIF) across these demographic variables. Therefore, the practical consequences of MI across these variables in the current sample do not seem to be excessive.

No significant DIF was identified across grade levels. This result provides evidence that the CDSE-SF functions similarly across Australian students in grades 10 through 12. Nonetheless, there were items in the subscales that could introduce bias, which could affect the validity of conclusions made when generalizing across gender, achievement level, and age. Specifically within the OI scale, Items 1 (“Find information at the library about occupations you are interested in”) and 15 (“Find out about the average yearly earnings of people in an occupation”) had significant DIF by gender. Girls were more likely to endorse Item 1 while boys were more likely to endorse Item 15, despite having the same level of the latent trait. These results indicate that girls focus more on information about occupations that are interesting, while boys focus more on salary when seeking occupational information. The practical consequence of this result for using the total score on the OI subscale is not substantial, as the two items with DIF were biased in opposite directions, which removed the effect. Consequently, the total score for the subscale does not introduce systematic bias for gender. Still, the results will be useful for career counselors and

researchers as they could be an indication of gender differences in motivation for collecting occupational information and thinking about a career. Item 1 is clearly a measure of intrinsic motivation, whereas Item 15 is a measure of extrinsic motivation (Deci & Ryan, 1985). DIF by gender was also identified for Item 5 (“Accurately assess your abilities”), where boys scored higher despite having the same level of the latent trait. The results are consistent with existing research that has found that female self-evaluations are more responsive to the valence of the evaluative feedback they receive compared to males (Roberts, 1991). Therefore, males are typically more confident in their ability to accurately assess their abilities.

Item 5 also displayed DIF by achievement level in favor of high achievement students. This indicates that high achieving students are more confident in their ability to accurately assess their own ability than their lower performing colleagues, despite having the same level of self-appraisal. A possible explanation for this result could be that a subjective, self-report measure of achievement was used in this study. Therefore, it is possible that students who are more likely to respond that they are high achievers are also more likely to respond that they are good at accurately assessing their own abilities. Future research should, therefore, investigate if DIF by achievement level is identified when objective measures of achievement are used. DIF by achievement level was also identified in favor of the high achievement group for Item 23 (“Find information about colleges or universities”) from the OI subscale. This could be a consequence that high achieving students expect to attend college or university while lower achieving students may not necessarily be considering a college or university degree.

Finally, there was one item in the CDSE-SF that displayed DIF by age. This was Item 21 (“Identify employers, firms, and institutions relevant to your career possibilities”) from the P subscale. The results are surprising because the group with 15 and 16 year old students scored significantly higher on the item compared to older (17-19 years of age) and younger (14 years of

age) students despite scoring the same level on the latent trait of planning. This finding may be related to a school-based activity that encourages students to explore work experience and tertiary education pathways at different points in their education.

Although few items were identified with DIF in this study, there are several options available for dealing with the lack of MI in the form of DIF. The first is to eliminate or revise the items. Another option is to use stratified norms and cut-off scores when using the CDSE-SF. A final option is to use statistical procedures that can account for the DIF by using group specific item parameters (e.g., Kreiner & Christensen, 2011; Makransky & Glas, 2013). These procedures correct for the differences at the item level by estimating ability with different item parameters for each demographic group.

Future research

Future research should use the Rasch model framework to investigate the generalizability of the results found in this study across different cultures and CDSE-SF language versions. The use of the Rasch model is important because it provides a direct test of the validity of using a total-score for the general CDSE-SF scale and the five subscales. This type of research would further generate evidence about the dimensionality of the CDSE-SF across different settings. Additionally, as there were few results in this study which seemed to be a consequence of statistical artifacts, future research should investigate if the results are replicated in other settings.

Further, Rasch or IRT models would be appropriate to provide information about MI across a broader spectrum of demographic variables. This would add to the validity and generalizability of the CDSE-SF. One example is to investigate MI across time points. There are items in the CDSE-SF where the item parameters may have changed across time. One example is Item 1 (“Find information at the library about occupations you are interested in”). Today, it may not be appropriate to ask about the library as most people use the internet to search for information.

Investigating MI by investigating item parameter drift over time within the framework of IRT (e.g., Makransky et al., in press) would provide information about the items that should be revised. This research could also provide valuable information about how the construct of career decision self-efficacy is changing. Additionally, in our study, MI was investigated across three grade levels (10th, 11th, and 12th grades) and across children from 14 to 19 years of age. Future research should investigate the MI of the scale across a broader range of grade or age levels. Future research should also investigate MI across different cultures and language versions of the test.

References

- Albion, M. J. (2000). *Developing and validating a model of career decision making*. Unpublished doctoral dissertation, University of Southern Queensland, Australia.
- Anderson, D. S., & Vervoorn, A. E. (1983). *Access to privilege: Patterns of participation in Australian post-secondary education*. Canberra, Australia: Australian National University Press.
- Andrich, D., Sheridan, B., & Luo, G. (2010). *Rasch models for measurement: RUMM2030*. Perth, Australia: RUMM Laboratory.
- Arnold, J., & Bye, H. (1989). Sex and sex role self-concept as correlates of career decision-making self-efficacy. *British Journal of Guidance and Counselling*, *17*, 201-206.
doi:org/10.1080/03069888908258053
- Bandura, A. (1986). *Social foundations of thought and action: A social cognitive theory*. Englewood Cliffs, NJ: Prentice Hall.
- Betz, N. E., Hammond, M. S., & Multon, K. D. (2005). Reliability and validity of five-level response continua for the Career Decision Self-Efficacy Scale. *Journal of Career Assessment*, *13*, 131-149. doi:10.1177/106907279600400103
- Betz, N. E., & Klein, K. (1996). Relationships among measures of career self-efficacy, generalized self-efficacy, and global self-esteem. *Journal of Career Assessment*, *4*, 285-298.
doi.org/10.1177/106907279600400304
- Betz, N. E., Klein, K. L., & Taylor, K. M. (1996). Evaluation of a short form of the Career Decision-Making Self-Efficacy Scale. *Journal of Career Assessment*, *4*, 47-57.
doi.org/10.1177/106907279600400103
- Betz, N. E. & Taylor, K. M. (2012). *Career Decision Self-Efficacy Scale and Short Form and Manual*. Menlo Park, CA: Mind Garden, Inc.

- Bond, T. G., & Fox, C. M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates. doi.org/10.1111/j.1745-3984.2003.tb01103.x
- Buyukgoze-Kavas, S. (in press). A psychometric evaluation of the Career Decision Self-Efficacy Scale – Short Form with Turkish university students. *Journal of Career Assessment*.
- Chaney, D., Hammond, M. S., Betz, N. E., & Multon, K. D. (2007). The reliability and factor structure of the Career Decision Self-Efficacy Scale-SF with African Americans. *Journal of Career Assessment*, 15, 194-205. doi:10.1177/1069072706298020
- Chung, B. Y. (2002). Career Decision-Making Self-Efficacy and Career Commitment: Gender and ethnic differences among college students. *Journal of Career Development* 28, 277-284. doi.org/10.1023/A:1015146122546
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Chicago, IL: Rand McNally.
- Creed, P. A., Patton, W., & Watson, M. B. (2002). Cross-cultural equivalence of the Career Decision-Making Self-Efficacy Scale - Short Form: An Australian and South African comparison. *Journal of Career Assessment*, 10, 327-342. doi:10.1177/10672702010003004
- Crites, J. O. (1961). A model for the measurement of vocational maturity. *Journal of Counselling Psychology*, 8, 255-259.
- Crites, J. O. (1978). *Career Maturity Inventory*. Monterey, CA: CTB/McGraw-Hill.
- Deci, E. L., & Ryan, R. N. (2010). *Self-Determination*. Hoboken, NJ: Wiley.
- Duffy, R. D., Diemer, M. A., & Jadidian, A. (2012). The development and initial validation of the Work Volition Scale – Student Version. *The Counselling Psychologist*, 40, 291-319. doi:10.1177/0011000011417147

- Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Gaudron, J.-P. (2011). A psychometric evaluation of the Career Decision Self-Efficacy Scale Short-Form among French university students. *Journal of Career Assessment, 19*, 420–430.
- Hackett, G., & Betz, N. E. (1981). A self-efficacy approach to the career development of women. *Journal of vocational behavior, 18*, 326-339. doi.org/10.1016/0001-8791(81)90019-1
- Hampton, N. Z. (2005). Testing for the structure of the Career Decision Self-Efficacy Scale-Short Form among Chinese college students. *Journal of Career Assessment, 13*, 98-113. doi.org/10.1177/1069072704270298
- Hong, S., Kim, B. S. K., & Wolfe, M. M. (2005). A psychometric revision of the European American Values Scale for Asian Americans using the Rasch model. *Measurement and Evaluation in Counseling and Development, 37*, 194-207.
- Kim, B. S. K., & Hong, S. (2004). A psychometric revision of the Asian Values Scale using the Rasch model. *Measurement and Evaluation in Counseling and Development, 37*, 15-27.
- Lent, R. W., Brown, S. D., & Hackett, G. (1994). Toward a unifying social cognitive theory of career and academic interest, choice, and performance. *Journal of Vocational Behavior, 45*, 79-122. doi.org/10.1006/jvbe.1994.1027
- Loevinger, J. (1957). Objective tests as instruments of psychological theory. *Psychological Reports, 3*, 635-694. doi.org/10.2466/pr0.1957.3.3.635
- Luzzo, D. A. (1995). The relative contributions of self-efficacy and locus of control to the prediction of career maturity. *Journal of College Student Development, 36*, 61-66.
- Makransky, G., & Glas, C. A. W. (2013). Modeling differential item functioning with group-specific item parameters: A computerized adaptive testing application. *Measurement, 46*, 3228-3237. doi.org/10.1016/j.measurement.2013.06.020

- Makransky, G., Schnohr, C., Torsheim, T., Currie, C. (in press). Equating the HBSC Family Affluence Scale across survey years: A method to account for item parameter drift using the Rasch model. *Quality of Life Research*.
- Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149-174.
doi.org/10.1007/BF02296272
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, *50*, 741-749. doi.org/10.1037//0003-066X.50.9.741
- Miller, M. J., Roy, K. S., Brown, S. D., Thomas, J., & McDaniel, C. (2009). A confirmatory test of the factor structure of the short form of the Career Decision Self-Efficacy Scale. *Journal of Career Assessment*, *17*, 507-519. doi.org/10.1177/1069072709340665
- Nam S. K., Yang E., Lee S. M., Lee S. H., & Seol H. (2011). A psychometric evaluation of the Career Decision Self-Efficacy Scale with Korean students: A Rasch model approach. *Journal of Career Development*. *38*, 147-166. doi.org/10.1177/0894845310371374
- Nilsson, J. E., Schmidt, C. K., & Meek, W. D. (2002). Reliability generalization: An examination of the Career Decision-Making Self-Efficacy Scale. *Educational and Psychological Measurement*, *62*, 647-658. doi.org/10.1177/0013164402062004007
- Osipow, S. H., & Gati, I. (1998). Construct and concurrent validity of the Career Decision-Making Difficulties Questionnaire. *Journal of Career Assessment*, *6*, 347-364.
doi.org/10.1177/106907279800600305
- Page, J., Bruch, M. A., & Haase, R. F. (2008). Role of perfectionism and five-factor model traits in career indecision. *Personality and Individual Differences*, *45*, 811-815.
doi:10.1016/j.paid.2008.08.013

- Pallant, J. F., & Tennant, A. (2007). An introduction to the Rasch measurement model: An example using the Hospital Anxiety and Depression Scale (HADS). *British Journal of Clinical Psychology, 46*, 1-18. doi.org/10.1348/014466506X96931
- Peterson, S. L., & del Mas, R. C. (1998). The component structure of career decision-making self-efficacy for underprepared college students. *Journal of Career Development, 24*, 209-225. doi:10.1177/089484539802400304
- Presti A. L., Pace F., Mondo M., Nota L., Casarubia P., Ferrari L., & Betz N. E. (2012). An examination of the structure of the Career Decision Self-Efficacy Scale (Short Form) among Italian high school students. *Journal of Career Assessment 21*, 337-347. doi.org/10.1177/1069072712471506
- Robbins, S. B. (1985). Validity estimates for the Career Decision-Making Self-Efficacy Scale. *Measurement and Evaluation in Counseling and Development, 18*, 64-71.
- Roberts, T. A. (1991). Gender and the influence of evaluations on self-assessments in achievement settings. *Psychological Bulletin, 109*, 297-308. doi.org/10.1037//0033-2909.109.2.297
- Shulman, L. S. (1970). Reconstruction of educational research. *Review of Educational Research, 40*, 371-3. doi.org/10.3102/00346543040003371
- Smith, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal Applied Measurement, 3*, 205-231.
- Taylor, K. M., & Betz, N. E. (1983). Applications of self-efficacy theory to the understanding and treatment of career indecision. *Journal of Vocational Behavior, 22*, 63-81.
- Taylor, K. M., & Popma, J. (1990). An examination of the relationships among career decision making self-efficacy, career salience, locus of control, and vocational indecision. *Journal of Vocational Behavior, 37*, 17-21. doi.org/10.1016/0001-8791(90)90004-L

- Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis & Rheumatism*, 57, 1358-1362. doi.org/10.1002/art.23108
- Wolfe, E. W., & Smith, E. V. Jr. (2007). Instrument development tools and activities for measure validation using Rasch models: Part II—Validation activities. *Journal of Applied Measurement*, 8, 204-233.
- Wright, B. D. (1996). Local dependency, correlations and principal components. *Rasch Measurement Transactions*, 10, 509-11.

Table 1

General Results of the PCM Analyses.

Scale	General fit	Reliability		Unid.	Item fit	LI	Use of response categories
		PS	Alpha				
Pre-test							
Global scale	$\chi^2 = \mathbf{0.001}$	0.91	0.91	13.53%	9, 13, 16	2-6 4-6 6-9 24-25	-
OI	$\chi^2 = 0.59$	0.69	0.68	1.37%	-	-	-
SA	$\chi^2 = \mathbf{0.005}$	0.65	0.66	4.71%	-	-	22
GS	$\chi^2 = 0.28$	0.72	0.73	6.86%	-	-	-
P	$\chi^2 = 0.09$	0.69	0.72	2.94%	-	-	-
PS	$\chi^2 = 0.89$	0.67	0.68	4.71%	-	-	-
Post-test							
Global scale	$\chi^2 = \mathbf{0.01}$	0.91	0.91	12%	1, 16	2-6 13-17	-
OI	$\chi^2 = 0.46$	0.68	0.68	4.10%	-	-	-
SA	$\chi^2 = 0.42$	0.69	0.70	2.51%	-	-	-
GS	$\chi^2 = 0.11$	0.70	0.72	3.87%	-	-	-
P	$\chi^2 = 0.43$	0.69	0.71	4.10%	-	-	-
PS	$\chi^2 = 0.96$	0.67	0.67	5.98%	-	-	-

Results in bold indicate miss-fit to the PCM

Table 2

Differential Item Functioning in the CDSE-SF.

Scale	Grade	Age	Gender	Achievement
OI	-	-	1, 15	23
SA	-	-	5	5
GS	-	-	-	-
P	-	21	-	-
PS	-	-	-	-

Figure Titles

Figure 1: Person-item distribution map for the five subscales in the CDSE-SF.

Figure 2: Category probability curves map for Item 22 “Define the type of lifestyle you would like to live” in the SA subscale.

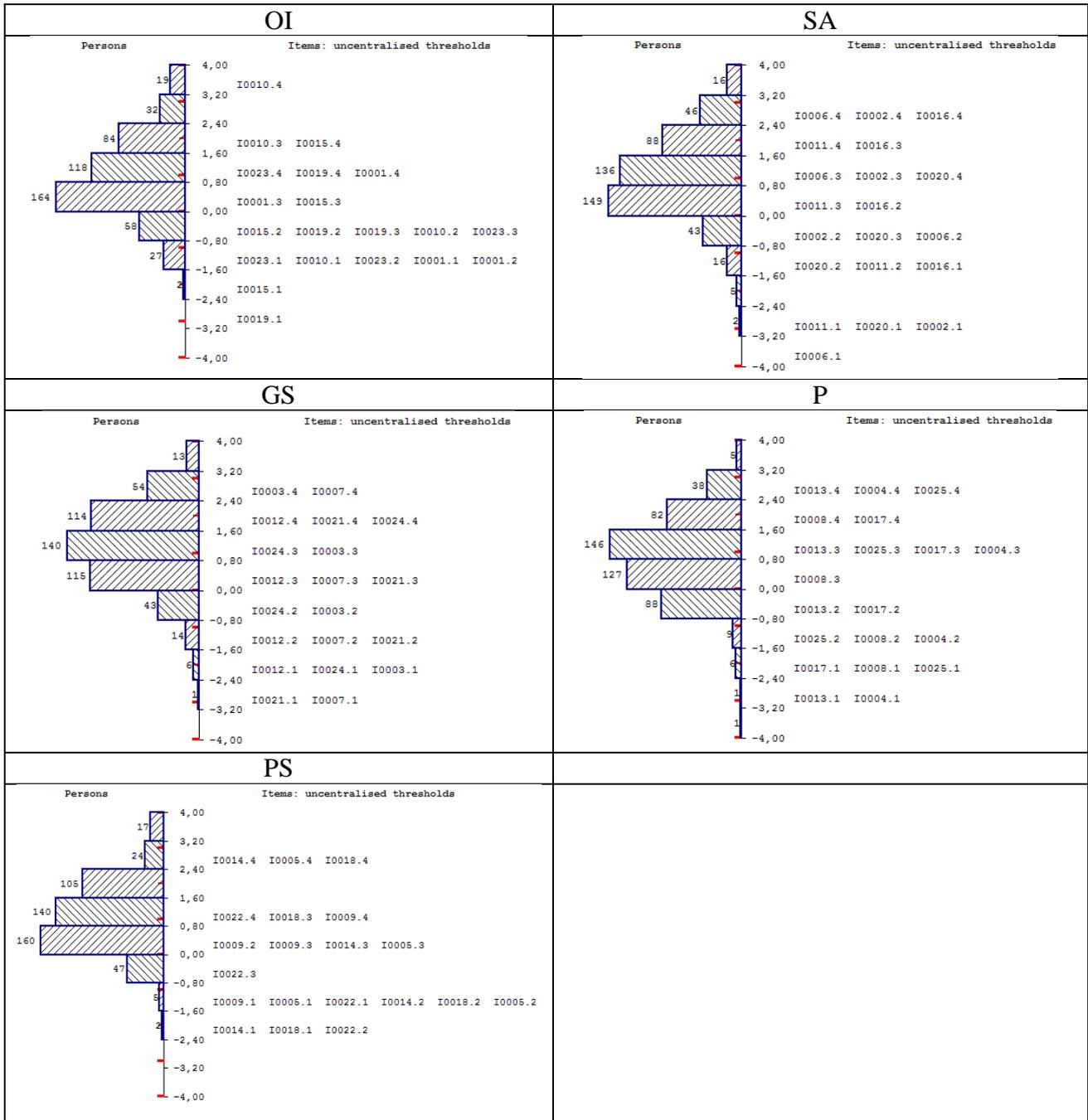


Figure 1: Person-item distribution map for the five subscales in the CDSE-SF.

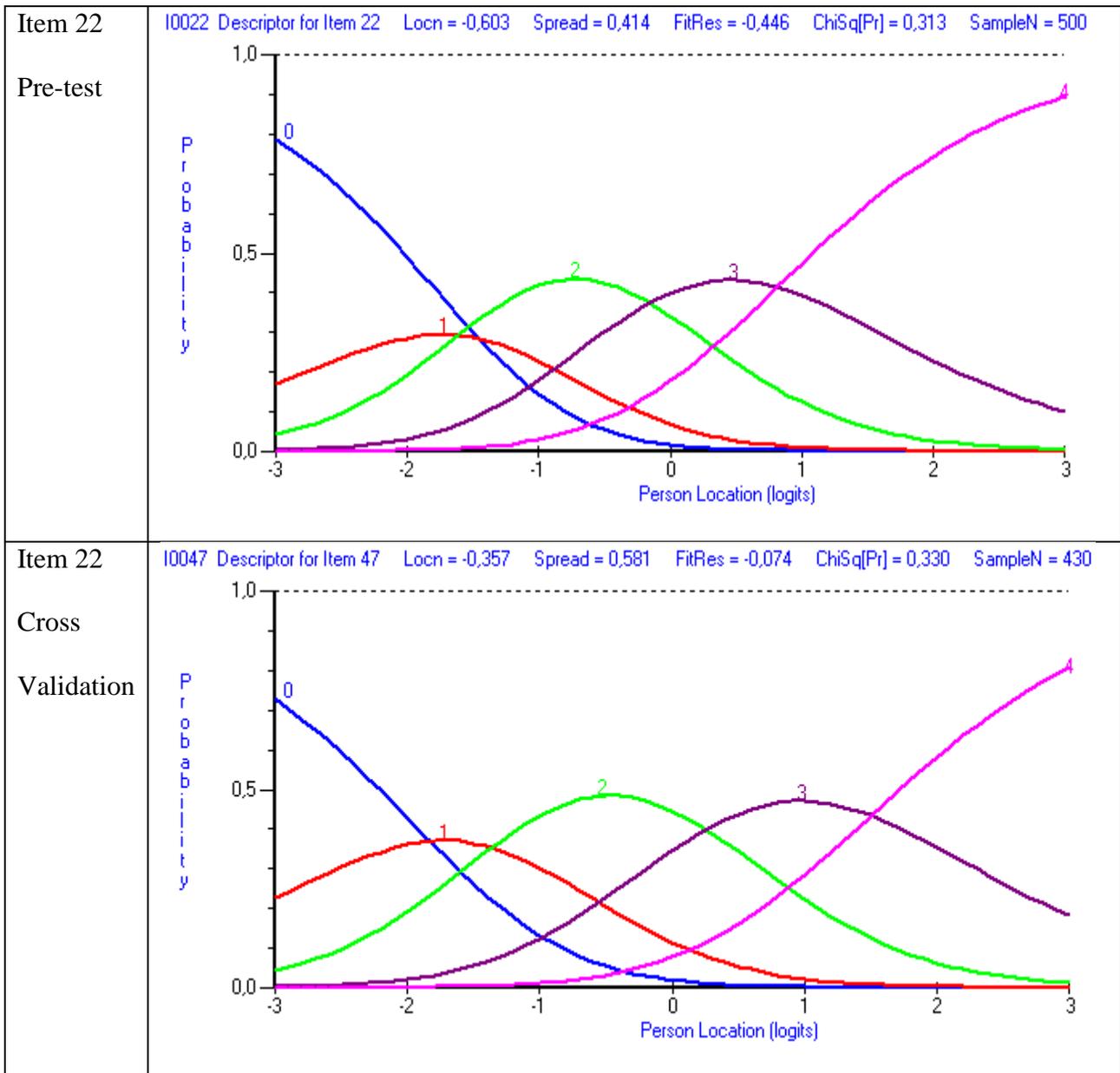


Figure 2: Category probability curves map for Item 22 “Define the type of lifestyle you would like to live” in the SA subscale.