

Running head: EDPS 591M Test Review

A Review of the Clifton Strengths Finder Instrument

Doug Kueker

Purdue University

Lopez, S., Hodges, T., & Harter, J. (2005). *The Clifton StrengthsFinder Technical Report: Development and Validation* [technical report]. Princeton, NJ: The Gallup Organization.

Overview

The Clifton Strengths Finder claims to measure personal talents. The instrument is based upon Don Clifton's notion of talent (Clifton & Nelson, 1992). Hodges and Clifton (2004) formally define the construct of personal talent as "naturally recurring patterns of thought, feeling, or behavior that can be productively applied" (Lopez, Hodges, & Harter, 2005, p. 257) and "manifested in life experiences characterized by yearnings, rapid learning, satisfactions, and timelessness" (Lopez, Hodges, & Harter, 2005, p. 3). Thus, talent is considered a trait-like individual difference that is the product of development from regular biological processes and successful experiences during childhood and adolescence. Though the instrument's label suggests that it aids in measuring "strengths" the authors draw a distinction between strengths and talents. Specifically, they define strengths as an extension of talents, "a strength construct combines talents with associated knowledge and skills" in such a way that the individual is able to deliver a near-perfect performance on a particular task (Lopez, Hodges, & Harter, 2005, p. 3). This review concerns the StrengthsFinder instrument, which actually measures Clifton's notion of talents as described in this section. Important issues for test users are considered including the test's construction and technical adequacy.

Test Description

Test Coverage and Use

The intended use of the test is clearly outlined in the technical manual provided. The test is primarily designed to provide feedback that can foster intrapersonal development. In addition, both appropriate and inappropriate applications of the instrument are further defined. Appropriate applications include administration to adolescents and adults with a reading level of 10th grade or higher for the purposes of personal development. While the construct being measured seems to have applicability to such uses as employee or student selection; the test publishers actually take a clear position that this is not an appropriate use of the instrument. The test is not designed or validated for use in employee selection or mental health screening. In addition, a normative comparison of talent profiles across multiple individuals is discouraged. Lastly the technical report notes that the Strengths Finder is not sensitive to change. It should not be used as a pre-post measure of an individual's personal growth. One should also note that the authors also identify the need for using related supporting materials (i.e. books by Buckingham & Clifton, 2000; and Clifton & Nelson, 1992) to assist individuals with maximizing the use and processing of individual results. Each administration of the test for students costs \$15 and \$25 and both costs include some form of access (online or print) to the supporting materials suggested to make the results more meaningful.

Administration

The Clifton StrengthsFinder is administered online to an individual in less than one hour. Individuals must have access to a secure Internet connection. At the beginning of the test the individual responds to demographic questions, however these have no bearing on test results and are only used by the test publisher to analyze the psychometric properties of the instrument among different groups of respondents. The demographic section is completely optional. One should note that there are no standardized conditions for test administration beyond the fact that the test is presented via the computer in a consistent manner. This condition does increase the potential for some of the external circumstances to influence results (i.e. distractions in the environment). The fact that instructions are delivered the same way each time through the computer delivery method does help to standardize the conditions for test taking.

Scale and Instrument Completion

The information presented describing the scale is fairly complete. The scale for each construction is similar to the Edwards Personal Preference Schedule (Anastasi & Urbana, 1997). Similar to Edwards (1959) assessment of needs, the Clifton StrengthsFinder instrument consists of 180 pairs of items presented in the preferred language of the user (17 languages are available and the test can be modified for individuals with disabilities). Each item lists a pair of potential self-descriptors on two ends of a polar continuum with a modified Likert-type agreement scale in between. An example item pairing is, “I read instructions carefully” and “I like to jump right into things” (Lopez, Hodges and Harter, 2005 p. 3). Items are grouped into 34 subscales, each with 6-9 items that represent a particular talent theme.

The test taker has 20 seconds to respond to an item pairing before the system moves on to the next item pair. The publisher states this is intended to produce a top-of-mind response, which relates to their operational definition of talent. The publisher does disclose that the 20-second limit results in a negligible item non-completion rate. However, they do not disclose the specific amount of this rate making it hard to evaluate the actual evidence.

As mentioned earlier, the scale is a modification of a forced choice item model. Test takers select their choice between the two descriptors along a continuum of five available options presented between the two polar ends. It should be noted that the items are not always polar ends of the same continuum for a particular concept. The options for each description are labeled (i.e. Strongly Agree, Agree, Neutral, Agree, and Strongly Agree). Usually, the forced-choice technique is employed to reduce socially desirable answers (Anastasi & Urbana, 1997). However, the authors do not discuss this as part of the rationale for constructing the scale as they did in the technical guide provided.

Scoring of Items

Based upon the scale used and past history from similar scales (i.e. Edwards, 1959) the test user should show concern for information regarding the scoring of the items. Information regarding scoring of the test presented in the technical manual is vague. Authors of the technical report state that, “scores are calculated on the mean of the

intensity of self-description” (Lopez, Hodges, & Harter, 2005, p. 6). A proprietary formula assigns a value to each of the 34 response categories. The 34 response categories represent 34 themes of talent that a person may exhibit in their regular behaviors with some degree of intensity. Values for items in each theme subscale are averaged to derive a mean theme score. The test technical document suggests that these theme scores can be reported as a mean, a standard score, and a percentile.

A concern with the format of test items lies in the potential for the creation of an ipsative matrix when items are scored. Ipsative scoring would suggest that the strength of each talent is expressed, not in absolute terms, but in relation to the strength of the individual’s other talents as measured by this instrument. Ipsativity is a concern because results would reflect nothing more than a ranking of the 34 themes for each individual rather than a measure of intensity regarding each theme for any one given individual (which, as posited by the test publishers is the intention of the instrument). The frame of reference in ipsative scoring is the individual rather than the normative sample, thus inter-individual comparisons are meaningless with an ipsative data matrix (Anastasi & Urbana, 1997). Thus, if the test is scored in such a way that it produces an ipsative matrix any further analysis regarding the technical adequacy of the test and generalization regarding the constructs measured beyond the individual would be meaningless. The test manual directly addresses this issue and provides some empirical evidence that suggests the test is not ipsatively scored. Plake (1990) found that less than 30 percent of the 180 item pairs are ipsatively scored (no more than one item for than of the 34 themes tested). A table of supporting evidence for this statement is not presented, for evaluation by the test user. This may be a potential area for further exploration regarding the capacity of the instrument.

Feedback from Instrument

The format for receiving feedback from the instrument is perhaps one of the most unique aspects of the test. Participants receive feedback after completion in various ways. The form of feedback mostly depends upon the purpose for completing the test. Numerical summary scores are not provided to anyone who takes the test. Based upon the scoring issues presented earlier it is possible that the scores may be difficult to interpret for an individual without the assistance of a professional. Thus, the respondent receives only a written report describing his or her top five talent themes, those in which they received the highest mean score, in order of intensity. When working with a consultant or in a supervised session the participant may receive a ranked order of all 34 themes along with developmental information. Feedback that includes instruction, experiential learning and mentoring is suggested for all test users, however it is only required if receiving the entire 34 theme-sequence. Coupling feedback with instruction (i.e. self-referenced instruction from reading a book or with an approved consultant from the test publisher) is a unique aspect of this personality type instrument. However, it does seem to reduce the pool of individuals who can get full value out of the instrument and its feedback to those who can afford the services and instructional materials necessary to understand the feedback provided.

Technical Adequacy

Standardization Sample, Norms and Standard Scores

The test manual provides a negligible amount of information regarding the standardization sample and norming processes used in developing the instrument. For example, pilot testing of the instrument is mentioned, but not described in detail. The authors do note, “a number of sets of items” (Lopez, Harter, & Hodges, 2005, p. 5) were pilot tested, but do not offer detailed information regarding the demographics of the sample used for this process. After pilot-testing a follow-up study was conducted with 601,049 respondents. Again, the authors do not provide any details about the demographic breakdown of this audience. The manual suggests that the test is created to be used with adolescents and adults with a reading level of 10th grade or higher. Not releasing information regarding the standardization sample makes interpreting whether or not the test is valid for this stated audience difficult. The authors seem to use large numbers (in the hundred’s of thousands) as a way to infer an appropriate standardization sample, however large numbers do not necessarily equate to an appropriate standardization sample for conducting the norming process.

The test manual also offers little detail regarding the standard scores produced from the instrument. The authors claim that “a proprietary formula” assigns a value to each response category (Lopez, Hodges, & Harter, 2005, p. 6). The values for items in each of the 34 response categories are averaged to derive 34 theme scores. Then the theme scores are presented as a mean, a standard score, and a percentile. Although, it should be noted, that these scores are only added to the Gallup Organization’s database, not presented to the test taker or user in any fashion. The authors could provide examples to help in understanding this particular matter, but no examples are provided. They also do not offer whether the standard score provided is a z-score or another type of standard score. Criteria for interpreting the standard scores and percentiles produced are not mentioned in the manual. However, the feedback provided following the test is not numerical, nor is it intended to be interpreted normatively. The feedback is only offered for personal development purposes and is based upon the mean score for each of the thirty-four themes measured by the instrument. Without information regarding the standardization sample the standard scores the authors claim the test produces lack depth in their meaning.

Item Development and Selection

The initial items for the Clifton Strengths Finder were written based upon the analysis of interviews conducted by Clifton and colleagues. The test manual states that “more than two-million” individuals were interviewed in the 1990s (Lopez, Hodges, & Harter, 2005, p. 4). The items for this particular instrument were then developed based upon results of the semi-structured interviews regarding work and academic success. Interviews also included observations on the job site or in an academic setting with outstanding performers in their particular roles. An initial set of more than 5,000 items were constructed on the basis of what the authors call “traditional validity evidence” (Lopez, Hodges, & Harter, 2005, p. 4). It should be noted that no real evidence or methods to explain what this statement means are outlined. The authors state that the

initial 5,000 items were developed with an eye toward construct validity – thus trying to develop a set of items that represented the domain of personal talents along with the hypothesized sub-themes that emerged from the interview results. However, this seems to describe more about content validity rather than any statistical processes typically used to validate construct validity such as confirmatory factor analysis.

The information presented regarding item analysis methods used to reduce the original set of items is very general and does not provide adequate details for a test user to evaluate the appropriateness of the process. The authors state that the smaller pool was reduced from “quantitative review of item functioning,” which may suggest some item analysis procedures. However, none of the specific item analysis procedures were explained. Since the instrument is based upon differentiating between people on a particular set of behaviors it seems that an item discrimination analysis would be one of the most appropriate forms of item analysis for this particular instrument. The authors do state that the evidence used to evaluate the item pairs was taken from a “database of criterion-related validity studies” that included over 100 predictive-validity studies (Lopez, Hodges & Harter, 2005, p. 4). However, no actual evidence is provided to support this statement or the relevance of the data produced to the development of this instrument. The authors also note that “factor and reliability analyses” were conducted in multiple samples to “assess the contribution of items to measurement of themes and the consistency and stability of theme scores” (p. 4). Again, no quantitative evidence is provided to support these statements.

The only solid piece of evidence provided to support the item development does not seem to be produced using the appropriate statistical method. The authors note that the item-to-proposed-theme correlation (corrected for part-whole overlap) is 6.6 times larger than the average item correlation to other themes on the final instrument released from the development process. However, the authors do not mention doing a confirmatory factor analysis to achieve these results. In fact, later in the manual the authors note that a confirmatory factor analysis is still planned, but has not yet been conducted as of 2005 (p. 11). It is inappropriate for the authors to make this statement without the use of the appropriate confirmatory factor analysis procedure to support this type of conjecture.

The authors do give some details regarding the end result of the item development and selection process. In 1999 a 35-theme version of the Clifton StrengthsFinder instrument was launched for public purchase and use. After initial data collection one theme was pulled from the instrument due to redundancy with other themes based upon “quantitative analyses” (p. 5). This resulted in a total of 180 item pairs with 360 total items on the instrument. 256 of these items are scored to produce the final results. Many of the statements throughout the item development and selection section in the test manual allude to potentially valuable methods for producing evidence that supports the decisions made during the test development process. However, the manual does not provide sufficient evidence to conclude that the instrument developed has technical adequacy to support the results and interpretations it provides.

Reliability

Reliability evidence, or evidence regarding the dependability of the scores, is reported in a special section of the test manual. The authors report on two primary measures of reliability for the instrument: internal consistency and stability. With regard to internal consistency the authors report Chronbach's Coefficient alpha for each of the "sub-scales" in the instrument. The alphas reported were based upon the responses of 706 Gallup employees who were asked to take the instrument as part of joining the Gallup Organization staff. The alpha coefficients for each theme range from .55 to .81. The standard expectation for commercial measures used in psychology practice is typically .70. Twenty-three of the 34 themes are at or above this particular level. It should be noted that only one of the themes below .70 is at the .55 alpha level. The rest of the themes (10 themes) below .70 are no less than .65. The alpha for the entire test is not reported, however it may make less sense to report this number since the test feedback is all based upon the theme scores rather than an aggregate test score. It should be noted that internal consistency measures only tell us about a single moment in time and are limited in making generalizations about the domain of personal talents or how the themes measured by the instrument generalize across domains.

The authors also report on the stability of items based upon a test-retest reliability coefficient. The authors do not present an actual table to evaluate the test-retest reliability. However, they only state that "almost all" of the themes have test-retest reliability over a six-month interval between .60 and .80. The authors note evidence from two studies regarding overall test-retest reliability. The first was conducted over a 3-week test-retest period and resulted in a .76 coefficient. A second study was conducted over a 17-month test-retest period resulted in a .74 test-retest coefficient. This evidence tells us how well test takers hold their overall scores on the test from test one to test two. This information may be a little misleading given the intended use of the test (measurement of theme intensity for each of the 34 themes). It may matter very little how well the test taker's overall score on the instrument holds up over time. Given that the purpose of the test and the feedback reported from the test is centered upon the themes and their intensity as demonstrated by the scores an individual achieves; it would seem more prudent to understand the specific test-retest reliability of the themes over time. We are left only with the vague statement from the authors regarding this matter. Do test takers hold the same scoring patterns with respect to their particular theme profile from time one to time two would make an excellent question to explore regarding this matter.

Validity

The authors report on the validity of the measure in a special section of the test manual. Specifically, the test manual claims to address construct validity most heavily. This seems to be most appropriate since the test is based on a hypothesized model for personal talents.

In an effort to establish construct validity, the authors report on the item-total correlations in support of the validity of the latent constructs measured by the test. It should be noted that item-total correlation is really another method of measuring item discrimination and ultimately helps us understand more about the internal consistency of a particular set of items rather than the overall validity. The higher all items correlate

with each other the higher the internal consistency. The authors adapted the typical item-total correlation used in item analysis to find the item-total correlation for each of the 34 themes. To do so they began by calculating an average correlation for the item-total correlations of all items in a particular theme grouping. This revealed, on average, how well a set of items correlated with all of the other items on the instrument. These values range from .45 to .25 suggesting that no one particular set of items, or theme is completely redundant when compared to the test as a whole. To further confirm construct validity the authors compare this information with a calculation of the grand average of the item-cross total correlations. The authors developed an average item-cross correlation by figuring the grand average of the correlations for the other 33 themes. The authors then compare the item-total correlation for a specific theme with the item-cross total correlation as a metric providing evidence for construct validity. This, however, is a weak way to establish construct validity.

The item-total correlation evidence presented does not reveal that any of the item-total correlations for any one theme is smaller than the item-cross total correlation. For example, if the correlation within a theme were smaller than the average correlation for all themes together suggest that the theme is not unique. Overall, this seems like a very weak measure of construct validity that may lead to incorrect assumptions regarding the distinctiveness of each trait measured by the instrument. More complex statistical methods could be employed to determine whether or not the constructs measured are distinguishable as hypothesized. For instance, a confirmatory factor analysis would be more appropriate in determining the construct validity of a particular measure such as the Clifton StrengthsFinder. The particular method reported does not provide adequate evidence of construct validity.

The authors also address construct validity by addressing convergent and discriminant validity for the StrengthsFinder instrument. Convergent validity was addressed by reporting on a study by Harter and Hodges (2003). This study explored the relationship between the Clifton StrengthsFinder and the five-factor model of personality (neuroticism, extroversion, openness/intellectence, agreeableness, and conscientiousness. Priori hypotheses regarding relationships between the 34 themes and each of the big-five personality factors were established and then explored through students taking the Strengths Finder instrument concurrently with an instrument developed to measure the five-factor model of personality. It should be noted that the actual instrument used to measure the five-factor model is not reported in the manual. This study revealed that selected themes were generally correlated as expected with the five-factor model. Convergence generally occurred as expected (either positively or negatively) on themes that were hypothesized to be similar and dissimilar. However the correlations among the hypothesized links were all very small and close to zero. Strong evidence of convergence (+/- .60 or higher correlations) were only achieved between three of the themes and the respective hypothesized factor from the big-five model. Though the title to this particular section of the manual also suggests that discriminant validity will be addressed no evidence or studies are reported where the instrument was found to be uncorrelated with other measures of constructs to which it should not be related.

Evidence presented in the description of the test development does support the content validity of the instrument. The test was derived from a well described content domain which was derived from systematically conducted interviews. The use of an

initial bank of 5000 items would suggest a solid representation of the domain of personal talents. The authors also note methods for systematically analyzing the test items after each revision to ensure that the remaining items are a representative sample of the domain being measured. Thus, though the authors do not address it specifically in this section of the test manual, evidence is provided regarding content validity in other areas of the documentation. No evidence regarding criterion-related validity is provided.

Test and Item Bias Analysis

The authors do report on the measurement properties of the StrengthsFinder as it relates to specific cultural and demographic variables such as culture, age and gender. The authors took a test-level approach to analyzing themes measured by the instrument across each of the cultural and demographic variables of interest. Culture was defined as the country of the respondent's current residence when they completed the instrument and the language in which the survey was administered. The authors chose to use the weighted means and standard deviations of average item-total correlations for each of the themes as the index used to examine the instrument for invariance. The authors suggest that finding relatively small weighted standard deviations and small, positive item-total correlations in a large sample of people (536,415) from 25 different countries of residence is evidence of the stability of the themes and ultimately evidence of invariance. While this does indicate a measure of internal consistency of them items as correlated with the test it does not provide evidence suitable for evaluating the issue of invariance. Analysis of these results does not examine for invariance among the different groups who took the test. Without disaggregating the results and use a more sophisticated statistical method like significance testing with regression or an appropriate analysis of variance technique it is difficult to tell if there are systematic differences in the way the test measures the proposed themes across the 25 different countries of residence involved.

The authors provide the same weighted means and standard deviations, as described earlier, across respondents for three other variables including: 13 different survey languages; four age groups ranging from 15-60 years old; and males versus female respondents. The same argument offered above remains true. This information does tell us about the relative stability as a whole. However, as Simpson's Paradox suggests, when data for various groups is aggregated we may misinterpret or overlook important differences between the groups (Moore & McCabe, 2004). Even the stability of the measure would need to be systematically investigated by disaggregating the data to be able to support the test manual authors' assertion that the instrument is truly stable across groups.

The aggregated information presented is not an appropriate source of evidence regarding the actual invariance of this instrument across groups on any of the variables outlined by the authors. The authors' state, "evidence provided by Gallup researchers suggest that the structure of talent measured by the Clifton StrengthsFinder does not vary across demographic variables" (Lopez, Hodges, & Harter, 2005, p. 19). The information presented does not support this statement. An appropriate measure of invariance as determined by external or internal methods needs to be conducted. For example, a regression analysis to determine if the various groups of interest statistically share a common regression line would provide some evidence to support the authors' statement.

Additionally, latent means analysis or differential item functioning might both reveal important information at both the test and item level regarding group differences.

Conclusion

The Clifton Strengths Finder instrument is widely used (the authors' note that over 1 million people have responded to the instrument as of April 2004). Despite wide-spread use of the instrument the technical manual provided by the company leaves many questions regarding the instrument's technical adequacy. There are several areas where the test maker needs additional evidence to support the knowledgeable use of this instrument. First, the standardization and norming procedures need to be presented with actual evidence from the process. While a systematic and scientific process seems to have been used to produce the instrument the authors provide no evidence to support decisions made throughout the process. Item analysis procedures and results could be used to support this particularly important aspect of the instrument. As well, more detailed information regarding the actual demographics of the standardization samples used to develop the standard score distributions currently used is needed. Second, the construct validity of the instrument needs to be explored with more complex and appropriate statistical procedures. The item-total correlation data provided is a weak form of evidence to suggest the distinctiveness of the 34 theme areas measured by the instrument. A confirmatory factor analysis needs to be performed on the instrument to provide an appropriate and adequate source of evidence to support the hypothesized structure of personal talents. Additionally, the issue of discriminant validity should be more closely addressed to establish that this test is not measuring the same constructs as other personality instruments that claim to measure personal talents. The construct validity evidence presented regards convergent, rather than discriminant, validity with the big-five personality trait model from the field of general psychology. Additionally, the relationships noted do not provide substantial evidence that this test measures unique constructs. Last, the issue of invariance needs careful attention through appropriate statistical methods. Though the authors present numerous and large tables showing the weighted means and standard deviations for various groups this does not provide evidence to support that the instrument functions the same across different audiences. The issue of invariance needs to be addressed with both external and internal methods to produce statistical evidence to support the authors' claim that the results do not vary across groups. In conclusion, the instrument presents a unique and potentially valuable measure of a construct that may be valuable to personal development for students and workers. However, more evidence needs to be produced regarding the technical adequacy of the instrument to support continued wide-spread use.

References

- Anastasi, A., & Urbina, S.J. (1997). *Psychological Testing* (7th Ed.). Upper Saddle River, NJ: Prentice Hall.
- Buckingham, M. & Clifton, D.O. (2000). *Now discover your strengths*. New York: Free Press.
- Clifton, D.O., & Nelson, P. (1992). *Soar with your strengths*. New York: Delacorte Press.
- Harter, J.K., & Hodges, T.D. (2003). *Construct validity study: StrengthsFinder and the Five Factor Model* [technical report]. Omaha, NE: The Gallup Organization.
- Hodges, T.D., & Clifton, D.O. (2004). Strengths-based development in practice. In A. Linley & S. Joseph (Eds.), *Handbook of positive psychology in practice*. Hoboken, New Jersey: John Wiley and Sons, Inc.
- Lopez, S., Hodges, T., & Harter, J. (2005). *The Clifton StrengthsFinder Technical Report: Development and Validation* [technical report]. Princeton, NJ: The Gallup Organization.
- Moore, D.S., & McCabe, G.P. (2006). *Introduction to the Practice of Statistics* (5th Ed.). New York: Freeman Press.
- Plake, B. (1999). *An investigation of ipsativity and multicollinearity properties of the StrengthsFinder Instrument* [technical report]. Lincoln, NE: The Gallup Organization.