Journal of Nursing Measurement, Volume 25, Number 3, 2017

An Evaluation of the Psychometric Properties of the Behavior Change Strategies for Healthy Eating Scale

Kenneth D. Royal, PhD, MSEd

North Carolina State University

Rachael A. Royal, BS

University of North Carolina at Chapel Hill

Background and Purpose: The Behavior Change Strategies for Healthy Eating Scale (BCSHES) has potentially broad implications for use by researchers, practitioners, and educators across various medical and allied health professions. To date, however, the psychometric properties of the BCSHES have not been rigorously evaluated, potentially limiting its use. The purpose of this study was to thoroughly evaluate the psychometric properties of the BCSHES. Methods: The Rasch Rating Scale Model was used to analyze BCSHES data on a sample of 191 graduate/professional program students. Results: Messick's framework for construct validity was used to discern validity evidence, which noted the BCSHES possessed very strong psychometric properties and is capable of yielding valid and reliable scores. Conclusions: Use of the scale is encouraged, where appropriate.

Keywords: eating behavior; health promotion; reliability and validity; instrumentation; psychometrics; education

In 2010, Norman and colleagues published a paper in which a series of scales were presented. One scale in particular that has broad implications for practitioners working in various medical and allied health professions was the Behavior Change Strategies for Healthy Eating Scale (BCSHES). The instrument was reported to assess the thoughts, feelings, and behaviors individuals may experience when attempting to eat healthy. Unfortunately, the BCSHES has received little attention in the research literature to date, presumably because validity evidence surrounding the measure is not well-documented. In fact, to our knowledge no validation studies exist on the BCSHES. Thus, the purpose of this work was to fill this void in the research literature by thoroughly evaluating the psychometric properties of the BCSHES by way of a powerful item response theory (IRT) measurement model.

METHODS

Participants

All students enrolled in a doctor of veterinary medicine (DVM) program at a large veterinary medical school in the United States were invited to participate in a study that

http://dx.doi.org/10.1891/1061-3749.25.3.411

TABLE 1. Demographic Characteristics of the Sample

	N	%
Year		
1st year	72	37.7
2nd year	32	16.8
3rd year	39	20.4
4th year	48	25.1
Sex		
Male	30	15.7
Female	161	84.3
Race/ethnicity		
White	163	85.3
Other	28	14.7

assessed students' strategies for making healthy eating choices. The population of students consisted of 393 students across all 4 years of the DVM program. In total, 191 students completed the survey, resulting in 48.6% response rate. The median age of students was 25 years. A summary of demographic characteristics are presented in Table 1.

Instrumentation

The BCSHES was developed by Norman and colleagues (2010) and was inspired by a scale created by Saelens and colleagues (2000) and based on the transtheoretical model for process change (Prochaska, Redding, & Evers, 1995). The BCSHES consists of 15 items that assess the thoughts, feelings, and behaviors individuals may experience when attempting to eat healthy. The BCSHES used a 5-point rating scale with the categories 1 (*never*), 2 (*almost never*), 3 (*sometimes*), 4 (*often*), and 5 (*many times*). Each item on the scale represents a unique strategy for healthy eating; therefore, the instrument does not contain any subscales. Recommendations for scoring the instrument include simply summing each of the items to produce a total score. Higher scores are recommended to be interpreted as indicative of greater use of change strategies. Previous research using the BCSHES included Norman and colleagues' original study in which the instrument was administered to 49 college students in one study, and 842 individuals enrolled in an online weight-loss intervention study. Results from both studies presented some limited validity evidence given the sample frames assessed.

Validation Framework

The psychometric literature has thoroughly discussed the limitations associated with traditional statistical procedures for validation studies. Royal (2010) lists six specific problems associated with traditional statistical approaches, including (a) the erroneous treatment of ordinal-level raw scores as interval-level measures, (b) the assumption that all items are equally important, (c) the assumption that error is equally distributed across all measures,

(d) results are sample-specific and sample-dependent, (e) the assumption that data must be normally distributed to be useful, and (f) the treatment of missing data. Several measurement scholars have noted the Rasch family of models is the "gold standard" for validation studies because they overcome each of the aforementioned limitations of traditional statistical models and are the only models that guarantee invariance within a standard error when data sufficiently fit the model's expectations (Royal, 2010; Salzberger, 2002; Wright, 1997).

Although thorough discussion of Rasch models is beyond the scope of this article, a quick overview is necessary. Rasch models are probabilistic, unidimensional measurement models stemming from the larger IRT family of models. Unlike traditional statistical models where a model is fitted to describe data, Rasch models are static models for which data must accord to its expectations. Rasch models assert that an individual with a greater amount of a latent trait will always have a greater probability of endorsing an item than someone with a lesser amount of the latent trait. Likewise, a more difficult item to endorse will always have a lower probability of being endorsed than a less difficult item (Rasch, 1960). Rasch models essentially convert raw scores to interval-level measures by converting data onto a logarithmic scale. Because both the latent trait for individuals (e.g., tendency to endorse or agree with a given item) and the difficulty of the item can be placed onto a common linear continuum, the relative distance between the person and item parameters indicates the probability that an individual is likely to endorse (e.g., agree with, select) the item. Rasch models require stringent quality control procedures, which are particularly useful for thoroughly evaluating the psychometric properties of an instrument. Readers are encouraged to see Bond and Fox (2015) and Engelhard (2013) for a more thorough overview and specifics.

The Rasch Rating Scale Model (RRSM) is a Rasch model designed for the analysis of survey data in which the rating scale categories are static (Andrich, 1978). The following equation presents the RRSM's formulation, where the probability of a person n responding in category x to item i, is given by:

$$P_{xni} = \frac{\exp \sum_{j=0}^{x} [\beta_n - (\delta_i + \tau_j)]}{\sum_{k=0}^{m} \exp \sum_{j=0}^{k} [\beta_n - (\delta_i + \tau_j)]} \qquad \chi = 0, 1, \ldots, m$$

where $\tau_0 = 0$ so that $\exp \sum_{j=0}^{0} [\beta_n - (\delta_i + \tau_j)] = 1$, β_n is the person's position on the variable, δ_i is the scale value (difficulty to endorse) estimated for each item i, and $\tau_1, \tau_2, \ldots, \tau_m$ are the m response thresholds estimated for the $m + 1$ rating categories. Data analyses were performed using Winsteps (Version 3.81) measurement software. Winsteps uses joint maximum likelihood estimation procedures (Wright & Masters, 1982) for estimating parameters.

RESULTS

Table 2 presents the items appearing on the BCSHES and provides a statistical summary for each of the 15 items. Readers should note the statistical summary is for illustrative purposes only because analyzing mean scores for the rating scale categories (1 = *never*; 5 = *many times*) may not be informative or appropriate.

TABLE 2. Results of Traditional Statistical Analysis

	M	*SD*
1. I look for information about eating healthy foods.	3.39	1.03
2. I keep track of what I eat.	2.95	1.14
3. I find ways to get around the things that get in the way of eating healthy foods.	2.89	0.91
4. I think about how my surroundings affect the foods I eat (surroundings are things like fast food restaurants, vending machines, and prepackaged foods in the store).	3.21	1.07
5. I put reminders around my home to eat healthy foods.	1.65	0.95
6. I reward myself for eating healthy foods.	2.31	1.16
7. I do things to make eating healthy foods more enjoyable.	3.08	1.14
8. I think about the benefits I will get from eating healthy foods.	3.89	1.03
9. I try to think more about the benefits of eating healthy foods and less about the hassles of eating healthy foods.	3.51	1.14
10. I say positive things to myself about eating healthy foods.	3.05	1.29
11. When I get off track from my healthy eating goals, I tell myself I can start again and get right back on track.	3.40	1.14
12. I have a friend or family member who encourages me to eat healthy foods.	3.21	1.20
13. I try different kinds of healthy foods so that I have more choices.	3.58	1.14
14. I set goals to eat healthy foods.	3.22	1.26
15. I make back-up plans to be sure I eat healthy foods.	2.53	1.15

Psychometric Properties of the Behavior Change Strategies for Healthy Eating Scale

Dimensionality. In the social and behavioral sciences, it is well-understood that no data set is perfectly unidimensional. However, data may still be sufficiently unidimensional for constructing high-quality measures. To investigate the dimensionality of the BCSHES data, a Rasch-based principal components analysis (PCA) of standardized residual correlations as described by Linacre (2016a) was performed. A total of 54.9% of the primary (Rasch) dimension was explained, with 22.9% being attributed to the items. A hint of multidimensionality was discernible as the largest secondary dimension explained 6.7% of the variance and had an eigenvalue of 2.23 (about 2 items). An eigenvalue less than two units (e.g., items) in magnitude typically is indicative of measurement noise. Thus, results indicate there was evidence of some multidimensionality, but given the secondary dimension was limited to 2 of the 15 total items, the primary (Rasch) dimension was sufficiently unidimensional for all practical purposes.

Reliability. Reliability was evaluated using both traditional statistical procedures and Rasch model measures. The Cronbach's alpha reliability coefficient was .914 for the collective 15-item scale. Rasch-based reliability estimates were .89 for "real," and

TABLE 3. Rating Scale Diagnostics

Rating Category	*n*	%	INFIT Mean Square	OUTFIT Mean Square	Structure Calibration	Category Measure
1 = Never	405	14	1.27	1.36	None	−2.85
2 = Almost never	530	19	0.90	0.89	−1.48	−1.29
3 = Sometimes	796	28	0.92	0.91	−0.77	−0.10
4 = Often	749	26	0.92	0.96	0.41	1.25
5 = Many times	379	13	0.96	0.98	1.84	3.09

.91 for "modeled," indicating the true measure of reliability is somewhere in between. All three measures of reliability indicate highly reproducible measures (Royal & Hecker, 2016). An additional Rasch-based statistic is the separation index, which refers to the number of statistically distinguishable levels (also known as *strata*) are discernible within the data. The separation statistic was 3.27, indicating approximately four strata were discernible.

Rating Scale Effectiveness. Rating scale effectiveness was evaluated to determine the extent to which participants made use of each rating scale category and the degree to which they were able to appropriately interpret what was meant by each rating scale option (Table 3). Results illustrate that participants made full use of the rating scale, calibration measures increased in a stepwise manner as anticipated (Linacre, 2002), and fit statistics were all within the acceptable range (0.60–1.40; Wright & Linacre, 1994). Collective evidence suggests the rating scale effectively captured participants' responses and functioned as intended.

Item and Person Measure Quality

Several statistical indicators were used to evaluate person and item measure quality (Table 4). First, item quality indicators were examined. Results indicate logit calibration measures ranged between −1.24 and 2.18 with a mean standard error of .09 (SD = .01). These measures indicate good variation with small standard errors to ensure statistically stable results. All fit statistics (INFIT and OUTFIT mean square values) were within acceptable range (0.60–1.40) with exception to Item 12 (I have a friend or family member who encourages me to eat healthy foods), which provided somewhat inflated statistics. Nonetheless, fit statistics for this item fell below the 2.00 criteria for "unproductive measurement" (Wright & Linacre, 1994); thus, the item does not necessarily warrant evidence for exclusion on future administrations of the instrument but should be monitored to determine how the item functions across samples. Finally, high (positive) point-measure correlations indicate each of the items possess strong discriminatory properties.

Person measures were evaluated using similar procedures and criteria. Person measures ranged from −3.30 to 4.11 with a mean standard error of .32 (SD = .06). These measures indicate good variation with relatively small standard errors. Only 13 (6.8%) participants from the sample of 191 yield fit statistics (either INFIT or OUTFIT mean square) exceeding 2.0, the criteria for unproductive measurement.

Global fit statistics were evaluated to discern how well the persons and items functioned overall. Mean square fit statistics for persons were 1.04 (INFIT) and 1.03 (OUTFIT) and

TABLE 4. Item Quality Indicators

Item	Difficulty Measure	Error	INFIT Mean Square	OUTFIT Mean Square	Point Measure Correlation
1	−0.47	.09	0.74	0.76	.68
2	0.14	.08	1.00	1.03	.64
3	0.23	.08	0.74	0.83	.64
4	−0.22	.09	1.14	1.20	.55
5	2.18	.11	1.37	1.15	.56
6	1.04	.09	1.24	1.25	.60
7	−0.04	.08	0.78	0.78	.72
8	−1.24	.09	0.77	0.69	.71
9	−0.65	.09	0.79	0.75	.73
10	0.01	.08	1.19	1.25	.66
11	−0.49	.09	0.90	0.87	.69
12	−0.22	.09	1.72	2.06	.43
13	−0.74	.09	1.07	1.08	.64
14	−0.23	.09	0.93	0.90	.74
15	0.71	.09	0.80	0.76	.73

1.01 (INFIT) and 1.02 (OUTFIT) for items. In addition, only six data points were missing which resulted in a 99.8% completed response vector.

A review of standardized residual correlations was performed to identify any item pairs with correlations ≥.30 because these items may exhibit local item dependence (Marais & Andrich, 2008; Smith, 2000). Items 8 (I think about the benefits I will get from eating healthy foods and 9 (I try to think more about the benefits of eating healthy foods and less about the hassles of eating healthy foods) exhibited a correlation of .43, making these items potentially dependent. However, it should be noted that these items appeared next to one another on the scale, and extant research has indicated that items with close proximity have a tendency to generate higher measures of local dependence and thus likely to generate higher instances of false-positive detections (Royal, 2015, 2016).

A differential item functioning (DIF) analysis was performed to assess if the construct remained invariant based on one's sex. The iterative-logit (Rasch-Welch) method presented in Linacre (2016b) was performed. Because multiple comparisons were made across 15 items, a Bonferroni correction was necessary to control for compounding family-wise error. The Bonferroni correction reduced the p value from .05 to .0033 as the criteria for detecting statistically significant differences. No statistically significant differences were discernible for sex with $p \leq .0033$.

Construct Hierarchy. The construct hierarchy, also known as an *item map* or *Wright Map*, is presented in Figure 1. The figure illustrates the psychometric ruler onto which both participants and items were placed and visualizes how the two parameters interact. The left side of the map illustrates participants, and the right side illustrates the items. Persons appearing at the top of the map indicate individuals with greater amounts of

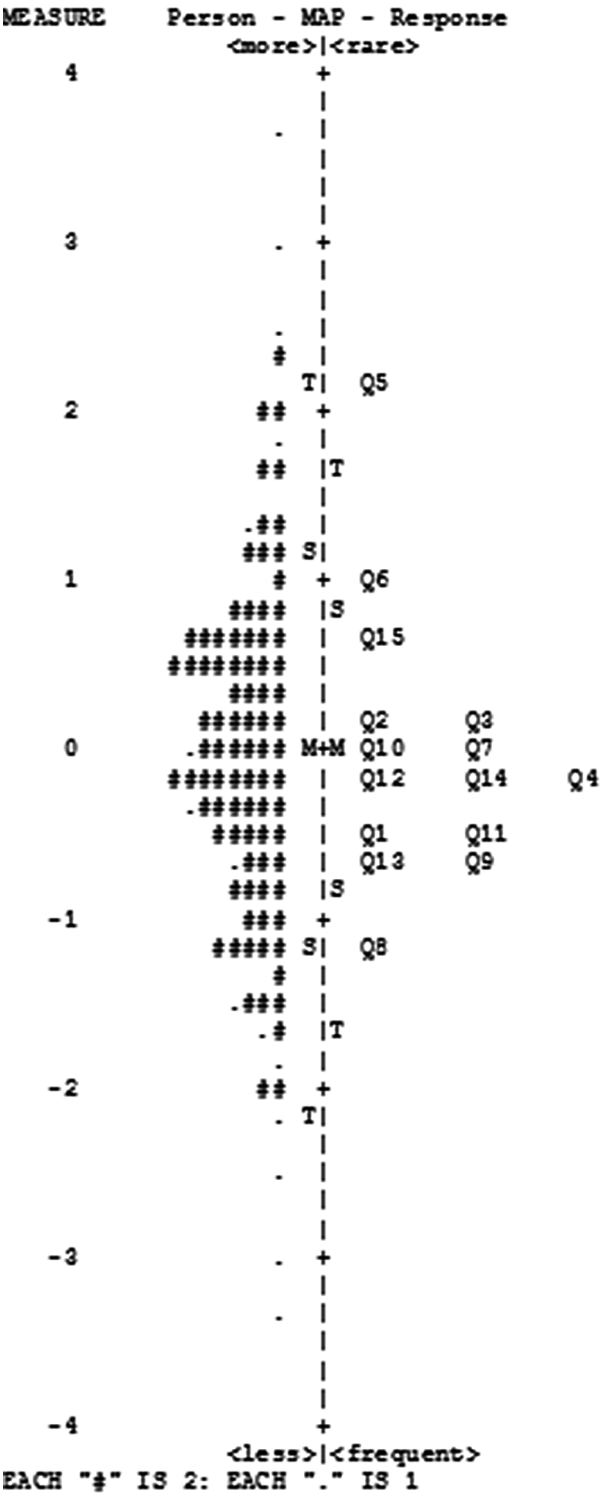

Figure 1. Construct Hierarchy for Items Appearing on the BCSHES.

the latent trait, which in this instance refers to persons who use more change strategies. Likewise, persons at the bottom of the map indicate individuals who use fewer change strategies for healthy eating. Similarly, items appearing at the top of the map indicate the most difficult to endorse, which in this case refers to the least selected change strategy. Likewise, items appearing at the bottom of the map indicate frequently selected change strategies.

DISCUSSION

Psychometric Properties of the Behavior Change Strategies for Healthy Eating Scale

Several frameworks exist for discussing and presenting validity evidence. For this study, Messick's framework for construct validity (Messick, 1989, 1995) was used to inventory validity evidence. Messick's framework is based on the notion that validity is an integrated concept and all validity evidence is essentially construct validity, which is comprised of various "aspects." The six aspects of Messick's framework include substantive, content, generalizability, structural, external, and consequential.

Results of the Rasch-based PCA of standardized residual correlations indicated the data primarily were unidimensional as the authors of the scale insinuate. This evidence speaks to the substantive aspect of validity. Various statistical indicators demonstrated highly reproducible measures, with speaks to the generalizability aspect of validity. Statistical measures also confirmed the rating scale functioned effectively which speaks to the structural aspect of validity. Various person and item measure indicators also were psychometrically sound, which speaks to the content aspect of validity. Results from the DIF analyses indicated invariance across the sex variable, which speaks to the systematic and generalizability aspects of validity. Given the limited use of the scale at present, scores have not been correlated with other studies with similar populations. Therefore, we present no evidence that speaks to the external aspect of validity. We also are unaware of any consequences (positive or negative) that may result from the use of the measures, so we cannot speak to the consequential aspect of validity (Royal & Puffer, 2014). Collectively, there is a tremendous amount of validity evidence that suggests the BCSHES is a psychometrically sound instrument capable of producing valid and reliable measures.

Future Research

Although a plethora of validity evidence was discernible there may be some ways in which the BCSHES can be improved. As noted previously, Item 12 yielded some inflated fit statistics. It is unknown if this is an artifact of the instrument's functioning relative to this sample or indicative of a "noisy" item across various adult populations. Future research should continue to monitor this item's functioning. In addition, Items 8 and 9 had a paired standardized residual correlation of .43, indicating participants' responses to one item might impact their response on the other item. However, it should be noted that the items appeared next to one another on the scale, so it is possible that the high correlation simply is a false-positive detection. Future administrations of the BCSHES should separate these two items to ensure local item dependency (also known as *statistical dependency*) is mitigated. Of course, future research also should evaluate the local item dependency once item ordering is changed.

CONCLUSION

The purpose of this study was to evaluate the psychometric properties of the BCSHES scale, a relatively new measure for assessing behavioral change strategies for healthy eating. Results indicate the BCSHES possesses very strong psychometric properties and is capable of yielding valid and reliable scores. The instrument should be of use to anyone interested in measuring behavior change strategies relating to healthy eating, and we encourage others to consider adopting the scale.

REFERENCES

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Abingdon, United Kingdom: Routledge.

Engelhard, G. (2013). *Invariant measurement: Using Rasch model in the social, behavioral and health sciences*. Abingdon, United Kingdom: Routledge.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement*, *3*(1), 85–106.

Linacre, J. M. (2016a). *Correlations: Point-biserial, point-measure, residual*. Retrieved from http://www.winsteps.com/winman/correlations.htm

Linacre, J. M. (2016b). *Differential item functioning DIF pairwise*. Retrieved from http://www.winsteps.com/winman/table30_1.htm

Marais, I., & Andrich, D. (2008). Effects of varying magnitude and patterns of response dependence in the unidimensional Rasch model. *Journal of Applied Measurement*, *9*(2), 105–124.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 13–103). New York, NY: Macmillan.

Messick, S. (1995). Standards of validity and the validity of standards in performance assessment. *Educational Measurement: Issues and Practices*, *14*(4), 5–8.

Norman, G. J., Carlson, J. A., Sallis, J. F., Wagner, N., Calfas, K. J., & Patrick, K. (2010). Reliability and validity of brief psychosocial measures related to dietary behaviors. *The International Journal of Behavioral Nutrition and Physical Activity*, *7*, 56.

Prochaska, J. O., Redding, C. A., & Evers, K. E. (1995). The transtheoretical model and stages of change. In K. Glanz, F. M. Lewis, & B. K. Rimer (Eds.) *Health behavior and health education: Theory research and practice* (2nd ed., pp. 60–84). San Francisco, CA: Jossey-Bass.

Rasch, G. (1960). *Probabilistic models for some intelligence and achievement tests*. Copenhagen, Denmark: Danish Institute for Educational Research.

Royal, K. D. (2010). Making meaningful measurement in survey research: A demonstration of the utility of the Rasch model. *IR Applications*, *28*, 1–16.

Royal, K. D. (2015). Does item sequence order impact local dependence in surveys? *Rasch Measurement Transactions*, *29*(1), 1507–1508.

Royal, K. D. (2016). The impact of item sequence order on local item dependence: An item response theory perspective. *Survey Practice*, *9*(4), 1–7. Retrieved from http://www.surveypractice.org/index.php/SurveyPractice/article/view/344

Royal, K. D., & Hecker, K. G. (2016). Understanding reliability: A review for veterinary educators. *Journal of Veterinary Medical Education*, *43*(3), 1–4.

Royal, K. D., & Puffer, J. C. (2014). The consequential validity of ABFM examinations. *Journal of the American Board of Family Medicine*, *27*(3), 430–431.

Saelens, B. E., Gehrman, C. A., Sallis, J. F., Calfas, K. J., Sarkin, J. A., & Caparosa, S. (2000). Use of self-management strategies in a 2-year cognitive-behavioral intervention to promote physical activity. *Behavior Therapy*, *31*(2), 365–379.

Salzberger, T. (2002). The illusion of measurement: Rasch versus 2-PL. *Rasch Measurement Transactions*, *16*(2), 882.

Smith, R. M. (2000). Fit analysis in latent trait measurement models. *Journal of Applied Measurement*, *1*(2), 199–218.

Wright, B. D. (1997). Fundamental measurement. *Rasch Measurement Transactions*, *11*(2), 558.
Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*, 370.
Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis: Rasch measurement*. Chicago, IL: MESA Press.

Correspondence regarding this article should be directed to Kenneth D. Royal, PhD, MSEd, North Carolina State University, College of Veterinary Medicine, 1060 William Moore Dr., Raleigh, NC 27607. E-mail: kdroyal2@ncsu.edu