Research in Psychology and Behavioral Sciences, 2019, Vol. 7, No. 1, 23-33
Available online at http://pubs.sciepub.com/rpbs/7/1/4
Published by Science and Education Publishing
DOI:10.12691/rpbs-7-1-4

Modern Psychometric Analysis of the Muscle Strengthening Activity Scale (MSAS) Using Item Response Theory

Peter D. Hart*

Health Promotion Research, Havre, MT 59501
*Corresponding author: pdhart@outlook.com

Received November 10, 2019; Revised December 12, 2019; Accepted December 23, 2019

Abstract **Background**: With the growing need to promote muscle strengthening activity (MSA) for improved health-related quality of life (HRQOL) comes the growing need for proper measurement of MSA behavior. The purpose of this study was to examine test and item functioning of the MSA scale (MSAS) using item response theory (IRT). **Methods**: The current research fit data from a sample of N = 400 respondents to two different graded response models (GRMs), a three-item muscular *strength* scale and a three-item muscular *endurance* scale. For each GRM, model-data fit was examined and IRT assumptions assessed. **Results**: An unconstrained GRM was found to fit the data better than the constrained model ($\Delta\ G^2_{\text{Strength}} = 10.3$, $p = .006$, RMSEA = .043 & $\Delta\ G^2_{\text{Endurance}} = 7.0$, $p = .031$, RMSEA = .021). GRM boundary location parameters covered the latent trait scale well for both *strength* (*b*s: -4.26 to 2.58) and *endurance* (*b*s: -3.86 to 1.79) scales with each item showing adequate fit to the data (all RMSEAs < .05). Test information was approximately evenly distributed around a theta of zero with summed information from theta ranges ±4 of 92.8 (*strength*) and 93.5% (*endurance*). Only 2.3 and 1.5% of persons misfit the *strength* and *endurance* GRMs, respectively. **Conclusion**: The MSAS has shown to be a valid tool for measuring MSA behavior in adults using modern psychometric theory.

Keywords: *Muscle strengthening activity (MSA), Item response theory (IRT), Graded response model (GRM), test information*

Cite This Article: Peter D. Hart, "Modern Psychometric Analysis of the Muscle Strengthening Activity Scale (MSAS) Using Item Response Theory." *Research in Psychology and Behavioral Sciences*, vol. 7, no. 1 (2019): 23-33. doi: 10.12691/rpbs-7-1-4.

1. Introduction

The current physical activity (PA) recommendations make it clear – adults should participate regularly in muscle strengthening activity (MSA) in order to gain the many different associated health benefits, including improved health-related quality of life (HRQOL) [1,2,3]. And like any health behavior, the ability to measure MSA would have vast implications for researchers, practitioners, clinicians, and educators. The muscle strengthening activity scale (MSAS) is a self-report assessment tool designed to measure MSA behavior in adults and has exhibited promising psychometric properties [4,5]. However, thus far, evidence supporting the MSAS has been limited to classical test theory (CTT) methods. CTT is the most prevalent and conventional model used by researchers to validate behavioral scales [6]. The focus of CTT is placed on the unweighted sum of responses across items of an instrument, otherwise known as the observed score (X). The CTT model states that an individual's observed score is a function of their true score (T) and random error (E) [7]. Where true score is the mean of an infinite number of independent observed scores from an infinite number of independent test administrations. Therefore, any given single observed score will differ from its mean due to error. Symbolically, the true score model for individual i is

$$X_i = T_i + E_i.$$

There are many flaws, however, in CTT that motivate researchers to search for other means of assessing the psychometric properties of a scale [8]. Firstly, CTT is not a testable model and provides no clear methods for assessing model-data fit. Secondly, under CTT, an observed score is influenced by the characteristics (e.g., difficulty) of the test (i.e., an easier test will result in higher observed and true scores). Lastly, conventional reliability coefficients are affected by characteristics of the individuals (i.e., observed scores with more variability will mathematically inflate the reliability of a test). This can be seen by viewing the symbolic form of reliability under CTT:

$$\rho_{XX'} = \frac{\sigma_X^2 - \sigma_E^2}{\sigma_X^2}.$$

Where true score variance is estimated from the difference between observed score and error variances (numerator).

A more modern approach to scale development and validation, which can complement CTT-based research, is item response theory (IRT). IRT provides a system of mathematical equations which can model the relationship between latent traits (e.g., ability) and observed responses to items [9]. IRT models then can assess the functioning of each item to determine how well they perform in measuring the trait of interest. IRT can also provide a measure of reliability similar to CTT but with an added advantage of measuring how that precision varies across the latent trait [10]. Another large benefit of IRT is that item difficulty is estimated on an interval scale and the same scale as person ability, called theta (θ) [11]. Finally, IRT has strengths where CTT has limitations. That is, with IRT: 1) models can be tested for appropriate fit to the data, 2) item parameters are invariant to changes in persons (i.e., regardless of sample population), and 3) person parameters are invariant to changes in the test (i.e., easy vs. difficult tests) [12].

In summary, the ability to properly measure MSA is of increasing interest to researchers and related professionals concerned with the associations between PA and health outcomes in adults. IRT is a modern psychometric approach to validating self-report behavioral scales by examining how well each item in the instrument functions. Therefore, the purpose of this study was to employ IRT to examine the functioning of the MSAS. Specifically, the graded response model (GRM) was used to examine the appropriateness of each item of the MSAS in measuring MSA behavior in adults.

2. Materials & Methods

2.1. Study and Scale Development Procedures

The development procedures related to the MSAS have been explained in detail elsewhere [4,5]. Briefly, a total of N=400 adults who indicated participating in regular MSA provided responses to the MSAS. After an item analysis, the initial version of the MSAS resulted in a final seven-item scale measuring three distinct MSA constructs: a three-item muscular *strength* construct, a three-item muscular *endurance* construct, and a single-item *body* weight exercise construct. The final version of the MSAS is enclosed in the appendix. Item stems for the three MSA scales consist of personalized statements regarding muscular strength training behavior, muscular endurance training behavior, and body weight exercise training behavior. For example, "I often exercise my muscles with heavy weight that I can lift 1 to 8 times". Each response scale contains the same five-category options ranging from "Never true" to "Always true". Two additional items are included in the MSAS that ask participants about their frequency and duration of MSA *participation*. These *participation* items are included to quantify amounts of MSA performed but are not evaluated in this study. Directions are given at the bottom of the MSAS to obtain *strength*, *endurance*, and *body* attribute scores as well as an MSA *participation* score.

2.2. Graded Response Model (GRM)

There are a number of different IRT models available for polytomous response items. Such options include the Rasch rating scale model (RSM) [13], Rasch partial credit model (PCM) [14], generalized versions of RSM [15] and PCM [16], and the nominal response model (NRM) [17]. This study, however, used the graded response model (GRM) [18,19]. The GRM is a generalization of the two-parameter logistic item response model (2PLM) for polytomous response items. The GRM was used over other IRT models because 1) MSAS has the same ordinal-level response options across its items, 2) both item difficulty as well as item discrimination parameters are estimated in the GRM, and 3) GRM allows the researcher to constrain (set equal) item discrimination parameters and perform a nested model test to examine the statistical usefulness of freely estimating the parameters - hence providing justification (or lack of) for separate item discrimination parameter estimates. As mentioned above, item discrimination was of interest in this study because of its ability to identify how strongly each item is associated with the MSAS latent traits [20]. Additionally, GRM item difficulty values relate to the level of the latent trait required where the respondent has a 50% chance of endorsing the current or higher response categories versus all lower categories [21]. Evaluating the extent to which location parameters within and between items cover the latent trait range is essential in determining how well the scale items function. In combination, both item parameters were sought in this study to properly assess MSAS scale functioning. The cumulative boundary response function (BRF) of the GRM is defined as

$$P_{Xj}^{*}(\theta) = \frac{e^{\alpha_j(\theta-\delta_{Xj})}}{1+e^{\alpha_j(\theta-\delta_{Xj})}}$$

Where θ (theta) is the latent trait, α_j is the discrimination parameter for item j, δ_{Xj} is the category boundary location (difficulty) for category Xj with k categories and k -1 category boundary locations for each item [22]. Plotting each cumulative BRF with respect to theta can be defined as an item's boundary characteristic curve (BCC) graph. In brief, the above equation specifies the probability of obtaining a category score of Xj or higher on item j. Therefore, to compute the probability of obtaining a particular category score of Xj or in a particular category k (P_k), the difference between cumulative probabilities for adjacent categories must be found. This is specifically shown as

$$P_k = P_k^{*} - P_{k+1}^{*}.$$

Where P^{*}_k and P^{*}_k are from the BRF above. Plotting P_k across theta for each category can be defined as an item's category characteristic curve (CCC) graph.

2.3. Statistical Analyses

The following procedures were the same and performed separately for the *strength* and *endurance* scales of the MSAS. The IRT analyses were divided into three categories: descriptive statistics, scale calibration and

assessment, and IRT assumption checking. For the descriptive statistics, item category response rates were reported, CTT reliability coefficients (i.e., Cronbach alpha) computed, and cumulative frequency histograms (i.e., Pareto charts) constructed for scale sum scores. For scale calibration and assessment, a series of six steps were followed. First, two competing GRMs were fit to the data, one with item discrimination parameters constrained (set equal). The GRMs were of homogenous class (item discrimination the same across category options) with parameters set to the logistic metric (scaling factor of 1.0). For both models Akaike's information criterion (AIC), sample size adjusted AIC, Bayesian information criterion (BIC), sample size adjusted BIC (SABIC), and root mean square error of approximation (RMSEA) were computed as measures of fit. Additionally, a likelihood ratio test between the two IRT models was conducted to determine if the estimation of extra parameters is statistically warranted. Models were fit using marginal maximum likelihood estimation (MMLE) with the Gauss-Hermite quadrature rule [23]. Second, parameter estimates were reported for the better fitting GRM, including item discrimination, category boundary location (difficulty), person ability, and their standard errors. Third, an item characteristic curve (ICC) graph was generated for each item to examine the probability of selecting an item category across the latent trait scale. When scale data are polytomous, an ICC technically becomes a BCC and hereafter referred to as such. Each BCC was evaluated to ensure item responses were in accord with the latent trait (theta). Fourth, an item response category characteristic curve (CCC) graph was generated for each item to examine the latent trait values at which the probability of selecting an item category or higher is 50%. Each CCC was evaluated for proper item functioning across the latent trait (theta). Fifth, test information (*I*) was computed across specific areas of the MSAS latent trait. Information tells us how certain we are about a person's location (theta) on the latent trait continuum and it has a reciprocal relationship with the standard error of estimate (SEE). More specifically, test information provides a way to quantify how well a scale discriminates across the latent trait. Based on the test information function provided by the fit GRMs, marginal reliability (MR) was computed and information graphs constructed. Test information was inspected to ensure consistent and wide coverage across the latent trait continuum. Sixth, summary statistics were computed and graphs constructed for MSAS construct person (theta) values.

For IRT assumption checking, three main assumptions were assessed: local independence, unidimensionality, and model fit. Local independence refers to a characteristic where responses to an item are independent of responses to any other item after controlling for person location (theta). This assumption was assessed using the local dependence (LD) chi-square statistic, standardized residuals, and Cramer's V coefficients [24]. Standardized residuals greater than the absolute value of 10.0 and Cramer's V of 0.40 or larger were considered problematic [25,26]. Unidimensionality refers to the notion that responses to items are solely a function of a single latent trait. This assumption was assessed using Velicer's minimum average partial (MAP) where a series of principal components are partialed out of the item correlation matrix to yield a series of partial correlation matrices [27,28]. The step that results in the lowest average squared (or 4th power) partial correlation determines the number of components to retain. Finally, model fit was assessed by examining model and item RMSEA statistics where values ≤ .05 indicate adequate fit [29]. Additionally, a standardized statistic (Zh) was computed for each person where values greater than the absolute value of 2.0 were considered misfit to the GRM. Negative values of Zh reflect person responses that are inconsistent (unlikely) given the GRM and positive values of Zh reflect person responses that are more consistent than the GRM predicts [30]. The percentage of persons misfitting the model, using the Zh statistic, was used as a measure of model fit. All IRT analyses were conducted using the R *ltm* and *mirt* packages [31,32,33,34].

3. Results

Table 1 contains item category endorsement distributions for the N=400 MSAS respondents. Each category received respondent endorsements. Although six categories saw an endorsement rate less than ten percent, this might be expected from a relatively small sample size. Additionally, reliability estimates shown ($\alpha_{Strength}$ = .62 & $\alpha_{Endurance}$ = .64) are acceptable for scales of this size. That is, using the Spearman-Brown Prophecy formula, we see that if each MSAS scale was doubled to a size of six items, the new scale reliability estimates would increase to .77 and .78 for *strength* and *endurance*, respectively. Figure 1 and Figure 2 both show the cumulative frequency distribution of the MSAS *strength* and *endurance* scale sum scores. Both graphs indicate sparse ceiling and floor sum scores from respondents. Table 2 contains model fit statistics for both the constrained and unconstrained GRMs. Although both models fit well (RMSEAs < .05), the likelihood ratio test indicates that the unconstrained GRM fits the sample data better ($\Delta\ G^2_{Strength}$ = 10.3, p = .006 & $\Delta\ G^2_{Endurance}$ = 7.0, p = .031). Therefore, hereafter, all results will summarize the unconstrained GRM.

Table 3 contains GRM item parameter estimates for the MSAS. Item 1 displays the greatest discrimination in the *strength* scale (a = 2.77) whereas item 5 displays the greatest discrimination in the *endurance* scale (a = 2.29). These findings are consistent with the item-to-theta correlation (r_{Theta}) values, which also serves as a measure of item discrimination. The GRM boundary location parameters appear to cover the latent trait well for both the *strength* (*b*s: -4.26 to 2.58) and *endurance* (*b*s: -3.86 to 1.79) scales. With each item showing adequate fit to the data (all RMSEAs < .05). Inspection of item category coverage across theta is enhanced by visual inspection of the CCCs in Figure 3. It can be seen that each item category has its own area on the theta scale where its probability of endorsement is greater than any other category, indicating adequate item functioning. Figure 3 also visually indicates the same boundary location parameters displayed in Table 3, as the intersection of each CCC. However, the visual inspection of these boundary location parameters are improved by viewing the BCCs in Figure 4. Two main characteristics are

noteworthy in Figure 4. One, the discrimination parameters are also visually noticeable. That is, each item category has the same discrimination (slope), indicative of the homogenous class of fitted models as well as the greatest slopes are again seen with item 1 and item 5 of the *strength* and *endurance* scales, respectively. Two, boundary locations (theta values where the probability of endorsing the category or higher is 0.50) are adequately spread across the latent trait scale when considering all scale items together.

Table 4 contains test information values for each MSAS scale. There are two main noteworthy comments regarding these values. One, a large percentage of the total information can be collected from theta ranges of -4 to +4 using the *strength* (92.8%) and *endurance* (93.5%) scales. Two, this total information appears roughly equal for higher MSAS trait and lower MSAS trait. The marginal reliability (MR) values capture the average information across the theta scale, with acceptable values (given only three item scales) of .72 and .69 for the *strength* and *endurance* scales, respectively. Figure 5 and Figure 6 display the test information functions as well as the standard error of estimate lines. These graphs reinforce the information from Table 4 with the added detail of showing where on the theta scale the MSAS information is greatest. That is, for both MSAS scales, measurement precision is greatest around the theta value of zero. Unlike CTT measures of reliability, IRT test information has the added advantage of indicating the measured location of the person trait where reliability is strong or possibly where improvement is needed. For example, the MSAS scale may require more difficult items if we wish to measure extreme MSA behavior with a high level of precision.

Table 1. Proportions for MSAS category responses and scale reliability measures

Labels		Never True	Rarely True	Sometimes True	Usually True	Always True	Reliability		
Categories		1	2	3	4	5	α	α_{del}	r_{SB}
Strength							.624		.768
	Item 1	.110	.168	.250	.215	.258		.421	
	Item 2	.123	.270	.355	.178	.075		.563	
	Item 3	.013	.080	.180	.358	.370		.560	
Endurance							.635		. 777
	Item 4	.058	.203	.380	.228	.133		.550	
	Item 5	.033	.130	.283	.375	.180		.460	
	Item 6	.018	.105	.333	.393	.153		.590	

Note. N=400. Category response values are relative frequencies. α is Cronbach alpha. α_{del} is alpha with row item deleted from scale. r_{SB} is Spearman-Brown Prophecy coefficient predicting reliability if scales were doubled. $r_{SB.Strength}$ = (2*0.624) / (1 + 0.624) = .768. $r_{SB.Endurance}$ = (2*0.635) / (1 + 0.635) = .777.

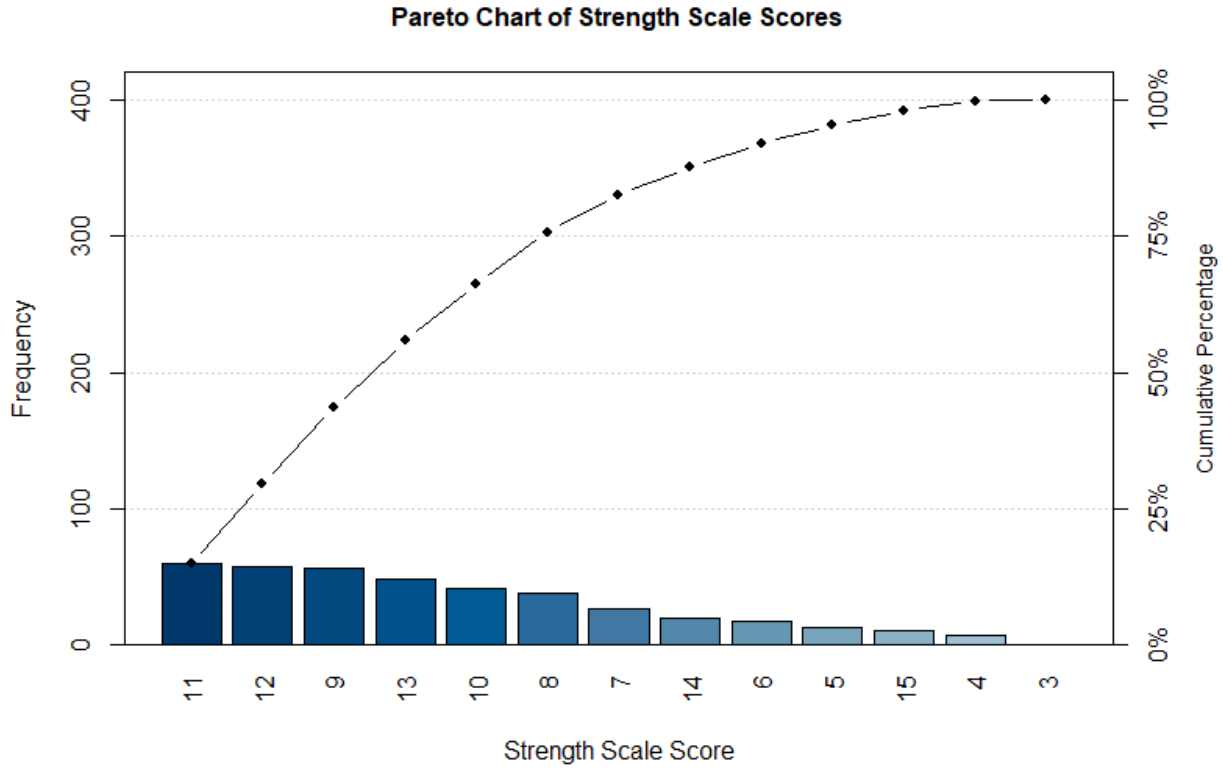

Figure 1. Pareto chart of summed MSAS strength scale scores

Table 2. Comparison of GRM with and without discrimination parameters freely estimated across MSAS items

	df	G^2	*AIC*	*AICc*	*BIC*	*SABIC*	*RMSEA*	*Δ* G^2	*Δ df*	*Δ p*
Strength										
GRM	109	188.72	3351.54	3352.79	3411.41	3363.81	.0428	10.30	2	.006
GRMC	111	199.02	3357.83	3358.78	3409.72	3368.47	.0446			
Endurance										
GRM	109	128.57	3230.58	3231.83	3290.45	3242.86	.0212	6.96	2	.031
GRMC	111	135.53	3233.54	3234.48	3285.43	3244.18	.0235			

Note. GRM is graded response model. GRMC is GRM with discrimination parameter constrained. G^2 is the likelihood ratio test statistic. RMSEA is the root mean-square error of approximation. AIC is the Akaike Information criterion. AICc is the sample size adjusted AIC. BIC is the Bayesian information criterion. SABIC is the sample size adjusted BIC. LL is the negative log-likelihood statistic. *df* is model degrees of freedom. Δ G^2 is the likelihood ratio test for nested IRT models.

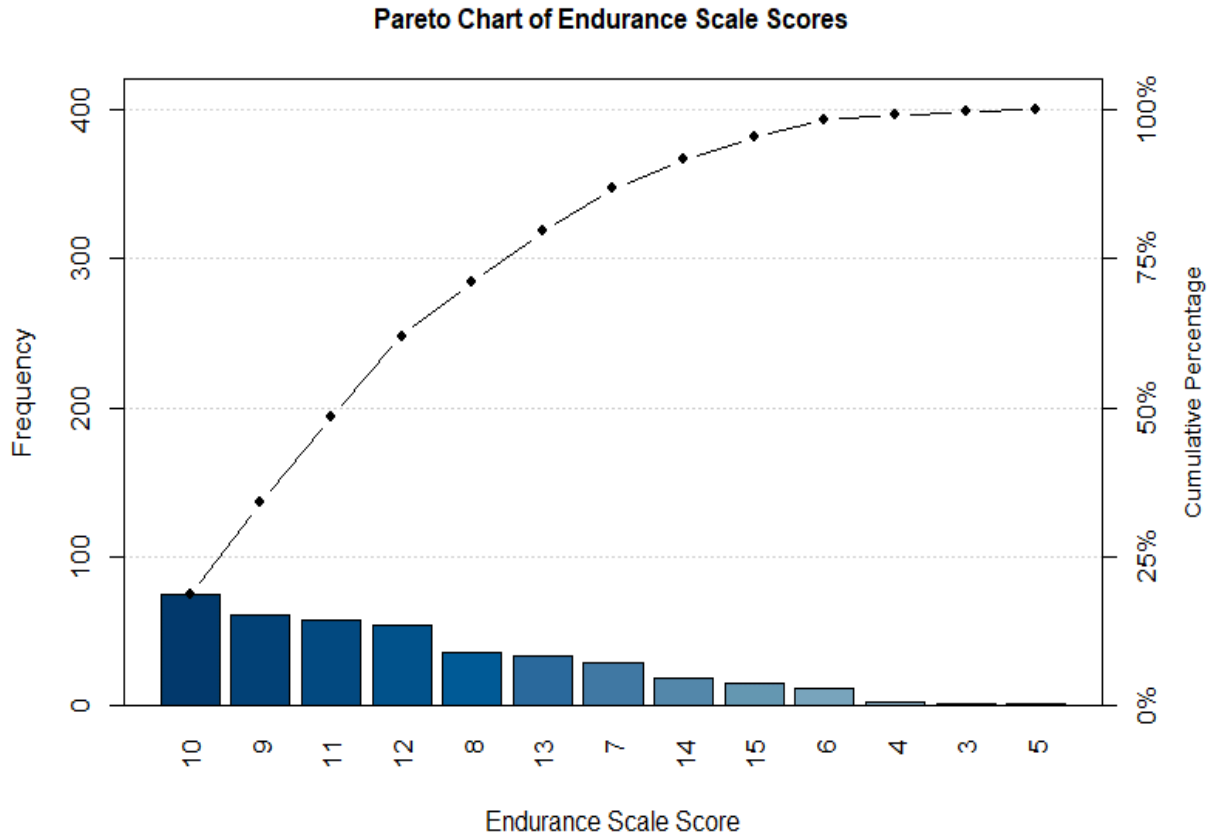

Figure 2. Pareto chart of summed MSAS endurance scale scores

Table 3. Graded response model (GRM) item parameter estimates for the MSAS

Parameter	*a*	*SEa*	*b1*	SE_{b1}	*b2*	SE_{b2}	*b3*	SE_{b3}	*b4*	SE_{b4}	r_{Theta}	RMSEA
Strength												
Item 1	2.77	0.79	-1.45	0.16	-0.69	0.39	0.08	0.20	0.76	0.52	.950	.027
Item 2	1.18	0.18	-2.04	0.26	-0.40	0.17	1.21	0.30	2.58	2.13	.588	.039
Item 3	1.18	0.19	-4.26	0.66	-2.33	0.57	-1.01	0.45	0.58	0.39	.589	.040
Endurance												
Item 4	1.34	0.20	-2.62	0.32	-1.05	0.27	0.54	0.17	1.79	0.90	.687	.019
Item 5	2.29	0.46	-2.34	0.26	-1.23	0.43	-0.19	0.29	1.14	0.69	.900	.022
Item 6	1.22	0.18	-3.86	0.55	-1.97	0.47	-0.18	0.33	1.77	0.64	.632	.025

Note. *a* is GRM discrimination parameter. *b* is GRM boundary location (difficulty) parameter. *SE* is standard error. r_{Theta} is Pearson correlation coefficient between each item response score and IRT theta estimate. RMSEA is the root mean-square error of approximation.

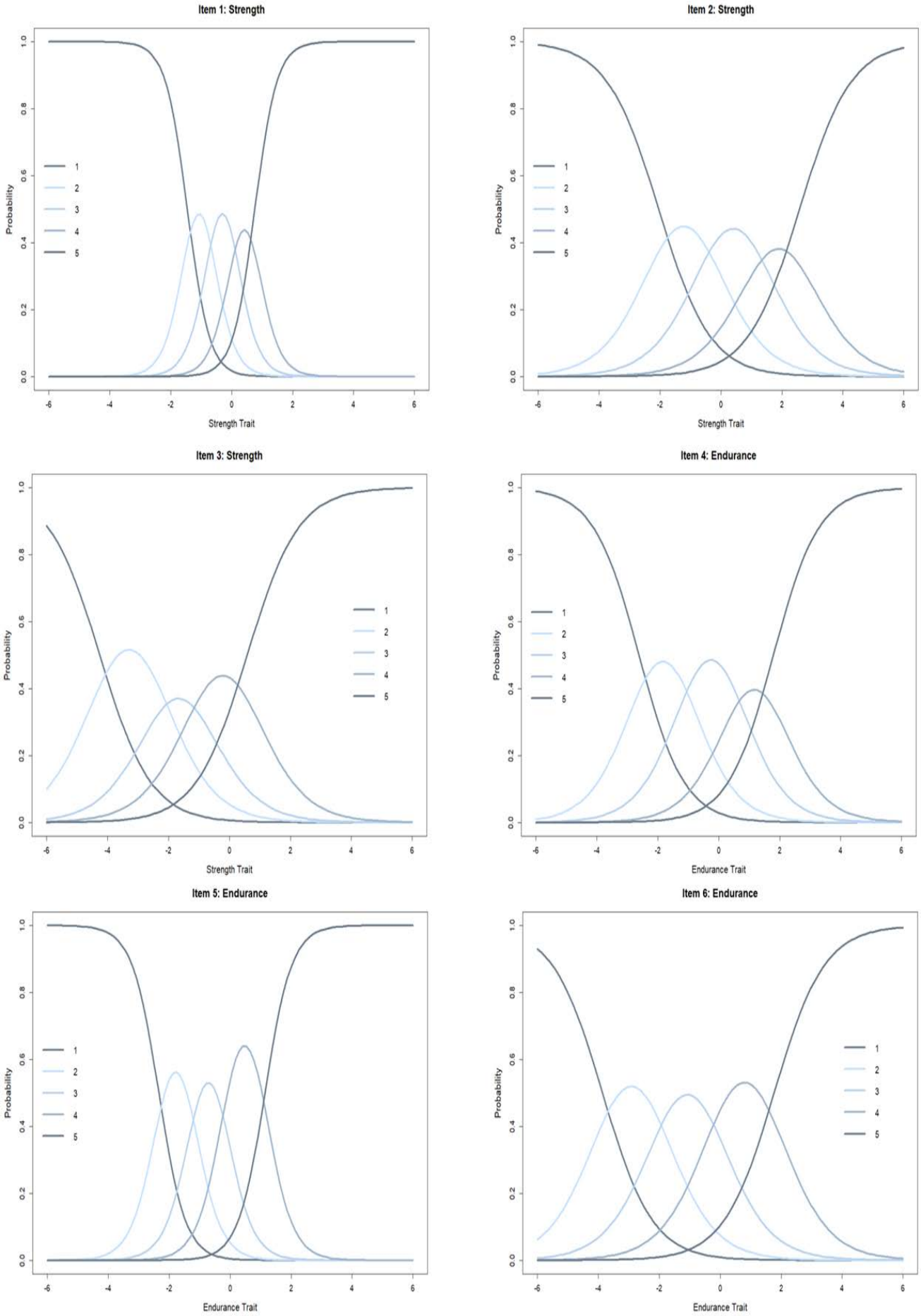

Figure 3. Category characteristic curve (CCC) graphs for each MSAS construct item

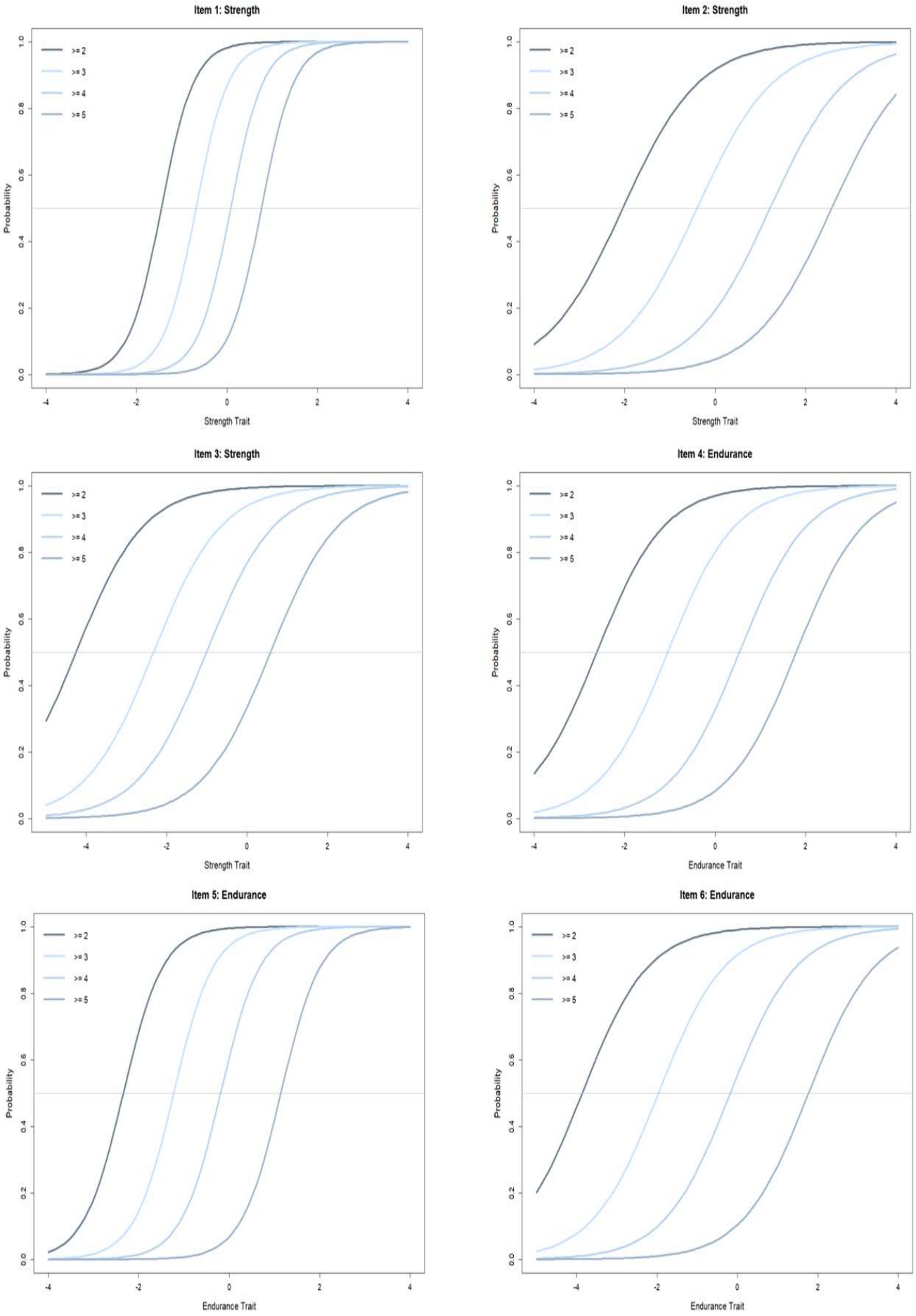

Figure 4. Boundary characteristic curve (BCC) graphs for each MSAS construct item

Table 4. Test information (*I*) for MSAS latent trait regions

	I_{total}	$I_{-4,4}$	$I_{-4,0}$	$I_{0,4}$	*MR*
Strength	14.01 (100)	13.00 (92.8)	7.58 (54.2)	5.41 (38.6)	.716
Endurance	14.43 (100)	13.49 (93.5)	8.00 (55.4)	5.49 (38.1)	.685

Note. *I* is test information (%). *MR* is marginal reliability coefficient of theta.

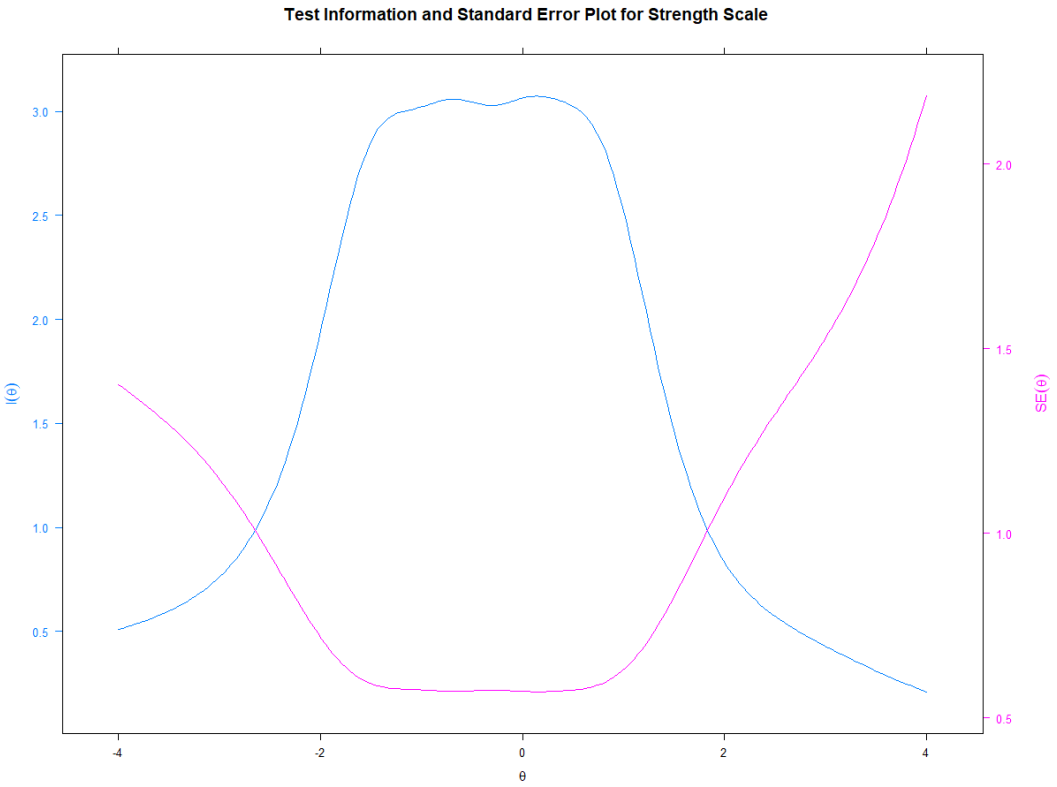

Figure 5. Test information curve with standard error of estimate line for the MSAS strength scale

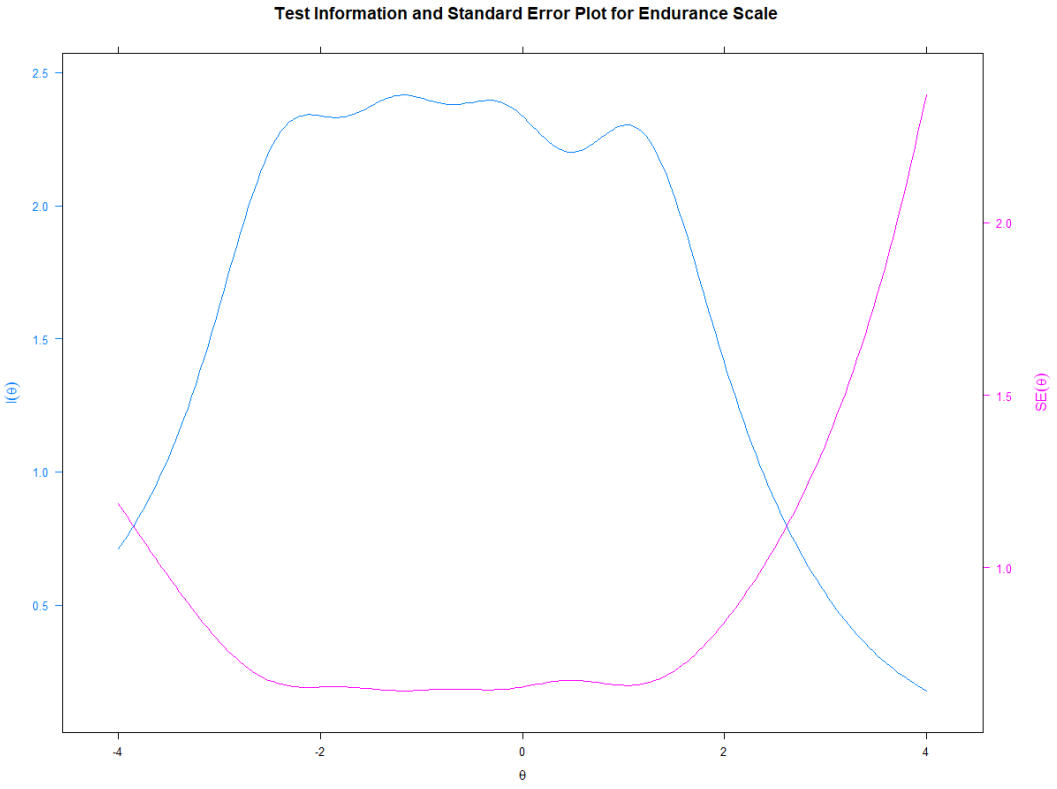

Figure 6. Test information curve with standard error of estimate line for the MSAS endurance scale

Table 5. MSAS person latent trait estimates with standardized fit statistics

	Min	*Q1*	*Median*	*Mean*	*Q3*	*Max*	r_{score}	Misfit #	Misfit %
Strength									
Theta	-2.085	-0.592	0.030	-0.018	0.596	1.515	.957	-	-
Zh	-3.719	0.031	0.695	0.403	1.034	1.281	-	9	2.3
Endurance									
Theta	-2.562	-0.518	-0.002	-0.010	0.527	1.753	.975	-	-
Zh	-4.253	0.042	0.505	0.350	0.986	1.229	-	6	1.5

Note. Zh is the standardized statistic for person fit to theta. r_{score} is Pearson correlation coefficient between scale scores and GRM theta estimate.

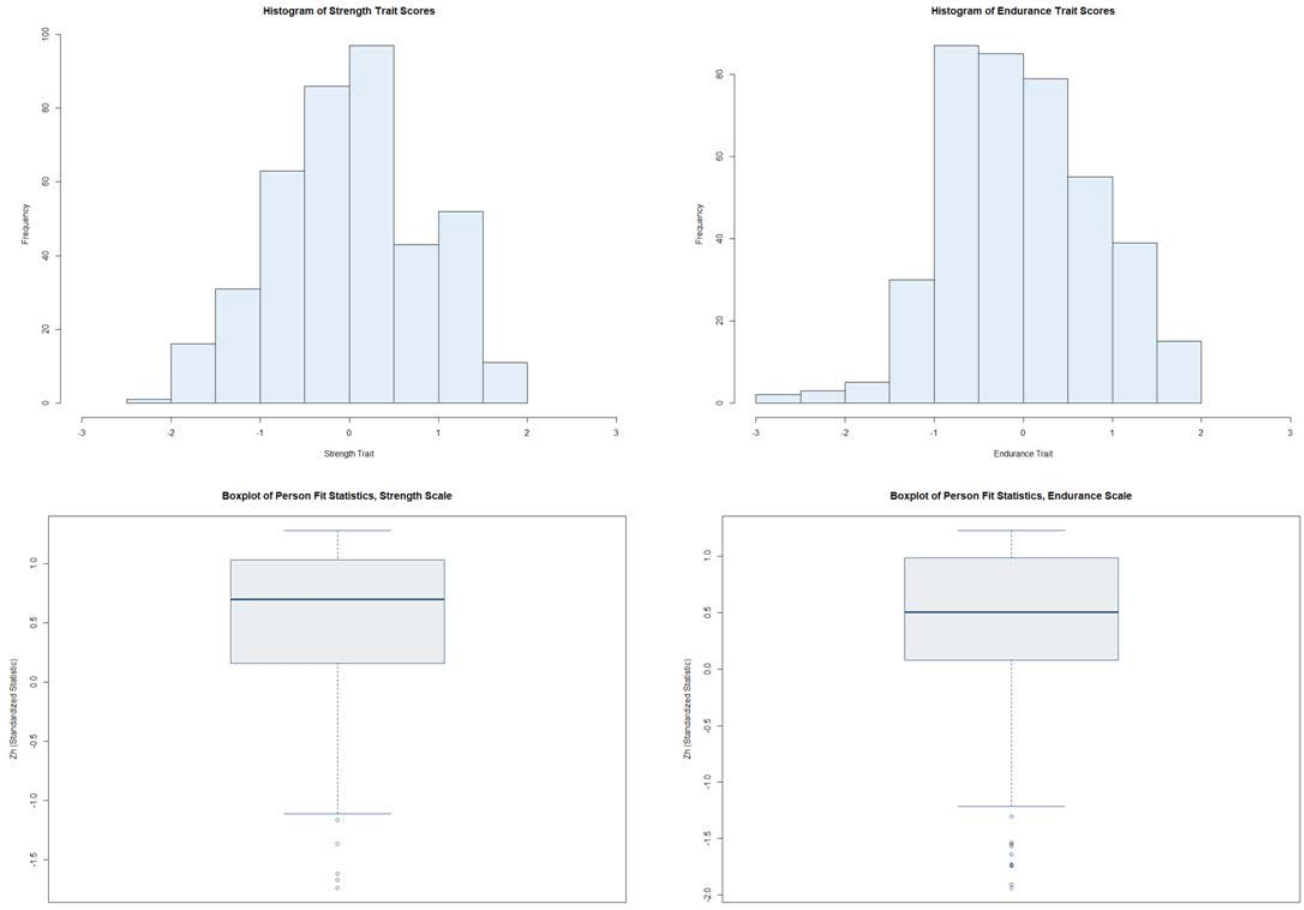

Figure 7. Distributions of MSAS person scores (histograms) and person fit standardized statistics (box plots)

Table 5 contains person estimates and indicate a moderately wide spread of MSAS trait from *strength* (theta: -2.1 to +1.5) and *endurance* (theta: -2.6 to 1.8) scales. Furthermore, MSAS scale sum scores showed a very strong association with the GRM person theta values for both *strength* (r_{score} = .96) and *endurance* (r_{score} = .98) scales. The top portion of Figure 7 displays the relative frequency histograms of theta. Both distributions indicate a roughly bell-shaped pattern with the endurance distribution showing slight left skewness.

Finally, the three IRT assumptions of local independence, unidimensionality, and model fit were assessed. The local independence assumption was met for both models with non-significant LD chi-square statistics, standardized residuals in acceptable range (*strength*: 0 to 9.97 & *endurance*: 1.4 to 3.3), and Cramer's V coefficients indicating approximate weak associations (*strength*: 0.11 to 0.21 & *endurance*: 0.12 to 0.14). The unidimensionality assumption was met for both MSAS scales with Velicer's MAP yielding a lowest average squared (and 4th power) partial correlation for the first principle component in each scale. Lastly, the model fit assumption was met with all model and item RMSEAs < .05 and only 2.3 and 1.5% of persons misfitting the *strength* and *endurance* GRMs, respectively. The bottom of Figure 7 displays standardized statistics (Zh) of person fit distributions in the form of box plots, with misfits removed. These graphs show a slight bias toward negative Zh values, indicating a slight bias toward inconsistent person responses to the GRM.

4. Discussion

The results from this modern psychometric analysis clearly support the valid measurement of muscular *strength* and muscular *endurance* behavior using the

MSAS. These results are backed by acceptable model fit, acceptable item fit, respectable item category functioning across theta, adequate distribution of test information, and the meeting of IRT model assumptions. Results from this study also complement previous research utilizing CTT methods and reinforce the strong psychometric evidence validating the MSAS [4,5]. Albeit, the current research does have some relatively weak aspects of noteworthy significance. One important comment, this study did present a wide item coverage across theta. However, this theta coverage did not extend to extreme levels of theta with high precision. Therefore, MSAS may not necessarily provide error free estimates for extreme MSA behavior. Further research may be warranted and should include more extreme MSA participants in the sample to determine if information collected increases at extreme theta values. A second important comment, with each MSAS scale, a single item presented as the dominant discriminating item. This also means that each scale includes two items with relatively lower item discrimination. A fact that also relates to lower test information. Despite this weak discrimination aspect, all items indeed fit the GRM. As well, the items displaying relatively lower discrimination also showed a response category with low endorsement (< 10%). Therefore, this weak aspect may be due to the relatively small sample size. Further research of the MSAS is suggested with larger samples (Ns > 1,000) where category endorsements, item discrimination, and test information may improve. Other research is also suggested to add to the body of psychometric evidence for the MSAS. Firstly, Rasch measurement is another set of IRT models that are indicated for polytomous response scales [35]. A Rasch measurement analysis of the MSAS could add to this body of evidence philosophically by assessing the extent to which the hypothesis that muscular *strength* and *endurance* are each quantitative and measurable traits. A Rasch measurement analysis could add to this body of evidence statistically by fixing the discrimination parameter equal across all items and assessing the extent to which the category difficulty parameters should be held constant across items (RSM) or freely estimated across items (PCM). Furthermore, with the ability to test the better fitting polytomous Rasch model (i.e., RSM vs. PCM), one can assess the extent to which the MSAS scale should be considered an interval-level (RSM) or ordinal-level (PCM) scale [36]. Secondly, an experimental study on the MSAS is suggested to further evaluate its ability to separate MSA participants with known trait differences. Lastly, an additional study is necessary to evaluate part II of the MSAS for its ability to quantify *participation* in MSA.

5. Conclusions

The MSAS is a seven-item self-report instrument that can measure three MSA constructs: a three-item muscular *strength* construct, a three-item muscular *endurance* construct, and a single-item *body* weight exercise construct. The results from this study provide modern psychometric evidence to support the valid measurement of muscular *strength* and muscular *endurance* behavior using the MSAS. Two additional items (nine items total) are included in the MSAS to quantify MSA *participation* and are in line for future validation. The MSAS is free to use without restrictions, providing proper citation.

Acknowledgements

No financial assistance was used to assist with this project.

References

[1] Piercy KL, Troiano RP, Ballard RM, Carlson SA, Fulton JE, Galuska DA, George SM, Olson RD. The physical activity guidelines for Americans. JAMA. 2018 Nov 20; 320(19): 2020-8.

[2] U.S. Department of Health and Human Services. Physical Activity Guidelines for Americans, 2nd edition. Washington, DC: U.S. Department of Health and Human Services; 2018.

[3] Hart PD, Buck DJ. The effect of resistance training on health-related quality of life in older adults: Systematic review and meta-analysis. Health promotion perspectives. 2019; 9(1): 1.

[4] Hart PD. Development and item analysis of a multidimensional scale to measure muscle strengthening behavior: The Muscle Strengthening Activity Scale (MSAS). EAS Journal of Psychology and Behavioural Sciences. 2019. 1(2): 29-35.

[5] Hart PD. Construct validity evidence for the Muscle Strengthening Activity Scale (MSAS). American Journal of Public Health Research. 2019. 7(5): 189-193.

[6] Haynes SN, Smith GT, Hunsley JD. Scientific foundations of clinical assessment. Routledge; 2018 Nov 8.

[7] DeVellis RF. Classical test theory. Medical care. 2006 Nov 1: S50-9.

[8] Magno C. Demonstrating the difference between classical test theory and item response theory using derived test data. The International Journal of Educational and Psychological Assessment. 2009 May 26; 1(1): 1-1.

[9] Edwards MC. An introduction to item response theory using the need for cognition scale. Social and Personality Psychology Compass. 2009 Jul; 3(4): 507-29.

[10] van der Linden WJ, Hambleton RK, editors. Handbook of modern item response theory. Springer Science & Business Media; 2013 Mar 9.

[11] Hambleton RK, Swaminathan H, Rogers HJ. Fundamentals of item response theory. Sage; 1991.

[12] Reise SP. Item Response Theory. The Encyclopedia of Clinical Psychology. 2014 Dec 29: 1-0.

[13] Andrich D. A rating formulation for ordered response categories. Psychometrika. 1978 Dec 1; 43(4): 561-73.

[14] Masters GN. A Rasch model for partial credit scoring. Psychometrika. 1982 Jun 1; 47(2): 149-74.

[15] Muraki E. Fitting a polytomous item response model to Likert-type data. Applied Psychological Measurement. 1990 Mar; 14(1): 59-71.

[16] Muraki E. A generalized partial credit model: Application of an EM algorithm. ETS Research Report Series. 1992 Jun; 1992(1): i-30.

[17] Bock RD. The nominal categories model. InHandbook of modern item response theory 1997 (pp. 33-49). Springer, New York, NY.

[18] Samejima F. Graded response models. In Handbook of item response theory, volume one 2016 Oct 14 (pp. 123-136). Chapman and Hall/CRC.

[19] Samejima F. Estimation of latent ability using a response pattern of graded scores. Psychometrika monograph supplement. 1969 Dec.

[20] Ma Z, Wu M. The Psychometric Properties of the Chinese eHealth Literacy Scale (C-eHEALS) in a Chinese Rural Population: Cross-Sectional Validation Study. Journal of medical Internet research. 2019; 21(10): e15720.

[21] Ostini R, Nering ML. Polytomous item response theory models. Sage; 2006.

[22] De Ayala RJ. The theory and practice of item response theory. Guilford Publications; 2013 Oct 15.

[23] Johnson MS. Marginal maximum likelihood estimation of item response models in R. Journal of Statistical Software. 2007 Feb 22; 20(10): 1-24.
[24] Chen WH, Thissen D. Local dependence indexes for item pairs using item response theory. Journal of Educational and Behavioral Statistics. 1997 Sep; 22(3): 265-89.
[25] Rea LM, Parker RA. Designing and conducting survey research: A comprehensive guide. John Wiley & Sons; 2014 Sep 9.
[26] Nguyen TH, Han HR, Kim MT, Chan KS. An introduction to item response theory for patient-reported outcome measurement. The Patient-Patient-Centered Outcomes Research. 2014 Mar 1; 7(1): 23-35.
[27] Courtney M, Gordon R. Determining the Number of Factors to Retain in EFA: Using the SPSS R-Menu v2 0 to Make More Judicious Estimations. Practical Assessment, Research, and Evaluation. 2013; 18(1): 8.
[28] Velicer WF. Determining the number of components from the matrix of partial correlations. Psychometrika. 1976 Sep 1; 41(3): 321-7.
[29] Maydeu-Olivares A, Cai L, Hernández A. Comparing the fit of item response theory and factor analysis models. Structural Equation Modeling: A Multidisciplinary Journal. 2011 Jun 30; 18(3): 333-56.
[30] Reise SP. A comparison of item-and person-fit methods of assessing model-data fit in IRT. Applied Psychological Measurement. 1990 Jun; 14(2): 127-37.
[31] Rizopoulos D. Latent trait models under IRT. 2018 Apr. R package version 1. 1-1.
[32] Rizopoulos D. ltm: An R package for latent variable modeling and item response theory analyses. Journal of statistical software. 2006 Nov 17; 17(5): 1-25.
[33] Chalmers P. Multidimensional Item Response Theory. 2019 Sep. R package version 1. 31.
[34] Chalmers RP. mirt: A multidimensional item response theory package for the R environment. Journal of Statistical Software. 2012 May 24; 48(6): 1-29.
[35] Bond TG, Fox CM. Applying the Rasch model: Fundamental measurement in the human sciences. Routledge; 3rd edition; 2015 June 16.
[36] Avian A, Messerer B, Frey A, Meissner W, Weinberg A, Ravekes W, Berghold A. Scaling properties of pain intensity ratings in paediatric populations using the Faces Pain Scale-revised: Secondary analyses of published data based on the item response theory. International journal of nursing studies. 2018 Nov 1; 87: 49-59.

Appendix: (download from: http://www.fitmetrics.org/MSAS.pdf)

Muscle Strengthening Activity Scale (MSAS)

Please circle your best response to each statement.

Do you regularly engage in muscle strengthening activity (such as push-ups, sit-ups, yoga or weight lifting) as a form of exercise?

Yes (continue to the rest of the survey)
No (you can stop here, thank you)

Think about your muscle strengthening exercise in a **typical week** when responding to each statement.

Muscle Strengthening Activity Scale (MSAS) Part I	Never True	Rarely True	Sometimes True	Usually True	Always True
1. I often exercise my muscles with heavy weight that I can lift 1 to 8 times.	1	2	3	4	5
2. I make sure to rest for long periods of time between muscle strengthening sets in order to lift heavy weight.	1	2	3	4	5
3. I make an effort to strengthen all major muscles of my body.	1	2	3	4	5
4. I often arrange several exercises in an order and quickly move from one exercise to the next.	1	2	3	4	5
5. I spend a lot of effort performing floor exercises for my stomach (like sit-ups, crunches, leg raises) lasting until I fatigue.	1	2	3	4	5
6. I often rest for short periods between muscle strengthening sets to build endurance.	1	2	3	4	5
7. I often exercise my muscles using only my body weight (like calisthenics, yoga, Pilates).	1	2	3	4	5

Muscle Strengthening Activity Scale (MSAS) Part II							
1. On average, how many **days per week** do you exercise your muscles?	1	2	3	4	5	6	7
2. On a typical day, how many **minutes** do you approximately spend exercising your muscles?	10	20	30	45	60	90	120+

Scoring Instructions

Part I: Add questions 1, 2 and 3 to get a ***Strength*** score. Add questions 4, 5 and 6 to get an ***Endurance*** score. Question 7 is a ***Body*** exercise score.
Part II: Multiply questions 1 and 2 to get a MSA ***Participation*** score. Minimum score is 10 minutes per week and maximum score is 840 minutes per week.