Original Article

Grit

Two Related but Independent Constructs Instead of One. Evidence From Item Response Theory

Yulia Tyumeneva, Elena Kardanova, and Julia Kuzmina

Institute of Education, National Research University Higher School of Economics, Moscow, Russian Federation

Abstract: The Grit scale is a popular measure of achievement-striving behavior. Consisting of two subscales, Consistency of Interests (CI) and Perseverance of Effort (PE), this scale has been repeatedly demonstrated to have high reliability and validity. At the same time, an increasing number of studies explicitly report a low correlation between the subscales and distinct patterns of associations with external measures that each subscale forms. We explored whether there is psychometric evidence that a substantive single grit construct underlies the scale. To answer this question, we investigated the scale structure in a more robust framework than the classical test theory and factor analyses could previously provide. The Russian version of the Grit scale was developed and implemented on a representative sample of high school students (n = 2,269), and different models of item response theory (IRT), both unidimensional and multidimensional, were compared to find the best fitting model. The results confirmed that the subscales reflect related but independent constructs rather than the whole grit construct. The psychometric properties of the subscales were analyzed with the two-dimensional Partial Credit Model. Both subscales of the Russian version of the Grit scale are unidimensional, have good psychometric properties, and can be used to estimate respondents' ability.

Keywords: construct validity, dimensionality, grit scale, IRT

Grit is a relatively new construct that joins a list of personal traits associated with goal-oriented behavior, such as persistence, hardiness, and tenacity. Duckworth, who proposed an instrument to assess grit, described the construct of grit as the ability to maintain long-term goals and overcome difficulties in the course of achieving them (Duckworth, Peterson, Matthews, & Kelly, 2007). Accordingly, the Grit scale consists of two subscales: Consistency of Interests (CI) and Perseverance of Effort (PE). Both the full and short versions of this scale (Duckworth, Quinn, & Seligman, 2009) have shown many advantages over other scales from a family of self-reported measures of attainment-oriented personal traits, such as the Tenacity Scale by Baum and Locke (2004) and the Motivational Persistence Scale (Constantin, Holman, & Hojbotă, 2012). As the authors of the Grit scale claim, the scale offers face validity for various populations, a low probability of ceiling effects in high-achieving groups, and a precise focus on the construct of grit (Duckworth et al., 2007).

Both versions of the Grit scale include the idea of grit as a whole construct and produce a total scale score, which is the sum of the PE and CI subscale scores. The developers of the Grit scale found that the Grit score can predict different outcomes better than the subscales individually (Duckworth et al., 2007). Later, they justified the use of the total Grit score based on the results of a hierarchical factor analysis where grit was interpreted as a second-order factor with PE and CI as first-order factors (Duckworth et al., 2009).

A number of studies have supported predictive validity of the Grit scale using the overall Grit score as a simple sum across the PE and CI items. The grit score has predicted educational attainment, the number of lifetime career changes, and the completion of some training programs (Duckworth et al., 2007); retention in the military, workplace, school, and marriage (Eskreis-Winkler, Shulman, Beal, & Duckworth, 2014); and teachers' efficiency (Duckworth et al., 2009; Robertson-Kraft & Duckworth, 2014). Grit has also demonstrated positive links with other theoretically associated psychological constructs, such as well-being in the domain of general surgery (Salles, Cohen, & Mueller, 2014) and job satisfaction among doctors (Reed, Schmitz, Baker, Nukui, & Epperly, 2012).

Despite the extensive use of grit as a whole construct, there is an increasing amount of evidence that the CI and PE subscales can reflect independent constructs rather than aspects of the single grit construct. Importantly, this evidence has been obtained based on different methodological approaches, such as internal structure analyses and criterion-related studies.

In their analysis of the internal structure of the short version of the Grit scale, Datu, Valdez, and King (2015)

tested a hierarchical model of grit wherein CI and PE were explored as first-order factors of a higher-order grit factor. They found that grit comprised two distinct dimensions rather than composing a hierarchical construct. Both dimensions had low reliability coefficients (α = .61 for CI and .58 for PE; .59 for the whole scale). The factors were poorly correlated with each other (r = .15). Generally, estimates of the strength of the relation between PE and CI factors appear to be rather mixed: some researchers have reported positive and strong correlations between these two components (Arslan, Akin, & Çitemel, 2013; Meriac, Slifka, & LaBat, 2015), whereas others have found quite low correlations ranging from .26 (Sheehan, 2014) to .45 (Duckworth et al., 2007; see also Bowman, Hill, Denson, & Bronkema, 2015; Chang, 2014; Rojas, 2015; Weston, 2014).

Several criterion-related studies have demonstrated that both subscales predict different outcomes. In the previously mentioned work of Datu et al. (2015), it was found that PE was a significantly better predictor of key psychological outcomes, academic engagement, and subjective well-being than CI. In contrast, CI did not predict such outcomes and was linked to lower behavioral and emotional engagement. Hatchimonji (2016) also showed that PE was a consistent predictor of academic achievement, whereas CI was not. It is worth noting that both studies were conducted on non-Western samples, and the authors appealed to cultural differences to explain the divergent patterns of the relationships of PE and CI with external outcomes. However, Wolters and Hussain (2014) reported similar results on a Western sample. They investigated grit and its relationships with students' self-regulated learning and past and present academic achievements. Their results showed that only PE was a consistent predictor of all the indicators of self-regulated learning, whereas CI was associated with only two facets of self-regulated learning: time and study environment management strategies and procrastination. Additionally, previous achievement was associated with PE but not CI.

Distinct patterns of correlations have also been found in studies using the short version of the Grit scale. Bowman et al. (2015) showed that PE predicted greater academic adjustment, college grade point average, college satisfaction, sense of belonging, faculty-student interactions, and intent to persist, and it was negatively related to intent to change majors. At the same time, CI was associated with less intent to change majors and careers, but it was not significantly associated with any other outcome in the expected direction. Chang (2014) found that CI was a significant and negative predictor of first-year GPA, while PE was a significant and positive predictor. CI and PE also showed a diverse pattern of correlations with certain personal traits (Weston, 2014).

Even at the psychophysiological level, different correlates of PE and CI have been found. Silvia, Eddington, Beaty, Nusbaum, and Kwapil (2013) examined whether PE and CI differently affect cardiac autonomic activity when people are solving mental tasks. They measured certain characteristics of heart and respiratory activity as indicators of sympathetic and parasympathetic impact during effortful engagement. Their results showed that people who were high in PE exhibited a pattern of autonomic co-activation: sympathetic and parasympathetic activity increased during the task. In contrast, people high in CI showed weaker sympathetic activity and no change in parasympathetic activity.

In summary, an increasing body of studies has repeatedly indicated that two factors, PE and CI, have mixed mutual correlations and form inconsistent nomological networks according to Cronbach and Meehl (1955). These findings call into question whether the Grit scale score reflects the single psychological construct of grit.

We know of several measures that not only are multidimensional but also have low-correlated components (e.g., Antisocial Behavior Scale; Kaat et al., 2015; Adaptive Behavior Scale-S; Watkins, Ravert, & Crosby, 2002) or low correlations of measures of Type A Behavior patterns (Edwards, Baglioni, & Cooper, 1990). The point, however, is that the composite overall score makes sense as an indicator of any substantial single construct. Obviously, a composite score may be a good predictor for certain important life outcomes, and it makes a score practically useful. However, to have a psychologically sensible interpretation, an overall score should reflect a single underlying construct. The score on the Adaptive Behavior Scale should reflect adaptive behavior as a single psychological reality (even multidimensional) or be reinterpreted as reflecting two different constructs: personal independence and social behavior (this point has received special attention for many years). The low correlations among measures addressing Type A behavior patterns are not ignored in the literature but are extensively studied as possible indicators of differences in underlying constructs (Edwards et al., 1990). In the same way, the low correlation between the PE and CI subscales of the Grit scale along with their distinct criterion validity indicates that two constructs underlie the scale instead of the single grit construct.

This hypothesis is verifiable at the psychometric level, but it requires reconsideration of the psychometric model of the Grit scale. Previously, psychometric properties of the Grit scale were assessed using the classical test theory (CTT). Although useful, these previous analyses fall short of providing a comprehensive psychometric analysis of the Grit scale. Item response theory (IRT) has the potential to supplement CTT in a number of important ways. In addition to the well-known advantages of IRT methods and

procedures (see Embretson & Reise, 2000), IRT can model the probability of response options as a function of latent trait levels, allowing the consideration of one or more traits simultaneously and thus giving a simple way to implement both unidimensional and multidimensional models. Additionally, from a practical point of view, IRT modeling offers the substantial advantage of providing a way to investigate the quality of response categories for Likert-type scales and to study dimensionality (Linacre, 2002; Smith, 2002). Therefore, we used the IRT approach to explore the psychometric foundation of the Grit scale.

The purpose of this paper was to present psychometric arguments that CI and PE do not form the whole measure and instead should be considered as independent, although still interrelated, unidimentional measures of different traits. To achieve this aim, an appropriate IRT model should be identified and then used to conduct the item analysis for the Grit scale and the subscales.

To accomplish this goal, we developed the Russian version of the Grit scale. Therefore, in the current study, we also aimed to develop the Russian-language scale and provide all necessary psychometric analysis. In addition, we believe that the development of the Russian-language version of the Grit scale, which is valuable in itself, will also allow for the collection of new data on this measure and improve our understanding of this construct.

Method

The Development of a Russian Version of the Grit Scale

The translation of the Grit scale into Russian was consistent with the guidelines of Van de Vijver and Hambleton (1996) and the International Test Commission (ITC) Guidelines on Test Adaptation (Bartram & Coyne, 1999). The procedure consisted of a primary translation into Russian by two independent translators (not authors) who did not know that their work would be translated back into English. Then, two authors reached a reconciled version of these translations, which were generally found to be very similar. However, sometimes awkward wording and farfetched constructions required revision. Then, we conducted a pilot study with a small group of respondents (n = 10) aged 16–17 years, corresponding to the intended sample age. We flagged the items where at least three of our respondents had consistently been confused by the complicated wording and vague meaning. We were also interested in their interpretation of the items as well as the appropriateness of response alternatives. During this process, some changes in item wording were implemented, where great attention was devoted to maintaining an analogous meaning with the original items. Finally, independent back translation into English was performed by another translator, and the original and back-translated versions were compared by the authors. After the comparison, we found a small number of discrepancies, but we were inclined to favor more readable and natural wording over a literal translation.

As a result of this process, two items ("*I am a hard worker*" and "*I am diligent*") were reformulated to be more consistent with cultural features of the perception of diligence as a personal trait. Although a diligent individual is held in respect and although diligence is an encouraged trait in Russian culture, referring to oneself as a diligent person or a hard worker is perceived as bragging. Therefore, these items were reformulated to reduce their straightforwardness and categoricalness: "I am a hard worker" → "Without irony, I am a hard worker" and "I am diligent" → "At work, I am diligent."

The item "Setbacks don't discourage me" confused some of our respondents from the pilot study because of its double negative wording in conjunction with the negative answer options ("*Do not agree at all*" and "*Do not fully agree*"). To avoid confusion, the item was written with positive wording. We also added "as a rule" to emphasize the regularity of the situation. The item was ultimately worded as follows: "As a rule, setbacks discourage me."

In all other respects, the back translation and the original form were remarkably similar.

Participants

A sample of Russian high school students (n = 2,269, 49% boys; the mean age was 16 years [SD = 0.47]) participated in a longitudinal national survey of high school graduates, in which the Grit Scale was implemented. The sample was representative of the national population of high school graduates in terms of gender, age, type of community (urban/rural), and school size.

Instruments and Procedures

The developed Russian version of the Grit scale included 12 items that referred to two subscales: six items for CI and six items for PE. Each item was rated on a 5-point Likert scale that ranged from 1 (= *do not agree at all*) to 5 (= *completely agree*); all 12 items of the Grit scale are presented in Table 1. The Grit scale was a part of a paper-based survey used in a longitudinal national research of school and university graduates. Students were asked to complete the questionnaire during their usual school class. Incentives were not used.

Analyses

The data analysis was performed in three stages. The purpose of the first stage was to investigate the dimensionality of the Grit scale. For this investigation, we used the Rating

Table 1. Standardized residual loading for the 12 items of the grit scale

Item number	Items	Loadings	
		Dimension 1	Dimension 2
Consistency of interests			
1	I often set a goal but later choose to pursue a different one*		−.27
3	New ideas and new projects sometimes distract me from previous ones*		−.47
5	I become interested in new pursuits every few months*		−.56
7	My interests change from year to year*		−.53
9	I have been obsessed with a certain idea or project for a short time but later lost interest*		−.39
10	I have difficulty maintaining my focus on projects that take more than a few months to complete*		−.23
Perseverance of effort			
2	Without irony, I am a hard worker	.52	
4	I have achieved a goal that took years of work	.59	
6	I have overcome setbacks to conquer an important challenge	.57	
8	As a rule, setbacks discourage me*		−.11
11	I finish whatever I begin	.62	
12	At work, I am diligent	.68	

Note. *Item was reverse scored. It should be noted that when we do PCA of standardized residuals, not of the original observations, the interpretation of PCA is different from usual factor analysis. The "first factor" (in the traditional factor analysis sense) is ability level and this component (dimension) is removed, and we look at secondary dimensions. If there are no other dimensions in the data, then the residuals are random noise and show no structure. Otherwise, there are contrasting patterns in the residuals, and the items with positive and negative loadings contribute to different factors (dimensions). So, in the PCA of residuals, the interpretation is based on the contrast between positive and negative loadings.

Scale Model (RSM), which is a Rasch-type model designed especially for Likert-type scales (Linacre, 2002). The purpose of the second stage was to identify an appropriate IRT model for further data analysis, taking into account the multidimensionality of the data. For this purpose, we compared different models (both unidimensional and multidimensional) to find the best fitting model. The purpose of the third stage was to conduct the item analysis of the Grit subscales under the model identified during the second stage as the best fitting model.

Stage 1. Dimensionality Study

To investigate the dimensionality of the Grit scale, we used the Rating Scale Model (RSM), which is a Rasch-type model designed specifically for Likert-type scales (Linacre, 2002; Wright & Masters, 1982). We examined the dimensionality of the Grit scale by conducting a principal component analysis (PCA) of the standardized residuals, which are the differences between the observed response and the response expected under the model (Ludlow, 1985; Smith, 2002). Winsteps software (Linacre, 2011) was used for this purpose. Theoretically, if the unidimensionality assumption is withheld, then correlations between item-level residuals should be close to zero. If there is no second dimension remaining in the residual variation, then the PCA should generate eigenvalues that are all close to 1, and the percentage of variance across the components should be uniform.

Stage 2. Model Selection

To identify an appropriate IRT model for further data analysis, taking into account the multidimensionality of the data, we compared different models (both unidimensional and multidimensional).

The two subscales of the Grit scale measure related (but supposedly different) latent examinees' characteristics. There are three approaches in the item response modeling to such scales. First, we can ignore the multidimensionality of the scale and apply a unidimensional model. Second, we can recognize multidimensionality and apply a unidimensional model to each dimension consecutively. Third, we can apply multidimensional models.

The unidimensional approach yields a composite score, that is, a single estimate of a respondent's traits and the associated standard error. Furthermore, the reliability of these respondents' estimates is usually higher than that with other approaches. The question is to what extent a composite overall score makes sense as an indicator of any substantial single construct taking into account the multidimensionality of the initial scale. In addition, a disadvantage of the unidimensional approach is the loss of information on the respondent's ability in different dimensions.

The consecutive approach implies that raw scores for each dimension are modeled independently as unidimensional constructs. The advantage of this approach is that it produces ability estimates and their standard errors for each dimension. However, if the number of items per

dimension is small, the standard errors of respondents' estimates can essentially be large, especially in comparison with the unidimensional approach. This can be explained by the fact that the consecutive approach ignores the possible interrelation of different variables.

Under the multidimensional approach, the raw scores for each dimension are treated as distinct information about each respondent, incorporating the correlation between the latent variables. Thus, the loss of reliability is less than that with the consecutive approach.

At the stage of model selection, all three approaches were applied.

Members of the Rasch family of item response models were employed. Although a single set of Likert-type rating categories was used for all the items and although the Rating Scale Model (RSM) seems to be the most appropriate model for data analysis, the Partial Credit Model (PCM) may be useful for exploring the use of the categories across items. Moreover, although the rating categories might work well with the majority of items, with other items, there may be indications that some categories are misunderstood or misused. The PCM is similar to RSM; however, within it, each item has its own threshold parameters (Andrich, 1988; Wright & Masters, 1982). This model thus takes into account possible differences in categories functioning. The RSM and the PCM were employed for unidimensional analysis. The Multidimensional Random Coefficients Multinomial Logit Model (MRCMLM; Briggs & Wilson, 2003) was applied for multidimensional analysis. For the purposes of our study, the MRCMLM was adjusted to a two-dimensional RSM or PCM. In these two-dimensional models, each item loaded on only one dimension, which is referred to as a between-item multidimensional model (Adams, Wilson, & Wang, 1997).

Unidimensional and multidimensional analyses were conducted with ConQuest software (Wo, Adams, & Wilson, 1998). The item parameters and population means and variances were estimated by using the Marginal Maximum Likelihood (MML) technique. A constraint was applied that each distribution of item difficulties had a mean of 0. Standard errors and fit statistics were produced for each parameter estimated.

We obtained several other indices relevant for our study. First, the reliability index was computed. Second, the goodness of fit of the model was evaluated using the deviance index. It is known that for nested models (two models, one being a special case of the other), the difference in deviances has approximately a chi-squared distribution with degrees of freedom equal to the difference between the numbers of parameters in the two models. In addition, the Akaike information criterion (AIC; Akaike, 1974) and Bayesian information criterion (BIC; Schwarz, 1978) were examined to determine which model best fitted the data. It is known that when a sample size is large, AIC tends to favor complex models, whereas BIC may favor more parsimonious models because of the incorporation of a penalty for additional components (Kang, Cohen, & Sung, 2009). Lower AIC and BIC values indicate better fit. Third, the correlation between various dimensions in multidimensional and consecutive approaches and between different approaches was analyzed.

Stage 3. Item Analysis of the Grit Subscales

We conducted a series of analyses that aimed to investigate the psychometric quality of the items and of the subscales. Item analysis was conducted by applying the multidimensional approach with the two-dimensional Partial Credit Model in ConQuest software. The parameters of the model were identified through the Marginal Maximum Likelihood (MML) method. A constraint was applied that the distribution of item difficulties for each subscale had a mean of 0.

We first examined the subscales using a model fit analysis. To measure the extent to which the items fit the PCM, we used the unweighted and weighted mean square statistics provided by ConQuest (in terms of ConQuest output: Unweighted MNSQ and Weighted MNSQ, respectively). These statistics rely on standardized residuals, which represent the differences between the observed response and the response expected under the model. In general, statistical values between 0.7 and 1.3 are considered to indicate acceptable fit between the data and the model PCM (Wright & Masters, 1982).

The next step was to investigate the quality of the response categories. In the PCM context, several criteria are useful for analysis of response categories: for each item, all categories should be used; the categories should demonstrate a good model fit; the step calibrations should advance monotonically with the categories; and the correlation between performance in the category on a single item and ability level should also advance monotonically with the categories. All these criteria were applied to check whether a five-category response system is adequate for all the items.

The next component of the analysis was devoted to the properties of the CI and PE subscales as a whole. We computed a reliability index and the standard error of the ability measurement, and we analyzed the variable map for each scale.

Results

Stage 1. Dimensionality Study

The analysis of the eigenvalues of the Grit scale residual correlation matrix for 12 components indicated that there

was one component with an eigenvalue of 2.9, whereas the eigenvalues for the other components ranged from 1.3 to .65 (with the exception of the eigenvalue for the last component, which was .012). In addition, the percentage of variance for the first component was 24%, whereas the variance for the other components was roughly evenly split across the components.

To obtain further evidence of the scale's multidimensionality, we performed "tailored" simulations on random data with the same person ability and item difficulty structure (Linacre, 2011; Ludlow, 1985). Five independent sets of data that fit the model were simulated. The result of those simulations revealed that the mean eigenvalues of the Grit scale residual correlation matrix for 11 of 12 components ranged roughly from 1.2 to 0.9 (and the mean of eigenvalues for the last component was 0.03), and the variance accounted for in the distribution was roughly evenly split across components, with a variance of 10.1% for the first component.

Based on these results, the scale was two-dimensional and consisted of two subscales. Table 1 shows the standardized residual loadings for all items.

Although our findings confirmed the two-factor structure, the item "As a rule, setbacks discourage me" moved from the PE subscale to the CI subscale. Thus, the PE subscale now consisted of five items, whereas the other seven items composed the CI subscale.

Further, we conducted a dimensionality study for the subscales using the same approach used for the total Grit scale. The eigenvalues of the CI subscale residual correlation matrix for six of the seven components ranged roughly from 1.5 to 0.9, and the eigenvalue for the last component was 0.01. The eigenvalues of the PE subscale residual correlation matrix for four of the five components ranged roughly from 1.4 to 1.1, and the eigenvalue for the last component was 0.02. In addition, for both subscales, the variance accounted for in the distribution was split roughly evenly across the components. These results indicate that the subscales are unidimensional.

Stage 2. Model Selection

The results of scaling the Grit scale using different approaches, namely, unidimensional, multidimensional, and consecutive ones, are shown in Table 2.

First, we compared the PCM and RSM under unidimensional and multidimensional approaches. The PCM is hierarchically related to the RSM, and the multidimensional approach is hierarchically related to the unidimensional approach. Therefore, the model fit can be compared to the change in the deviance value. As Table 2 indicates, the difference in deviance between the models was very large, and the multidimensional PCM fits the data much better than the unidimensional models and multidimensional RSM. Additionally, the comparison of the models by AIC and BIC criteria confirmed the same conclusion: the multidimensional PCM provided the best explanation of the data.

It can be noted that under the multidimensional approach, the reliability of each dimension comes closer to the unidimensional reliability estimates, despite the small number of items in each subscale. Moreover, the correlation between respondents' abilities under the multidimensional PCM was low (.24). We can suggest that two dimensions can reflect independent constructs, and a consecutive approach should also be appropriate for the data. The results of the test data scaling using the consecutive approach are shown in the bottom of Table 2.

The conclusion drawn from the consecutive analysis is that for both subscales, the PCM fits the data much better than the RSM. Therefore, for further analysis, we selected the PCM. A choice must be made between the multidimensional approach (two-dimensional PCM) and the consecutive approach (two unidimensional PCMs for two subscales). The analysis of Table 2 revealed that under both approaches, the reliability coefficients for each dimension were almost the same: .7 for both approaches for the CI subscale (Dimension 1) and .81 (multidimensional approach) and .79 (consecutive approach) for the PE subscale (Dimension 2). The correlation between respondents' abilities under the two approaches was .99 for both subscales, and the correlation between the items' difficulties was 1.00.

In conclusion, our analysis suggests that both the multidimensional approach with the two-dimensional PCM and the consecutive approach with two unidimensional models provide good explanations for the Grit scale data. This provides statistical support for our hypothesis that the CI and PE subscales reflect related but independent constructs rather than compose the single grit construct. Second, this finding indicates that for further analysis, we can select any of these approaches. We decided to apply the multidimensional approach with the two-dimensional Partial Credit Model, as it incorporates the correlation between the latent variables.

Stage 3. Item Analysis of the Grit Subscales

Table 3 shows item statistics for both scales, including item difficulty, the standard error of the item difficulty measurement, fit statistics (Unweighted and Weighed MNSQ), and an index of item discrimination as correlation between performance on a single item and ability level.

Table 2. Summary of scaling

Approach	Model	Number of parameters	Deviance	Reliability	AIC	BIC
Unidimensional	RSM	16	71,784.83	.74	71,816.83	71,908.51
	PCM	49	71,293.49	.73	71,391.49	71,672.26
Multidimensional	RSM	18	70,185.76		70,221.76	70,324.90
	Dimension 1			.77		
	Dimension 2			.73		
	PCM	51	69,260.31			69,654.54
	Dimension 1			.70		
	Dimension 2			.81		
Consecutive	RSM					
	Dimension 1	11	42,866.75	.70	42,888.75	42,951.78
	Dimension 2	9	26,900.78	.78	26,918.78	26,970.35
	PCM					
	Dimension 1	29	42,632.60	.70	42,690.60	42,856.77
	Dimension 2	21	26,763.13	.79	26,805.13	26,925.46

Notes. RSM = rating scale model; PCM = partial credit model; deviance = −2 log likelihood; reliability = person reliability index (which is close in value and interpretation to reliability under CTT); AIC = Akiake information criterion; BIC = Bayesian information criterion.

Table 3. Item statistics

Item number	Difficulty	Error	Unweighted MNSQ	Weighted MNSQ	Index of item discrimination
Consistence of interests subscale					
1	−0.41	.02	0.96	0.96	.61
3	0.28	.02	1.02	1.02	.57
5	0.53	.02	1.10	1.10	.55
7	0.36	.02	1.00	1.00	.65
8	−0.57	.02	1.02	1.03	.57
9	−0.26	.02	0.89	0.89	.68
10	0.06*	.04	1.05	1.05	.56
Perseverance of effort subscale					
2	1.07	.02	1.07	1.07	.71
4	−0.29	.02	1.10	1.09	.72
6	−0.08	.02	1.03	1.02	.72
11	−0.16	.02	0.92	0.94	.77
12	−0.54*	.04	0.91	0.91	.78

Note. An asterisk next to a parameter estimate indicates that it is constrained.

Our analysis demonstrates that both the Unweighted and Weighted MNSQ item statistics for all items are within the acceptable range. These results indicate that all of the items in both subscales fit the model in accordance with the chosen criteria.

The next step was to investigate the quality of the response categories. The Electronic Supplementary Material, ESM 1 provides a summary of the subscale category structure and demonstrates that a five-category response system is adequate for all the items. Indeed, all categories are used; the categories demonstrate a good model fit; the step calibrations advance monotonically with the categories; and the correlation between performance in the category on a single item and ability level also advances monotonically with the categories.

The next component of the analysis was devoted to the properties of the CI and PE subscales as a whole. Table 4 shows the results. For example, for the CI subscale, Rasch analysis provided a reliability index of .70, which indicates that the proportion of observed person variance considered to be true is 70% (Stone, 2003; Wright & Masters, 1982). This index is close to classical reliability in terms of value and interpretation (α = .75 for the subscale). The mean of ability distribution was .22 (*SD* = 0.86), and the standard error of ability measurement for the CI subscale was .44 (*SD* = 0.09). Similarly, the results for the PE subscale are presented.

In Figure 1, the variable map for both subscales is presented, which shows the relative distribution of items and respondents in a common metric. The left column is

Table 4. Properties of subscales

Subscale	Rasch reliability	Classical reliability (α)	Ability distribution mean (*SD*)	Standard error of ability measurement mean (*SD*)
Consistency of interests	.70	.75	.22 (0.86)	.44 (0.09)
Perseverance of effort	.81	.83	.90 (1.47)	.67 (0.13)

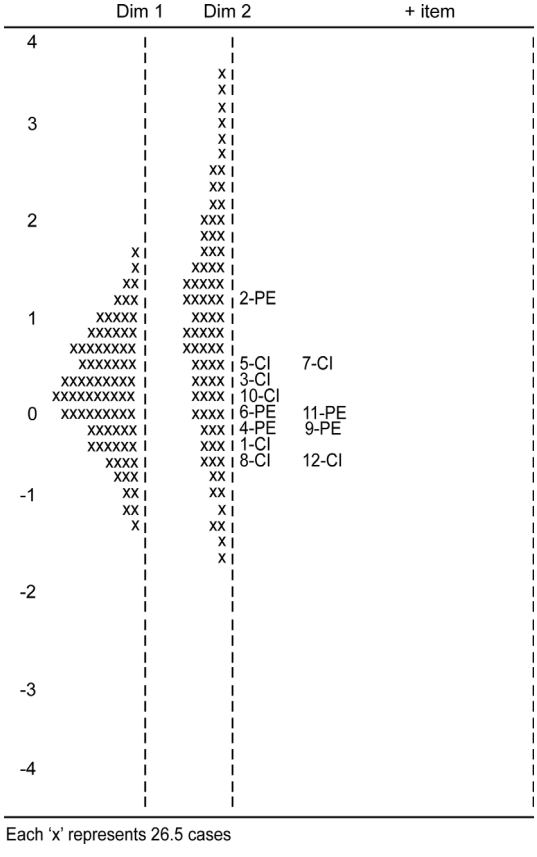

Figure 1. Variable map of the two subscales. Dimension 1 indicates the CI subscale, and Dimension 2 indicates the PE subscale. For each item, the first figure indicates the item number in the scale, and CI and PE indicate the subscale.

the "logit" unit of measurement scale (Wright & Masters, 1982). On the map, persons are separately represented on the left side for the two subscales, and the items are represented on the right. Items that were more difficult to endorse and higher-performing persons are located in the upper part of the map (positive logits), whereas items that were easier to endorse and lower-performing persons are placed in the lower part of the map (negative logits).

The distribution of persons for both subscales is wide and represents good differentiation between high ability and low ability persons for measurement purposes. The analysis of the distribution of item locations shows that although the person sample is well distributed relative to the items for both subscales, there is a lack of difficult items that are appropriate for high-performing persons.

In summary, the IRT analysis of the Grit scale demonstrated that the scale consists of two subscales: CI and PE. Both subscales are unidimensional, have high reliability and good psychometric properties, and can be used to estimate respondents' ability.

Discussion

The first purpose of this study was to obtain psychometric evidence that both factors of the Grit scale should be separately considered to reflect different constructs. We also had a second goal: to develop the Russian version of the Grit scale and explore its psychometric properties using the best fitting model. Both goals were realized.

The results of Stages 1 and 2 are clearly consistent with our hypothesis that the Grit scale deals, in fact, with two different traits: consistency of interests and perseverance of effort. Both subscales analyzed in the multidimensional approach exhibited good psychometric properties and quality in the 5-item response categories. Thus, they can be used as separate measures of the different traits. We have not found psychometric indicators that favor a single construct underlying the scale. On the contrary, the results of the comparison between the multidimensional approach with the two-dimensional Partial Credit Model and the consecutive approach with two unidimensional models provided clear statistical support that both subscales reflect related but independent constructs rather than compose the whole Grit scale as an integrated measure.

To illustrate how two constructs can be related but independent, Duckworth and Gross (2014) referred to relations between grit and self-control. They assumed that these constructs operate at different levels: "self-control . . . corresponds to a goal that is more valued in the moment, and another (grit) . . . corresponds to a goal that is of a greater enduring value" (Duckworth & Gross, 2014, p. 321). The authors provided multiple examples of how people who are high in self-control at a routine level can fail long-term goals because of a lack of grit.

We can use somewhat similar logic to explain how different patterns of behavior may be based on different compositions of CI and PE. People high in PE and low in CI assign all resources to a goal, in spite of failures and fatigue, but later become interested in a new goal, and they pursue it as intensively as the previous one. People high in

CI and low in PE have stable interests in some domain but do not have the ambition to achieve any significant goal in that domain. Finally, people high in both PE and CI, so called "gritty persons," strive to achieve long-term goals.

Interestingly, our findings concur with the results of a recent meta-analysis of the grit literature (Credé, Tynan, & Harms, 2016). The authors found that (a) correlations between CI and PE are substantially moderated, (b) structure models with grit as a higher-order factor poorly fit actual data or are unidentifiable, and (c) CI and PE are inconsistent within a network of outcomes. A general conclusion has been drawn that the grit construct requires a critical reappraisal. It is especially important that we arrived at a similar conclusion about CI and PI as separate measures using modern test theory, which goes beyond the studies included in the meta-analytic review.

One more methodological point has to be discussed here. It concerns the migration of the item "As a rule, setbacks discourage me" from the PE to the CI subscale. We could have explained the migration by two changes that this item underwent in comparison with the original version "Setbacks don't discourage me." First, after the positive rewording, the item became reversed with regard to the other PE subscale items and was therefore recoded. This might have caused its migration to the CI subscale, where all items were reversed and therefore recoded. This tendency to pool reversed items in a separate factor has been extensively discussed in previous decades (Bagozzi, 1993; Marsh, 1996). Second, adding the "as a rule" time statement emphasized the trait's regularity and associated this item with the CI subscale, where all items contain adverbs of time (regularity), such as "usually" or "sometimes."

However, we found that the same migration was observed in two previous studies with no rewording; thus, these changes hardly mattered (Hatchimonji, 2016; Sheehan, 2014). Interestingly, however, participants in these studies and ours were all middle and high school students, and thus, their young age might have affected their interpretation of the item. An intended interpretation of "Setbacks don't discourage me" is that a person keeps pursuing a goal in defiance of setbacks, and thus, the aim is to reflect PE. It is likely that young people do not consider their goals as worthwhile as adults, and therefore, *setbacks do not discourage* them not because they are strong-willed achievers but because the achievement per se is not very important. Instead, agreement with the item actually means that the person puts the goal into a long-term perspective, and therefore, the item loads with the CI subscale. Certainly, this interpretation of the item content can occur in adults as well, but it seems to be especially relevant to younger people.

In addition, young age could affect the need for rewording per se. The items may likely be perceived more clearly by adults. On the other hand, to initially validate the grit scale, the developers used even younger students ranging in age from 7 to 15 years. They did not report any effect on the scale functioning due to such a young sample. Thus, the issue of age dependence of the scale clearly requires additional research including a more nuanced qualitative evaluation at the item level. In addition, more research replicating the current methodology on different samples is necessary to explore the psychological reality underlying the Grit scale.

Acknowledgments

Financial support from the Basic Research Program of the National Research University "Higher School of Economics" is gratefully acknowledged.

Electronic Supplementary Material

The electronic supplementary material is available with the online version of the article at http://dx.doi.org/10.1027/1015-5759/a000424

ESM 1. Table (PDF).
Item category statistics.

References

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*, 1–23. doi: 10.1177/0146621697211001

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control, 19*, 716–723. doi: 10.1109/TAC.1974.1100705

Andrich, D. (1988). *Rasch models for measurement*. Newbury Park, CA: Sage.

Arslan, S., Akin, A., & Çitemel, N. (2013). The predictive role of grit on metacognition in Turkish university students. *Studia Psychologica, 55*, 311–320.

Bagozzi, R. P. (1993). An examination of the psychometric properties of measures of negative affect in the PANAS-X scales. *Journal of Personality and Social Psychology, 65*, 836–851. doi: 10.1037/0022-3514.65.4.836

Bartram, D., & Coyne, I. (1999). *The ITC international guidelines for test use*. Hull, UK: University of Hull, Psychology Department.

Baum, J. R., & Locke, E. A. (2004). The relationship of entrepreneurial traits, skill, and motivation to subsequent venture growth. *The Journal of Applied Psychology, 89*, 587–598. doi: 10.1037/0021-9010.89.4.587

Bowman, N. A., Hill, P. L., Denson, N., & Bronkema, R. (2015). Keep on truckin' or stay the course? Exploring grit dimensions as differential predictors of educational achievement, satisfaction, and intentions. *Social Psychological and Personality Science, 6*, 639–645. doi: 10.1177/1948550615574300

Briggs, D. C., & Wilson, M. (2003). An introduction to multidimensional measurement using Rasch models. *Journal of Applied Measurement, 4*, 87–100.

Chang, W. (2014). *Grit and academic performance: Is being grittier better?* (Doctoral dissertation). Miami, FL: University of Miami.

Constantin, T., Holman, A., & Hojbotă, M. (2012). Development and validation of a motivational persistence scale. *Psihologija, 45*, 99–120. doi: 10.2298/PSI1202099C

Credé, M., Tynan, M. C., & Harms, P. D. (2016). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and Social Psychology*. doi: 10.1037/pspp0000102

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*, 281–302. doi: 10.1037/h0040957

Datu, J. A. D., Valdez, J. P. M., & King, R. B. (2015). Perseverance counts but consistency does not! Validating the short grit scale in a collectivist setting. *Current Psychology, 35*, 121–130. doi: 10.1007/s12144-015-9374-2

Duckworth, A., & Gross, J. J. (2014). Self-control and grit: Related but separable determinants of success. *Current Directions in Psychological Science, 23*, 319–325. doi: 10.1177/0963721414541462

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology, 92*, 1087–1101. doi: 10.1037/0022-3514.92.6.1087

Duckworth, A. L., Quinn, P. D., & Seligman, M. E. P. (2009). Positive predictors of teacher effectiveness. *The Journal of Positive Psychology, 4*, 540–547. doi: 10.1080/17439760903157232

Edwards, J. R., Baglioni, A. J., & Cooper, C. L. (1990). Examining the relationships among self-report measures of the Type A behavior pattern: The effects of dimensionality, measurement error, and differences in underlying constructs. *The Journal of Applied Psychology, 75*, 440–454. doi: 10.1037/0021-9010.75.4.440

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Hillsdale, NJ: Erlbaum.

Eskreis-Winkler, L., Shulman, E. P., Beal, S. A., & Duckworth, A. L. (2014). The grit effect: Predicting retention in the military, the workplace, school and marriage. *Frontiers in Psychology, 5*, 36. doi: 10.3389/fpsyg.2014.00036

Hatchimonji, D. R. (2016). *Grit in Latino middle school students: Construct validity and psychometric properties of the short grit scale* (Master's thesis). New Brunswick, NJ: Rutgers University, Graduate School – New Brunswick.

Kaat, A. J., Farmer, C. A., Gadow, K. D., Findling, R. L., Bukstein, O. G., Arnold, L. E., … Aman, M. (2015). Factor validity of a proactive and reactive aggression rating scale. *Journal of Child and Family Studies, 24*, 2734–2744. doi: 10.1007/s10826-014-0075-5

Kang, T., Cohen, A. S., & Sung, H.-J. (2009). Model selection indices for polytomous items. *Applied Psychological Measurement, 33*, 499–518. doi: 10.1177/0146621608327800

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

Linacre, J. M. (2011). *A user's guide to WINSTEPS program manual 3.71.0*. Retrieved from http://www.winsteps.com/a/winsteps.pdf

Ludlow, L. H. (1985). A strategy for the graphical representation of Rasch model residuals. *Educational and Psychological Measurement, 45*, 851–859. doi: 10.1177/0013164485454015

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology, 70*, 810–819. doi: 10.1037/0022-3514.70.4.810

Meriac, J. P., Slifka, J. S., & LaBat, L. R. (2015). Work ethic and grit: An examination of empirical redundancy. *Personality and Individual Differences, 86*, 401–405. doi: 10.1016/j.paid.2015.07.009

Reed, A. J., Schmitz, D., Baker, E., Nukui, A., & Epperly, T. (2012). Association of "grit" and satisfaction in rural and nonrural doctors. *The Journal of the American Board of Family Medicine, 25*, 832–839. doi: 10.3122/jabfm.2012.06.110044

Robertson-Kraft, C., & Duckworth, A. L. (2014). True grit: Trait-level perseverance and passion for long-term goals predicts effectiveness and retention among novice teachers. *Teachers College Record, 116*, 030302.

Rojas, J. P. (2015). *The relationships among creativity, grit, academic motivation, and academic success in college students* (Doctoral dissertation) Lexington, KY: Universtity of Kentucky. Retrieved from http://uknowledge.uky.edu/edp_etds/39/

Salles, A., Cohen, G. L., & Mueller, C. M. (2014). The relationship between grit and resident well-being. *The American Journal of Surgery, 207*, 251–254. doi: 10.1016/j.amjsurg.2013.09.006

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics, 6*, 461–464. doi: 10.1214/aos/1176344136

Sheehan, K. (2014). *Storm clouds in the mind: A comparison of hope, grit, happiness, and life satisfaction in traditional and alternative high school students* (Doctoral dissertation). Hempstead, NY: Hofstra University.

Silvia, P. J., Eddington, K. M., Beaty, R. E., Nusbaum, E. C., & Kwapil, T. R. (2013). Gritty people try harder: Grit and effort-related cardiac autonomic activity during an active coping challenge. *International Journal of Psychophysiology, 88*, 200–205. doi: 10.1016/j.ijpsycho.2013.04.007

Smith, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement, 3*, 205–231.

Stone, M. H. (2003). Substantive scale construction. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement* (pp. 282–297). Maple Grove, MN: JAM Press.

Van de Vijver, F., & Hambleton, R. K. (1996). Translating tests. *European Psychologist, 1*, 89–99. doi: 10.1027/1016-9040.1.2.89

Watkins, M. W., Ravert, C. M., & Crosby, E. G. (2002). Normative factor structure of the AAMR adaptive behavior scale-school, second edition. *Journal of Psychoeducational Assessment, 20*, 337–345. doi: 10.1177/073428290202000402

Weston, L. C. (2014). *A replication and extension of psychometric research on the grit scale* (Master's thesis). College Park, MD: University of Maryland.

Wo, M. L., Adams, R. J., & Wilson, M. R. (1998). *ConQuest: Generalized item modelling software* [Computer program]. Camberwell, VIC, Australia: ACER Press.

Wolters, C. A., & Hussain, M. (2014). Investigating grit and its relations with college students' self-regulated learning and academic achievement. *Metacognition and Learning, 10*, 293–311.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago, IL: MESA Press.

Received March 30, 2016
Revision received January 24, 2017
Accepted January 27, 2017
Published online August 1, 2017
EJPA Section / Category: Personality

Yulia Tyumeneva
Institute of Education
National Research University "Higher School of Economics"
16 Potapovskiy Pereulok, Building 10
Moscow, 101000
Russian Federation
jutu@yandex.ru