Student Grittiness: A Pilot Study Investigating Scholarly Persistence in EFL Classrooms

Brandon Kramer, Stuart McLean, & Eric Shepherd Martin

学生の Grit：EFL 教室での継続的な意欲

クレーマー・ブランドン、マクリーン・スチュアート、&
シェパード・マーティン・エリック

Abstract

In this paper, we review this concept of grit, operationalized by Duckworth, Peterson, Matthews, and Kelly (2007), and discuss an initial correlational investigation of how well grit predicted performance on two tasks, vocabulary learning (*n* = 21) and extensive reading (ER) (*n* = 58), that were thought to require Japanese university EFL students to demonstrate grit over a long period of time. A modified version of Duckworth et al.'s (2007) original 12-item Grit Scale was administered in Japanese and examined using Rasch analysis (1960), followed by a correlational analysis with the dependent variables of summed vocabulary quiz scores (over one semester) and words read through extensive reading (over one year). Both results were statistically insignificant, with a moderate effect size for the relationship between grit and weekly vocabulary quiz scores, and a weak effect size between grit and the amount of extensive reading.

Keywords: grit, extensive reading, vocabulary, Rasch, EFL

(Received September 26, 2017)

抄　　録

本研究では、grit（継続的な意欲）という概念をもとに、grit と語彙学習および多読との相関関係分析を実施した。日本人学習者にとって、長期の学習期間における grit は必要であり、grit がいかにパフォーマンスの良し悪しを予測する指標となるかを考察する。Duckworth ら（2007）の 12 項目の質問紙を日本語版に修正して実施し、そのデータをラッシュモデル（1960）により検証した。1 学期間に実施した語彙テストの点数（*n* = 21）と 1 年間の多読で読まれた単語数（*n* = 58）との相関分析を行った結果、grit と語彙学習には中程度の相関が認められたが、grit と多読との相関は弱かった。

キーワード： グリット、多読、語彙、ラッシュ、EFL

（2017 年 9 月 26 日受理）

Have you ever wondered why some students will push to the end of any task they are given, but others quickly get distracted and pull out their cell phones? Some researchers, attempting to label what underpins the difference between these two groups, have hypothesized the existence of a quality called *grit*, or a willingness to push on, avoid distractions, and continue a task. In this paper, we review this concept of grit, operationalized by Duckworth, Peterson, Matthews, and Kelly (2007) and popularized by authors such as Malcolm Gladwell (2008), and discuss a correlational investigation of how well grit predicted performance on two tasks which required Japanese university EFL students to demonstrate this *grittiness* over the span of a semester or year.

Background

There are many affective factors and constructs which have been used to explain student L2 performance in the classroom such as ideal L2 self (Dörnyei, 2009), willingness to communicate (MacIntyre & Charos, 1996), and international posture (Yashima, 2002). Looking past the L2 literature, however, and into educational psychology, one idea that is gaining traction is the notion of personal grittiness, defined as a "perseverance and passion for long-term goals" (Duckworth et al., 2007, p. 1087). Duckworth and her co-researchers proposed that "the gritty individual approaches achievement as a marathon; his or her advantage is stamina" (p. 1088).

Grit has been promoted for its ability to predict performance success in a wide variety of contexts. For example, Duckworth et al. (2007) found through six different studies that in the short-term, grit was a better predictor of the completion of a strenuous summer training program among cadets at the United States Military Academy, West Point, than candidates' reports of self-control and their Whole Candidate Score (a composite of high school rank, SAT score, and other factors used to evaluate students for admittance to the academy). Gritty students at elite universities were also found to generally outperform their peers, even though the construct was also found to be associated with lower SAT scores, pointing to its role as a factor independent of aptitude.

Similarly, researchers have examined grit's influence on children and adolescent behavior. Duckworth found that finalists of the 2005 and 2006 Scripps National Spelling Bee competitions generally worked harder and longer than their peers, spending more private time engaged in practice, activities which are, in theory, associated with the trait of grit (Duckworth et al., 2007; Duckworth, Kirby, Tsukayama, Berstein, & Ericsson, 2011). In academic contexts, grit was cited as an important instructional dimension, along with curiosity and self-control, academic tenacity, and self-efficacy, associated with reading outcomes among a group of disabled adolescent learners from different linguistic backgrounds in America (Proctor, Daley, Louick, Leider, & Gardner, 2014).

The Grit Survey

The results of all previous grit studies depend on the measurement of this target latent construct of grit. The original Grit Scale published by Duckworth used 12 items to measure two subconstructs which were thought to make up the higher-order construct of grit. The first subconstruct was *perseverance of effort*, and the other was *consistency of interests* (Duckworth, et al, 2007). Within the original scale, the items measuring consistency of interests were written in such a way that agreeing with the statements are thought to be reflective of *a lack of* grit, and thus their scores on the Likert scale need to be reversed when calculating the total score. The original scale used a 5-point Likert scale response pattern with two positive, two negative, and one neutral response.

Criticisms of Previous Grit Studies

In their 2017 meta-analysis of grit studies, Credé, Tynan, and Harms argued that grit, as operationalized by Duckworth et al. (2007), seems to be problematic for several reasons. First, they found little evidence to support the conceptualization of grit as a higher-order construct comprised of the two subconstructs, perseverance of effort and consistency of interests. This seemed to be particularly true for studies within Asian collectivist cultures such as Korea (Hwang, Lim, & Ha, 2017) and the Philippines (Datu, Valdez, & King, 2016), where consistency of interests did not show much psychometric or predictive value. A second criticism of grit was that although Duckworth claims that grit fits within the Big Five personality factor of conscientiousness, the evidence for any divergent validity is limited. Despite the claim that grit differs primarily in "its emphasis on long-term stamina rather than short-term intensity" (Duckworth & Eskreis-Winkler, 2013, p. 1), if the measurement of the two constructs are indistinguishable, this clarification does not have much value. To this end Credé et al. concluded that grit is most likely simply a repackaging and relabeling of the well-studied construct, conscientiousness (2017). Third, Credé took issue with the claimed predictive ability of the Grit Scale. The scale as a whole showed only a modest predictive ability, especially when compared with other well-known constructs such as study habits or cognitive ability, raising questions of its practical utility. Furthermore, previous studies have shown that the perseverance subconstruct predicts future success better than the combined score of all items, with consistency of interests showing little predictive ability. To what degree Credé's comments apply to the grit construct itself or are simply relevant to its operationalization in Duckworth's commonly used Grit Scale is unknown.

Purpose of the Study

This study is intended to address four gaps in previous research on grit. First, there has been very little investigation into the grit scale itself based on item-response theory, so Rasch analysis will be used to analyze item-level functioning. Second, studies in non-English speaking contexts are still quite limited, with more research needed into Datu et al.'s claim that consistency of interests plays a non-significant role in the success of those in collectivist cultures such as the Philippines (2016), or presumably Japan. Third, previous literature is largely "based on concurrent designs" (Credé, et al., 2017, p. 12), so more studies are needed which show grit's ability to predict future performance on tasks which require sustained effort. Finally, research into grit's relationship with outcomes besides general measures of academic success such as grade point average and retention are limited, so this study will use other criteria for academic success.

This research is intended as an exploratory analysis to investigate both the performance of the grit survey with Japanese university students and whether or not the construct of grit as measured by the original 12-item scale (Duckworth et al., 2007) has any detectable relationship with the studying behavior of first language (L1) Japanese university students studying English in Japan. Specifically, the dependent variables investigated were vocabulary learning over one semester and extensive reading (ER) throughout one school year. These variables were chosen as they were thought to represent the type of studying that could be defined as *gritty*, meaning they require persistent effort and a perseverance over time to accomplish the stated goals.

Methods

Participants

All participants (N = 102) were 19-20 years old and in their second year of study at two separate co-ed universities in Western Japan. The students at both universities had two English classes each week: one with a native English-speaking instructor (the authors) who taught an oral communication class and one with a native Japanese teacher of English which focused on reading and grammar. All students were informed of the purpose of the study and given the option of withdrawing their data without penalty in accordance with the policies of both universities. As shown in Table 1, while the participants from the vocabulary study made up an entire intact class (n = 29) from their university, those from the ER study were from three separate classes who remained active in both semesters, with four students cut due to absenteeism, producing a collected sample of 73 students. The *hensachi* (a department ranking among all universities in Japan based on standardized pre-matriculation test scores) for the two schools was substantially different, with the participants in the ER study ranking about average (set at 50) and those from the vocabulary study ranking almost two standard deviations above the national average at 69 (10 *hensachi* points represent one standard deviation).

Table 1. ***Descriptive Information for the Participants in the Two Studies (N = 102)***

Study	*n*	Department *hensachi*	Major	Classes	Length
Vocabulary	29	69	Sciences	1	1 semester (Spring)
ER	73	49	Humanities	3	2 semesters (Spring and Fall)

Measuring the Independent Variable: The Adapted Grit Survey

The instrument used in this study was a Japanese translation of an adapted 12-point Grit Scale originally created by Duckworth et al. (2007). A Japanese translation was used to reduce the possibility of participants misinterpreting any of the items on the survey. The text was translated and back-translated until it was deemed sufficiently accurate by two native Japanese university lecturers. Six items measured perseverance of effort (e.g., 大事な課題を克服するために困難を乗り越えたことがあります [*daiji na kadai wo kokufuku suru tame ni konnan wo norikoeta koto ga arimasu*, "I have overcome setbacks to conquer an important challenge"]), and six measured consistency of interests (e.g., 新しいアイデアや計画を思いついたために、現在の課題を投げ出してしまうことがあります [*atarashii aidea ya keikaku wo omoitsuita tame ni, genzai no kadai wo nagedashiteshimau koto ga arimasu*, "New ideas and projects sometimes distract me from previous ones"]). Rather than using the 5-point Likert scale for response choices, a 6-point scale was created to prevent the neutral option from being interpreted ambiguously due to social desirability bias (Garland, 1991) or situationally specific item-response patterns (Kulas & Stachowski, 2013). The six response options for each item ranged from 全く当てはまらない (*mattaku atehamaranai*, "Not like me at all") to とても当てはまる (*totemo atehamaru,* "Very much like me"). The adapted scale was administered via Survey Monkey <http://www.surveymonkey.com>, and the original scale can be found at <angeladuckworth.com/research>.

Measuring the Dependent Variables

In addition to the Grit Scale scores, one measurement variable was included for each group: summed vocabulary quiz scores over one semester for the first group and total words read from simplified graded reading texts over one year for the second group.

Vocabulary quiz scores. Vocabulary quizzes were given at the beginning of each class and consisted of 10 randomly selected target items. Five of the items required the students to read a short definition written in English and produce the target English form, while the other five items required students to write a short definition of the target words produced in English. During the first week of the course students were given 15 words to learn. Then, each week 15 new words were added to the total set, from which 10 quiz items were randomly selected. By the final week of the semester the pool of possible quiz items consisted of 210 items. As a result, students were required to maintain a high and increasing level of effort to ensure that they could correctly answer the possible vocabulary quiz items each week. This method was used to encourage spaced rehearsal as adapted sequencing has been shown to facilitate the development of lexical knowledge more efficaciously than studying lexical items only once for each quiz, or not reviewing previously studied items while continuing to study new items (for a detailed discussion see Nation, 2013, pp. 451-458). These quizzes were given and their scores were recorded throughout the Spring semester of 2014. Students stated that they understood the benefits of this assessment method, and also that it required them to maintain a high degree of effort each week. Individual total scores were established by calculating the sum of all quiz scores during the semester.

Extensive reading word count. Extensive reading (ER) is a study method intended to increase reading development which is characterized by fluent comprehension, high reading speed, reading large amounts of text, and a focus on the meaning of text (Waring & McLean, 2015). From the third week of the Spring semester the second group of students was required to read 4,000 words each week from graded readers contained at the university's library. Their reading was recorded and tracked over the Spring and Fall semesters of the 2014-2015 school year using the learner management system M-Reader <www.mreader.org>. On M-Reader students take a 10-item quiz for each book they read, with books (and their word counts) not counted for any book with a quiz score below 60% to ensure student compliance and adequate comprehension. University policy stated that the students should not be assigned more than one hour of homework per week, so a weekly target of 4,000 words was assigned. At a mean reading rate of 77 words per minute (as measured in class in their first two weeks), participants could be expected to read for a mean of only 52 minutes a week. However, it was expected that when completing quizzes some re-reading of material would take place. Therefore, a weekly reading goal of a minimum of 4,000 words was set for the ER group, resulting in a total reading requirement of 52,000 words for each semester (104,000 total words for the year). The benefits of reading this amount were clearly explained at the beginning of the Spring semester, and the students' reading was checked weekly with advice given on their progress at the beginning of each class.

Analysis

Grit survey data. To assess how the survey was functioning, the data were first analyzed together (from both studies) with Winsteps software v. 3.73 (Linacre, 2011) using the Rasch

Rating Scale model for categorical data (Andrich, 1978). Before entering the data, the valence of the Likert ratings was reversed to account for the items expressing negative statements (statements 2, 3, 5, 7, 8, and 11). In addition to item and person reliability indices, the analyses consisted of Rasch person and item fit analysis, and Rasch PCA of item residuals to investigate the unidimensionality of the construct (Apple, 2013).

Although previous researchers using the Grit Scale (Duckworth et al., 2007) simply counted the scores for each question and used the total for correlational analyses, in this study the Rasch model was used due to several benefits it offers over the more traditional approach. First and foremost, utilizing this approach allowed us to examine the interaction of items and people to determine the degree to which they fit the Rasch model, and by extension the hypothesized construct, grit. Rasch analysis, a probabilistic psychometric measurement model, allows us to place test items and test-takers on the same interval scale of difficulty and ability for direct comparison (Bond & Fox, 2015; Rasch, 1960). Within normal instrument creation contexts, we would seek to write items which fit the model, thought to be ideal, producing better quality instruments and therefore better-quality measurements. As the items were already created, however, this study should be considered an exploratory analysis. As the focus of this first analysis is the interaction of Grit Scale items and Japanese university students rather than on the external criterion measures of vocabulary learning and extensive reading, the Rasch-based analyses of the survey behavior were conducted with all participants together.

Correlational relationships with vocabulary studying and extensive reading. Using Winsteps, Rasch person measures for participant grittiness were recalculated for the participants within each subgroup and a correlation was found using SPSS (v. 23) with either the total vocabulary scores or total number of words read. The data were checked for normality using both the descriptive statistics (Table 2) and a visual check of the histograms (Figure 1). Due to the large degree of skewness and kurtosis within the ER word counts, the more conservative non-parametric Spearman correlation was used for both analyses. Finally, to investigate the claims by Credé et al. (2017), the same calculations were then made for each study using the two subconstructs of grit as predictor variables.

Table 2. ***Descriptive Statistics for the Two Dependent Variables***

	N	Min	Max	*M*	*SD*	Skewness	*SES*	Kurtosis	*SEK*
Vocab. total	29	16	110	59.4	30.0	.22	.43	−1.30	.85
ER word count	73	16,095	276,246	97,392.5	31,169.2	2.08	.28	14.88	.56

Note. *M* = mean; *SD* = standard deviation; *SES* = standard error of skewness; *SEK* = standard error of kurtosis.

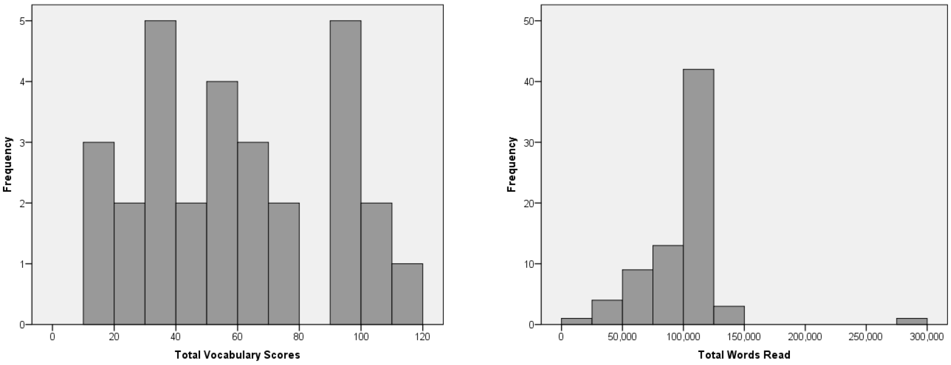


***Figure 1*. Distributions of total vocabulary scores and total words read.**

Results

Item and Person Reliability

The technical quality of the items can be determined by inspecting the Rasch reliability estimates. Both person reliability and item reliability should be inspected, but we would expect item reliability to be higher than person reliability in most circumstances, as items are more reliable and stable. Person reliability can be thought of as the degree with which the items reliably separate the people by ability, and likewise item reliability is the degree with which the people reliably separate the items by difficulty. Rasch reliability estimates can be thought of as similar to the more traditional Cronbach's alpha estimate, but are preferable because they are based on data that fit the Rasch model (Bond & Fox, 2015). For both reliability estimates, we would ideally strive for values above .90, but anything above .80 could be considered acceptable for uses that are not high stakes.

Along with the person and item reliabilities, the person and item separation indices were investigated. While the reliability estimates are often sufficient, they can suffer from a ceiling effect as improvements in reliability above .90 are not reflected in the estimate. They also suffer from a lack of linearity, meaning that an improvement of .70 to .80 is not twice the amount of improvement in reliability as .90 to .95 (Smith, 2001). The separation indices, however, are linear and have no upper limit, so are thought to better represent the degree to which either the item or people are being reliably separated into their respective difficulty or ability hierarchies (Bond & Fox, 2015). Values above 2.00 for both estimates are considered sufficient (Linacre, 2013).

Item analyses were conducted first to ensure that non-fitting person responses were not removed from the data set because of non-fitting item responses. Initial item reliability was .95, with a Rasch item separation of 4.20, both well above our previously mentioned criteria of .80 and 2.00, respectively. The Rasch reliability of person responses was estimated at .75, with a

Rasch person separation of 1.73. These are both below our criteria of .80 and 2.00, respectively, indicating that the instrument did not reliably separate the participants according to the construct of grit.

Item and Person Fit Analyses

Person and item fit to the target construct of grit was determined through the production of both *infit* and *outfit* statistics (Linacre, 2002). Infit statistics are weighted to provide information about persons who are at or near item endorsability levels. Weighted outfit statistics are more easily influenced by respondents who endorse difficult items too easily or easy items too harshly. Infit values are of greater interest than outfit values when trying to establish the quality of items (Bond & Fox, 2015). For this study, misfit was categorized as less than 0.60 mean-squares (MNSQ), or greater than 1.40 MNSQ as recommended by Linacre (1994).

The Infit and Outfit MNSQ (see Table 3) statistics for the survey items range from .69 to 1.48. Those items which display *overfit* to the Rasch model (meaning the response patterns match the prediction of the model too well, represented by lower values) are not normally considered to be problematic, as they are simply redundant. Those items which *underfit* (represented by higher values), however, are more problematic in that they highlight unpredicted and random response patterns. Two items seem to be slightly problematic in this regard: items 10 and 11. Ideally these items would be replaced with items that are more productive for measurement, but as they are part of the original Grit Scale and are not erroneous enough to degrade measurement (Linacre, 1994), their data were retained.

Table 3. ***Descriptive Statistics for the Items on the 12-Item Grit Scale***

Item	Statement	Measure	Infit MNSQ	Outfit MNSQ
1	I have overcome setbacks to conquer an important challenge.	−.69	.92	.92
2*	New ideas and projects sometimes distract me from previous ones.	−.09	1.01	1.02
3*	My interests change from year to year.	.20	1.31	1.34
4	Setbacks don't discourage me.	−.42	1.09	1.09
5*	I have been obsessed with a certain idea or project for a short time but later lost interest.	.92	.99	1.00
6	I am a hard worker.	−.35	.71	.73
7*	I often set a goal but later choose to pursue a different one.	.30	.75	.76
8*	I have difficulty maintaining my focus on projects that take more than a few months to complete.	.56	.81	.81
9	I finish whatever I begin.	−.39	.70	.69
10	I have achieved a goal that took years of work.	.25	1.48	1.47
11*	I become interested in new pursuits every few months.	−.16	1.41	1.42
12	I am diligent.	−.13	.74	.75

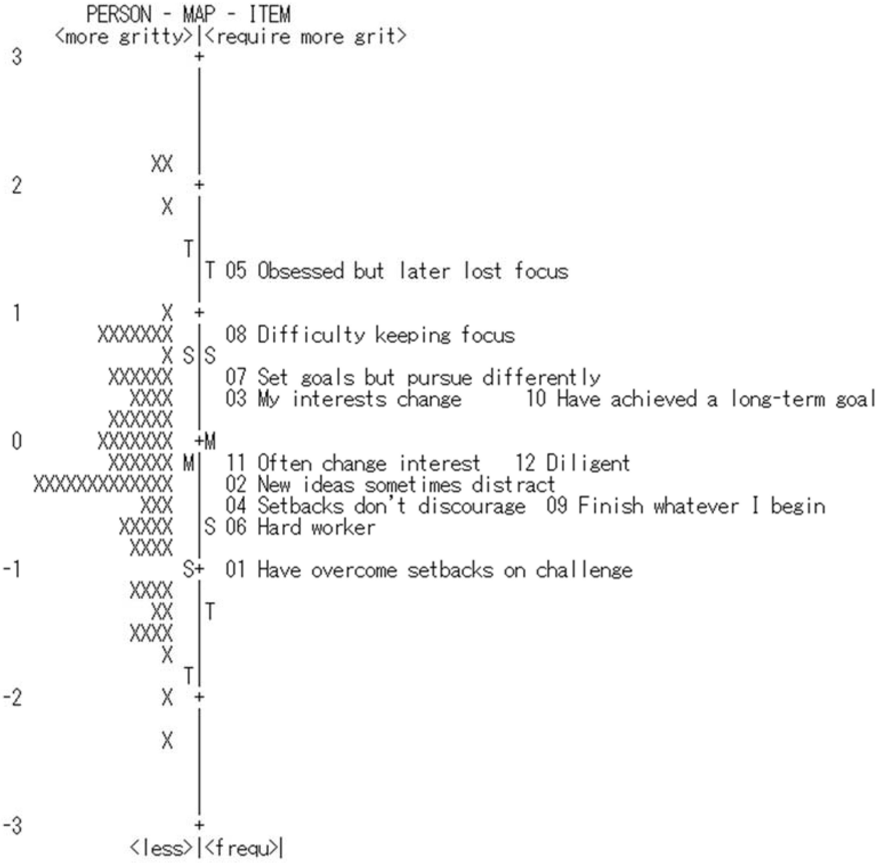
Note. MNSQ = Mean-Square fit statistics. *Items which measure *consistency of interests*, reflective of *a lack of* grit, and thus their scores on the Likert scale were reversed.

Based on the fit criteria, the responses from 23 of the 102 total persons were found to systematically misfit the Rasch model. Extreme scores add measurement error and adversely affect item fit and the unidimensionality of the construct (Apple, 2013; Linacre, 2013) and so it is recommended that such responses be excluded. These participants (8 from the vocabulary study and 15 from the ER study) were therefore removed from all subsequent analyses. The resulting Rasch person reliability of the 79 remaining participants was .82 and the person separation was 2.11 (previously .75 and 1.73, respectively), and the item reliability and separation were .95 and 4.53, respectively (previously .95 and 4.20), indicating that the removed students were indeed problematic for measurement due to their erratic answering patterns.

Item-Person Map

In addition to examining the data fit to the Rasch model, constructing an Item-Person map allows us to visually inspect the interaction of survey items (along with the performance of each Likert sub-scale) and the students who answered the survey questions on the same linear scale. Specifically, we want to look for any floor or ceiling effects, where the participants score minimally or maximally. When this happens Rasch is unable to accurately estimate their person ability, which can negatively influence later analyses. To avoid these problems there should be items on the instrument which every person can correctly answer (or strongly endorse in the case of survey instruments), and similarly items which even the highest ability person has trouble with (Weaver, 2004, p. 399). When checking if there are enough items, or looking for gaps in the data, we looked for a gap between item difficulties where person measurement error would be expected to rise if their ability fell between those values, estimated to be half a logit or more.

The item-person map (Figure 2) shows the spread of item endorsability, covering the range of participants without any ceiling or floor effects. There also do not appear to be any noticeable gaps in the item hierarchy. Although the students in the vocabulary study, who were from a higher ranked school, had a higher mean Rasch-person ability estimate (.13, *SD* = .93) compared with those who engaged in ER (-.22, *SD* = .66), the difference was not statistically significant at the .05 alpha level (*t*(100) = -2.14, *d* = .43, *p* = .071) when calculated using an independent samples Student *t*-test, although it showed a small effect size by the standards recommended by Plonsky and Oswald (2014).



```
         PERSON - MAP - ITEM
      <more gritty>|<require more grit>
 3                    +
                      |
                      |
                      |
                  XX  |
 2                    +
                   X  |
                     T|
                      |T 05 Obsessed but later lost focus
                      |
 1                 X  +
             XXXXXXX  |  08 Difficulty keeping focus
                   X S|S
              XXXXXX  |  07 Set goals but pursue differently
                XXXX  |  03 My interests change      10 Have achieved a long-term goal
              XXXXXX  |
 0           XXXXXXX  +M
              XXXXXX M|  11 Often change interest   12 Diligent
       XXXXXXXXXXXXX  |  02 New ideas sometimes distract
                 XXX  |  04 Setbacks don't discourage  09 Finish whatever I begin
               XXXXX  |S 06 Hard worker
                XXXX  |
-1                   S+  01 Have overcome setbacks on challenge
                XXXX  |
                  XX  |T
                XXXX  |
                   X  |
                     T|
-2                 X  +
                      |
                   X  |
                      |
                      |
-3                    +
                <less>|<frequ>|
```

Figure 2. **Item-person (Wright) map for all participants and the grit survey items. The participants are represented on the left with X's, with those participants measuring higher on the Grit Scale towards the top. The items are on the right of the item-person map, with those items requiring more grit to endorse placed higher on the scale.**

Unidimensionality of the Data

Rasch Principal Components Analysis (PCA) allows us to look at the degree to which the data contribute to the construct of grit. It is important to remember that this is not an attempt to "uncover" or "reveal existing underlying dimensions, but rather *construct* dimensions for the purposes of measurement on the basis of test performance" (emphasis ours) (McNamara, 1991, p. 143). It is therefore not important if survey or test items represent slightly different content; it is only important whether, together, they construct a single construct through their interaction with participants.

We can investigate the dimensionality of an instrument through the standardized residuals in Winsteps. Looking at the amount of unexplained variance in the different contrasts found, we might expect values up to about 2.0 by chance due to random variation in the data. Any contrasts above this value, however, warrant further analysis. If the items which have the strongest positive and negative loadings seem to make up sets of items which logically seem to group together, it might be the case that the measurement is multidimensional (Linacre, 2013; Wolfe & Smith, 2006).

In these data, the Rasch model accounted for 39.0% of the variance (eigenvalue = 7.7). The degree to which the data demonstrates unidimensionality, however, is not dependent on the amount of variance in this first dimension, but instead depends on the size and composition of the residual contrasts. The first residual contrast accounted for 14.2% of the variance (eigenvalue = 2.8). This was greater than the previously established criteria, and a further investigation uncovered a non-random pattern among the positive and negative loadings. As seen in Table 4, except for questions 2 and 8, all the statements with valence-reversed Likert ratings loaded positively above .40, a common threshold for inspection (Stevens, 2002). Items 1, 4, and 6 similarly loaded negatively below -.40. Despite their inclusion on the same survey as the positively worded items, the Rasch PCA analysis found that these two sets of questions, measuring the two subconstructs of grit, produce distinct answering patterns and thus likely indicate a second dimension within the data.

Table 4. ***Rasch Principal Components Analysis of Item Residuals of the Principal Contrast to the 12-Item Grit Scale***

Item	Loading	Measure	Infit MNSQ	Outfit MNSQ
3*	**.69**	**.20**	**1.31**	**1.34**
11*	**.69**	**−.16**	**1.41**	**1.42**
5*	**.50**	**.92**	**.99**	**1.00**
7*	**.41**	**.30**	**.75**	**.76**
2*	.27	−.09	1.01	1.02
8*	−.25	.56	.81	.81
12	−.28	−.13	.74	.75
9	−.34	−.39	.70	.69
10	−.38	.25	1.48	1.47
4	**−.45**	**−.42**	**1.09**	**1.09**
1	**−.52**	**−.69**	**.92**	**.92**
6	**−.69**	**−.35**	**.71**	**.73**

Note: MNSQ = mean-squared. *Items which measure *consistency of interests*, reflective of *a lack of* grit, and thus their scores on the Likert scale were reversed. Measures are in Rasch logits. Items above the dotted line indicate items which loaded positively onto the first contrast and items below the line indicate items which loaded negatively. Items with loadings greater than .40 or less than -.40 are bolded.

Correlations with Student Performance

Spearman correlation coefficients were computed between the Rasch person measures of grittiness for the participants in each study (excluding those 23 participants who were cut due to extreme misfit to the Rasch model) and their respective dependent variable, either the sum of vocabulary quiz scores or total number of words read from graded readers over a year. Using the Bonferroni approach to control for Type I errors across the two correlations, a *p*-value of less than .025 (.05/2 = .025) was required for significance. By the correlational interpretations suggested by Plonsky and Oswald (2014), the correlation between participant grittiness and vocabulary quiz scores (n = 21, r = .39, p = .035) showed a medium effect size and the correlation between grittiness and words read (n = 58, r = .23, p = .056) showed a small effect size. Neither relationship was found to be statistically significant under the p < .025 adjusted criteria.

In order to investigate the claim that the perseverance subconstruct predicts future success better than combined grit scores, Spearman's correlations were calculated between the respective dependent variables and the Rasch person ability estimates produced from only those items which measured each respective subconstruct. As shown in Table 5, the correlation strength of each subconstruct, either consistency of interests or perseverance of effort, was not as strong as the correlation produced from the whole 12-item Grit Scale for either dependent variable, suggesting that the combined scores provided the highest predictive ability.

Table 5. ***Correlations Between the 12-Item Grit Scale Person Measures and the Dependent Variables***

	Grit Scale Section	Person Reliability	Person Separation	Item Reliability	Item Separation	r_s	p
ER (N = 58)	Overall	.76	1.77	.94	4.06	.23	.056
	Consistency	.67	1.41	.94	4.00	.21	.079
	Perseverance	.75	1.71	.90	3.04	.19	.106
Vocab (N = 21)	Overall	.88	2.73	.86	2.51	.39	.035
	Consistency	.80	2.02	.74	1.68	.37	.047
	Perseverance	.81	2.09	.74	1.69	.37	.051

Note. ER = Extensive Reading; Vocab = Vocabulary Studying; r_s = Spearman's Rho correlations with the dependent variable of each study.

Discussion

Implications of the Findings

In this study the items within the 12-item Grit Scale were examined using Rasch analysis and were found to fit the Rasch model reasonably well, determined by looking at the item fit statistics, with only two items showing a misfit slightly outside the desirable range. The degree with which the items reliably separated the participants, however, could be considered weak, which could be the result of either a lack of variance in item endorsability within the survey itself, or a lack of variance in the latent construct among the participants. This was particularly noticeable in the muted reliability and separation estimates produced for the subpopulation who took part in each study, shown in Table 5. Improving the variance in these data would be expected to strengthen any correlations between participant grittiness and the dependent variables, if indeed they represent latent constructs which exist as predicted.

Through the analysis of Rasch PCA residuals, it was apparent that the data were not unidimensional in nature. The two uncovered dimensions aligned with the two subconstructs of the overall grit construct, consistency of interests and perseverance of effort. It is also possible, however, that the performance differences between the two sub-constructs was simply the result of the different answering styles of items measuring them, with those targeting consistency of interests utilizing an opposing valence to the others. Rewriting the scale such that all items are answered using the same Likert scale and valence would help to resolve this issue and possibly produce a more unidimensional data set.

The correlational analyses provided mixed evidence of the predictive power of the Grit Scale. While the results of the vocabulary analysis showed a medium-sized relationship between grit and weekly vocabulary studying as measured by students' weekly quizzes (here *medium-sized relationship* is defined relative to other language learning studies, as discussed in Plonsky & Oswald, 2014), the relationship between grit and the amount of extensive reading showed a weak relationship, with the *p*-values of both analyses lacking statistical significance. Further research with larger samples of participants would help determine if these relationships are generalizable. Furthermore, despite the evidence of multi-dimensionality, the correlational analyses by subconstructs showed that the combination of both dimensions together produced a stronger relationship with the dependent variables than separately. This supports the decision by the creators of the Grit Scale to include both of these dimensions within the same scale, and provides negative evidence to Credé et al.'s claim that only the perseverance of effort dimension predicts future success (2017). In particular, the results of this study are in contrast to those of other collectivist Asian cultures where consistency of interests was found to have limited success (Datu et al., 2016; Hwang et al., 2017), with this study showing it to be at least equal in value to the persistence of effort dimension.

Study Limitations and Future Research

Greater correlations with the target studying practices (vocabulary instruction and extensive reading) might have been found if more variance had been achieved in these constructs, an issue raised by Credé et al. (2017). Although the vocabulary scores seemed to show a relatively normal distribution, the participants' reading amounts clustered around the total number of words required for homework over the year by their instructors. While many students fell short of that goal, very few read above and beyond what they were required, limiting the variance and therefore the statistical power available for the correlational analysis.

The grit instrument itself, while it was able to detect a weak to moderate relationship between the investigated variables, lacks any items which directly reference language education. This could be important, for example, because one can easily conceive of a student who would be labelled as *gritty* in other ventures such as sports or music but who shows less persistence and consistency when studying English. More research is necessary to investigate the distinction that might exist between grit as a global trait, and perhaps a more domain-specific conceptualization of the construct. There is precedence for this distinction in other affective traits such as self-efficacy, which is thought to be domain-specific (Bandura, 1986).

Finally, although in this study the original 12-point Grit Scale (Duckworth et al., 2007) was used, there is a more recently produced 8-point Short Grit Scale (Grit-S) (Duckworth & Quinn, 2009). This newer version boasts slightly improved psychometric properties and greater efficiency with four fewer items, so it is possible that it would show different results.

These initial results show that while the 12-item Grit Scale was able to provide some insight into how grit related to success on language learning tasks such as daily vocabulary studying and extensive reading, future research could benefit from improvements and revisions to the instrument to obtain data which better fit to the Rasch model and provide greater variance. Furthermore, if grit can be more robustly found to be an indicator of the effort put into foreign language learning, an instrument such as this might be useful for identifying students that might need extra assistance or special attention in the classroom to successfully achieve long-term learning goals. Also, if grit can be shown to predict future behavior, it is worth investigating whether attempts to increase students' *grittiness* would increase the frequency of certain behaviors, particularly those which are known to lead to success. At this point the 12-item Grit Scale shows a limited ability to provide this function.

Conclusion

In this study, we investigated the use of grit as a predictive measure of Japanese university students' foreign language studying practices. We hope to see further investigations in the future which build on these results and help us to better understand how grit plays a part in foreign language education.

Acknowledgements

The authors would like to thank the anonymous reviewers for their constructive and insightful critiques and comments that have improved this work. Any remaining faults are the sole responsibility of the authors.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

Andrich, D. (1978). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement*, *38*(3), 665-680. https://doi.org/10.1177/001316447803800308

Apple, M. T. (2013). Using Rasch analysis to create and evaluate a measurement instrument foreign language classroom speaking anxiety. *JALT Journal*, *35*(1), 5-28.

Bandura, A. (1986). Social foundations of thought and action: A social cognitive theory. Englewood Cliffs, NJ: Prentice-Hall.

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

Credé, M., Tynan, M. C., & Harms, P. D. (2017). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and Social Psychology*, *113*(3), 492-511. https://doi.org/10.1037/pspp0000102

Datu, J. A. D., Valdez, J. P. M., & King, R. B. (2016). Perseverance counts but consistency does not! Validating the Short Grit Scale in a collectivist setting. *Current Psychology*, *35*(1), 121-130. https://doi.org/10.1007/s12144-015-9374-2

Dörnyei, Z. (2009). The L2 motivational self system. In Z. Dörnyei & E. Ushioda (Eds.), *Motivation, language identity and the L2 self* (pp. 9-42). Bristol, UK: Multilingual Matters.

Duckworth, A. L., & Eskreis-Winkler, L. (2013). True grit. *The Observer*, *26*(4), 1-3.

Duckworth, A. L., Kirby, T. A., Tsukayama, E., Berstein, H., & Ericsson, K. A. (2011). Deliberate practice spells success: Why grittier competitors triumph at the National Spelling Bee. *Social Psychological and Personality Science*, *2*(2), 174-181. https://doi.org/10.1177/1948550610385872

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*(6), 1087. https://doi.org/10.1037/0022-3514.92.6.1087

Duckworth, A. L., & Quinn, P. D. (2009). Development and validation of the Short Grit Scale (Grit-S).

Journal of Personality Assessment, *91*(2), 166-174. https://doi.org/10.1080/00223890802634290

Duckworth, A. L., & Seligman, M. E. P. (2005). Self-discipline outdoes IQ in predicting academic performance of adolescents. *Pyschological Science*, *16*(12), 939-944. https://doi.org/10.1111/j.1467-9280.2005.01641.x

Fulcher, N. G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge.

Garland, R. (1991). The mid-point on a rating scale: Is it desirable? *Marketing Bulletin*, *2*, 66-70. https://doi.org/10.1.1.462.1083

Gladwell, M. (2008). *Outliers: The study of success*. New York, NY: Little, Brown and Company.

Hwang, M. H., Lim, H. J., & Ha, H. S. (2017). Effects of grit on the academic success of adult female students at Korean Open University. *Psychological Reports*. Advance online publication. https://doi.org/10.1177/0033294117734834

Kulas, J. T., & Stachowski, A. A. (2013). Respondent rationale for neither agreeing nor disagreeing: Person and item contributors to middle category endorsement intent on Likert personality indicators. *Journal of Research in Personality*, *47*(4), 254-262. https://doi.org/10.1016/j.jrp.2013.01.014

Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions*, *8*(3), 370.

Linacre, J. M. (2002). What do infit, outfit, mean-square, and standardized mean? *Rasch Measurement Transactions*, *16*(1), 878.

Linacre, J. M. (2011). WINSTEPS (Version 3.73). Retrieved from http://www.winsteps.com/

Linacre, J. M. (2013). *A user's guide to WINSTEPS*. Chicago, IL: MESA. Retrieved from http://www.winsteps.com/

MacIntyre, P. D., & Charos, C. (1996). Personality, attitudes, and affect as predictors of second language communication. *Journal of Language and Social Psychology*, *15*(1), 3-26. https://doi.org/10.1177/0261927X960151001

McNamara, T. F. (1991). Test dimensionality: IRT analysis of an ESP listening test. *Language Testing*, *8*(2), 139-159. https://doi.org/10.1177/026553229100800204

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge: Cambridge University Press.

Plonsky, L., & Oswald, F. L. (2014). How big is "big"? Interpreting effect sizes in L2 research. *Language Learning, 64*(4), 878-912. https://doi.org/ 10.1111/lang.12079

Proctor, C. P., Daley, S., Louick, R., Leider, C. M., & Gardner, G. L. (2014). How motivation and engagement predict reading comprehension among native English-speaking and English-learning middle school students with disabilities in a remedial reading curriculum. *Learning and Individual Differences*, *36*, 76-83. https://doi.org/10.1016/j.lindif.2014.10.014

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Denmarks Paedogogiske Institute.

Stevens, J. (2002). *Applied multivariate statistics for the social sciences* (4th ed.). Mahwah, NJ: Erlbaum.

Wolfe, E. W., & Smith, E. V., Jr. (2006). Instrument development tools and activities for measure validation using Rasch models: Part II-validation activities. *Journal of Applied Measurement*, *8*(2), 204-234.

Yashima, T. (2002). Willingness to communicate in a second language: The Japanese EFL context. *The Modern Language Journal*, *86*(1), 54-66. https://doi.org/10.1111/1540-4781.00136