OPEN ACCESS

Evaluating the psychometric properties of the Grit scale in Marine recruits using Rasch analysis

Iris Dijksma ,[1,2] M Stuiver,[1] C Lucas,[1] R Lindeboom[1]

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjmilitary-2021-001813).

[1]Epidemiology and Data Science, Master Evidence Based Practice in Health Care, Amsterdam UMC Locatie AMC, Amsterdam, The Netherlands
[2]Defense Health Care Organization, Royal Netherlands Army, Utrecht, The Netherlands

Correspondence to
Maj Iris Dijksma, Epidemiology and Data Science, Master Evidence Based Practice in Health Care, Amsterdam UMC Locatie AMC, 3584 AB Amsterdam, North Holland, The Netherlands; i.dijksma@amsterdamumc.nl

Received 17 February 2021
Accepted 16 September 2021

To cite: Dijksma I, Stuiver M, Lucas C, *et al*. *BMJ Mil Health* Epub ahead of print: [*please include* Day Month Year]. doi:10.1136/bmjmilitary-2021-001813

ABSTRACT

Introduction Successful completion of initial military training has been suggested to be predicted by physical abilities, cognitive abilities and non-cognitive abilities such as hardiness and grit. This study aimed to assess the psychometric properties of a Dutch version of a grit measurement scale: the NL-Grit scale.

Methods We assessed the factor structure, unidimensionality of the subscales, discriminative quality of the rating scale and investigated to what extend the items together can reliably measure the entire range of grit levels in Dutch Marine recruits. We used data of Marine recruit training platoons of the Royal Netherlands Marine Corps.

Results Principal component analysis reflected two subscales: 'consistency of interests' and 'perseverance of effort'. Rasch analysis confirmed the unidimensionality of the intended subscales. Rasch rating scale analysis indicated that the five-point response scale was not used as intended by respondents. Disordered rating scale categories were collapsed to obtain ordered rating categories. The item and person parameters (grit levels) largely overlapped, indicating that the item spread was sufficient for measuring the entire range of grit trait levels. However, larger gaps between item location parameters suggested a low discriminative capacity of the NL-Grit scale for respondents with trait levels within the gaps.

Conclusion Our evaluation of the NL-Grit scale suggests sound psychometric quality of the NL-Grit in Dutch Marine recruits. Reliability could be improved by adding items to fill the observed gaps in item content.

Key messages

- We assessed the psychometric properties of a Dutch version of the Grit scale: the NL-Grit scale.
- We assessed the discriminative quality of the rating scale and to what extent the entire range of grit levels in Dutch Marine recruits can be measured.
- Principal component analysis reflected two subscales: 'consistency of interests' and 'perseverance of effort'.
- The item spread was sufficient for measuring the entire range of grit trait levels.
- This study suggests sound psychometric quality of the NL-Grit in Dutch Marine recruits.

INTRODUCTION

Graduation from Marine recruit training requires both physical and mental toughness. While some people get discouraged in case of adversity, setbacks or failure, others seek for learning, progress and opportunities to persevere and follow through to achieve their goal.[1] Moreover, while some have the conviction that success is result-oriented and determined by innate personality traits, others believe that the amount of effort should be both the focus and the end itself.[2] Since mental fitness can be trained, much like physical fitness, early recognition of recruits' mental profile provides opportunities for timely optimisation of training and education to help recruits unlock their potential and prevent dropout from military training.

Successful completion of initial military training has been suggested to be predicted by physical abilities (ie, strength, stamina), cognitive abilities (ie, memory, attention, reasoning) and non-cognitive abilities (ie, effort, self-efficacy), based on a study in over 10 000 US Army cadets.[3] The term *non-cognitive* has become an inclusive term for traits not captured by cognitive ability and knowledge tests.[4] Non-cognitive abilities encompass traits as self-regulation, conscientiousness, problem-solving skills and grit.[4] Grit is defined as passion and perseverance for long-term goals of personal significance.[5] In the aforementioned study, the authors found that grittier—but not necessarily more cognitively or physically able—cadets were more likely to complete a six-week physically and mentally highly demanding initiation training (known as 'Beast Barracks') and then continue to their training.[3] This underscores the importance of measuring the level of grit pre-entry or at the arrival of military training, both to explain and predict the individual likelihood of successful completion, as well as to provide the opportunity of individualised training and mentoring.

Scales measuring grit were developed relying on the assumptions of classical test theory.[6] The internal consistency reliability (coefficient α) of the original English 12-item Grit scale (Grit-O), ranged from 0.77 to 0.85.[6] Modifications of this scale, such as a short version (Short Grit scale: Grit-S), and language modifications were also developed and, in some cases, validated by using Rasch analysis from the framework of item response theory.[7–11] Rasch analysis is a useful and, compared with classical test theory, more robust method to examine and test unidimensionality and reliability of scales that measure latent constructs, such as grit.[12] Factor analyses of the original and short-form versions provided evidence for a two-factor structure, namely 'consistency of interests'

and 'perseverance of effort'.[6 7 11] While 'perseverance of effort' refers to the quality of working hard and progressively towards a set goal despite adversity, 'consistency of interests' implies the persistence of focus and passion for a set goal.[13]

There is also a language modification of the Grit scale in Dutch available, from now on referred to as the NL-Grit scale. The psychometric properties of the scale in Dutch marine recruits, however, have not been examined, and no attempts at Rasch analysis have been made before. The objectives of this study were (1) to examine the factor structure of the NL-Grit scale, (2) to assess the unidimensionality of the subscales, (3) to assess the discriminative quality of the rating scale and (4) to assess to what extend the NL-Grit scale items together can reliably measure the entire range of grit levels in the target population.

METHODS

Participants

For this study, we used data of four consecutive platoons Marine recruit training of the Royal Netherlands Marine Corps (RNLMC), Rotterdam, the Netherlands, starting military training between 14 January 2019 and until 28 October 2019. The minimum age for employment is 17 years and 6 months, with a maximum age of 27 years and 11 months. Minimal height for employment in the RNLMC is 1.65 m and minimal weight is 65 kg. Both men and women could sign up.

Measurements

At pre-entry attendance—eight weeks before entering Marine recruit training—recruits undergo several physical tests, questionnaires and interviews. With the ultimate aim to study the association between the level of grit and dropout from Marine recruit training, we added the NL-Grit scale to the usual procedure. The NL-Grit scale contains 10 items, comprises two subscales and was derived from the Dutch translation of the book written by Duckworth and professionally translated by Henk Popken.[14 15] Recruits were asked to score the items on a five-point Likert scale to the extent to which they believed the statement applied to them, ranging from 'strongly agree' (= 'very much like me') to 'strongly disagree' (= 'not like me at all'). Two examples of the 10 items are: '*New ideas and new projects sometimes distract me from previous ones*', and: '*Setbacks don't discourage me. I don't give up easily*'. (see Table 1 for all 10 items.) To calculate the Grit score, all the points of the items are added up to a sum score and divided by the number of items—five per subscale. Five positively worded items constituting the 'perseverance of effort' subscale (ie, like the second item example) were rescored so that higher scores indicate higher grit levels for both subscales. Thus, the maximum score is five (extremely gritty), and the lowest possible score is one (not gritty at all).

Statistical analysis

For statistical analysis, R V.3.6.1 was used (packages psych and eRm).[16] *Please see online supplemental file 2, for the full statistical analysis paragraph.*

Factor structure

The intended factor structure of the NL-Grit scale was examined by a principal component analysis (PCA) with Varimax rotation ('structural validity'). We used a parallel analysis scree plot to determine the number of factors to retain.[17] We calculated internal consistency reliability (coefficient alpha, also known as Cronbach's α), as well as item means and SD of the two subscales separately.

Unidimensionality of the subscales

We assessed the unidimensionality of both subscales by examining their fit to the Rasch model. For that goal, we dichotomised the items on their mean value. The Rasch model is a probabilistic model that describes the interaction of respondents with the questionnaire items and is governed by two parameters: item difficulty and person ability. Fit to the Rasch model of individual items was examined by infit and outfit statistics.[18] The overall fit of the items together to the unidimensional Rasch model was tested using Andersen's likelihood ratio test.[19] A p

Table 1 NL-Grit scale with subscales, coefficient α, means and SD, and factor loadings

Subscale/item	Coefficient α (95% CI)	Mean	SD	Factor loadings
Consistency of interests	0.68 (0.62 to 0.73)	3.9	0.62	**Factor 1**
Item 1	*New ideas and new projects sometimes distract me from previous ones*. Nieuwe ideeën en projecten leiden me soms af van vorige.	3.4	1.00	0.534
Item 3	*I often set a goal but later choose to pursue a different one*. Ik stel mezelf vaak een doel, maar kies later voor een ander doel.	4.0	0.99	0.584
Item 5	*I have difficulty maintaining my focus on projects that take more than a few months to complete*. Ik heb moeite om me te blijven concentreren op dingen die langer dan een paar maanden duren.	4.1	0.90	0.683
Item 7	*My interests change from year to year*. Mijn interesses verschillen van jaar tot jaar.	3.7	1.01	0.689
Item 9	*I have been obsessed with a certain idea or project for a short time but later lost interest*. Ik ben korte tijd geobsedeerd door iets nieuws, maar verlies snel mijn belangstelling.	4.2	0.77	0.793
Perseverance of effort	0.67 (0.61 to 0.72)	4.5	0.39	**Factor 2**
Item 2	*Setbacks don't discourage me. I don't give up easily*. Tegenslagen ontmoedigen me niet. Ik geef niet snel op.	4.6	0.64	0.681
Item 4	*I am a hard worker*. Ik ben een harde werker.	4.7	0.49	0.666
Item 6	*I finish whatever I begin*. Ik maak af waar ik aan begin.	4.6	0.61	0.658
Item 8	*I am diligent. I never give up*. Ik ben toegewijd. Ik geef nooit op.	4.5	0.62	0.723
Item 10	*I have overcome setbacks to conquer an important challenge*. Ik heb tegenslagen overwonnen bij het aangaan van belangrijke uitdagingen.	4.4	0.64	0.496

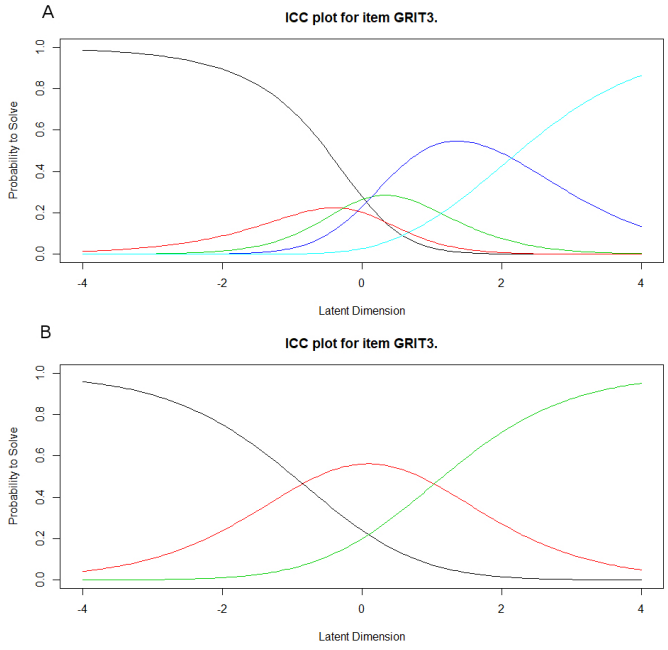

Figure 1 Category probability curves of disordered (A) and ordered (B) in item response function plots. (A and B) Item category curves (ICCs) indicate the probability of selecting an item category (y-axis) as a function of grit trait level in logit units (x-axis). (A) A disordered rating scale where the red and green categories both have a low probability to be selected by respondents, regardless trait level. (B) Ordered item category curves after collapsing the red and green categories with the adjacent lower category. Now, at each trait level, a single category is most probable. A category threshold difficulty measure is the point on the latent scale where adjacent categories are equally probable.

value >0.05 indicates that the Rasch model is accepted for an item set ('Rasch homogeneous').

Discriminative quality of the rating scale

To investigate the discriminative quality of the item rating scale, we performed Rasch rating scale analysis using the partial credit model.[12] We plotted item category response curves to examine whether the item category measures ('threshold difficulties') were ordered (ie, that at each point on the latent Grit scale, a single-item category score is the most probable category). In that case, threshold difficulties should increase when moving from lower to higher categories. Disordered rating scale categories and cells including <10 observations were collapsed in such a way that an ordered and logical rating category emerged (Figure 1A,B).

Comprehensiveness

To assess to what extend NL-Grit scale items together can reliably measure the entire range of grit levels present in the sample, we plotted the person-item map based on the amended item scoring. The person-item map displays the location of person measures and item category difficulty, respectively, along the same latent dimension. We verified whether there was sufficient overlap between item measures and NL-Grit person measures. Furthermore, we examined whether there were substantial gaps between the item category measures along the total range of the Grit scale, indicating less discriminative capacity within that range.

Table 2 Sample characteristics, n=354

Variable	
Age (years), mean±SD	21±2.4
Gender, male	100%
Height (m), mean±SD	1.81±6.5
Weight (kg), mean±SD	77.7±8.7
Educational level, n (%)	
University education	2 (0.6%)
Higher professional education	10 (2.8%)
Secondary vocational education	223 (63%)
Secondary education	119 (34%)

RESULTS

Data were available from 354 Marine recruits of whom 27 had missing data, resulting in n=327 complete observations for analysis. Sample characteristics are presented in Table 2. Excluded individuals did not differ in demographics from included individuals.

Parallel analysis scree plot confirmed the intended two-factor structure of the NL-Grit. PCA indicated that all items loaded predominantly on their intended scale; there were no cross-loaders. Cumulative explained variance by the factors was 45%. Internal consistency coefficient α of the 'consistency of interests' and 'perseverance of effort' subscales were 0.68 and 0.67, respectively. There were no items harming the reliability of the subscales.

Item infit and outfit statistics for each subscale fell within the acceptable range of 0.5–1.5 (please see online supplemental material, appendix 1). Anderson's likelihood ratio test p-value was p=0.255 for 'consistency of interests' and p=0.004 for 'perseverance of effort', indicating that the Rasch model is accepted only for the NL-Grit subscale 'consistency of interests'. For the 'perseverance of effort' subscale, the items 8 and 10 had an estimated item difficulty parameter that was notably different between low and high scoring respondents (Figure 2A,B). PCA of the residuals that remained after Rasch analysis showed no other significant dimension present for both subscales. Eigenvalues of the first two components were below 1.5.

Rasch rating scale analysis indicated that the five-point response scale was not used as intended by respondents. Items in both subscales had disordered item category measures. Before collapsing categories, 7 out of 10 items (except for items 1, 7 and 9) had <10 observations in cells and/or disordered category thresholds. Disordered rating scale categories were collapsed into three levels (0=not like me at all/not much like me/neutral, 1=mostly like me, 2=very much like me), which resulted in ordered thresholds for all items (see Table 3 and Figure 1A,B).

Figure 3A,B shows the item and person parameters of the amended NL-Grit scale. Regarding the subscale 'consistency of interests', the distribution of person scores was approximately normally distributed. There was sufficient overlap between item measures and person measures. Regarding the subscale 'perseverance of effort', the distribution of person scores was skewed-left, but the item parameters were also overlapping with most sum scores, indicating that the item spread was sufficient for measuring the entire range of grit trait levels. However, there was a large gap in item content (gap in small vertical lines below the grey bars) between zero and one logit, indicating a low discriminative capacity of the NL-Grit for these respondents.

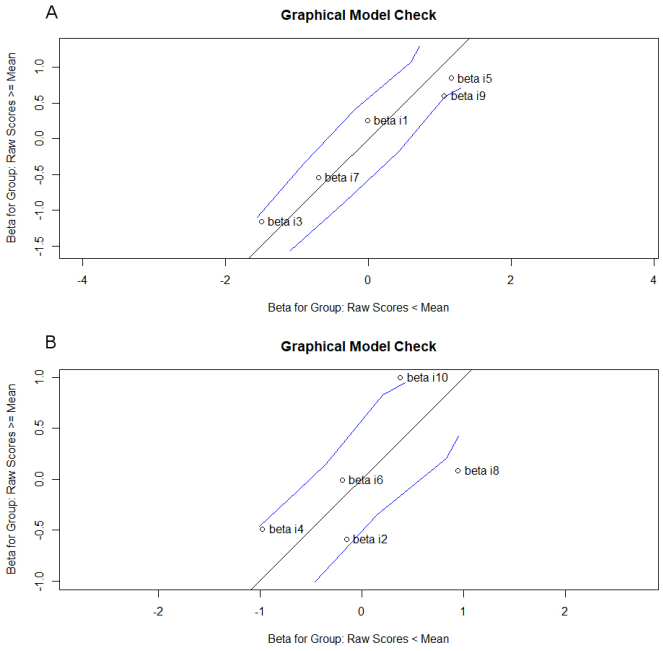

Figure 2 Graphical model goodness of fit plot. (A) 'Consistency of interests', (B) 'Perseverance of effort'. (A and B) Anderson's likelihood ratio test of item fit to the Rasch model. The dichotomised items 8 and 10 had an estimated item difficulty parameter that was notably different between low and high scoring respondents (below and above the mean score).

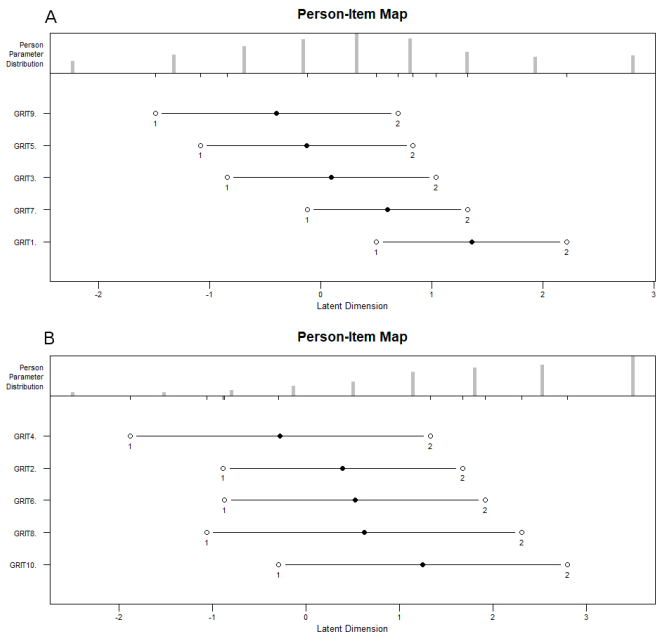

Figure 3 Person-Item maps of the subscales of the NL-Grit scale, Partial Credit Model, collapsed item categories. (A) 'Consistency of interests', (B) 'Perseverance of effort'. The grey bars represent frequencies for sum score groups from low (left) to high scores, excluding minimum and maximum scores. There was sufficient overlap between item measures and person measures.

DISCUSSION

Our study evaluated the psychometric properties of the NL-Grit scale in Marine recruits. PCA supported the intended two-factor structure of the instrument. Rasch analysis confirmed the unidimensionality of the subscales although two items of the 'perseverance of effort' subscale functioned differentially for low and high scoring respondents. Rasch rating scale analysis indicated that the five-point response scale was not used as intended by respondents and an amended scoring system is proposed to improve the reliability of a NL-Grit score. The item spread was sufficient for measuring the entire range of grit trait levels in the population, indicated by the overlap between item measures and person measures.

Table 3 Item difficulties and item thresholds per subscale, three response options

	Difficulty	Threshold 1	Threshold 2
Consistency of interests			
Item 1	1.36	0.50	2.21
Item 3	0.10	−0.85	1.04
Item 5	−0.13	−1.09	0.83
Item 7	0.60	−0.12	1.32
Item 9	−0.40	−1.49	0.69
Perseverance of effort			
Item 2	0.39	−0.89	1.67
Item 4	−0.28	−1.89	1.33
Item 6	0.52	−0.87	1.91
Item 8	0.62	−1.06	2.30
Item 10	1.25	−0.30	2.79

Thresholds represent the location on the logit scale of the latent dimension were each adjacent response category becomes the highest probability to solve—or to be chosen. Item difficulties are the average threshold, which also can be seen in the person-item maps (Figure 3A,B).

In our sample, we observed a positive skew towards the favourable end, particularly on the 'perseverance of effort' subscale, for example, (strongly) agree to '*I am a hard worker*'. This could, partly, explain the relatively low α coefficients in this sample. In other military studies, using the original 12-item scale, reported total scale means were lower at 3.61 and 3.75.[6 20] The 'true level of grit' is, like any other latent construct, unknown, also for the respondents themselves. Self-discrepancy theory states that individuals compare their 'actual' self to internalised standards, known as the 'ideal vs ought self' discrepancy.[21] Particularly individuals who feel the need to prove themselves may unintentionally report inflated levels of grit. Furthermore, samples that have been selected on prior performance, such as in our sample in which all respondents have already passed physical, medical and mental screening, may yield some restriction on the range of grit levels. To conclude, we cannot fully account whether our respondents indeed were more gritty than respondents in aforementioned studies, or just reported inflated grit levels, or that subtle changes due to the Dutch language modification of the scale are responsible for the higher mean scores in our sample.

PCA supported the intended two-factor subscale structure of the NL-Grit scale in our sample of Marine recruits. Additionally, analysis of the Rasch residuals of both subscales showed no other significant dimension governing responses to the items. However, the subscale 'perseverance of effort' contained two questionable items. Item difficulties for low and high grit level subjects for item 8 and item 10 differed with 0.9 and 0.6 logit, respectively. Considering the half-logit rule for meaningful 'differential item functioning', these differences may be considered substantial.[22] To determine whether these items need to be revised or removed from the scale, further analysis regarding differential item functioning is required.[23]

With regard to investigating grit in relation to successful completion of military training, it has been suggested that

perseverance is a better predictor of performance than either consistency of effort or overall grit.[13] Therefore, and because of the confirmed two-factor structure of the scale, for explanatory research and recruit selection purposes, we suggest to work with subscale scores rather than using the overall sum score.

Limitations and implications

Several limitations of this study are worth mentioning. First, given that all participants in this study were male with an average age of 21 years old, internal consistency reliability, rating scale ordering and item and person measures are not necessarily generalisable to female military service members. However, to date, female Marine recruits are very scarce in the RNLMC. Second, the original 12-item Grit scale and Grit-S are scored on a five-point Likert using 'not like me at all' to 'very much like me'. In the Dutch version, 'strongly agree' (~ 'very much like me') to 'strongly disagree' (~ 'not like me at all') is used. We preserved the professionally translated and published version of the Dutch language modification of the Grit scale, instead of making further modifications to the scale. However, this may have had consequences to how the statements and response scale options were interpreted and scored by respondents. However, to enable (international) comparisons on grit levels of recruits, future research should focus on the measurement invariance of the Grit scale to examine whether items responses are exclusively caused by grit level and not by subtle translations effects or the specifics of a sample.

CONCLUSION

Our evaluation of the NL-Grit scale suggests collapsing item response categories in such a way that the variation in sum scores of the subscales represents true variation in trait levels. Reliability could be improved by adding items to fill the observed gaps in item content, which proposes going back to the phase of item development and generation. This study contributed to the cross-cultural description of the psychometric properties and soundness the NL-Grit scale. Future research could build on our evaluation by investigating criterion related validity, for example, the usefulness of the scale in screening recruits to predict successful completion of military training controlling for physical abilities and injuries.

Acknowledgements Special thanks to the participants. We thankfully acknowledge Marc Duineveld from the MOC Rotterdam for helping us gathering the data. We also thank Remco Blom, Surgeon General of the Netherlands Armed Forces, for his support of this research project.

Contributors ID, MS, CL and RL designed the study; ID collected and prepared the data; ID and RL analysed the data; ID, MS, CL and RL wrote the article.

Funding The authors have not declared a specific grant for this research from any funding agency in the public, commercial or not-for-profit sectors.

Disclaimer The opinions or assertions contained herein are the private views of the authors and are not to be construed as official or reflecting the views of the Department of Defense or Dutch government.

Competing interests None declared.

Patient consent for publication Consent obtained directly from patient(s).

Ethics approval The Medical Research Ethical Committee of the University Medical Centre Utrecht, the Netherlands, confirmed that the Medical Research Involving Human Subjects Act does not apply to this study (protocol number: 18-790/C) and waived the study from formal approval. Participants all signed informed consent, authorising the use of their anonymised data for scientific purposes.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement No data are available.

ORCID iD

Iris Dijksma http://orcid.org/0000-0002-3372-5070

REFERENCES

1 Burnette JL, Knouse LE, Vavra DT, *et al*. Growth mindsets and psychological distress: a meta-analysis. *Clin Psychol Rev* 2020;77:101816.
2 Dweck CS, Yeager DS. Mindsets: a view from two eras. *Perspect Psychol Sci* 2019;14:481–96.
3 Duckworth AL, Quirk A, Gallop R, *et al*. Cognitive and noncognitive predictors of success. *Proc Natl Acad Sci U S A* 2019;116:23499–504.
4 West MR, Kraft MA, Finn AS. Promise and paradox: measuring students' non-cognitive skills and the impact of schooling. *Educ Eval Policy Anal* 2016;38:148–70.
5 Tyumeneva Y, Kardanova E, Kuzmina Y. Grit: two related but independent constructs instead of one. Evidence from item response theory. *Eur J Psychol Assess* 2017.
6 Duckworth AL, Peterson C, Matthews MD, *et al*. Grit: perseverance and passion for long-term goals. *J Pers Soc Psychol* 2007;92:1087–101.
7 Duckworth AL, Quinn PD. Development and validation of the short grit scale (grit-s). *J Pers Assess* 2009;91:166–74.
8 Arco-Tirado JL, Fernández-Martín FD, Hoyle RH. Development and validation of a Spanish version of the Grit-S scale. *Front Psychol* 2018;9:96.
9 Nishikawa K, Okugami S, Amemiya T. Development of the Japanese short Grit scale (Grit-S). *Japanese J Pers* 2015;24:167–9.
10 Areepattamannil S, Khine MS. Evaluating the psychometric properties of the original Grit scale using Rasch analysis in an Arab adolescent sample. *J Psychoeduc Assess* 2018;36:856–62.
11 Tyumeneva Y, Kuzmina J, Kardanova E. Irt analysis and validation of the Grit scale: a Russian investigation. *SSRN J* 2014.
12 Wright BD, Masters GN, Analysis RS. *Rating scale analysis*. Chicago, IL: MESA Press, 1982. https://research.acer.edu.au/measurement/2/
13 Credé M, Tynan MC, Harms PD. Much ado about grit: a meta-analytic synthesis of the grit literature. *J Pers Soc Psychol* 2017;113:492–511.
14 Duckworth A. De Gritfactor - De Kracht van Passie En Doorzettingsvermogen. *Translated Lev* 2016.
15 Duckworth A. *Grit: the power of passion and perseverance*. Scribner/Simon & Schuster, 2016. ISBN: 978-1-5011-1110-5.
16 R Core Team. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2015. http://www.r-project.org/
17 Field A, Miles J, Field Z. Exploratory Factor Analysis. In: *Discovering statistics using R*. London, England: AGE Publications, 2012: 749–811.
18 Linacre JM. What Do Infit and Outfit, Mean-Square and Standardized Mean?In: *Rasch measurement transactions*. 16, 2002.
19 Andersen EB. A goodness of fit test for the Rasch model. *Psychometrika* 1973;38:123–40.
20 Lovering ME, Heaton KJ, Banderet LE, *et al*. Psychological and physical characteristics of U.S. marine recruits. *Military Psychology* 2015;27:261–75.
21 Higgins ET. Self-discrepancy theory: what patterns of self-beliefs cause people to suffer. *Adv Exp Soc Psychol* 1989;22:93–136.
22 Draba RE. *The identification and interpretation of item bias (research memorandum No. 25)*. Chicago, IL: The University of Chicago, Department of Education, Education Statistics Laboratory, 1977.
23 Martinková P, Drabinová A, Liaw Y-L, *et al*. Checking equity: why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE Life Sci Educ* 2017;16:rm2.

Online supplemental material, Appendix 1: *Fit to the Rach model, per subscale*

	Outfit MSQ	Infit MSQ
Consistency of interests		
Item 1	1.041	1.024
Item 3	1.153	0.966
Item 5	0.789	0.838
Item 7	1.002	0.961
Item 9	0.745	0.848
Perseverance of effort		
Item 2	0.875	0.888
Item 4	1.017	1.041
Item 6	0.996	1.005
Item 8	0.757	0.808
Item 10	1.116	1.117

Legend: *MSQ= Mean sum of squares. Infit and outfit statistics within the range of 0.5 - 1.5 indicate that an item effectively contributes to the latent construct the instrument purports to measure.*

Statistical analysis

For statistical analysis, R version 3.6.1 was used (packages psych and eRm).[1]

Factor structure

The intended factor structure of the NL-Grit scale was examined by a principal component analysis (PCA) with Varimax rotation ("structural validity"). We used a parallel analysis scree plot to determine the number of factors to retain.[2] Factor loadings >0.40 for individual items were considered satisfactory. We calculated internal consistency reliability (coefficient alpha, also known as Cronbach's α), as well as item means and standard deviations of the two subscales separately.

Unidimensionality of the subscales

We assessed the unidimensionality of both subscales by examining their fit to the Rasch model. For that goal, we dichotomized the items on their mean value. The Rasch model is a probabilistic model which describes the interaction of respondents with the questionnaire items and is governed by two parameters: item difficulty and person ability. Fit to the Rasch model of individual items was examined by infit and outfit statistics.[4] Outfit is based on a sum of squared standardized residuals between observed and expected responses, infit is an information-weighted form of outfit.[5] Infit and outfit statistics within the range of 0.5 to 1.5 indicate that an item effectively contributes to the latent construct the instrument purports to measure.[4] The overall fit of the items together to the unidimensional Rasch model was tested using Andersen's likelihood ratio test.[6] The rationale of the likelihood ratio test is that if the Rasch model holds in the population, equivalent item parameter estimates should be obtained in low and high scoring subgroups (below and above the mean NL-Grit score, respectively). A p-value >0.05 indicates that the Rasch model is accepted for an item set ("Rasch homogeneous"). Unidimensionality was further examined by a PCA of the residuals that remain after Rasch analysis. The rationale here is that there is only one dimension, therefore residuals do not

contain other significant dimensions. Reference values for the first two eigenvalues of the PCA with 100 observations are 1.8 and 1.6 and are 1.4 and 1.3 with 500 respondents, respectively.[7]

Discriminative quality of the rating scale

To investigate the discriminative quality of the item rating scale, we performed Rasch rating scale analysis using the partial credit model (PCM).[8] We plotted item category response curves to examine whether the item category measures ("threshold difficulties") were ordered (i.e., that at each point on the latent grit scale, a single item category score is the most probable category). In that case, threshold difficulties should increase when moving from lower to higher categories. With a disordered rating scale, on the other hand, this is not the case and item measures are reversed (i.e. lower item categories having an item measure exceeding that of an adjacent higher item category). Disordered rating scale categories and cells including <10 observations were collapsed in such a way that an ordered and logical rating category emerged (*Figure 1: A and B in the article*).

Comprehensiveness

To assess to what extend NL-Grit scale items together can reliably measure the entire range of grit levels present in the sample, we plotted the person-item map based on the amended item scoring. The person-item map displays the location of person measures and item category difficulty respectively along the same latent dimension. The person item map relates the distribution of estimated person measures (grit trait levels) to the distribution of item category measures on a common log-odds unit scale. By using logit value, person-item maps are able to show simultaneously the hierarchies of both person and items at the same scale. We verified whether there was sufficient overlap between item measures and NL-Grit person measures. Furthermore, we examined whether there were substantial gaps between the item category measures along the total range of the grit scale, indicating less discriminative capacity within that range.

References

1. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2015. http://www.r-project.org/.
2. Field A, Miles J, Field Z. Exploratory Factor Analysis. In: *Discovering Statistics Using R*. London, England: AGE Publications.; 2012:749-811.
3. Hayes AF, Coutts JJ. Use Omega Rather than Cronbach's Alpha for Estimating Reliability. But…. *Communication Methods and Measures*. 2020;14(1):1-24. doi:10.1080/19312458.2020.1718629
4. Linacre JM. *What Do Infit and Outfit, Mean-Square and Standardized Mean? Rasch Measurement Transactions 16:2*.; 2002.
5. Masters GN, Wright BD. The Partial Credit Model BT - Handbook of Modern Item Response Theory. In: van der Linden WJ, Hambleton RK, eds. New York, NY: Springer New York; 1997:101-121. doi:10.1007/978-1-4757-2691-6_6
6. Andersen EB. A goodness of fit test for the rasch model. *Psychometrika*. 1973;38(1):123-140. doi:10.1007/BF02291180
7. Brentani E, Golia S. Unidimensionality in the Rasch model: how to detect and interpret. *Statistica; Vol 67, No 3 (2007)DO - 106092/issn1973-2201/3508* . September 2007. https://rivista-statistica.unibo.it/article/view/3508.
8. Wright BD, Masters GN. *Rating Scale Analysis*. Chicago, IL: MESA Press.; 1982. https://research.acer.edu.au/measurement/2/.

Online supplemental material, Appendix 1: *Fit to the Rach model, per subscale*

	Outfit MSQ	Infit MSQ
Consistency of interests		
Item 1	1.041	1.024
Item 3	1.153	0.966
Item 5	0.789	0.838
Item 7	1.002	0.961
Item 9	0.745	0.848
Perseverance of effort		
Item 2	0.875	0.888
Item 4	1.017	1.041
Item 6	0.996	1.005
Item 8	0.757	0.808
Item 10	1.116	1.117

Legend: *MSQ= Mean sum of squares. Infit and outfit statistics within the range of 0.5 - 1.5 indicate that an item effectively contributes to the latent construct the instrument purports to measure.*

Statistical analysis

For statistical analysis, R version 3.6.1 was used (packages psych and eRm).[1]

Factor structure

The intended factor structure of the NL-Grit scale was examined by a principal component analysis (PCA) with Varimax rotation ("structural validity"). We used a parallel analysis scree plot to determine the number of factors to retain.[2] Factor loadings >0.40 for individual items were considered satisfactory. We calculated internal consistency reliability (coefficient alpha, also known as Cronbach's α), as well as item means and standard deviations of the two subscales separately.

Unidimensionality of the subscales

We assessed the unidimensionality of both subscales by examining their fit to the Rasch model. For that goal, we dichotomized the items on their mean value. The Rasch model is a probabilistic model which describes the interaction of respondents with the questionnaire items and is governed by two parameters: item difficulty and person ability. Fit to the Rasch model of individual items was examined by infit and outfit statistics.[4] Outfit is based on a sum of squared standardized residuals between observed and expected responses, infit is an information-weighted form of outfit.[5] Infit and outfit statistics within the range of 0.5 to 1.5 indicate that an item effectively contributes to the latent construct the instrument purports to measure.[4] The overall fit of the items together to the unidimensional Rasch model was tested using Andersen's likelihood ratio test.[6] The rationale of the likelihood ratio test is that if the Rasch model holds in the population, equivalent item parameter estimates should be obtained in low and high scoring subgroups (below and above the mean NL-Grit score, respectively). A p-value >0.05 indicates that the Rasch model is accepted for an item set ("Rasch homogeneous"). Unidimensionality was further examined by a PCA of the residuals that remain after Rasch analysis. The rationale here is that there is only one dimension, therefore residuals do not

contain other significant dimensions. Reference values for the first two eigenvalues of the PCA with 100 observations are 1.8 and 1.6 and are 1.4 and 1.3 with 500 respondents, respectively.[7]

Discriminative quality of the rating scale

To investigate the discriminative quality of the item rating scale, we performed Rasch rating scale analysis using the partial credit model (PCM).[8] We plotted item category response curves to examine whether the item category measures ("threshold difficulties") were ordered (i.e., that at each point on the latent grit scale, a single item category score is the most probable category). In that case, threshold difficulties should increase when moving from lower to higher categories. With a disordered rating scale, on the other hand, this is not the case and item measures are reversed (i.e. lower item categories having an item measure exceeding that of an adjacent higher item category). Disordered rating scale categories and cells including <10 observations were collapsed in such a way that an ordered and logical rating category emerged (*Figure 1: A and B in the article*).

Comprehensiveness

To assess to what extend NL-Grit scale items together can reliably measure the entire range of grit levels present in the sample, we plotted the person-item map based on the amended item scoring. The person-item map displays the location of person measures and item category difficulty respectively along the same latent dimension. The person item map relates the distribution of estimated person measures (grit trait levels) to the distribution of item category measures on a common log-odds unit scale. By using logit value, person-item maps are able to show simultaneously the hierarchies of both person and items at the same scale. We verified whether there was sufficient overlap between item measures and NL-Grit person measures. Furthermore, we examined whether there were substantial gaps between the item category measures along the total range of the grit scale, indicating less discriminative capacity within that range.

References

1. R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. 2015. http://www.r-project.org/.
2. Field A, Miles J, Field Z. Exploratory Factor Analysis. In: *Discovering Statistics Using R*. London, England: AGE Publications.; 2012:749-811.
3. Hayes AF, Coutts JJ. Use Omega Rather than Cronbach's Alpha for Estimating Reliability. But…. *Communication Methods and Measures*. 2020;14(1):1-24. doi:10.1080/19312458.2020.1718629
4. Linacre JM. *What Do Infit and Outfit, Mean-Square and Standardized Mean? Rasch Measurement Transactions 16:2*.; 2002.
5. Masters GN, Wright BD. The Partial Credit Model BT - Handbook of Modern Item Response Theory. In: van der Linden WJ, Hambleton RK, eds. New York, NY: Springer New York; 1997:101-121. doi:10.1007/978-1-4757-2691-6_6
6. Andersen EB. A goodness of fit test for the rasch model. *Psychometrika*. 1973;38(1):123-140. doi:10.1007/BF02291180
7. Brentani E, Golia S. Unidimensionality in the Rasch model: how to detect and interpret. *Statistica; Vol 67, No 3 (2007)DO - 106092/issn1973-2201/3508* . September 2007. https://rivista-statistica.unibo.it/article/view/3508.
8. Wright BD, Masters GN. *Rating Scale Analysis*. Chicago, IL: MESA Press.; 1982. https://research.acer.edu.au/measurement/2/.