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 Lindeboom1

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 hardness 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.

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’ 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.\textsuperscript{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.\textsuperscript{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).\textsuperscript{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.\textsuperscript{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.\textsuperscript{18} The overall fit of the items together to the unidimensional Rasch model was tested using Andersen’s likelihood ratio test.\textsuperscript{19} A p

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

<table>
<thead>
<tr>
<th>Subscale/item</th>
<th>Coefficient α (95% CI)</th>
<th>Mean</th>
<th>SD</th>
<th>Factor loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Consistency of interests</strong></td>
<td>0.68 (0.62 to 0.73)</td>
<td>3.9</td>
<td>0.62</td>
<td>Factor 1</td>
</tr>
<tr>
<td>Item 1: New ideas and new projects sometimes distract me from previous ones.</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 2: I often set a goal but later choose to pursue a different one.</td>
<td>3.4</td>
<td>1.00</td>
<td>0.534</td>
<td></td>
</tr>
<tr>
<td>Item 3: My interests change from year to year.</td>
<td>4.0</td>
<td>0.99</td>
<td>0.584</td>
<td></td>
</tr>
<tr>
<td>Item 5: I have been obsessed with a certain idea or project for a short time but later lost interest.</td>
<td>4.1</td>
<td>0.90</td>
<td>0.683</td>
<td></td>
</tr>
<tr>
<td>Item 7: I have difficulty maintaining my focus on projects that take more than a few months to complete.</td>
<td>3.7</td>
<td>1.01</td>
<td>0.689</td>
<td></td>
</tr>
<tr>
<td><strong>Setbacks don’t discourage me. I don’t give up easily.</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Item 9: I am a hard worker.</td>
<td>4.2</td>
<td>0.77</td>
<td>0.793</td>
<td></td>
</tr>
<tr>
<td>Item 10: I am diligent. I never give up.</td>
<td>4.5</td>
<td>0.39</td>
<td>Factor 2</td>
<td></td>
</tr>
<tr>
<td><strong>Perseverance of effort</strong></td>
<td>0.67 (0.61 to 0.72)</td>
<td>4.6</td>
<td>0.64</td>
<td>0.681</td>
</tr>
<tr>
<td>Item 1: I am a hard worker.</td>
<td>4.7</td>
<td>0.49</td>
<td>0.666</td>
<td></td>
</tr>
<tr>
<td>Item 6: I finish whatever I begin.</td>
<td>4.6</td>
<td>0.61</td>
<td>0.658</td>
<td></td>
</tr>
<tr>
<td>Item 8: I am diligent. I never give up.</td>
<td>4.5</td>
<td>0.62</td>
<td>0.723</td>
<td></td>
</tr>
<tr>
<td>Item 10: I have overcome setbacks to conquer an important challenge.</td>
<td>4.4</td>
<td>0.64</td>
<td>0.496</td>
<td></td>
</tr>
</tbody>
</table>

Please see online supplemental file 2, for the full statistical analysis paragraph.
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.\(^\text{12}\) 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-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 extent 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.

**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 $\alpha$ 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.
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.

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

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

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

Thresholds represent the location on the logit scale of the latent dimension where 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).
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. 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 modulation 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 item 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 Duneveld from the MOC Rotterdam for helping us gathering the data. We also thank Rencio 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/I) 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.

Supplemental material

This content has been supplied by the author(s). It has not been vetted by BMJ Publishing Group Limited (BMJ) and may not have been peer-reviewed. Any opinions or recommendations discussed are solely those of the author(s) and are not endorsed by BMJ. BMJ disclaims all liability and responsibility arising from any reliance placed on the content. Where the content includes any translated material, BMJ does not warrant the accuracy and reliability of the translations (including but not limited to local regulations, clinical guidelines, terminology, drug names and drug dosages), and is not responsible for any error and/or omissions arising from translation and adaptation or otherwise.

Open access

This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/

ORCID iD

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

REFERENCES

18 Linacre JM. What Do Infit and Outfit, Mean-Square and Standardized Mean? In: Rasch measurement transactions. 16, 2002.
22 Dabná BE. The identification and interpretation of item bias (research memorandum No. 25). Chicago, IL: The University of Chicago, Department of Education, Education Statistics Laboratory, 1977.