Psikohumaniora: Jurnal Penelitian Psikologi
Vol 10, No 1 (2025): 1–20. DOI: 10.21580/pjpp.v10i1.23214
ISSN 2527-7456 (online); 2502-9363 (printed)

Psychometric properties of the 18-item Indonesian Mental Toughness Questionnaire using the Rasch model and Machine Learning

Ananta Yudiarso [1*], **Ista Wirya Ardhiani** [1], **Roy Surya** [1], **Ferry Yohannes Watimena** [2], **Mami Kanzaki** [3]

[1]Department of Psychology, Faculty of Psychology, Universitas Surabaya, Surabaya – Indonesia; [2]Department of Sports Coaching, Faculty of Sport Science, Universitas Negeri Jakarta, Jakarta – Indonesia; [3]Faculty of Education, Kyoto University of Education, Kyoto – Japan

Abstract: The psychometric properties of the Indonesian version of the 18-item Mental Toughness Questionnaire (MTQ-18) remain vague. This study uses the Rasch model to elucidate these properties. In addition, boosting classification was adopted to assess the predictive validity of athletes' achievement. The sample size comprised 400 athletes. According to the Martin-Loef likelihood-ratio test = 482, p = 1.0 and factor analysis of the Rasch residuals, the questionnaire tends to make unidimensional assumptions. The MADaQ3 = 0.074 shows the overall tendency of local independency across all items, with the majority clustered in moderate to low-level measures. Q11, Q15, and Q18 were clearly identified as showing gender bias, with significant effect sizes. According to the boosting classification, the performance between national vs no achievement (*F1* = 0.7, *AUC* = 0.56) and international vs no achievement (*F1* = 0.62, *AUC* = 0.58) was flagged as unsatisfactory predictive performance. In conclusion, the abridged questionnaire is not preferable for determining an individual's future performance or achievement. Future studies are needed to develop a better version that is more unimpeded by gender bias, and to resolve the variability of the items.

Keywords: differential item functioning; gradient boosting classification; Mental Toughness Questionnaire; rating scale model; Wright Map

To cite this article (APA Style):
Yudiarso, A., Ardhiani, I. W., Surya, R., Watimena, F. Y., & Kanzki, M. (2025). Psychometric properties of the 18-item Indonesian Mental Toughness Questionnaire using the Rasch model and Machine Learning. *Psikohumaniora: Jurnal Penelitian Psikologi*, *10*(1), 1-20. https://doi.org/10.21580/pjpp.v10i1.23214

*__Corresponding Author:__ Ananta Yudiarso (ananta@staff.ubaya.ac.id), Faculty of Psychology, Universitas Surabaya, Jl. Tenggilis Mejoyo, Kali Rungkut, Surabaya, Jawa Timur 60293 – Indonesia.

Introduction

The importance of mental toughness (MT) is clearly seen in various contexts such as sports, education or office workplaces. A probable explanation for this is because mental toughness is considered to be an essential resource in achieving optimal mental health and performance (Gerber et al., 2013; Lin et al., 2017; Papageorgiou et al., 2019). In addition, the use of MT in sports has attracted research interest, as shown in the study of Hsieh et al. (2024), who used a systematic review approach to explore the implications of MT for athletes' performance. Regarding the variability of the questionnaires, they highlighted the need to use updated definitions of MT and performance. Another study by Nicholls et al. (2016) by the quality of coaching and task-involving climate.

The original version of MTQ was developed using the 4Cs model (challenge, commitment, control, and confidence). The instrument aimed to evaluate an individual's level of mental toughness and consisted of 48 items. Clough et al. (2002) showed that mental toughness is a factor that influences peoples' friendliness and outgoing nature, helping them be calm and relaxed in competitive situations. According to Kobasa (1979), challenges reflect the degree to which a person sees obstacles and trials as opportunities for personal growth, while commitment characterizes the determination and ability to complete a task successfully. Control shows a person's level of confidence in their ability to influence their course of life, and confidence reflects self-confidence in one's abilities, especially in completing tasks.

In particular, the psychometric properties of the MTQ-48 have given rise to debate regarding the dimensionality of the construct. Based on their second study, Gucciardi et al. (2015) proposed the unidimensional idea of mental toughness rather than the 4Cs model, regarding the indication of overlap between the scales when treated as a multidimensional test. On the contrary, Perry et al. (2013) and Perry et al. (2021) suggested a multidimensional model, but noted that MT could also be considered as an umbrella representing the general trait of associated constructs that influence performance.

In the development of the original questionnaire also gained some attention for the short version generation, with fewer items. Kawabata et al. (2021) created two abridged versions of the MTQ, the short MTQ (S-MTQ) and very short MTQ (VS-MTQ), with support for the multidimensional model.

In addition, Denovan et al. (2021) demonstrated that while the MTQ-18 had acceptable psychometric qualities in their Russian sample, it showed a slight problem with the factorial structure based on confirmatory factor analysis. Therefore, Dagnall et al (2019) recognized the MTQ-18 as an effective test, but preferred the MTQ-10 as it was more concise for practical purposes and tended to be unidimensional, rather than making multi-dimensional assumptions. Concerning the internal reliability of the MTQ-18, previous results showed the highest to lowest Cronbach's α reported by Brand et al. (2017) at $\alpha = .91$; Sabouri et al. (2016) at $\alpha = .84$; and Lang et al. (2019) at $\alpha = .70$.

The findings discussed above relate to the properties of the English version of the MTQ-18, and were obtained in a well-developed manner, but several issues have not been dealt with. This is considered to be a gap of knowledge and motivation to conduct this adaptation study with an Indonesian sample. The first issue is reliability; the most common way to explain this the extent to which the items would behave in a similar way if they were administered to another sample from the same population (Schmidt et al., 2000). The uses of Cronbach's alpha in previous studies have successfully demonstrated the internal consis-

tency of the MTQ-18, but reliability was not considered from other perspectives, such as the separation of items and individuals. Second, previous studies have not clarified the agreement level of items compared to individuals, including their relationship, for practical purposes. Constructing an effective questionnaire requires understanding the difficulties of the items and the magnitude of individuals' latent traits. Third, the performance of the MTQ-18 rating scale also remains unknown. This is concerning, because the distance between rating scales is critical for the validity of the measurement (Pornel & Saldaña, 2013; Wakita et al., 2012).

Moreover, we highlight the inconsistent findings on the ability of the MTQ to predict performance and achievement. A systematic review by Guszkowska dan Wójcik (2021) revealed that among 18 studies, 16 were found to indicate a strong relationship between mental toughness and athletes' performance. For instance, Meggs et al. (2019) discovered that mental toughness is strongly correlated to athletes' subjective performance, as well as being antecedent of dispositional flow. However, recent research by Stimson et al. (2022) concluded that there was a minimal contribution of mental toughness measured by the MTQ-48 to performance or achievement. The aforementioned studies have used conventional statistics (linear regression) to predict mental toughness to performance. For this reason, a study to investigate the implications of the extent to which the MTQ, especially the short form, can predict athletes' performance using another method such as machine learning (ML) is required.

In machine learning, the algorithms are typically designed to deal with regression or classification problems. One of the key differences between ML and traditional statistics concerns the assumptions made. The traditional approach is top-down, with a predefined assumption or rigid premise, together with the use of the *p*-value, while ML is bottom-up and approaches the data as largely unknown, with prioritization of metrics such as accuracy and predictive performance (Orrù et al., 2020). Therefore, the main drawback of the traditional approach is if an inappropriate assumption is made to investigate the data, this may potentially lead to misleading results (Ley et al., 2022); for example, using linear regression on non-linear data.

Researchers believe that the rationale behind the use of ML in psychological studies, such as in validation studies, relies on the assumption that the conventional approach may not be capable of standing alone in comprehensively executing all the problems in psychological data. This belief is supported by Fokkema et al. (2022), who emphasize the use of ML in psychological studies, especially to leverage the ability to predict. In addition, the utilization of ML is also valuable for the criterion and construct validity of the scales (Gonzalez, 2021; Trognon et al., 2022).

As briefly mentioned above, some studies have already delivered convincing results. Methodologically, Dagnall et al. (2019) and Denovan et al. (2021) used factor analysis and structural equation modeling (SEM) to study the quality of the MTQ-18. Notwithstanding the merit of their procedures, researchers argue that the maximum likelihood estimation (MLE) type of factor analysis is flawed compared to weighted least squares estimators (WLS). Marôco (2024) indicates that polychoric correlation with diagonal WLS is preferable in normally distributed data compared to Pearson correlation with MLE. Consequently, our study does not emphasize using the same method (factor analysis), instead preferring to use the Rasch model with conditional maximum likelihood estimation (CML). The rationale for this regards the logit score to promote a more linear and objective measure, as it noted by Boone (2016) and Bond and Fox (2015).

With respect to the methods used in previous research, this study offers three novelties. The first concerns the adaptation of the MTQ-18 for an Indonesian sample, while the second is the amalgamation of the internal structure and predictive validity to study the quality of the MTQ-18. The third is the use of the Rasch model and ML (gradient boosting machine) as the main techniques. This study is the first to propose a combination of the Rasch model and ML as the main approach to adapting and validating the Indonesian version of the MTQ-18.

Regarding the standard of psychological testing by AERA, APA and NCME (2014) this study aims to use the Rasch model to gauge the evidence of internal structure by unidimensionality, local independence, fit statistics, rating scale performance, and the Wright map. For predictive validity, the research uses ML to determine the extent to which the questionnaire scale is able to predict overall athletes' achievement.

Methods

Participants

Ethical clearance was given by the ethical committee of the University of Surabaya, 68/KE/IV/2022. The study involved 400 participants, 194 male (48.5%) and 206 female (51.5%), all Indonesian athletes, with ages varying from 13 to 56 (M = 22.12, SD = 5.9). Their sports fields were swimming (55.25%); athletics (9.5%); diving and underwater hockey (7.75%); basketball (6.25%); chess (3%); *pencak silat* (2.5%); softball (2.25%); football and futsal (2.25%); calisthenics (1.75%); badminton (1.25%) and others (8.25%), comprising volleyball, dance sport, e-sport, golf, handball, judo, karate, rock climbing, petanque, *sepak takraw*, water skiing, taekwondo, *tarung derajat*, tennis, triathlon, and *wushu*. The researchers used non-random sampling through Google forms as the data collection tool, after obtaining informed consent.

Mental Toughness Questionnaire – 18

This study focused on the MTQ-18, by Dagnall et al. (2019). The Likert scales used were: 1 = strongly disagree, 2 = disagree, 3 = neither agree or disagree, 4 = agree, 5 = strongly agree. The forward and back translation of the MTQ-18 was performed by an English-Indonesian translator, then reviewed by four independent raters, researchers, and postgraduate students with a background in psychological studies. Table 1 shows the final form of the Indonesian version of the MTQ-18. Items Q11, Q6, Q3, Q17, Q16, Q12, Q2, Q8 and Q9 should be administered in reverse. The length of time to finish this test is approximately 5-8 minutes, according to the pre-trial test involving 10 participants, and none of them were bewildered by the instructions and items.

Rasch Model

The Rasch model uses a logit-based analytical paradigm to examine items and participants' raw data (Linacre, 1989). Equation 1 represents the ground form of the polytomous Rasch, the rating scale model (RSM). The parameter estimation process usually uses joint maximum likelihood, marginal or conditional. The study employed conditional maximum likelihood (CML) estimation. $P(X_{si})$ is the probability that a person selects a category (x) from item (i) (Andrich, 1978). The summation ensures the calculation of probabilities of every possible response of category (h), from 0 to the total number of categories (m_i). θ_s is the latent trait of persons and shows the location in the latent continuum. δ_s is the difficulty parameter of item (i). τ_k is the threshold that explains the transition point between adjacent categories.

Table 1

Mental Tough Questionnaire 18 Indonesian Version

MTQ-18 (Original Version)	MTQ-18 (Indonesian Version)	Item
I am generally able to react quickly when something unexpected happens	*Saya umumnya mampu bereaksi dengan cepat saat sesuatu yang tidak terduga terjadi*	Q13
I generally cope well with any problems that occur	*Saya biasanya mengatasi dengan baik setiap masalah yang terjadi*	Q4
I often wish my life was more predictable	*Saya berharap hidup saya lebih bisa diprediksi*	Q11*
"I just don't know where to begin" is a feeling I usually have when presented with several things to do at once	*"Saya tidak tahu harus mulai dari mana" adalah perasaan yang biasa saya rasakan ketika dihadapkan pada beberapa hal yang harus dilakukan sekaligus*	Q6*
I usually find it hard to summon enthusiasm for the tasks I have to do	*Saya biasanya sulit membangkitkan semangat terhadap tugas-tugas yang harus dikerjakan*	Q3*
I usually find it difficult to make a mental effort when I am tired	*Saya biasanya merasa sulit untuk melakukan usaha mental ketika lelah*	Q17*
I generally find it hard to relax	*Saya biasanya sulit untuk rileks*	Q16*
When I am feeling tired I find it difficult to get going	*Ketika lelah, saya merasa kesulitan untuk memulai sesuatu*	Q12*
Even when under considerable pressure I usually remain calm	*Bahkan ketika berada dibawah tekanan yang besar, saya biasanya tetap tenang*	Q1
I tend to worry about things well before they actually happen	*Saya cenderung mengkhawatirkan segala sesuatunya jauh sebelum hal itu terjadi*	Q2*
I generally feel in control	*Saya biasanya merasa memegang kendali*	Q10
When I make mistakes, I usually let it worry me for days after	*Ketika saya membuat kesalahan, saya cenderung merasa khawatir selama beberapa hari*	Q8*
In discussions, I tend to back down even when I feel strongly about something	*Dalam diskusi, saya sering mundur meskipun merasa yakin tentang sesuatu*	Q9*
If I feel somebody is wrong, I am not afraid to argue with them	*Jika saya merasa seseorang salah, saya tidak takut untuk berdebat dengan mereka*	Q18
I generally feel that I am a worthwhile person	*Secara umum, saya merasa bahwa saya adalah orang yang berharga*	Q5
I usually speak my mind when I have something to say	*Saya selalu berbicara jujur ketika memiliki sesuatu untuk disampaikan*	Q7
I generally look on the bright side of life	*Saya biasanya melihat sisi positif kehidupan*	Q15
However bad things are, I usually feel they will work out positively in the end	*Meskipun keadaannya buruk, saya yakin semuanya akan berakhir baik*	Q14

Equation 1
Rasch Model Rating Scale

$$P(X_{si}) = \frac{exp[\sum_{k=0}^{x}(\theta_s - \delta_s + \tau_k)]}{\sum_{h=0}^{m_i} exp\left[\sum_{k=0}^{h}(\theta_s - \delta_i + \tau_k)\right]} \quad (1)$$

It is important that several metrics are understood and thorough guidelines into it can be found elsewhere (Bond & Fox, 2015; Linacre & Wright, 2012; Wolins et al., 1983). Infit-Outift statistics for both MNSQ (mean-square) and ZSTD (*z*-standardized) of residuals scores are expected to be from .5 ≤ *x* ≤ 1.5 (MNSQ) to -2 ≤ *x* ≤ 2 (ZSTD). The reliability of the study was based on separation with a desirable value of ≥ .7. The dimensionality measure of this approach indicated the construct validity of the instrument. The raw variance is not the sole method of estimating dimensionality; it should be followed by factor analysis of the Rasch residuals if there is an indication of eigenvalues higher than 1.5 or 2 in each contrast (Smith, 2002). The likelihood ratio Martin-Loef test was also used to support the unidimensionality assumption, (Christensen et al., 2002). For local independence, Q3 statistics were used. Q3 is a non-parametric method to gauge the between-item residual correlation (Debelak & Koller, 2020).

The rating scale analysis focused on the Rasch-Thurstone thresholds. This is a cumulative-probability measure approach that representing a value based on a 50% chance of favoring a certain rating scale or category, (Linacre, 1998). Despite previously tested in a dichotomous model, later this threshold is also viable to be applied in the rating scale model (Linacre, 2009).

For differential item functioning (DIF), this study employed Raju's area method to check the difference between groups (Raju, 1988, 1990). Following the adjustment of the significance value, the Benjamini-Hojberg (BH) test was used, as it is false discovery rate (FDR) control technique that minimize false positive and recommended by Kim and Oshima (2013).

A thorough analysis of rating scale model and differential item functioning was made using eRm (Mair et al., 2024); WrightMap package (Irribarra & Freund, 2022), and difR package (Magis et al., 2020) on RStudio, version 2024.4.2.

Boosting Classification

As an integral part of the study, the authors adopted machine learning, specifically the gradient boosting approach (GBM), to solve the classification problem of discerning participants' achievement with respect to their MTQ-18 scores. The gradient boosting conceptualization was previously proposed by Friedman (2001). The rationale for using this technique was due to its usability or flexibility in dealing with more imbalanced data compared to the other machines, such as support vector machines or random forests (Benkendorf et al., 2023). The algorithm (step 1 – 6) for the GBM is shown below (Algorithm 1).

Algorithm 1
Gradient Boosting Classification

Input: $\{(x_i, y_i)\}_{i=1}^{n}$ and $L(y_i, F(x))$

(1) $F_0(x) = \arg\min_{\gamma} \sum_{i=1}^{N} L(y_{i,}\gamma)$

(2) $r_{im} = -\left[\frac{\partial L(y_i, F(X_i))}{\partial F(x_i)}\right]_{F(x)=F_{m-1}(x)}, for\ i = 1, .., N$

(3) Fit a regression tree to the r_{im} and create R_{jm}, for $j = 1 \ldots J_m$

(4) $\gamma_{jm} = \underset{\gamma}{argmin} \sum_{x_i \in R_{ij}} L(y_i, F_{m-1}(x_i) + \gamma)$

(5) $F_m(x) = F_{m-1}(x) + v \sum_{j=1}^{J_m} \gamma_{jm} I(x \epsilon R_{jm})$

(6) Output $F_M(x)$

In the input, x_i refers to the features and y_i the target (classification outcome), while $L(y_i, F(x))$ is the transformation of the -log (likelihood), the differentiable loss function. Likewise, in the regression gradient boosting machine, step 1 in

the classification (GBM) is required to find the initial leaf or $F_0(x)$, which consists of a constant value. γ (gamma) represents the log(odds), while the summation requires us to sum up the loss function for each y_i (observed score); it is important to measure the log (odds) or γ that minimizes the sum ($\overset{argmin}{\gamma}$).

Step 2 involves calculating the pseudo residuals; $\partial L\left(y_i, F(X_i)\right)$ is the derivative of the loss function, with respect to the predicted log(odds). For a better understanding, the pseudo residual is obtained from the observed score subtracted by the predicted probability. More importantly, the $F(x) = F_{m-1}(x)$ require us to use the most recent or updated predicted log(odds).

Step 3 is the regression tree construction to predict the residuals (r_{im}), while Step 4 is concerned with calculating the output of the new tree, for $j = 1 \ldots J_m$. The output value of each leaf is naturally the score of gamma (γ) obtained by dividing the residuals by the second derivatives of the loss function or, for simplification, is residuals divided by p (1-p).

Step 5 involves creating a new prediction $(F_m(x))$ for each sample, given the new information from earlier measures $F_0(x))$, $F_1(x))$, $F_2(x))$ and so on. υ is the learning rate (usually set at a small amount, such as 0.1 or 0.01), which is preeminent to avoid over-optimistic results, whereas $\sum_{j=1}^{J_m} \gamma_{jm} I\left(x \epsilon R_{jm}\right)$ is the output values from the previous tree. $F_M(x)$ is the final product.

Several metrics need to be focused on for the classification: recall or the true positive rate (TPR); the false positive rate (FPR); the F1 score; the Matthew coefficient correlation (MCC); the Youden index (J); the area under curve (AUC); and and Andrew curves. Complete guidelines on these metrics are available elsewhere (Chicco et al., 2021; Chicco & Jurman, 2023; Hossin & Sulaiman, 2015; Vujovic, 2021). FPR are the measures of true and false observations of the data, such as a true observation that is literally true (TPR), or a false observation that is recognized as positive (FPR). MCC is the formula to indicate that the measurement is not simply a random guess. This score ranges from -1 to 1, with a value approaching 1 indicating perfect predictive performance; otherwise, any value approaching -1 shows total disagreement. The F1 score is the measure of the accuracy of the model using information from recall and precision. J is a metric to measure the effectiveness of the diagnostic tests.

Andrews curves are a data visualization method to elucidate how well the classification process distinguishes between the classes (e.g., binary outcome). The curves were derived from the Fourier series as a projection of high dimensional data, with the x-axis denoting the Fourier coefficient, which ranges from $-\pi$ to π (π = 3.14). In a well-performing boosting classification model, Andrews curves for data points from the same class should cluster together, indicating effective class separation (Moustafa, 2011). Our analysis was run using the GBM package in the Jeffreys's Amazing Statistic Program (JASP) 0.19.1.

Logistic Regression

We also provided the results of the logistic regression as a comparison to the previous ML approach (GBM). This approach is well known as a way to model the relationship between the target (category classes) and the predictor variables (Grömping, 2016). The main conceptualization of this technique is an estimation of the probability of the occurrence of an event, with respect to the given predictors.

Evaluation of the goodness of fit is made using the Akaike Information Criterion (AIC) and the Bayesian Information Criterion (BIC), with better models having lower AIC and BIC. In addition, a pseudo R^2 was also used, with R^2 which has a different interpretation than classic regression. It is a measure of fit that typically compares the

likelihood of the models (McFadden R^2) or the measure of the difference in the mean predicted probabilities between the classes (Tjur R^2), (Grömping, 2016). This analysis was run in JASP 0.19.1.

Results

Dimensionality Analysis

The results of the Martin-Loef likelihood ratio test with median split were: likelihood-ratio = 482, *DF* = 1,295, p = 1.0, which indicate no violation of the unidimensionality assumption. In addition, according to the Rasch residuals factor analysis, the eigenvalues for the first four components were 1.8, 1.5, 1.3 and 1.2 respectively, showing no indication of the secondary significant construct. However, as an additional report, we presented the highest loading score of the residuals of the items in the first component: Q8 = .39, Q11 = .35, and Q13 = -.35. For local independence, the MADaQ3 yielded .074, with the majority of pairs of Q3 across all items ≤ .3, except for Q13 – Q14 = .4.

Reliability Analysis

The result of Cronbach's α was .83, while person separation reliability was 0.352, with squared standard deviation (SSD) = .115, and mean squared error (MSE) = .074. Low separation reliability demonstrates the sufficiency of the test length that considerably low or it would be challenging to separate the item difficulties in the questionnaire into a wide range of different levels because they are clustered to relatively the same level of difficulty.

Item Analysis

The descriptive statistics were 7,200 data points, *M* = 1,717.8, *SD* = 207.8, log-likelihood X^2 = 10,3513, *DF* = 6,780, global root-mean-square (RMS) residuals = .579, and p = .00. As shown in Table 2, items Q8 and Q11 were the most challenging to agree on, which indicates that participants were less likely to disagree or choose category 1 (strongly disagree) or 2 (disagree) for these particular items. On contrary, Q14, Q13, and Q3 were relatively more inclined towards agreement. The point biserial correlation ranges from 0.24 – 0.48. All of the items showed fit indices.

Rating Scale Analysis

As can be seen in Table 3, and subsequently confirmed in Figure 1 the distance between *l*1 and *l*2 was relatively shorter compared to that between *l2* and *l3* and *l*3 and *l*4, indicating that categories 1, 2, and 3 were more likely to be chosen by the participants for all items, as they are closer to each other. On the contrary, the distances between categories 3, 4, and 5 were greater, indicating that individuals may behave more cautiously in agreeing on most of the items, especially Q11, Q8, and Q12.

Differential Item Functioning

DIF analysis using Raju's area approach and adjusted significance using Benjamini-Hochberg (BH) was conducted to check bias items between groups (male and female). The reference was set as males, with the focal group set as females. As shown in Table 4, several items were flagged as being biased towards male and female participants. Respecting the Δ Raju (≥ 1.5) as the effect sizes, Q11, Q15 and Q18 showed significantly ($p < .05$) different functioning across the two groups. Q8, however, exhibited a moderate effect size. The negative statistics show that the trend of bias is towards the focal group, with the positive statistics indicating that the items appeared to function in a way that was more favorable to the mental toughness trait of the reference group.

Table 2

Item Analysis

Item	Mean	Measure	*SE*	Infit MNSQ	Outfit MNSQ	Point biserial
Q11	1.82	1.2333	.0428	1.1	1.098	.268
Q8	2.67	.5697	.0464	1.218	1.227	.246
Q12	3.05	.2063	.0524	1.126	1.106	.35
Q15	3.06	.1897	.0527	.876	.899	.36
Q4	3.06	.1869	.0528	.765	.774	.377
Q9	3.07	.1813	.0529	1.122	1.107	.214
Q16	3.09	.1587	.0534	1.252	1.242	.307
Q7	3.11	.1299	.054	.887	.909	.331
Q17	3.19	.0391	.0561	.96	.919	.297
Q5	3.19	.0328	.0562	.911	.912	.431
Q2	3.31	-.1228	.0602	.997	.986	.395
Q1	3.34	-.1786	.0618	.874	.848	.459
Q18	3.35	-.1978	.0623	1.008	.992	.329
Q6	3.44	-.3387	.0666	1.309	1.286	.348
Q10	3.46	-.3748	.0677	.759	.735	.433
Q14	3.47	-.398	.0685	.797	.767	.485
Q13	3.59	-.6464	.0774	.821	.808	.45
Q3	3.6	-.6706	.0783	1.079	1.067	.345

Figure 1

Wright Map

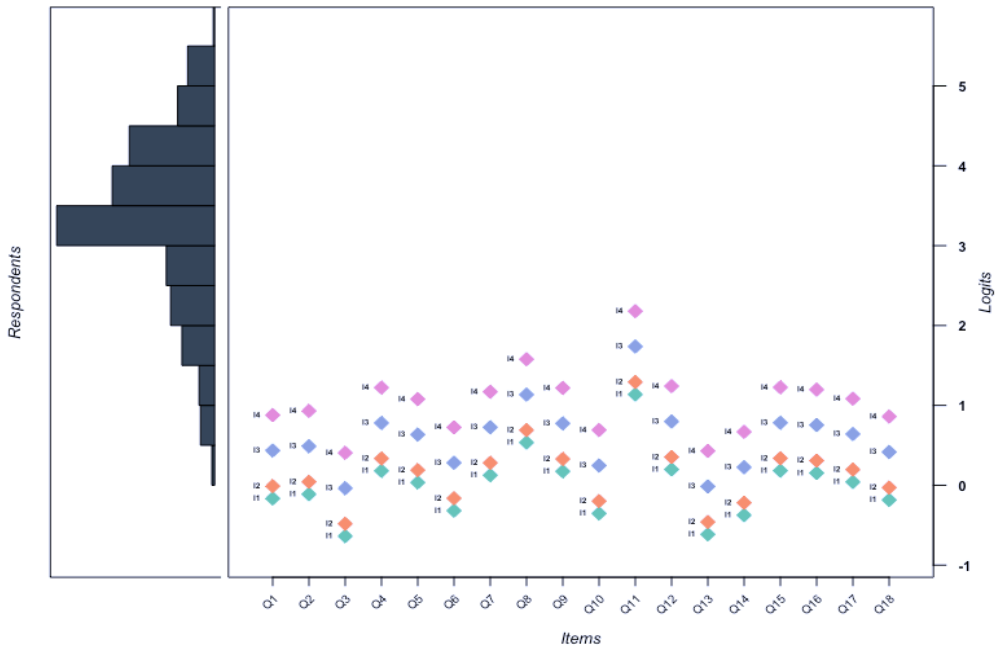


Table 3

Rating Scale Thresholds

Item	*l*1	*l*2	*l*3	*l*4
Q1	-.1654	-.0103	.4347	.878
Q2	-.1124	.0427	.4877	.931
Q3	-.6377	-.4826	-.0376	.406
Q4	.1796	.3347	.7796	1.223
Q5	.0348	.1899	.6348	1.078
Q6	-.3182	-.1631	.2818	.725
Q7	.1261	.2812	.7262	1.17
Q8	.5344	.6895	1.1344	1.578
Q9	.1743	.3294	.7744	1.218
Q10	-.3528	-.1977	.2472	.691
Q11	1.1365	1.2916	1.7365	2.18
Q12	.1977	.3528	.7977	1.241
Q13	-.6142	-.4591	-.0142	.429
Q14	-.3751	-.2199	.225	.669
Q15	.1822	.3373	.7822	1.226
Q16	.1532	.3083	.7532	1.197
Q17	.0407	.1958	.6408	1.084
Q18	-.1837	-.0286	.4163	.86

Table 4

Differential Item Functioning

Item	Statistic	Adjusted *p*	Δ Raju	Effect size
Q1	-.392	.736	.926	A
Q2	-1.519	.238	1.883	C
Q3	-2.051	.09	1.33	B
Q4	-2.089	.085	.979	A
Q5	.827	.516	-2.18	C
Q6	-1.682	.185	1.74	C
Q7	-1.541	.238	.981	A
Q8	-3.086	.007	1.443	B
Q9	.571	.659	-1.904	C
Q10	-2.043	.09	4.045	C
Q11	-2.712	.019	2.387	C
Q12	-1.257	.32	.638	A
Q13	.511	.686	-1.475	B
Q14	-1.475	.253	1.648	C
Q15	-4.961	<.001	2.695	C
Q16	-1.453	.257	.702	A
Q17	1.082	.41	-2.092	C
Q18	-3.549	.002	4.082	C

Note: A (< 1, negligible effect), B (> 1, moderate effect),
C (> 1.5, large effect)

Prediction of Achievement

The objective of the analysis was to evaluate the influence of MTQ-18 on prediction. We formulated this by classifying the ordinal hierarchy of athletes' achievements based on the level and competitiveness of the tournament. Achievement was assessed on the basis of their best accomplishment. International tournaments were labeled as 2; national competitions as 1; and if the achievement did not correspond to either national or international level, it was labeled as 0.

The data pre-processing was conducted by eliminating participants who possessed the highest unexpected responses, and was also based on their fit statistics. The final data for this GBM analysis came from 381 participants. These were assumed to be data outliers that may affect the integrity of the study and the classification process. Subsequently, they separated into two groups. The first comprised the national achievement class (214 athletes, 68%) vs the no achievement class (102 athletes, 32%), with a total of 316 individuals. The second group related to international achievement (65 athletes, 39%) vs no achievement (102 athletes, 61%), with a total of 167 individuals. The data processing was based on 20% as a sample and 5 folds for training-validation. The minimum number of observations in each node was 10, with 50% of the training data used per tree.

As shown Table 5, the model summary and evaluation metrics of the first model (national vs no achievement) using 253 persons as training—validation data and 63 as test data were: validation accuracy = .66; test accuracy = .78; trees = 7; shrinkage = .1, and Youden index (*J*) = .25. On the other hand, the evaluation of the second model (international vs no achievement) using 134 persons as training-validation data and 33 persons as test data was: validation accuracy = .63; test accuracy = .67; trees = 3; shrinkage = .1, and Youden index (*J*) = .29.

In particular, the process of classification was more distinguishable in the national vs no achievement group (Figure 2 - A), with higher AUC for national competitions. According to Figure 2 - C, the GBM model shows relatively inferior performance in predicting international achievement.

In comparison to the boosting classification performance, conventional logistic regression with the enter method was also conducted. Participants were eliminated based on the fit statistics of the Rasch model. Instead of using the raw ordinal scores, the analysis was run using the supervised data from the parameter estimation or the logit of participants from the Rasch (rating scale) model.

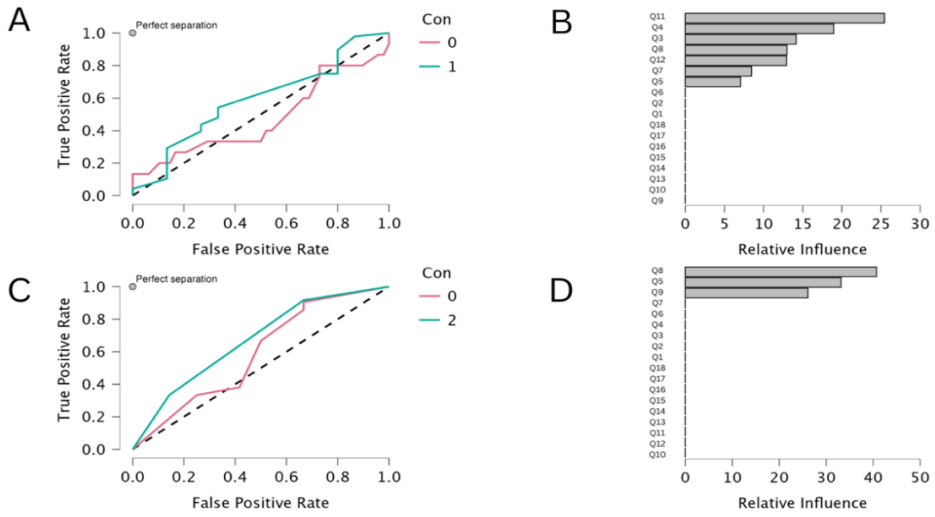
Table 5

Evaluation Metrics

Metrics	Achievement		Average/Total	Achievement		Average/Total
	0	1		0	2	
Accuracy	.78	.78	.78	.67	.67	.67
Precision	1	.77	.82	.68	.6	.65
Recall	.07	1	78	.90	.25	.67
FPR	.0	.27	.47	.75	.09	.62
F1 Score	.13	.87	.7	.77	.35	.62
MCC	.28	.28	.28	.20	.20	.20
AUC	.54	.59	.56	.75	.41	.58

Note: 0 = No achievement, 1 = National achievement, 2 = International achievement.

Figure 2
Receiver Operating Curves and Relative Influence Plots of Boosting Classification



Note: A - ROC between no achievement (0) and national achievement (1); B - Relative influence between 0 and 1; C - ROC between no achievement (0) and international achievement (2); D – Relative influence between 0 and 2.

The model summary of the classification between national vs no achievement was M_0 > M_1, with AIC_0= 399.497, BIC_0= 403.253, AIC_1= 400.555, BIC_1 = 408.067, DF = 314, p = .321, McFadden R^2 = .002, and Tjur R^2 = .003. The estimation of participant logit = .117, SE = .120, and odds ratio = 1.124. For the evaluation metrics, accuracy = .67 and area under curve (AUC) = .54.

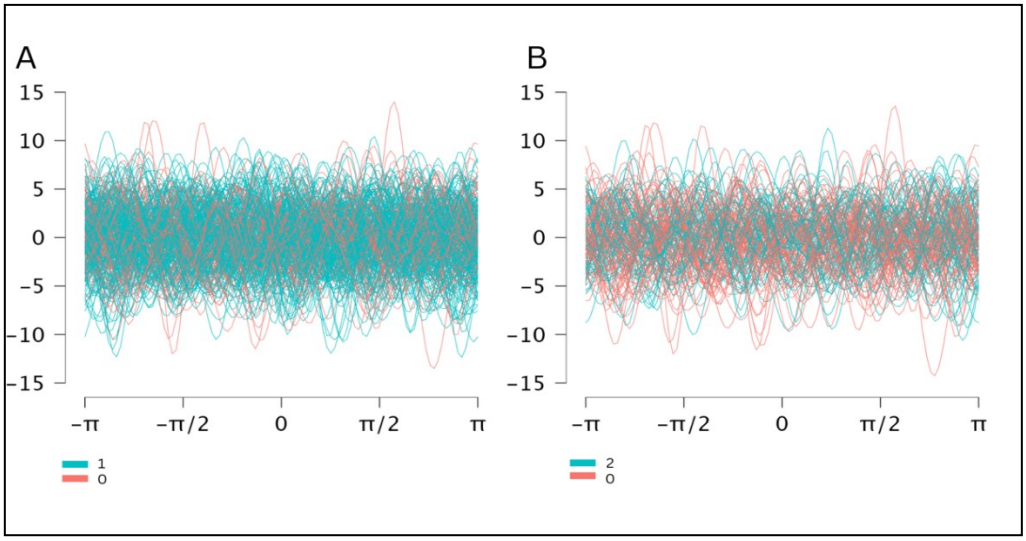
On the other hand, the classification between international vs no achievement was M_0 > M_1, with AIC_0= 225.245, BIC_0= 228.363, AIC_1= 226.953, BIC_1 = 233.189, DF = 314, p = .589, McFadden R^2 = .001, and Tjur R^2 = .002. The estimation of participant logit = .085, SE = .157, and the odds ratio = 1.089. For the evaluation metrics, accuracy = .61 and AUC = 0.5.

In summary, for the logistic regression, there was no evidence of significant results from the p-value or from the model through the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) in either group. Since the M_0 > M_1 indicating that the M_1 (participants logit) did not improve the prediction compared to the null model (M_0). However, the model's positive value of estimation indicated a positive relation between performance and the logit scores; the higher the logit, the higher the estimation of achievement.

Andrews curves (Figure 3) are a visualization technique that is used to interpret high-dimensional data, and can be particularly useful in set machine learning, including the boosting approach, when dealing with classification tasks to assess class separation. Each data point is represented by a curve, with similar points (typically from the same class) having similar curve shapes. The data from Figure 3 - A and B for each class overlapped with each other. This indicated that the model was struggling to distinguish between these classes, possibly due to noise or insufficient features. However, referring to Figure 3 - B, the distinction was more noticeable between no achievement vs international achievement, compared to Figure 3 – A.

Figure 3

Andrews Curves of Boosting Classification



Note: 0 = no achievement, 1 = national achievement, 2 = international achievement 3

Discussion

According to the Martin-Loef test, local independence, factor analysis of residuals, and based on our study sample, the MTQ-18 Indonesian version is a fine-tuned questionnaire in a unidimensional construct. However, a report by Denovan et al. (2021) also investigated the MTQ-18 and found some inconsistent factor loadings for the challenge and control dimensions. Their results may possibly be because to some extent the nature of mental toughness is more suited to unidimensional latent traits. Additionally, the tendency of unidimensionality in the shorter versions of the MTQ was also highlighted in research by Gerber et al. (2015). Moreover, along with the addition of more items, it turns the MTQ into a more multidimensional construct, despite Perry et al. (2021) recommending a bifactor model that is proven to be more suitable for MTQ-48.

Items Q14, Q13 and Q3 were the easiest items to agree on, while Q11 and Q8 were those that participants found more difficult to agree on. According to the mean-square (MNSQ), that followed the threshold of .5 to 1.5, all the items provided ideal fit indices. Moreover, according to the Wright map (Figure 1), some items were grouped into similar levels of agreement, so can be considered as redundant items; for instance, the pairs of Q13 and Q3, Q15 and Q4, and Q17 and Q5. As a consequence, it is legitimate to argue that these items actually measure the same level of difficulty of the latent trait. It is also suggested to eliminate one of them if an abridged version of the MTQ-18 is needed (e.g., MTQ-10 or MTQ-8).

The rating scale of the questionnaire performed well, and it is important to ensure that we have a constant range of category scales, regardless of the range of the rating, such as from 1-5, or 1-7. Pornel and Saldana (2013) explain that asymmetric verbal anchors may affect the validity of the rating scale used in research. Therefore, based on the threshold measure across every item, we found that the test items shared similar trends. The distance from *l*1 to *l*2 was closest indicating that the rating scales "strongly disagree" to "disagree" to "neither agree or disagree" were not well-differentiated in terms of meaning for the

participants. While this pattern did not affect overall scale performance, it indicates the potential benefit of revising category labels or even providing clearer definitions to enhance differentiation and interpretability

A previous study of the MTQ-48, by measuring the construct consistently across gender and age groups, found that the questionnaire functioned well (Perry et al., 2021). However, our study found some notable items that demonstrated bias towards gender (both male and female), with a large effect size from the Δ Raju calculations. These items were Q11, Q15, and Q18, with a large effect size (> 1.5), and Q8, with a more moderate size. This suggests that males and females may interpret or respond differently to these items, which could reflect inherent differences in how mental toughness is experienced or expressed across genders. Therefore, a comparison of the performance of males and females should be carefully made. Moreover, this finding has also been indicated in previous studies, which emphasize that males tend to have higher mental toughness scores (Nicholls et al., 2009; Yarayan et al., 2024).

Related to the use of the Rasch model in psychometric practice, this is not a new emerging approach. Wright (1996) discussed the comparison between factor analysis and the Rasch approach, indicating that the most problematic issue in the use of Likert-type data in factor analysis is the poor reproducibility of the factor sizes and loadings. Therefore, he believed that logit transformation was an alternative to overcome this issue. Jamieson (2004) also explained that it should be clear that any ordinal data, including Likert types, should be treated with non-parametric analysis. However, Sullivan and Artino (2013) presented contrasting arguments, claiming that if the normality distribution and adequate sample size hold, it is not necessary to treat Likert data as ordinal. Carifio and Perla (2007) also contend that although Likert as a response (consisting of one item) may behave in an ordinal fashion, as a scale (comprising several items) it exhibits interval-level measurement, referring to their terms of atom-molecule-scale.

For a comparison of the classification process, the performance of logistic regression in this study was inferior to the boosting classification. The dominance of boosting is reasonable, regarding the approach when dealing with classification problems, especially with more non-linear and complex data. Robustness is achieved by starting the training data analysis with a weak learner, which would result a false prediction. The earlier false prediction ($F_0(x)$) then becomes the new or updated subset ($F_m(x)$), with the same step as before of using a weak learner, producing an updated false prediction of the pseudo residuals; gradually narrowing the gap in the residuals in the correct direction; and later producing the final prediction (Ferreira & Figueiredo, 2012; Schapire, 2003). The advantages of boosting compared to traditional logistic regression have also been discussed in previous studies, (Belsti et al., 2023; Zheng et al., 2023).

The performance of MTQ-18 to predict athletes' achievement in this study was under the satisfactory level. This is in line with the findings of Stimson et al. (2022), who assessed the MTQ-48 and found minimum evidence of the capability of mental toughness to predict performance. Furthermore, the deficient performance of our model to some extent resembles the evidence of the low separation reliability of individuals and logit measure of items. This is confirmed in the Wright map (Figure 1), which shows that the MTQ-18 is more sensitive to measuring individuals with moderate to low agreement, so it would not be ideal to effectively distinguish between those with high and low agreements. However, previous studies by Meggs et al. (2019) and Cowden (2017) did demonstrate the importance of mental toughness in producing athletes with high performance and achievement.

Regarding the thorough process of our study, the authors emphasize the type of psychometric property construction process combining traditional approaches with machine learning. Internal structure validity, together with reliability measures, are important as preliminary steps towards ensuring the quality of the data before commencing predictive validity analysis using machine learning. These steps are important, in light of the GIGO or *garbage in garbage out* notion. The quality of the input data prior to statistical analysis could undoubtedly affect the output (Kilkenny & Robinson, 2018). Moreover, in terms of human annotation in machine-learning studies, researchers are advised to be fully responsible for ensuring the validity of the data for training before commencing prediction (Geiger et al., 2020).

As implicitly stated above, there are several implications of this study. The main argument was that the MTQ-18 displayed poor predictability of athletes' achievement. As also already noted, even the full version (MTQ-48) of the questionnaire in English version did not demonstrate notable performance in predicting achievement. Therefore, researchers may need to use a series of properties in order to predict achievement accurately. Additionally, practitioners should treat the male and female norm-scores separately, as some items in the questionnaire behave differently with males or females.

The results of the study were limited by the sample size and characteristics, as we focused on only 400 individual athletes. Moreover, the field category of the sports was imbalanced (dominated by swimming), and the boosting classification and regression logistic data were also imbalanced. Therefore, it is recommended that balanced data is used for the categorization of each group, together with a test for other sample characteristics such as students in education or employees in organizational settings. In addition, the Martin-Loef dimensionality test works better with more participants, such as > 600.

Further studies are needed to develop a better version of the MTQ for the Indonesian culture, which is less unimpeded by gender bias. Other studies are also encouraged to resolve the variability of the items, which should be capable of a wide array of different levels, rather than clustering in the middle and low levels.

Conclusion

In conclusion, the MTQ-18 Indonesian version is a unidimensional questionnaire with a positive internal consistency of items. However, the separation level of items is very poor and more appropriate for measuring individuals with moderate to low traits of mental toughness. Performance across gender should be cautiously understood, as some items favor gender bias. The predictive validity of the questionnaire is insufficient; therefore, this short test is not preferable for predicting individuals' future achievement or performance in a precise manner.[]

Acknowledgments

The authors would like to thank the English-Indonesian translator and the four raters who assisted in the translation process and evaluation of the item content.

Author Contribution Statement

Ananta Yudiarso: Conceptualization; Formal Analysis; Investigation; Methodology; Project Administration; Resources; Validation; Visualization; Writing Original Draft; Writing, Review & Editing.

Ista Wirya Ardhiani: Data Curation; Formal Analysis; Investigation; Methodology; Project Administration; Resources; Validation; Writing Original Draft. **Roy Surya:** Formal Analysis; Investigation; Methodology; Validation; Visualization; Writing, Review & Editing. **Ferry Yohannes Watimena:** Conceptualization; Investigation; Project Administration; Resources; Writing, Review & Editing. **Mami Kanzaki:** Conceptualization; Formal Analysis; Investigation; Writing Original Drafts; Writing, Review & Editing.

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing / American Educational Research Association, American Psychological Association, National Council on Measurement in Education*. American Educational Research Association.

Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement*, *2*(4), 581–594. https://doi.org/10.1177/014662167800200413

Belsti, Y., Moran, L., Du, L., Mousa, A., De Silva, K., Enticott, J., & Teede, H. (2023). Comparison of machine learning and conventional logistic regression-based prediction models for gestational diabetes in an ethnically diverse population; the Monash GDM Machine learning model. *International Journal of Medical Informatics*, *179*, 105228. https://doi.org/10.1016/j.ijmedinf.2023.105228

Benkendorf, D. J., Schwartz, S. D., Cutler, D. R., & Hawkins, C. P. (2023). Correcting for the effects of class imbalance improves the performance of machine-learning based species distribution models. *Ecological Modelling*, *483*, 110414. https://doi.org/10.1016/j.ecolmodel.2023.110414

Bond, T., & Fox, C. M. (2015). *Applying the Rasch Model: Fundamental measurement in the human sciences*. Routledge. https://doi.org/10.4324/9781315814698

Boone, W. J. (2016). Rasch analysis for instrument development: Why, when, and how? *CBE—Life Sciences Education*, *15*(4), rm4. https://doi.org/10.1187/cbe.16-04-0148

Brand, S., Kalak, N., Gerber, M., Clough, P. J., Lemola, S., Sadeghi Bahmani, D., Pühse, U., & Holsboer-Trachsler, E. (2017). During early to mid adolescence, moderate to vigorous physical activity is associated with restoring sleep, psychological functioning, mental toughness and male gender. *Journal of Sports Sciences*, *35*(5), 426–434. https://doi.org/10.1080/02640414.2016.1167936

Brand, S., Sabouri, S., Gerber, M., Sadeghi Bahmani, D., Lemola, S., Clough, P., Kalak, N., Shamsi, M., & Holsboer-Trachsler, E. (2016). Examining Dark Triad traits in relation to mental toughness and physical activity in young adults. *Neuropsychiatric Disease and Treatment*, 229. https://doi.org/10.2147/NDT.S97267

Carifio, J., & Perla, R. J. (2007). Ten common misunderstandings, misconceptions, persistent myths and urban legends about Likert Scales and Likert response formats and their antidotes. *Journal of Social Sciences*, *3*(3), 106–116. https://doi.org/10.3844/jssp.2007.106.116

Chicco, D., & Jurman, G. (2023). The Matthews Correlation Coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining*, *16*(1), 4. https://doi.org/10.1186/s13040-023-00322-4

Chicco, D., Warrens, M. J., & Jurman, G. (2021). The coefficient of determination R-squared is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation. *PeerJ Computer Science*, *7*, e623. https://doi.org/10.7717/peerj-cs.623

Christensen, K. B., Bjorner, J. B., Kreiner, S., & Petersen, J. H. (2002). Testing unidimensionality in Polytomous Rasch Models. *Psychometrika*, *67*(4), 563–574. https://doi.org/10.1007/BF02295131

Clough, P., Earle, K., & Sewell, D. (2002). Mental toughness: The concept and its measurement. In *Solutions in Sport Psychology* (pp. 32–46). Thomson Learning.

Cowden, R. G. (2017). On the mental toughness of self-aware athletes: Evidence from competitive tennis players. *South African Journal of Science*, *113*(1/2), 6. https://doi.org/10.17159/sajs.2017/20160112

Dagnall, N., Denovan, A., Papageorgiou, K. A., Clough, P. J., Parker, A., & Drinkwater, K. G. (2019). Psychometric assessment of shortened Mental Toughness Questionnaires (MTQ): Factor structure of the MTQ-18 and the MTQ-10. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.01933

Debelak, R., & Koller, I. (2020). Testing the local independence assumption of the Rasch Model with Q3 - based nonparametric model tests. *Applied Psychological Measurement*, *44*(2), 103–117. https://doi.org/10.1177/0146621619835501

Denovan, A., Dagnall, N., Hill-Artamonova, E., & Musienko, T. (2021). Mental Toughness Questionnaire (MTQ18): A Russian version. *National Security and Strategic Planning*, *2021*(3), 47–59. https://doi.org/10.37468/2307-1400-2021-3-47-59

Ferreira, A. J., & Figueiredo, M. A. T. (2012). Boosting algorithms: A review of methods, theory, and applications. In C. Zhang & Y. Ma (Eds.), *Ensemble Machine Learning* (pp. 35–85). Springer New York. https://doi.org/10.1007/978-1-4419-9326-7_2

Fokkema, M., Iliescu, D., Greiff, S., & Ziegler, M. (2022). Machine learning and prediction in psychological assessment. *European Journal of Psychological Assessment*, *38*(3), 165–175. https://doi.org/10.1027/1015-5759/a000714

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5). https://doi.org/10.1214/aos/1013203451

Geiger, R. S., Yu, K., Yang, Y., Dai, M., Qiu, J., Tang, R., & Huang, J. (2020). Garbage in, garbage out? *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 325–336. https://doi.org/10.1145/3351095.3372862

Gerber, M., Brand, S., Feldmeth, A. K., Lang, C., Elliot, C., Holsboer-Trachsler, E., & Pühse, U. (2013). Adolescents with high mental toughness adapt better to perceived stress: A longitudinal study with Swiss vocational students. *Personality and Individual Differences*, *54*(7), 808–814. https://doi.org/10.1016/j.paid.2012.12.003

Gerber, M., Feldmeth, A. K., Lang, C., Brand, S., Elliot, C., Holsboer-Trachsler, E., & Pühse, U. (2015). The relationship between mental toughness, stress, and burnout among adolescents: A longitudinal study with Swiss vocational students. *Psychological Reports*, *117*(3), 703–723. https://doi.org/10.2466/14.02.PR0.117c29z6

Gonzalez, O. (2021). Psychometric and machine learning approaches to reduce the length of scales. *Multivariate Behavioral Research*, *56*(6), 903–919. https://doi.org/10.1080/00273171.2020.1781585

Grömping, U. (2016). Practical guide to logistic regression. *Journal of Statistical Software*, *71*(3), 1–5. https://doi.org/10.18637/jss.v071.b03

Gucciardi, D. F., Hanton, S., Gordon, S., Mallett, C. J., & Temby, P. (2015). The concept of mental toughness: Tests of dimensionality, nomological network, and traitness. *Journal of Personality*, *83*(1), 26–44. https://doi.org/10.1111/jopy.12079

Guszkowska, M., & Wójcik, K. (2021). Effect of mental toughness on sporting performance: Review of studies. *Baltic Journal of Health and Physical Activity*, *Supplement*(2), 1–12. https://doi.org/10.29359/BJHPA.2021.Suppl.2.01

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 1–11. https://doi.org/10.5121/ijdkp.2015.5201

Hsieh, Y.-C., Lu, F. J. H., Gill, D. L., Hsu, Y.-W., Wong, T.-L., & Kuan, G. (2024). Effects of mental toughness on athletic performance: A systematic review and meta-analysis. *International Journal of Sport and Exercise Psychology*, *22*(6), 1317–1338. https://doi.org/10.1080/1612197X.2023.2204312

Irribarra, D. T., & Freund, R. (2022). *Package 'WrightMap.'* /https://cran.r-project.org/web/packages/WrightMap/WrightMap.pdf

Jamieson, S. (2004). Likert Scales: How to (ab)use them. *Medical Education*, *38*(12), 1217–1218. https://doi.org/10.1111/j.1365-2929.2004.02012.x

Kawabata, M., Pavey, T. G., & Coulter, T. J. (2021). Evolving the validity of a mental toughness measure: Refined versions of the Mental Toughness Questionnaire-48. *Stress and Health*, *37*(2), 378–391. https://doi.org/10.1002/smi.3004

Kilkenny, M. F., & Robinson, K. M. (2018). Data quality: "Garbage in – garbage out." *Health Information Management Journal*, *47*(3), 103–105. https://doi.org/10.1177/1833358318774357

Kim, J., & Oshima, T. C. (2013). Effect of multiple testing adjustment in differential item functioning detection. *Educational and Psychological Measurement*, *73*(3), 458–470. https://doi.org/10.1177/0013164412467033

Kobasa, S. C. (1979). Stressful life events, personality, and health: An inquiry into hardiness. *Journal of Personality and Social Psychology*, *37*(1), 1–11. https://doi.org/10.1037/0022-3514.37.1.1

Lang, C., Brand, S., Colledge, F., Ludyga, S., Pühse, U., & Gerber, M. (2019). Adolescents' personal beliefs about sufficient physical activity are more closely related to sleep and psychological functioning than self-reported physical activity: A prospective study. *Journal of Sport and Health Science*, *8*(3), 280–288. https://doi.org/10.1016/j.jshs.2018.03.002

Ley, C., Martin, R. K., Pareek, A., Groll, A., Seil, R., & Tischer, T. (2022). Machine learning and conventional statistics: Making sense of the differences. *Knee Surgery, Sports Traumatology, Arthroscopy*, *30*(3), 753–757. https://doi.org/10.1007/s00167-022-06896-6

Lin, Y., Mutz, J., Clough, P. J., & Papageorgiou, K. A. (2017). Mental toughness and individual differences in learning, educational and work performance, psychological well-being, and personality: A systematic review. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.01345

Linacre, J. M. (1989). *Rasch models from objectivity: A generalization*. ERIC.

Linacre, J. M. (1998). Estimating measures with known polytomous item difficulties. *Rasch Measurement Transactions*, *12*(2), 638.

Linacre, J. M. (2009). Dichotomizing rating scales. *Rasch Measurement Transactions*, *23*(3), 1228.

Linacre, J. M., & Wright, B. (2012). *Winsteps help for Rasch analysis*. Winsteps, USA.

Magis, D., Beland, S., & Raiche, G. (2020). *Package 'difR.'* https://cran.r-project.org/web/packages/difR/difR.pdf

Mair, P., Rusch, T., Hatzinger, R., Maier, M. J., & Debelak, R. (2024). *Package 'eRm.'* https://cran.r-project.org/web/packages/eRm/eRm.pdf

Marôco, J. (2024). Factor analysis of ordinal items: Old questions, modern solutions? *Stats*, *7*(3), 984–1001. https://doi.org/10.3390/stats7030060

Meggs, J., Chen, M. A., & Koehn, S. (2019). Relationships between flow, mental toughness, and subjective performance perception in various triathletes. *Perceptual and Motor Skills*, *126*(2), 241–252. https://doi.org/10.1177/0031512518803203

Moustafa, R. E. (2011). Andrews curves. *WIREs Computational Statistics*, *3*(4), 373–382. https://doi.org/10.1002/wics.160

Nicholls, A. R., Morley, D., & Perry, J. L. (2016). Mentally tough athletes are more aware of unsupportive coaching behaviours: Perceptions of coach behaviour, motivational climate, and mental toughness in sport. *International Journal of Sports Science & Coaching*, *11*(2), 172–181. https://doi.org/10.1177/1747954116636714

Nicholls, A. R., Polman, R. C. J., Levy, A. R., & Backhouse, S. H. (2009). Mental toughness in sport: Achievement level, gender, age, experience, and sport type differences. *Personality and Individual Differences*, *47*(1), 73–75. https://doi.org/10.1016/j.paid.2009.02.006

Orrù, G., Monaro, M., Conversano, C., Gemignani, A., & Sartori, G. (2020). Machine learning in psychometrics and psychological research. *Frontiers in Psychology*, *10*. https://doi.org/10.3389/fpsyg.2019.02970

Papageorgiou, K. A., Denovan, A., & Dagnall, N. (2019). The positive effect of narcissism on depressive symptoms through mental toughness: Narcissism may be a dark trait but it does help with seeing the world less grey. *European Psychiatry*, *55*, 74–79. https://doi.org/10.1016/j.eurpsy.2018.10.002

Perry, J. L., Clough, P. J., Crust, L., Earle, K., & Nicholls, A. R. (2013). Factorial validity of the Mental Toughness Questionnaire-48. *Personality and Individual Differences*, *54*(5), 587–592. https://doi.org/10.1016/j.paid.2012.11.020

Perry, J. L., Strycharczyk, D., Dagnall, N., Denovan, A., Papageorgiou, K. A., & Clough, P. J. (2021). Dimensionality of the Mental Toughness Questionnaire (MTQ48). *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.654836

Pornel, J. B., & Saldaña, G. A. (2013). Four common misuses of the Likert scale. *Philippine Journal of Social Sciences and Humanities*, *18*(2), 12–19.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika*, *53*(4), 495–502. https://doi.org/10.1007/BF02294403

Raju, N. S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement*, *14*(2), 197–207. https://doi.org/10.1177/014662169001400208

Schapire, R. E. (2003). The boosting approach to machine learning: An overview. In *Nonlinear Estimation and Classification* (pp. 149–171). Springer. https://doi.org/10.1007/978-0-387-21579-2_9

Schmidt, F. L., Viswesvaran, C., & Ones, D. S. (2000). Reliability is not validity and validity is not reliability. *Personnel Psychology*, *53*(4), 901–912. https://doi.org/10.1111/j.1744-6570.2000.tb02422.x

Smith, E. V. (2002). Detecting and evaluating the impact of multidimensionality using item fit statistics and principal component analysis of residuals. *Journal of Applied Measurement*, *3*(2), 205–231. http://www.ncbi.nlm.nih.gov/pubmed/12011501

Stimson, J. R., Cromwell, E. K., & Cromer, L. D. (2022). The predictive validity of the MTQ48 for academic and athletic success in student-athletes. *Journal for the Study of Sports and Athletes in Education*, 1–13. https://doi.org/10.1080/19357397.2022.2143148

Sullivan, G. M., & Artino, A. R. (2013). Analyzing and Interpreting data from Likert-type scales. *Journal of Graduate Medical Education*, *5*(4), 541–542. https://doi.org/10.4300/JGME-5-4-18

Trognon, A., Cherifi, Y. I., Habibi, I., Demange, L., & Prudent, C. (2022). Using machine-learning strategies to solve psychometric problems. *Scientific Reports*, *12*(1), 18922. https://doi.org/10.1038/s41598-022-23678-9

Vujovic, Ž. Đ. (2021). Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications*, *12*(6). https://doi.org/10.14569/IJACSA.2021.0120670

Wakita, T., Ueshima, N., & Noguchi, H. (2012). Psychological distance between categories in the Likert Scale. *Educational and Psychological Measurement*, *72*(4), 533–546. https://doi.org/10.1177/0013164411431162

Wolins, L., Wright, B. D., & Masters, G. N. (1983). Rating scale analysis: Rasch measurement. *Journal of the American Statistical Association*, *78*(382), 497. https://doi.org/10.2307/2288670

Wright, B. D. (1996). Comparing Rasch measurement and factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *3*(1), 3–24. https://doi.org/10.1080/10705519609540026

Yarayan, Y. E., Solmaz, S., Aslan, M., Batrakoulis, A., Al-Mhanna, S. B., & Keskin, K. (2024). Sex differences in athletic performance response to the imagery and mental toughness of elite middle- and long-distance runners. *Sports*, *12*(6), 141. https://doi.org/10.3390/sports12060141

Zheng, X., Wang, B., Liu, H., Wu, W., Sun, J., Fang, W., Jiang, R., Hu, Y., Jin, C., Wei, X., & Chen, S. S.-C. (2023). Diagnosis of Alzheimer's disease via resting-state EEG: Integration of spectrum, complexity, and synchronization signal features. *Frontiers in Aging Neuroscience*, *15*. https://doi.org/10.3389/fnagi.2023.1288295