Psychometric Properties of the Mental Toughness Questionnaire 48 (MTQ48) in Elite, Amateur and Non-athletes.

Robert Vaughan, Ph.D.
York St John University, UK

Donncha Hanna, Ph.D.
Queens University Belfast, UK

Gavin Breslin, Ph.D.
Ulster University, UK

Word Count: 7883

Contact Details: Robert Vaughan, Room DG311, School of Psychological and Social Sciences, York St John University, Lord Mayors Walk, York, UK, YO31 7EX, E-mail: r.vaughan@yorksj.ac.uk

Abstract

The purpose of this study was to assess the psychometric properties of the Mental Toughness Questionnaire 48 (MTQ48) and assess the measurement invariance across elite, amateur and non-athletes. In total, 1096 participants aged between 18 and 58 years with a range of athletic experience - elite (*n* = 181), amateur (*n* = 577) and non-athletes (*n* = 338) - from various sports completed the MTQ48. The internal consistency of the scale was gauged through Omega for the total and relevant subscales. Factorial validity was assessed using exploratory structural equation modeling in order to provide a comprehensive estimation of the scales dimensionality. Overall, results offered support for the scales reliability with acceptable internal consistency reported at the total and subscale level. However, the validity of the MTQ48 for the use with athletes of different levels may be questioned. The MTQ48's hypothesised four-factor model did not fit the data well, whereas the six-factor model produced acceptable levels of fit with large degrees of misspecification in the factor structures across elite, amateur and non-athletes. The results caution the use of the scale with elite athletes and call for refinement of the measure at the subscale level.

Key Words: *Mental Toughness; Exploratory Structural Equation Modeling; Psychometric Properties; & Elite Athletes*.

Introduction

Mental toughness (MT) has been conceptualised as a multi-dimensional construct characterised by unshakeable belief, coping effectively with pressure and adversity, being resilient, thriving on pressure, being committed, and having superior concentration skills (Connaughton, Hanton, & Jones, 2010; Crust, 2008; Clough, Earle, & Sewell, 2002). Research has indicated that athletes who score high on the mental toughness questionnaire 48 (MTQ48; Clough et al., 2002) exhibit higher pain tolerance (Crust & Clough, 2005), improved problem-task-orientated coping (Nicholls, Polman, Levy, & Backhouse, 2008), better attendance at injury rehabilitation clinics (Levy et al., 2006), more effective use of psychological strategies (Crust & Azadi, 2010), and an enhanced ability to prevent unwanted information interfering with cognition (Dewhurst, Anderson, Cotter, Crust, & Clough, 2012). Despite the importance given to MT in sport and calls in the literature to validate MT measurement with athletes, no study to date has directly examined the invariance of test scores between elite, amateur and non-athletes (Crust, 2008; Gerber et al., 2012; Golby & Sheard, 2004). If differences are to be attributed to athletic expertise rather than methodological reasons, then the assumption of measurement invariance with MT scales will be important. Additionally, the utility and psychometric properties of current measurement have yet to be evaluated in this context.

Clough and colleagues (2002) proposed a theoretical model of MT similar to the health concept of hardiness (Kobasa, 1979). However, Clough et al. added confidence to their framework so that MT could be conceptualised more accurately. Clough et al. coined this conceptualisation the 4Cs model which consists primarily of trait-like features albeit considered malleable over time with training (Lin, Mutz, Clough & Papageorgiou, 2017). The 4Cs model consists of four separate components, namely control, commitment, challenge and confidence. Clough and colleagues later augmented their theoretical model to better conceptualise the control and confidence components. Therefore, the model could manifest as either four or six components: *challenge*, which describes the degree to which individuals view difficulties as opportunities for personal development; *commitment*, which reflects deep involvement in pursuits and activities; *control of emotions*, which reflects control of anxieties and arousal in pressure situations; *control of life*, which reflects the belief that one is influential in determining outcomes; *confidence in abilities*, which involves a high sense of self-belief to achieve ones goals and less dependency on external influences; and *interpersonal confidence,* which reflects the ability to be assertive when interacting with others.

The hypothesised four and six-factor models (i.e. control and confidence is subdivided into two nested components) have formed the basis of research that has reviewed the psychometric properties of Clough and colleagues work. Clough et al.'s preliminary research adopted an abductive approach utilising the hardiness construct to propose the 4Cs model. This research bore resemblance to early MT research in that it was qualitatively driven. It is theorised that this emphasis resulted in less attention being given to measurement, that is, a lack of rigorous psychometric evaluation via quantitative methods (Crust & Swann, 2011). Furthermore, a recent analysis of this work has highlighted insufficient distinctiveness of Clough et al.'s conceptualisation, that is, whether the 4Cs model of MT is a distinct concept, or an extension of hardiness, thus clouding the uniqueness and operationalisation of the model (Gucciardi, 2017).

In order to operationalise MT and its components, Clough et al. (2002) developed the MTQ48 from a sample of 963 mixed student, athlete and occupational based participants (AQR, 2007). This 48-item measure has been used extensively within the MT literature, with high scores representing higher MT. The psychometric development of the MTQ48 involved principal components analysis with orthogonal varimax rotation. The most parsimonious model was a six-factor structure, thus supporting the conceptual six-factor model. Moreover, in order to facilitate use in applied settings Clough et al. developed a short 18-item version (MTQ18) alongside the MTQ48 representing a unidimensional interpretation which has yet to be psychometrically evaluated. Clough et al. provided evidence for the construct validity of the MTQ48 in terms of significant relationships with optimism (r = .48), self-image (r = .42), life satisfaction (r = .56), self-efficacy (r = .68), and trait anxiety (r = -.57). The MTQ48 has also been found to correlate with pain tolerance (Crust & Clough, 2005) and injury rehabilitation compliance via the MTQ18 (Levy et al., 2006). Clough et al. reported that participants with high MT reported lower ratings of exertion and the ability to bounce back after negative feedback during a physically demanding 30 minute cycling task across three trials controlling for fitness (e.g. VO2 Max), thus demonstrating the criterion validity of the MTQ48.

Despite the MTQ48's popularity it has been criticised due to insufficient conceptual distinctiveness, and poor psychometric evaluations resulting in a confusing narrative regarding the scales reliability and validity (Gucciardi et al., 2012; Gucciardi, Hanton, & Mallett, 2013). With regards to reliability, research has offered support for its stability at the total and subscale level. For example, internal consistency of the overall scale has been reported at α = .90, with its subscales reported at α = .71 - .91 (Nicholls et al., 2008). Test-retest coefficients have also been reported at .80 – .90 for both the total scale and subscales in a six-week interval assessment (Clough & Strycharczyk, 2012). However, studies have reported inadequate levels of internal consistency across the MTQ48 subscales (Crust & Keegan, 2010; Kaiseler, Polman & Nicholls, 2009; Levy et al., 2006; Nicholls, Levy, Polman & Crust, 2011), with the emotion and life control subscales frequently considered problematic (Crust & Swann, 2011). Despite the widespread use of the MTQ48 numerous studies have failed to test and substantiate the reliability of the scale, therefore conclusions on the reliability of the questionnaire lacks veracity (Gucciardi et al., 2012; Gucciardi et al., 2013).

The factorial validity of the MTQ48 has received mixed support with many investigations failing to provide data on the factor structure of the measure (Connaughton, Hanton, Jones, & Wadey, 2008). Factorial validity evidence provides insight into the adequacy of the operationalisation of a theoretical construct (Marsh, 2002). Perry et al. (2013) provided mixed support for the MTQ48 in a sample 8,207 participants. However, only 422 participants were athletes, with the remainder consisting of students (*n* = 978), and business staff (*n* = 6,786). Perry and colleagues reported good absolute fit to the data in a single, four and six-factor model using confirmatory factor analysis (CFA) and exploratory structural equation modelling (ESEM) with the latter reporting the best fit. It should be noted that some of the incremental fit measures fell below acceptable levels (Hu & Bentler, 1999), and several item factor loadings were poor or cross-loaded with unintended factors (Comrey & Lee, 1992). As a result, the authors called for refinement of the measure and warranted caution with some of the subscales. For example, the control emotion subscale consistently produced low internal consistency scores across samples which may have been a result of some negatively loaded items. Further research has provided support for the four-factor model utilising ESEM but not CFA (Gerber et al., 2013). However, despite good model fit and largely satisfactory loadings, several cross-loadings were reported in the sample of 424 physically active adolescents and young adults.

Gucciardi et al. (2012) examined the factor structure of the MTQ48's four and six-factor models in a sample of 1,325 participants consisting of athletes (*n* = 686) and managers (*n* = 639) utilising both CFA and ESEM techniques. The resultant analyses did not confirm the four or six-factor models proposed in the literature nor did the authors offer any alternative models moving forward. Nonetheless, these findings were questioned on the basis of inadequate sampling, over reliance on statistical methods, and narrow review of the literature (Clough et al., 2012). A recent review of the MTQ48's factorial validity utilising moderate (*n* = 480) and large (*n* = 1206) athletic samples also failed to provide support for the hypothesised four or six-factor models via CFA (Birch, Crampton, Greenless, Lowry & Coffee, 2017). The authors concluded that caution was warranted with use of the MTQ48 in athletic samples.

Previous research has assessed the psychometric properties of the MTQ48 in sport samples with limited support (Birch et al., 2017; Clough et al., 2002; Connaughton et al., 2008; Crust, 2008; Gucciardi et al., 2012; Perry et al., 2013). However, research examining the individual differences in MT across sport level has received little to no attention to date (Crust, 2008; Golby & Sheard, 2004). Research has indicated nominal differences with regard to the psychometric properties in participants from different achievement contexts (Perry et al., 2013). Equally, research has reported insignificant mean differences in MTQ48 scores across achievement level in 677 athletes (Nicholls, Polman, Levy & Backhouse, 2009). Nonetheless, Gerber et al. (2012) reported a positive relationship between MTQ48 scores and physical activity. This relationship differentiated those who engaged in no moderate physical activity and those who engaged in moderate physical activity five to seven days a week. Furthermore, the authors speculated that the MTQ48 items may be interpreted differently by elite athletes compared to non-athletes. An implicit assumption underlying previous research is that the same test items are appropriately interpreted across athletic groups. No study to date has rigorously tested the assumption that responses to the MTQ48 are reasonably invariant over athletic expertise. In order to corroborate previous conclusions based on athletic expertise it is important to clarify that mean differences are attributable to theoretical rather than methodological reasons (Marsh et al., 2013).

Construct validation should be viewed as a continuing process, therefore all measures must be subject to a thorough psychometric examination before they can be adopted as a useful measurement tool. In order to continue to assess the psychometric properties of the MTQ48, a substantial body of research supporting the dimensionality of the scale must be collected. Re-examination of the psychometric properties is therefore important in order to corroborate findings and conclusions of MT research. Research that has subjected the MTQ48 to rigorous psychometric examination across sport is scarce. Marsh et al. (2011) warn that the widespread use of a measure before establishing its properties can lead to in-construct problems that characterise many psychological measures. Nonetheless, Hopwood and Donnellan (2010) argued that one poor CFA result is not a legitimate reason to discredit all previous findings using the measure, and that a measure should be evaluated equally by confirming and falsifying results.

Following the recommendations of Gucciardi and colleagues (2012) this research will utilise a more flexible approach to psychometric evaluation by adopting the ESEM technique. Exploratory structural equation modelling is a relatively new methodological approach that combines the strengths of both CFA and exploratory factor analysis (EFA). For example, ESEM avoids the strict requirements of CFA (e.g. only certain items can load onto certain factors) by allowing cross-loadings of items on non-intended factors like in EFA, and providing robust indicators of model fit (e.g. goodness-of-fit statistics) that are available with CFA procedures. Recent research has advocated the use and benefits of ESEM over CFA, such as improved model accuracy, as it is less likely to distort model adequacy through constraining cross-loadings to zero (Marsh et al., 2011).

Golby and Sheard (2004) called for future studies to adopt larger and more inclusive samples to better understand the sport related individual differences in MT. It is therefore the aim of this study to answer calls in the literature to re-examine the psychometric properties of the MTQ48 and MTQ18 using robust flexible methods in a sample of elite, amateur and non-athletes in order to determine the utility of the scale in sport and across athlete profiles via invariance testing. It is hypothesised that the MTQ48 data would map onto both the four and six-factor models of Clough and colleagues theory of MT. Furthermore, we predict that the assumption of measurement invariance will hold across athlete expertise.

Methods

Participants

The sample consisted of 1096 participants predominantly from a large university in Northern Ireland (691 males & 405 females). A wide range of elite (*n* = 181), amateur (*n* = 577) and non-athletes (*n* = 338) from various team and individual sports (e.g. soccer, rugby, golf, karate, volleyball, basketball, hockey, athletics, boxing and tennis) aged 18 – 58 years (*M* = 23.11 & *SD* = 6.52) completed the questionnaire. Classification of athlete status was based on Swann, Moran and Piggott's (2015) inclusion criteria from a review of 91 studies on elite sports performance.

Myers, Ntoumanis, Gunnell, Gucciardi and Seungmin (2017) recommend the use of Monte Carlo simulation for estimation of sample size in structural equation modelling, however, no guidelines exist for parameter estimation in ESEM. Using Monte Carlo simulation, applying CFA estimations with no missing data, standard error biases that do not exceed 10%, and coverage of confidence intervals set at 95% indicated that sufficient power (80%) could be achieved with a sample size of 825 (see Muthén & Muthén 2009 for an overview of this analysis). Furthermore, general 'rules of thumb' regarding minimum sample sizes for factor analysis were used as guidelines for participant recruitment. For example, a minimum of 1000 cases required for an 'excellent' factor analysis (MacCallum, Widaman, Preacher, & Hong, 2001).

Procedure

Ethical approval was granted from the Ethics Committee at a university in Northern Ireland. A request was made to sport coaches and lecturers for permission to attend training sessions and classes to ask for participants to take part in the study. Data was collected at designated laboratories or training facilities using a questionnaire gauging biographical information and the MTQ48 items. Participants were briefed prior to data collection and informed of their ethical rights and provided informed consent to participate. After survey completion, participants were debriefed and thanked for their participation. Data collection was discontinued once the a priori numbers of cases were collected. Analyses were conducted using SPSSv23 (e.g. descriptive data and to prepare the dataset) and Mplus 7.4 (e.g. modelling techniques) statistical software programs (Muthen & Muthen, 2014).

Materials

Mental toughness was measured using the MTQ48 which theoretically taps the 4Cs model (Clough et al., 2002). Responses are made to 48-items on a 5-point Likert scale ranging from 1 (strongly disagree) to 5 (strongly agree). Scores are polarised with high scores representing higher levels of MT and vice versa with low scores. The self-report scale provides a total and four or six subscales representing Clough et al.'s (2002) model. Example items for each of the subscales is as follows: challenge (e.g., "I usually enjoy a challenge"), Commitment (e.g., "I usually find something to motivate me"), control emotion (e.g., "I tend to worry about things well before they actually happen"), control life (e.g., "I generally feel that I am in control of what happens in my life"), confidence abilities (e.g., "I generally feel that I am a worthwhile person"), and confidence interpersonal (e.g. I usually take charge of a situation when I feel it is appropriate"). A short form of the scale can be configured using 18-items to reflect a total MT score. Completion time of the scale ranges form 10 – 15 minutes for the 48-item version (Crust & Clough, 2005). The scale utilises reverse scoring to combat acquiescent responding. Finally, demographic information was collected for descriptive and grouping purposes.

Design & Data Analytic Strategy

Data were screened for outliers and missing data, and checked for univariate and multivariate normality. Only a small number of cases (1.2%) contained random missing data; therefore, listwise deletion was employed in line with the recommendations of Tabachnick and Fidell (2007). Subsequently, descriptive statistics and internal consistency was computed for the total 48 and 18-item scales and relevant subscales. Multivariate normality was checked using multivariate skewness and kurtosis coefficients to assess whether the data departed from normality. Cronbach's alpha has recently received criticism due to biases of over and under estimation, unsuitability with non-unidimensional scales, and issues with error (Dunn, Baguley & Brunsden, 2014). On the other hand, omega (McDonald, 1999) is much more sensitive to multidimensional scales and more accurate at estimating internal consistency in the congeneric model where error variances are allowed to vary, ergo more suitable for data generated for psychological constructs (Dunn et al., 2014). Therefore, Omega will be used to calculate internal consistency with coefficients of .70 or higher considered sufficient (Tabachnick & Fidell, 2007).

The dimensionality of the scale was assessed using ESEM (for an overview see Gucciardi & Zyphur, 2016). The initial analysis tested the short and long unidimensional models and the hypothesised four and six-factor models suggested in the literature to determine the most appropriate baseline model (Perry et al., 2013), followed by an assessment of measurement invariance with latent means analysis across elite, amateur and non-athletes. For tests of invariance, competing models will be subjected to successive equivalence constraints in the model parameters across groups until the most parsimonious fit is achieved. Measurement invariance will be examined using the Mplus procedure proposed by Muthen and Muthen (2014) where invariance is tested between the configural model (i.e., the same pattern of factors and loadings across groups), metric model (i.e., invariant loadings), and scalar model (i.e., invariant factor loadings and intercepts).

The analyses utilised the robust maximum likelihood (MLR) estimator which can handle instances of missing data, non-normality (Beauducel & Herzberg, 2006) and categorical variables when there are at least five response categories (Bandalos, 2014). As an a priori hypothesised model exists, albeit with conflicting evidence regarding the factor structure of the MTQ48, an exploratory oblique target rotation was used to estimate how the a priori 48-items and latent factors of the MTQ48 are interrelated (Muthen & Muthen, 2014). An epsilon value of .50 was adopted which enables as many items as possible to be optimally identified within one component while minimising the potential number of doublets (Comrey & Lee, 1992).

Model fit was determined by using a combination of fit-indices along with the likelihood ratio statistic - chi-square (χ^2) - as suggested by Hu and Bentler (1999). A model is deemed acceptable if the root mean square error of approximation (RMSEA) with 90% confidence intervals (CI) and standardised root mean residual (SRMR) is .06 or less, and each of the comparative fit index (CFI) and Tucker Lewis index (TLI) is .90 or greater (Marsh, Hau, & Grayson, 2005; Marsh, Hau, & Wen, 2004). In order to select the most parsimonious model, the Bayes information criterion (BIC) and Akaike's information criteria (AIC) was used to compare competing models. The AIC and BIC assign a greater penalty to model complexity and therefore has a better propensity to select more efficient models (Chen, 2007). Chen (2007) suggested that changes less than .01 and .015 in the CFI and RMSEA, respectively, would be supportive of an invariant model in relation to the previous model. Finally, due to the exploratory nature of ESEM standardised solutions were examined to evaluate the significance and strength of parameter estimates. Standardised factor loadings were interpreted using Comrey and Lee's (1992) recommendations (e.g. > .71 = excellent, > .63 = very good, > .55 = good, > .45 = fair, > .32 = poor).

Results

Preliminary Analyses

Measures of central tendency, distribution, and dispersion were tabulated for the total and subscale scores of the MTQ48 and MTQ18. The scores produced fall within the upper percentiles of the scale with no outliers. A partially negative distribution with slight nonkurtotic values was found, although not problematic for psychometric analysis (Tabachnick & Fidell, 2007). Multivariate skewness and kurtosis coefficients (Muthen & Muthen, 2014) indicated no departure from normality ($p > .05$). The internal consistency (Ω) for the MTQ48 ranged from **Ω =** .72 - .84, therefore indicating a good level of composite reliability (see Table 1). Finally, a strong positive correlation was found between the MTQ18 and MTQ48 ($r = .91$) demonstrating the utility of the MTQ18 as a global measure of MT.

Insert Table .1 Here.

ESEM Models

The one-factor model for the 18 and 48-item scales was tested first and indicated a poor fit to the data (see Table 2). The four-factor model represented a better fit to the data, albeit still inadequate based on the recommendations of Hu and Bentler (1999). Analysis of the modification indices indicated that good model fit could be achieved by allowing two sets of error terms to correlate. However, as the initial phase of analysis was aimed at identifying a parsimonious baseline model these options were not explored. The six-factor model indicated good fit to the data χ^2 (3153) = 5513.736, $p < .01$, RMSEA = .043 with 90% CI (.040 - .045), SRMR = .046, TLI = .938, CFI = .947.

Insert Table 2. Here

The factor structure of the six-factor model indicated a largely parsimonious fit with Clough and colleagues' hypothesised model. However, several instances of misspecification existed (i.e., weak intended loadings and cross-loading items) for all factors with the confidence subscales containing the least with only one item cross-loading, and the control subscales containing the most with three items failing to load on their intended factor (see Table 3).

Insert Table 3. Here

Invariance Testing

Measurement invariance was tested comparing the six-factor configural model (e.g. all parameters allowed to be unequal across groups) to the weak invariance model (e.g. by holding loadings equal across groups) which produced fit that was significantly poorer ($\Delta\chi^2$ (236) = .544.764, $p < .05$). Comparison of the metric against the scalar model which imposed additional constraints of strong invariance (e.g. by constraining factor loadings and intercepts across groups) also produced poorer fit ($\Delta\chi^2$ (588) = 1159.986, $p < .01$). Therefore, suggesting that measurement of the six-factor model differs across elite, amateur and non-athletes. Furthermore, the parsimony corrected AIC and BIC produced lower values for the configural model. Nonetheless, all models produced adequate fits to the data with no significant change in cut-offs suggested by Chen (2007) (see Table 2).

Parameter Estimates for Invariance Measurement Models

The next stage of the analysis was to examine the factor structure of the six-factor model across elite, amateur and non-athletes (see supplementary material). The analysis of the latent means across groups were all freely estimated and produced factor matrixes that were partially representative of Clough et al.'s six-factor model of MT. The factor solution from the non-athletes produced the matrix with the least misspecification. Further inspection of the factor loadings revealed a degree of inconsistency between the hypothesised structure, according to the correlated six-factor model proposed by Clough et al. and the current data in the athlete groups. The factor loadings and residual variances produced values that indicated strong representations of their latent factors with most loadings producing scores ranging from excellent to poor on their intended subscale (Comrey & Lee, 1992). Nonetheless, the confidence subscales (abilities and interpersonal) contained three (6.25%) misloading items (e.g. items 32, 36 and 38), which was typical across elite, amateur and non-athlete groups. Furthermore, three (6.25%) items (e.g. 18, 19 and 33) had poor factor loadings (< .32) across elite, amateur and non-athletes (Comrey & Lee, 1992). The factor structure produced from the amateur athletes indicated the poorest fit with more cases of weak and improper cross-loadings, thus not representative of Clough and colleagues' six-factor model. The latent factor correlations (see supplementary material) indicate independence amongst the subscales (r = -.01 to .52) with the confidence subscales (abilities and interpersonal) displaying the weakest correlations.

Discussion

The aim of this research was to assess the structure of the MTQ48 and MTQ18 in a sample of elite, amateur and non-athletes. The findings indicated that the scale possesses high scores of internal consistency for all scales of Clough et al.'s (2002) MT model. Results from ESEM indicated that the six-factor model produced acceptable and better fit to the data compared to the four and one-factor models proposed in the literature (Perry et al., 2013). Moreover, the four-factor model did not produce a good fit the data similar to the Gucciardi et al. (2013) and Birch et al. (2017) findings. Nonetheless, several instances of weak and cross-loading items were noted in the six-factor model thus detracting from the models psychometric quality. Next, invariance testing suggested measurement invariance across elite, amateur and non-athletes. Furthermore, the factor structures indicated a large degree of misspecification with many instances of unacceptable loadings across all three groups.

The factor structure from the non-athletes produced the best fit and lowest χ^2 value, whereas the amateur athletes produced the worst fit and highest χ^2 value. Nonetheless, acceptable model fit was achieved in Clough et al.'s (2002) six-factor conceptualisation of MT thus partially supporting its factorial validity. However, analysis of the latent factor correlations indicated that the confidence subscales were not as strongly correlated with some of the other MT factors across groups, particularly non-athletes. This finding is noteworthy given that Clough et al. added confidence to the hardiness construct (Kobasa, 1979) in order to conceptualise their model of MT. Therefore, Clough et al.'s extension of the hardiness construct may not be as theoretically important as the other hardiness components (Gucciardi, 2017). Interestingly, the confidence subscales rotated with the least amount of misspecification in the overall sample. It is possible that context non-specific challenge, control and commitment is more subject to interpretation compared to confidence which may have resulted in its lack of congruence with the other components.

The total scale and subscales internal consistency was above the pre-determined .70 cut-off (Tabachnick & Fidell, 2007), therefore the scale can be considered reliable. However, these scores may have been a result of the limitations associated with the MTQ48's psychometric stability. For example, as the MTQ48 provides scores for overall aggregated MT and individual subscales, the composite reliability may have become inflated due to the high residual variances and factors loadings used to calculate Omega. Although internal consistency was achieved for all sub-scales, the lowest reliability was associated with the emotional control sub-scale. Research has indicated that this factor is problematic possibly due to the increased variability associated with emotional differences between individual's personalities (Crust & Swann, 2011). Nonetheless, it remains an important theoretical component of the MT model and was found to be internally reliable in this research. Researchers should note caution when using reliability estimates as the sole indicator of a scales utility with a particular sample. Although important in establishing consistency in results, researchers should also consider the practical aspects of what the scores from the data represent (Marsh et al., 2004; Marsh et al., 2005).

These findings do not support the work of Gucciardi et al. (2012) who also adopted an ESEM approach, but reported that the MTQ48 produced a poor fit to the data. Conversely, they also do not corroborate Perry and colleagues (2013) results as only the six-factor model produced acceptable fit as opposed to a one and four-factor solution. The findings of this research coincide with the literature in that the Clough and colleagues model of MT requires refinement (e.g. the data did not fit the 4Cs model with athletes). These findings raise concerns at two levels, first, the inability to fit the hypothesised four-factor model and second, the inconsistency in the factor structures across elite, amateur and non-athletes. Research has cautioned the use of confirmatory factor analytic techniques as a singular method for determining the psychometric properties of a measure (Hopwood & Donnellan, 2010; Marsh et al., 2011). However, it is believed that establishing factorial validity should be critical in assessing the robustness of a measure as this will provide evidence for a theory strong operationalisation (Gucciardi et al., 2013).

ESEM adopts a flexible approach to instrument evaluation however, as in all EFA techniques, its rotation procedures are numerically driven and negate theory, and different rotation procedures may produce different factor solutions but similar fit statistics (Asparouhov & Muthen, 2009; Tomas et al., 2014). Considering that the MTQ48 is an aggregate and multidimensional scale, providing overall and individual subscale scores, the scale must have moderate inter-correlations in order to obtain suitable internal consistency at the scale development phase. Thus, some misspecification may arise in ESEM. Future research should test this theory by examining bi-factor structures assessing the adequacy of the overall and subscale framework.

Although the MTQ48 has been previously evaluated in athletic samples, the current study is the first to examine the scale across expertise levels via invariance testing. Research has reported that elite athletes typically score higher than amateur (Golby & Sheard, 2004) and non-athletes (Gerber et al., 2012) with the latter suggesting that elite athletes may interpret the MTQ48 items differently to non-athletes. In comparison with previous psychometric research, the current findings are encouraging when considering the degree of misspecification. For example, Birch et al. (2017), Gucciardi et al. (2012) and Perry et al. (2013) reported unacceptable levels of fit and large degrees of misspecification in their data, whereas the current investigation found relatively acceptable levels of misspecification in an ESEM framework (Perry, Nicholls, Clough & Crust, 2015). However, this misspecification in the factor structure became unacceptable at the group level with each component of the six-factor model containing at least three instances of misspecification and three items failing to load on their intended factors across groups.

The current findings warrant caution regarding use of the MTQ48 with athletic populations as the largest degree of misspecification in the factor structure was found in the athlete groups. Furthermore, the MTQ48 is a general measure of MT (Clough et al., 2002), which may result in difficulties in item interpretation across samples different to its validation data (e.g. participants largely from business settings). For example, Gucciardi and Gordon (2009) developed a psychometrically sound measure of MT in cricket, however, the application and generalisability of data developed from this measure is inconclusive to the MT literature as a whole. Other researchers have successfully modified existing measures for domain specific purposes, for example, the COPE for measuring coping in sport (see Gaudreau & Blondin, 2002). Therefore, future research may wish to refine item wording of the MTQ48 to suit samples from different domains.

Psychometric evaluation should be based on both theoretical and empirical evidence (Hopwood & Donnellan, 2010). However, building a consensus and progressing with the MTQ48 is difficult due to the multidimensional framework proposed by Clough et al. (2002) i.e. competing one, four and six-factor models. As much research substantiates the scales reliability, but less so with regards to its validity, it is clear the scale measures something consistently (Clough & Strycharczyk, 2012; Crust & Clough, 2005; Crust & Swann, 2011), however what that is appears conceptually vague (Birch et al., 2017; Gucciardi et al., 2012; Perry et al., 2013). Thus, more empirical evidence is required to refine and corroborate Clough and colleagues' operationalisation of MT. Therefore, this research does not discredit the psychometric properties of the scale, but calls for Clough and colleagues to substantiate a direction for future research utilising either the hypothesised four or six-factor models. Doing so will enable MT researchers to develop a clear body of evidence underpinned by the same theoretical understanding which will help progress and develop the study of MT. In the meantime, researchers should interpret the data generated from the MTQ48 with caution in samples of elite athletes because the factor structure resembled the six-factor model in non-athletes but was less convincing in the elite and amateur athlete groups.

A strength of the aforementioned research is the size and coverage of the sample which offers a comprehensive domain of expression of MT in a sports context. Nonetheless, the current research findings should be interpreted in light of several limitations. First, cut-offs adopted for the ESEM fit indices were recommended for CFA procedures with no ESEM specific indicators developed. Second, the elite athlete sample size may have been inadequate for ESEM; future research should endeavour to increase the sample size of elite athletes or conduct simulation analyses to determine what may be considered sufficient within the context of the MTQ48 and other multidimensional scales.

In conclusion the MTQ48 achieved psychometric robustness on measures of composite reliability. However, ESEM techniques were unable to fit the hypothesised four-factor model to the data thus questioning the factorial validity of the MTQ48. Nonetheless, the alternative six-factor model did produce fit to the data but the factor solution contained instances of misspecification (e.g., poor intended factor loadings and cross-loading items). Measurement invariance models produced acceptable fit however the factor structure for elite, amateur and non-athletes differed indicating poor representations of their latent factors. It should be noted that just as one study cannot discredit a scale, one assessment cannot provide conclusive evidence for its reliability and validity (Hopwood & Donnellan, 2010). Therefore, this study does not reject the MTQ48 as a measure of MT at a general level; however, it calls for Clough and colleagues to refine the measure for use with different samples, and researchers should be aware of this when using the scale with elite athletes. Finally, Clough et al. are encouraged to revisit the theoretical basis of the MTQ48 and clarify its stance as a one, four or six-factor model so that future research can develop a consensus on the MT construct.

References

AQR. (2007). Mental Toughness Questionnaire 48 Technical Manual. Retrieved September 29th from http://civilsociety360.co.uk/sites/default/files/The%20MTQ48%20Technical%20Manual%20-%20Final%202015.pdf

Asparouhov, T. & Muthen, B. (2009). Exploratory Structural Equation Modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16(3)*, 397 – 438.

Bandalos, D. L. (2014). Relative performance of categorical diagonally weighted least squares and robust maximum likelihood estimation. *Structural Equation Modeling, 21*, 102 – 116.

Beauducel, A., & Herzberg, P. Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling, 13,* 186 – 203.

Birch, P.D. J., Crampton, S., Greenless, I., Lowry, R., & Coffee, P. (2017). The mental toughness questionnaire-48: A re-examination of factorial validity. *International Journal of Sport Psychology, 48(3),* 331-355.

Chen, F. F. (2007). Sensitivity of goodness of fit indices to lack of measurement invariance. *Structural Equation Modeling, 14,* 464 – 504.

Clough, P., Earle, K., & Sewell, D. (2002). Mental toughness: the concept and its measurement. In I. Cockerill (ed.), *Solutions in Sport Psychology* (pp. 32 – 43). London: Thomson.

Clough, P., Earle, K., Perry, J. L., & Crust, L. (2012). A response to progressing measurement in mental toughness: A case example of the mental toughness questionnaire 48 by Guccairdi, Hanton, & Mallett. *Sport, Exercise, and Performance Psychology, 1(3),* 283 – 287.

Clough, P., & Strycharczyk, D. (2012). *Mental Toughness: Improving performance, wellbeing, and positive behaviour in others*. London: Kogan Page Limited.

Comrey, A. L., & Lee, H. B. (1992). *A First Course in Factor Analysis*. Hillsdale, NJ: Erlbaum.

Connaughton, D., Hanton, S., & Jones, G. (2010). The development and maintenance of mental toughness in the world's best performers. *The Sport Psychologist*, 24, 168–193.

Connaughton, D., Hanton, S., Jones, G., & Wadey, R. (2008). Mental toughness research: Key issues in this area. *International Journal of Sport Psychology, 39,* 192 – 204.

Crust, L. (2008). A review and conceptual re-examination of mental toughness: implications for future researchers. *Personality and Individual Differences, 45,* 576 – 583.

Crust, L., & Azadi, K. (2010). Mental toughness and athletes' use of psychological strategies. *European Journal of Sports Science*, *10,* 43 – 51.

Crust, L., & Clough, P. J. (2005). Relationship between mental toughness and physical endurance. *Perceptual and Motor Skills*, *100*, 192 – 194.

Crust, L., & Keegan, R. (2010). Mental toughness and attitudes to risk-taking. *Personality and Individual Differences, 49,* 164–168.

Crust, L., & Swann, C. (2011). Comparing two measures of mental toughness. *Personality and Individual Differences, 50*, 217 – 221.

Dewhurst, S., Anderson, R., Cotter, G., Crust, l., & Clough, P. (2012). Identifying the cognitive bias of mental toughness: Evidence from the directed forgetting paradigm. *Personality and Individual Differences, 53*, 587 – 590.

Dunn, T. J., Baguley, T., Brunsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, *105*, 399 – 412.

Gaudreau, P. & Blondin, J. P. (2002). Development of a questionnaire for the assessment of coping strategies employed by athletes in competitive sport settings. *Psychology of Sport and Exercise, 3,* 1–34.

Gerber, M., Kalak, N., Lemola, S., Clough, P. J., Perry, J. L., Puhse, U., Elliot, C., Holsboer-Trachsler, E., Brand, S. (2013). Are adolescents with high mental toughness levels more resilient against stress? *Stress & Health, 29(2),* 164-171.

Gerber, M., Kalak, N., Lemola, S., Clough, P. J., Puhse, U., Elliot, C., Holsboer-Trachesler, E., & Brand, S. (2012). Adolescents' exercise and physical activity are associated with mental toughness. *Mental Health and Physical Activity, 5(1),* 35-42.

Golby, J. & Sheard, M. (2004). Mental toughness and hardiness at different levels of rugby league. *Personality and Individual Differences, 37,* 933 – 942.

Gucciardi, D. (2017). Mental toughness: progress and prospects. *Current Opinion in Psychology, 16,* 17-23.

Gucciardi, D., & Gordon, S. (2009). Development and preliminary validation of the Cricket Mental Toughness Inventory (CMTI). *Journal of Sports Sciences, 27,* 1293 – 1310.

Gucciardi, D., Hanton, S., & Mallett, C. (2012). Progressing measurement in mental toughness: A case example of the mental toughness questionnaire 48. *Sport, Exercise and Performance Psychology, 1(3),* 194 – 214.

Gucciardi, D., Hanton, S., & Mallett, C. (2013). Progressing measurement in mental toughness: A Response to Clough, Earle, Perry, and Crust. *Sport, Exercise and Performance Psychology, 2(3),* 157 – 172.

Gucciardi, D. F., & Zyphur, M. J. (2016). *Exploratory structural equation modelling and Bayesian estimation*. In N. Ntoumanis & N.D. Myers (Eds.), An introduction to intermediate and advanced statistical analyses for sport and exercise scientists (pp. 172-194). Chichester, West Sussex: Wiley.

Hopwood, C. J. & Donnellan, M. B. (2010). How should the internal structure of personality inventories be evaluated? *Personality and Social Psychology Review, 14*, 332 – 346.

Hu, L. T., & Bentler, P. M. (1999). Cut off criteria for fit indexes in covariance structural analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*, 1 -55.

Levy, A. R., Polman, R. C. J., Clough, P. J., Marchant, D. C., & Earle, K. (2006). Mental toughness as a determinant of beliefs, pain, and adherence in sport injury rehabilitation. *Journal of Sports Rehabilitation, 15*, 246 – 254.

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: Personality and Social Psychology*. 8:1345. doi: 10.3389/fpsyg.2017.01345

Kaiseler, M., Polman, R., & Nicholls, A. (2009). Mental toughness, stress, stress appraisal, coping and coping effectiveness in sport. *Personality and Individual Differences, 47*, 728–733.

Kobasa, S. C. (1979). Stressful life events, personality, and health – inquiry into hardiness. *Journal of Personality and Social Psychology, 37(1*), 1 – 10.

MacCallum, R. C., Widaman, K. F., Preacher, K. J., & Hong, S. (2001). Sample size in factor analysis: The role of model error. *Multivariate Behavioral Research, 36*, 611 – 637.

Marsh, H. W. (2002). A multidimensional physical self-concept: A construct validity approach to theory, measurement, and research. *Psychology: The Journal of the Hellenic Psychology Society, 9*, 459-493.

Marsh, H. W., Hau, K. T., & Grayson, D. (2005). Goodness of Fit Evaluation in Structural Equation Modelling. In A. Maydeu-Olivares & J. McArdle (Eds.), *Contemporary*

Psychometrics. A Festschrift for Roderick P. McDonald (pp. 275-340). Mahwah NJ: Erlbaum.

Marsh, H. W., Hau, K. T., & Wen, Z. (2004). In search of golden rules: Comment on hypothesis testing approaches to setting cut-off values for fit indexes and dangers in over-generalising Hu & Bentler's (1999) findings. *Structural Equation Modeling*, *11*, 320–341.

Marsh, H. W., Liem, G. A. D., Martin, A. J., Morin, A. J. S., & Nagengast, B. (2011). Methodological measurement fruitfulness of exploratory structural equation modeling (ESEM): New approaches to key substantive issues in motivation and engagement. *Journal of Psychoeducational Assessment*, *29*, 322 – 346.

Marsh, H. W., Vallerand, R. J., Lafreniere, M. K., Parker, P., Morin, A. J. S., Carbonneau, N., Jowett, S., Bureau, J. S., Fernet, C., Guay, F., Abduljabbar, A. S., & Paquet, Y. (2013). Passion: does one scale fit all? Construct validity of two-factor passion scale and psychometric invariance over different activities and languages. *Psychological Assessment*, *25*, 796 – 809.

McDonald, R. P. (1999). *Test theory: A unified treatment*. Mahwah, NJ: Lawrence Erlbaum.

Muthén, L. K. & Muthén, B. O. (2009). How to use a Monte carlo study to decide on sample size and determine power. Structural Equation Modeling: A Multidisciplinary Journal, 9(4), 599-620.

Muthén, L. K. & Muthén, B. O. (2014). *Mplus User's Guide*. Los Angeles, CA: Muthén & Muthén.

Myers, N. D., Ntoumanis, N., Gunnell, K. E., Gucciardi, D. F., & Seungmin, L. (2017). A review of some emergent quantitative analyses in sport and exercise psychology. International review of Sport and Exercise Psychology. doi: 10.1080/1750984X.2017.1317356

Nicholls, A. R., Levy, A. R., Polman, R. C. J., & Crust, L. (2011). Mental toughness, coping self-efficacy, and coping effectiveness among athletes. *International Journal of Sport Psychology*, *42*, 513-524.

Nicholls, A. R., Polman, R. C., Levy, A. R., & Backhouse, S. H. (2008). Mental toughness, optimism, and coping among athletes. *Personality and Individual Differences, 44*, 1182 – 1192.

Nicholls, A. R., Polman, R. C., 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*, 73 – 75.

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*, 587 – 592.

Perry, J. L., Nicholls, A. R., Clough, P. J., & Crust, L. (2015). Assessing Model Fit: Caveats and recommendations for confirmatory factor analysis and exploratory structural equation modeling. *Measurement in Physical Education and Exercise Science, 19(1)*, 12-21.

Swann, C., Moran, A., & Piggott, D. (2015). Defining elite athletes: Issues in the study of expert performance in sport psychology. *Psychology of Sport and Exercise 16(1)*, 3 – 14.

Tabachnick, B.G., & Fidell, L.S. (2007). *Using Multivariate Statistics*. Boston: Allyn and Bacon.

Tomas, I., Marsh, H. W., Gonzalez-Roma, V., Valls, V. & Nagengast, B. (2014). Testing measurement invariance across Spanish and English versions of the Physical Self-Description Questionnaire: An application of exploratory structural equation modelling. *Journal of Sport & Exercise Psychology, 36*, 179 – 188.

Table 1.

Means, Standard Deviations, Skewness, Kurtosis, and Reliability (Ω) Scores for components of the MTQ48 Total, Subscale, and Short Scale Scores across Elite, Amateur and Non-Athletes.

Scale (Items)	**M (SD)**				**Skewness**	**Kurtosis**	**Ω**
	Total	**Non-Athlete**	**Amateur**	**Elite**			
Total (48)	174.17 (42.50)	138.25 (35.09)	179.79 (27.93)	182.01 (25.83)	-.69	-1.43	.84
Total (18)	62.30 (14.66)	51.20 (12.64)	66.18 (12.78)	67.41 (12.30)	-.91	-1.68	.81
Challenge (8)	30.15 (9.08)	24.13 (6.53)	31.12 (5.16)	31.51 (4.88)	-.70	-1.21	.80
Commitment (11)	41.77 (12.01)	33.83 (8.56)	42.92 (6.91)	44.32 (6.46)	-.69	-1.29	.83
Control (14)	50.51 (15.06)	38.85 (9.38)	49.89 (9.68)	50.49 (8.71)	-.70	-1.31	.82
Control Emotion (7)	24.49 (6.64)	20.22 (5.06)	24.90 (4.38)	25.53 (4.31)	-.74	-1.05	.72
Control Life (7)	26.01 (26.01)	20.47 (6.10)	27.22 (4.56)	27.09 (4.50)	-.71	-1.24	.83
Confidence (15)	55.47 (18.53)	41.85 (12.30)	56.45 (10.77)	56.83 (10.61)	-.71	-1.33	.83
Confidence Abilities (9)	33.27 (11.10)	24.34 (7.45)	32.01 (7.35)	33.66 (7.18)	-.71	-1.26	.82
Confidence Interpersonal (6)	22.20 (7.66)	17.50 (5.76)	23.45 (4.30)	23.84 (5.56)	-.73	-1.19	.83

Table 2.

Global Fit Indices of the One, Four and Six Factor MTQ48 and MTQ18 Models.

Model	X^2	***df***	**RMSEA (ULCI-LLCI)**	**SRMR**	**TLI**	**CFI**	**AIC**	**BIC**
1 Factor Long	11737.59	3428	.084 (.082-.087)	.144	.702	.711	135393.97	136613.82
1 Factor Short	1657.46	135	.101 (.104-.099)	.067	.836	.855	53357.050	53627.019
4 Factor	6787.04	3266	.053 (.056-.050)	.061	.870	.881	130767.41	132797.17
6 Factor	5513.74	3153	.043 (.045.040)	.046	.938	.947	129710.11	132305.81
Configural	4353.75	2565	.045 (.048-.043)	.041	.925	.933	129729.12	132333.48
Metric	4898.51	2801	.046 (.049-.043)	.044	.919	.927	129833.52	132398.69
Scalar	5513.74	3153	.047 (.050-.044)	.045	.917	.924	129794.18	132389.41

Note. X^2 = Chi-Square, RMSEA = Root Mean Square Error of Approximation, ULCI = Upper Limit Confidence Interval, LLCI = Lower Limit Confidence Interval, SRMR = Standardised Root Mean Residual, Tucker Lewis Index, CFI = Comparative Fit Index, AIC = Akaike Information Criteria, BIC = Bayes Information Criterion. N = 1096.

Table 3.

Parameter Estimates for Total Sample on the Six Factor MTQ48 Model.

Item	**Factor 1**	**Factor 2**	**Factor 3**	**Factor 4**	**Factor 5**	**Factor 6**
Challenge						
MT4	**.752**	.054	.031	.132	-.043	-.021
MT6	**.319**	.196	.004	.191	.149	-.014
MT14	.148	**.406**	.122	.044	.094	.258
MT23	**.614**	.335	.002	-.025	.033	-.053
MT30	**.502**	.060	.131	.181	.004	-.192
MT40	.204	**.354**	.313	.066	.099	-.124
MT44	**.839**	-.066	.005	.172	-.027	-.033
MT48	**.508**	.061	.078	.085	-.083	-.032
MT1	**.594**	.054	.020	-.026	-.069	.086
Commitment						
MT7	**.742**	-.025	.003	.057	.009	.153
MT11	.064	.186	**.620**	.061	-.177	.094

MT19	-.289	.014	**.720**	.082	.081	.168
MT22	.328	.091	**.347**	.015	.048	.290
MT25	.127	-.244	**.632**	-.100	-.012	.082
MT29	-.122	-.004	**.648**	.104	.043	.303
MT35	.060	.029	**.338**	-.091	-.082	.033
MT39	.264	.015	**.637**	.093	-.026	.045
MT42	.243	.027	**.555**	-.024	-.022	.307
MT47	**.518**	.053	.067	.119	.028	.260
Control Emotion						
MT21	-.248	**.879**	.145	-.011	.086	.016
MT26	.615	.098	.078	-.149	**.794**	.066
MT27	-.106	**.657**	.558	.185	.087	-.076
MT31	**.573**	.354	.154	-.001	.085	-.129
MT34	.609	-.031	-.038	.001	**.617**	-.080
MT37	.351	-.016	.283	.050	**.384**	.068
MT45	.140	.006	-.287	-.040	**.582**	-.159

Control Life						
MT2	.131	.322	-.077	.106	-.015	**.487**
MT5	.307	.057	.086	**.396**	.002	-.048
MT9	.222	.113	.213	.042	.101	**.323**
MT12	.336	**.514**	-.074	.043	.002	.095
MT15	.080	.433	-0.22	.124	.068	**.310**
MT33	.033	.447	.100	-.033	.057	**.525**
MT41	.255	.398	.033	.091	-.092	**.545**
Confidence Ability						
MT3	.307	**.637**	-.176	.044	-.118	.194
MT8	.378	**.314**	.040	.142	-.142	.060
MT10	.044	**.547**	.416	.045	.066	.154
MT13	.502	**.625**	-.110	-.147	-.017	-.014
MT16	.369	**.561**	.004	-.210	-.112	-.034
MT18	.059	**.701**	.182	-.036	-.186	.220
MT24	-.015	.476	**.559**	.079	.180	.015

MT32	.003	**.615**	.117	-.033	-.094	.472
MT36	-.020	**.486**	.471	.282	.050	0.31
Confidence Interpersonal						
MT17	.223	.035	-.014	**.713**	-.057	.065
MT20	.332	.014	-.102	**.593**	-.019	.026
MT28	.074	**.378**	.192	.346	-.006	.121
MT38	.117	.150	.225	**.386**	-.154	-.019
MT43	.062	.014	.046	**.832**	.018	.091
MT46	.058	.122	-.001	**.382**	-.003	.202

Note. Values in bold indicate highest loading on that factor. Values underlined are interpreted as a factor. N = 1096.