Using Genetic Algorithms to Abbreviate the Mindfulness Inventory for Sport: A Substantive-Methodological Synthesis

Michael Noetel[a,b], Joseph Ciarrochi[b], Baljinder Sahdra[b], & Chris Lonsdale[b]

[a] School of Behavioural and Health Sciences, Faculty of Health Sciences, Australian Catholic University, 25A Barker Rd, Strathfield, NSW, Australia;

[b] Institute for Positive Psychology and Education, Faculty of Health Sciences, Australian Catholic University, 33 Berry St, North Sydney, NSW, Australia;

Corresponding Author: michael.noetel@acu.edu.au; ORCID 0000-0002-6563-8203

joseph.ciarrochi@acu.edu.au; ORCID 0000-0003-0471-8100

baljinder.sahdra@acu.edu.au; ORCID 0000-0001-5064-7783

chris.lonsdale@acu.edu.au; ORCID 0000-0002-2523-5565

Abstract

Objectives. To demonstrate the use of machine-learning for reducing questionnaire response burden, we created multiple, shorter versions of the Mindfulness Inventory for Sport. We then tested the reliability and validity of scores derived from these shorter versions in athletic populations.

Design. We used genetic algorithms to shorten the measure, and both cross-sectional and longitudinal data to test psychometric properties.

Method. We collected data from 859 undergraduate exercise science students and 118 golfers. We used 75% of the student sample to shorten the measure, and the rest of the data to test the internal consistency, test-retest reliability, content validity, and factorial validity. For criterion validity, we explored relationships between the subscales and other measures of mindfulness, golf handicaps, and an objective measure of putting accuracy.

Results. Genetic algorithms efficiently generated stable solutions to shortening the measure. Reliability decreased as the measure become shorter—especially between three and two items per subscale—but remained acceptable. Validity metrics for shorter versions were as good, and sometimes better, than the full questionnaire. Awareness and refocusing subscales demonstrated weak associations with golf handicap for long and short versions. Non-judgment showed no significant associations, and no subscales significantly predicted putting performance.

Conclusions. Genetic algorithms provide efficient solutions to reducing questionnaire response burden for athletes.

Keywords

Psychometric validity, reliability, machine learning, athletes, mindfulness, acceptance

Highlights

- Genetic algorithms efficiently shortened the Mindfulness Inventory for Sport
- Shorter measures demonstrated very high correlations with the full measure
- Most shorter measures still met acceptable standards for reliability
- Shorter measures demonstrated equivalent validity compared with longer measures
- The Mindfulness Inventory for Sport demonstrated weak associations, if any, with measures of golfing performance

Introduction

Self-report measures are the most common form of assessment in athletic populations (Halson, 2014). These measures are not perfect, with many implementations being subject to biases such as social desirability, arbitrary metrics, and common-method variance (for a longer discussion of these limitations, see: Blanton & Jaccard, 2006; Moorman & Podsakoff, 1992; Podsakoff, MacKenzie, & Podsakoff, 2012). Nevertheless, self-report measures are simple and inexpensive ways of measuring many constructs (Saw, Main, & Gastin, 2015). As a result, athletes complete many questionnaires. For example, of 55 high-performance sport programs in Australia and New Zealand, 55% administered questionnaires on a daily basis, with 97% doing it at least once per week (Taylor, Chapman, Cronin, Newton, & Gill, 2012). However, when coaches, athletes, and sport scientists were surveyed about faithful completion of questionnaires, one of the most common barriers was questionnaire length (Saw et al., 2015). There are many ways of reducing the response burden imposed at any given time-point: questionnaires could be distributed across time-points; researchers could use matrix sampling where questionnaire items are deliberately left missing at random (Kaplan & Su, 2016); or, as we will describe, researchers can use short measures that explain the majority of variance in the longer version.

Participants may be more likely to complete questionnaires, and complete them faithfully, when measures are brief. In a meta-analysis of randomised studies that assed different questionnaire formats, no studies found differences in data quality between long and short versions of questionnaires (Rolstad, Adler, & Rydén, 2011). Instead, they found those participants who were randomly assigned to the shorter questionnaire were more likely to complete it and return it to the experimenter (Rolstad et al., 2011). In another systematic review of survey design factors, Fan and Yan (2010) reported an inverse relationship between questionnaire length and completion rate. Incentives can increase response rates, but longer questionnaires often require commensurate incentives (Sahdra, Ciarrochi, Parker, & Scrucca, 2016). Shorter measures, therefore, can make data collection more affordable. In a randomised controlled trial comparing incentives, reminders, and shortened questionnaires, the only cost-effective strategy was an abridged questionnaire following a reminder (Glidewell et al., 2012). When time with clients is limited, short measures can reduce the in-session time required for clients to complete measures, and they can increase compliance with measures assigned between sessions (Basarkod, Sahdra, & Ciarrochi, 2018). Overall, short measures can allow for more flexible, efficient, affordable, and accurate data collection.

Short measures also facilitate many core aims of science, including description, prediction, and causal inference. Short questionnaires allow for more constructs to be assed in the same battery (Basarkod et al., 2018). Assessing multiple constructs using poor empirical designs can be fraught, because it may increase the probability of spurious correlations or Type I error, but assessing multiple constructs may answer more interesting questions, when appropriate controls are in place (e.g., Bonferroni corrections for multiple comparisons). This process has helped to describe patterns across nations through international surveys like Programme for International Student Assessment (Kaplan & Su, 2016) and the European Social Survey (Davidov, Schmidt, & Schwartz, 2008). It has helped with prediction, where outcomes are often predicted more accurately by including a large number of heterogeneous constructs than a small number from within one theory (Parker, Jerrim, Chmielewski, & Marsh, 2017). Short measures also help determine causality. While experimental design is crucial for internal validity, statistical controls provide alternate methods of supporting causal claims (Miller, Henry, & Votruba-Drzal, 2016), and shorter measures facilitate these controls. For example, causal models can be strengthened if researchers assess the gamut of variables that influence an outcome so researchers can statistically control for confounds and alternate explanations (Pearl, 2009). In intervention research, for example, brief measures allow researchers to more easily assess a larger number of baseline characteristics, mediators, and outcomes. So, short measures are not just pragmatic, but helpful for many of the core goals of sport and exercise psychology research.

Although they are useful tools for achieving important goals, many short measures fail to meet established psychometric standards (e.g., reliability, factor structure, correlation with full measure; Smith, McCarthy, & Anderson, 2000). They “have frequently been developed without careful, thorough examination of the new form's validity” (Smith et al., 2000, p. 102). Often, the final size of the measure is arbitrary, rather than being based on theoretical or empirical grounds. One problem is that traditional approaches to shortening measures require a delicate balance of competing factors. These approaches find the items that balance high item-total correlations, low cross-loadings, low correlated uniqueness, low chance of missingness, high face validity of construct coverage, and high internal consistency of the resulting scale (Marsh, Ellis, Parada, Richards, & Heubeck, 2005). This process asks researchers to consider seven different elements simultaneously, meaning that the possible combinations are beyond human calculation, even for relatively small measures. When reducing a 15-item scale to six items, there are over 2,500 possible combinations to consider. The complexity means researchers have to rely on unwritten idiosyncratic heuristics, frequently leading to poor measures (Smith et al., 2000). Even when using Item Response Theory to shorten questionnaires, researchers must make a series of judgments on the basis of a large number of parameters (Fletcher & Hattie, 2004).

Genetic Algorithm Approach

An efficient alternative to these approaches is to use machine learning to find the items that explain the most variability in the full measure (Sandy, Gosling, & Koelkebeck, 2014). Genetic algorithms were first introduced to solve pattern recognition problems (e.g., finding the best match for a visual pattern) and to computationally simulate biological processes (Holland, 1975). More recently, machine-learning techniques based on genetic algorithms have been used to create reliable and valid shortened measures of personality (Yarkoni, 2010), values (Sandy et al., 2014), psychopathy (Eisenbarth, Lilienfeld, & Yarkoni, 2015), experiential avoidance (Sahdra et al., 2016), and body image-acceptance (Basarkod et al., 2018). These algorithms function via principles of evolution and natural selection (Yarkoni, 2010). Items are analogous to the genes, and a series of items that form a scale are analogous to the chromosomes. The algorithm first generates a sample of chromosomes: it creates a random set of subscales from the whole measure. These chromosomes then compete for fitness. For shortening a questionnaire, the fitter chromosomes are those that explain more variability in the full measure. As in natural selection, the least fit chromosomes are removed from the gene pool, and the fittest items—usually the top 5%— reproduce to create a new pool of chromosomes (Yarkoni, 2010).

During reproduction, chromosomes mutate and genes crossover. In mutation, items are randomly substituted into the chromosome. In crossover, genes from one fit measure are swapped with genes from another fit measure. Following mutation and crossover, the new pool of chromosomes is tested for fitness, and the process repeats until the algorithm reaches a stable solution. That is, short measures are created, mutated, and recombined until the algorithm finds the short measure that explains the most variability in the full measure (Yarkoni, 2010). The shortened measures from this efficient process often produce scores that are as valid as those derived from traditional psychometric approaches (Sandy et al., 2014).

In software packages that use genetic algorithms to shorten questionnaires, the length of these shortened measures are usually determined in two key ways (Scrucca & Sahdra, 2016). Researchers can either specify the desired length of the short measures, or they can determine an item-cost parameter. In the first method, the algorithm will constrain the length of each chromosome to that provided by the researcher. This process will aim for the measure of that length that best explains the variability in the full measure (e.g., the best 3-item measure). Alternatively, researchers can specify the item-cost parameter. This parameter determines the weight a researcher places on having a shorter measure versus one that explains more variance. These two factors are in competition with each other, because longer measures will naturally explain more variance in the full measure, but shorter measures are desirable for the reasons described earlier. By setting the item-cost parameter, researchers can specify how much variance they are willing to lose by removing an item from the questionnaire. This parameter can be pre-specified, or it can be ‘tuned’ during pilot testing. Tuning occurs when researchers pilot-test machine learning parameters until the algorithm starts generating a desired solution. Tuning parameters in this way allows researchers to aim for a shortened measure, but set a limit on the amount of variance they are willing to sacrifice in the process. Choosing either a target length or item-cost parameter still requires subject matter expertise so arbitrary decisions about scale length do not compromise the psychometric integrity of the measure. Shortened measures should still aim to meet standard of reliability and validity (Smith et al., 2000), but using a genetic algorithm to shorten the measure may reduce the complexity of designing measures that meet those standards.

To our knowledge, no machine learning approach to questionnaire abbreviation has been demonstrated in sport and exercise psychology. Given the aforementioned response burden placed upon athletes, this study aimed to test the viability of this approach for generating abbreviated athletic questionnaires that meet reliability and validity standards (Smith et al., 2000). We chose to apply this method to the assessment of mindfulness in sport for a few reasons. There has been growing interest in the potential of mindfulness for performance enhancement (Bühlmayer, Birrer, Röthlin, Faude, & Donath, 2017; Noetel, Ciarrochi, Van Zanden, & Lonsdale, 2017; Sappington & Longshore, 2015). Researchers have often used non-sporting measures of mindfulness and there have been nascent explorations of causal relationships between mindfulness and positive athletic outcomes performance (Noetel et al., 2017). It would be easier to assess mindfulness as a predictor of performance, or as a mediator or outcome of interventions, if researchers had access to a shorter measure of mindfulness in athletes.

Mindfulness in Athletes

"I acknowledge the negative thoughts and let them slide by. This lets me focus on what is really important." (Djokovic, 2013, p. 88)

Novak Djokovic's approach to thoughts is characteristic of many mindfulness and acceptance approaches: notice thoughts and feelings, accept those feelings non-judgmentally, and refocus on the task at hand. The definition of mindfulness is a controversial topic (Chiesa, 2012), but for this paper we will use the most common, modern interpretation of the term: non-judgemental awareness of the present moment. This definition covers both the dispositional trait and the state induced through practice (Birrer, Rothlin, & Morgan, 2012; Chiesa, 2012). These practices have become increasingly popular among coaches, athletes, and sport psychology researchers. In the last ten years, authors have published over fifty papers on mindfulness in the sporting domain (Noetel et al., 2017). Both experimental and observational research has found mindfulness may help athletes perform better, suffer less anxiety, and experience flow more frequently (Bühlmayer et al., 2017; Noetel et al., 2017; Sappington & Longshore, 2015).

In previous studies, mindfulness has been shown to be associated with better performance (Blecharz et al., 2014; Röthlin, Horvath, Birrer, & Holtforth, 2016), lower anxiety (Röthlin et al., 2016), higher flow (Cathcart, McGregor, & Groundwater, 2014; Kaufman, Glass, & Arnkoff, 2009; Kee & Wang, 2008; Pineau, Glass, Kaufman, & Bernal, 2014), higher self-efficacy (Blecharz et al., 2014), fewer intrusive thoughts (Thienot et al., 2014), and less attachment to those thoughts (Zhang, Chung, Si, & Gucciardi, 2016).

Randomised trials of mindfulness interventions have found increases in performance (Gross et al., 2018; Ojaghi, Gholizade, & Mirheidari, 2013; Zhang, Si, et al., 2016), flow (Aherne, Moran, & Lonsdale, 2011; Zhang, Si, et al., 2016), and reductions in burnout (Moen & Wells, 2016). But, other randomised trials have failed to find significant effects of mindfulness on strength (Stocker, Englert, & Seiler, 2018) and shooting performance (John, Kumar, & Lal, 2012; Solberg, Berglund, Engen, Ekeberg, & Loeb, 1996).

Other concerns that may temper enthusiasm are the methodological limitations of the research thus far. Noetel and colleagues (2017) found no studies that met the Cochrane Collaboration's criteria for low risk of bias (Higgins, Altman, Sterne, & on behalf of the Cochrane Statistical Methods Group and the Cochrane Bias Methods Group, 2011). Similar methodological problems have been identified by two other systematic reviews (Bühlmayer et al., 2017; Sappington & Longshore, 2015). Although high-quality randomised trials control for most confounds (Green, 2008), few have controlled for these confounds in mindfulness interventions (Bühlmayer et al., 2017; Noetel et al., 2017; Sappington & Longshore, 2015). Similarly, it is important to examine potential mediators from mindfulness to the increase of adaptive and decrease of maladaptive outcomes. Doing so would bolster causal inferences (Pearl, 2009). Josefsson and colleagues (2017) found mindfulness was associated with decreased rumination and increased self-regulation, leading to better coping. Gustafsson and colleagues (2015) tested a different model, measuring trait mindfulness, exhaustion, perceived stress, and affect. They found mindfulness was associated with lower exhaustion, which was partially mediated by reduced negative affect and increased positive affect.

One problem with the latter model, as an example, is that it contradicts a common model proposed by mindfulness and acceptance theories. Mindfulness is not designed to target positive outcomes by changing the frequency and intensity of negative experiences (e.g., affect). Instead, it aims “to promote a modified *relationship* with internal experiences”

(Gardner & Moore, 2012, p. 309). This approach usually manifests in at least one of three ways (Thienot et al., 2014): increased ability to become aware of thoughts and feelings before they influence behaviour (acceptance); increased tendency to accept oneself despite negative thoughts and feelings (non-judgment); and increased ability to refocus attention where it is important (refocusing). Assessing all three of these constructs allow researchers to assess the unique benefits of each component of mindfulness for athletes. With current measures, it is possible to assess mediators alongside outcomes and possible confounders, but doing so would be less burdensome with shorter methods of assessment.

Often, researchers in sport use measures of trait mindfulness from non-sport domains (Five Facet Mindfulness Questionnaire; Baer, Smith, Hopkins, Krietemeyer, & Toney, 2006; e.g., Mindful Attention Awareness Scale; Brown & Ryan, 2003). Compared with general measures of psychological principles, sport-specific questionnaires tend to better assess psychological constructs in athletes (Moritz, Feltz, Fahrbach, & Mack, 2000). Two measures of athletic mindfulness are the Mindfulness Inventory for Sport (15 items; Thienot et al., 2014) and the Athlete Mindfulness Questionnaire (16 items; Zhang, Chung, & Si, 2017). Both questionnaires assess three constructs, and while each measure is not overly burdensome on its own, reducing the length would decrease response burden and increase the number of constructs that could be measured in parallel, thereby enhancing the quality of research that could be conducted.

As suggested by Smith and colleagues (2000), we started with one of the few mindfulness measures specifically validated for athletic populations: the Mindfulness Inventory for Sport (Thienot et al., 2014). We expected reliability to reduce with length (Tavakol & Dennick, 2011), but had no hypotheses about the effects of measure length on validity. We aimed to outline the trade-off between brevity, validity, and reliability for measures of various lengths, so researchers may choose the best set of items for their needs.

In doing so, the methodological aim of this research was to determine whether genetic algorithms might be a useful tool for reducing response burden in sport and exercise psychology.

Method

We conducted this study in three steps. First, we collected data from two samples of athletic populations, one studying exercise science and one playing golf regularly. Students completed the Mindfulness Inventory for Sport at two time points and the Child and Adolescent Mindfulness Measure at one time point. Golfers completed the Mindfulness Inventory for Sport, the state State Mindful Attention Awareness Scale, and a putting performance task at one time point. Second, we used data from 75% of the exercise science students to shorten the Mindfulness Inventory for Sport using a genetic algorithm. Using separate samples for training and testing machine learning models is considered best practice (Hastie, Tibshirani, & Friedman, 2009). This is because training the model on the testing sample means the error estimates are biased downwards, and errors would end up being higher when used on any new sample. As a result, our final step was to test these shortened measures on the golfers and the other 25% of the exercise science students to establish reliability and validity. Ethics approval was provided by the Australian Catholic University Human Research Ethics Committee.

Participants

Sample 1: Undergraduate Exercise Science students. One cohort of first-year students (N = 859, M_{age} = 19.57 years; SD_{age} = 3.84 years) studying "Psychology of Sport" in Australia completed the questionnaire battery as a learning exercise, using the cohort's data to write a research report. They provided informed consent online and were allowed to opt out of the study without consequence. Students reported a range of abilities in both athletic (15% elite, 36% sub-elite, 38% club, 11% social, <1% inactive) and academic domains (university entry score out of 100: M = 65.25; SD = 10.9). Three months later, students were invited to complete the questionnaire battery again. Both administrations were completed in English, during scheduled tutorials.

Sample 2: Golf sample. Over the space of 18 months, we recruited 118 Australian golfers (M_{age} = 51.19 years, SD_{age} = 16.48 years; 89 male, 29 female) of whom 112 had a golf handicap ($M_{handicap}$ = 14.48; $SD_{handicap}$ = 10.09). Six golfers did not have a handicap but had more than three years golfing experience**.** Golfers voluntarily signed up for the study in exchange for free biomechanical analysis of their technique using a Science and Motion PuttLab (Science & Motion, 2016), which a research assistant provided upon completion. Measures were administered in English and in a quiet area at each participant's golf club.

Measures

Demographics. Students were asked their age, level of competitive performance, the degree in which they were enrolled, and their university entry score. The golfer demographic questionnaire included age, gender, golf handicap, and years of playing experience.

Mindfulness Inventory for Sport. Our primary measure of interest was the Mindfulness Inventory for Sport (Thienot et al., 2014). This 15-item questionnaire is intended to measure three components of athletic mindfulness: present moment awareness; non-judgementality; and the ability to refocus (see items in Table 1). Participants respond on a six-point likert scale from *1 = "Not at all like me"* to *6 = "Very much like me."* It was initially validated with undergraduate students and elite athletes (Thienot et al., 2014).

Child and Adolescent Mindfulness Measure. Eighty-five percent of undergraduates were adolescents (21 years or younger), so to test concurrent validity we asked them to complete the Child and Adolescent Mindfulness Measure (CAMM; Greco, Baer, & Smith, 2011). This 10-item, unidimensional measure asks participants about their daily mindfulness (e.g., "I keep myself busy so I don't notice my thoughts or feelings" [reverse scored]).

Participants respond on a five-point likert scale from 0 = "*Never true*" to 4 = "*Always true*". The Child and Adolescent Mindfulness Measure demonstrated good internal consistency in previous studies (α = 0.81; Greco et al., 2011) and in our undergraduate sample (α = .83).

State Mindful Attention Awareness Scale. As an indicator of concurrent validity in the golf sample, we used the 5-item state Mindful Attention Awareness Scale (Brown & Ryan, 2003). Golfers reported their current levels of mindful awareness (e.g., "I found myself doing things without paying attention" [reverse-scored]) on a seven-point likert scale (from *0 = "Not at all"* to *6 = "Very much"*). This measure has shown significant positive associations with trait mindfulness, pleasant affect, perceived autonomy, and negative correlations with unpleasant affect (Brown & Ryan, 2003), and had good internal consistency in our golf sample (α = .86).

Putting performance. We measured acute putting performance by asking golfers to hit balls as close as possible to specified targets at various distances (Beilock, Jellison, Rydell, McConnell, & Carr, 2006). After a familiarisation round of 10 putts and completion of the questionnaires, golfers completed 10 putts aiming at five targets in a quasi-random order. Targets were at one-foot intervals from 8 to 12 feet away. A camera was suspended above the targets and accuracy was calculated by importing the footage into Kinovea software (Kinovea, 2016). Performance was operationalised as mean radial error across the 10 putts.

Analysis

Missing data. For cases where there was more than 20% missing data (2.57% of cases), we excluded the participant under the assumption that they abandoned the protocol. For cases with less than 20% missing data, we created 25 multiple imputations to fill the missing data using the Expectation-Maximisation procedure with bootstrapping in Amelia II software (Honaker, King, & Blackwell, 2012). This method is likely to produce unbiased results when data are missing at random (Jakobsen, Gluud, Wetterslev, & Winkel, 2017). All predictors and outcomes of interest passed a test of missingness (Jamshidian, Jalal, & Jansen, 2014), such that missing data appeared to be random.

Genetic algorithm procedure. We used the GAabbreviate package (Version 1.3; Scrucca & Sahdra, 2016) in R software (R Core Team, 2018) to implement the genetic algorithm procedure. As described earlier, the algorithm first generated a population of shorter scales by randomly sampling items from the full scale. We chose a population size of 200 short scales based on recommendations from previous authors (Basarkod et al., 2018; Sahdra et al., 2016; Yarkoni, 2010).

Each short scale was then tested for fitness by assessing the percentage of variance it explained in the full scale. The most fit measures were then subjected to crossover and mutation so a new pool of 200 short scales was developed. Following previous authors, we set the probability of crossover and mutation to the defaults of the GA package (Basarkod et al., 2018; Sahdra et al., 2016; Scrucca, 2013).

To choose the best short-measure of a certain length (e.g., nine item measure with three items for each of three subscales), we ran the algorithm 25 times for each of the 25 imputed data sets. Being a stochastic approach, not all runs resulted in identical subscales, so we tallied a vote each time an item was selected at the end of a run. Items were included when they were selected the most frequently at the end of the 625 runs. To get short measures of various lengths (i.e., 5-items per subscale, 4-items, 3-items, 2-items, 1-item), we completed the 625 runs four times, setting the maximum length of the subscale each time. Instead of using cross-validation, we separated the undergraduate population into a training sample (75%) and a testing sample (25%) to mitigate the chance of overfitting the models to our data (James, Witten, Hastie, & Tibshirani, 2013). Item-selection was performed on the training sample, but reliability and validity measures were examined in the testing sample and the sample of golfers.

Reliability and validity. Smith and colleagues (2000) suggested a series of steps to avoid the 'sins of short form development.' The key steps involve: establishing the reliability of each subscale of the short form; looking for overlapping variance with the long form; and confirming the factor structure in the short form. We also aimed to extend some of the validity checks conducted by Thienot and colleagues (2014) by assessing test-retest reliability over three months and predictive validity using objective putting performance.

Reliability. Despite the limitations of Cronbach's α (see Sijtsma, 2009), we calculated it for all subscales because it is the legacy measure of internal consistency. For a more definitive assessment of reliability, we relied on McDonald's omega because it requires fewer assumptions of the data and is more generalisable (Zinbarg, Revelle, Yovel, & Li, 2005). We used test-retest reliability to test temporal stability.

Content validity. For each subscale, we calculated a correlation between the participants' scores on the full scale and their scores on the shorter versions.

Factorial validity. For each short measure, we conducted a confirmatory factor analysis using the *lavaan* package (Rosseel, 2012) in R. We used robust maximum likelihood estimation because it is less sensitive to non-normality and non-independence of observations (Basarkod et al., 2018). To evaluate the fit of the models, we used the χ^2 goodness-of-fit statistics, Tucker-Lewis index (TLI), Comparative Fit index (CFI), root mean square error of approximation (RMSEA), standardised root mean square residual (SRMR) and Aitken's Information Criteria (AIC). We used the criteria from Marsh and colleagues (2005) for close (RMSEA $\leq$ 0.04; SRMR $\leq$ 0.08; TLI $\geq$ 0.99; CFI $\geq$ 0.99) and acceptable fit (RMSEA $\leq$ 0.08; SRMR $\leq$ 0.05; TLI $\geq$ 0.90; CFI $\geq$ 0.90).

Criterion-related validity. In the adolescent undergraduate sample, we conducted correlations between the short measures and the Child and Adolescent Mindfulness measure.

In the golf sample, we examined correlations between the short measures and the state Mindful Attention Awareness Scale, and examined correlations between the golfers' responses on the measures of mindfulness, their handicaps, and their putting performance.

Results

Consistent with open science principles (Tamminen & Poucher, 2018), we uploaded the raw data, analysis code, and results when eliminating social athletes to the Open Science Framework: https://osf.io/hv5dy/?view_only=31376b1ba4bb48a093c491598efe4002

Missing Data

Of students (Sample 1) who completed the initial battery of questionnaires (i.e., Time 1), 34% (n = 296) voluntarily completed the battery three months later (i.e., Time 2). Fewer may have completed Time 2 because tutorial attendance is notoriously lower at the end of the semester than at the beginning. Also, where students used Time 1 data for their own research reports, Time 2 data was used for research purposes only, so they had a less compelling rationale for participation. We explored differences in Time 1 responses between those who did and did not complete the questionnaire at Time 2. We found no significant differences on any questionnaires, on university entry scores, on their enrolled degree, or their level of athletic competition (ds < .125; ps > 0.05). Those who did complete the follow-up questionnaire were about six-months younger than those who did not ($t_{771} = 2.01$, $p = 0.04$). Ten percent of participants had missing data on more than 20% of the items. These responses were excluded under the presumption that students abandoned the procedure. Of those remaining, less than 1% of data was missing.

Genetic Algorithm-Derived Short Measures

We ran the genetic algorithm procedure using the training subsample of undergraduate students (N = 644, 75% of undergraduate sample). The selection process for an example run through the algorithm is depicted in Figure 1. The results from the 2,500 runs of

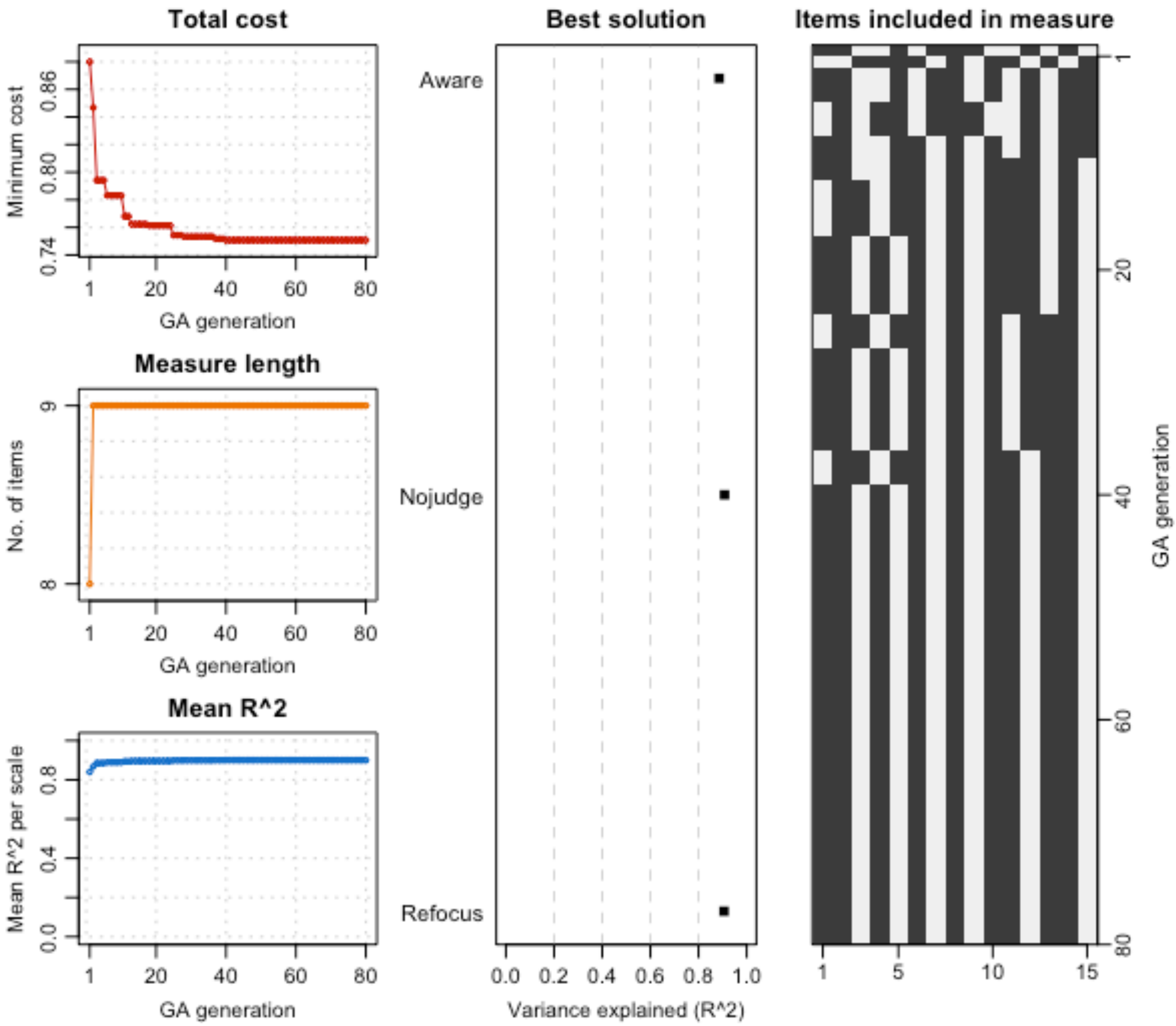

Figure 1. Plots demonstrating output from example genetic algorithm run. The left thee plots display improvements across generations for: cost of shortening measure (top), measure length (middle), and variance explained by short measure (bottom). The central plot demonstrates the variance explained by the best solution for each subscale. The right hand plot shows the selected items (in black), with high variation in the first runs (top of plot) and a stable solution emerging after 40 generations (bottom half of plot).

Table 1

Percentage of Runs in which Item Was Selected by Genetic Algorithm Procedure

Item	Number of items per subscale 1	2	3	4
Awareness subscale				
I am able to notice the intensity of nervousness in my body.	0%	0%	3%	**100%**
I pay attention to the type of emotions I am feeling.	**100%**	**100%**	**100%**	**100%**
I am aware of the thoughts that are passing through my mind.	0%	0%	**97%**	**99%**
I am able to notice the sensations of excitement in my body.	0%	**100%**	3%	1%
I am able to notice the location of physical discomfort when I experience it.	0%	0%	**97%**	**100%**
Non-judgementallity subscale (reverse scored)				
When I become aware that I am thinking of the final result, I blame myself for not being focused on relevant cues for my performance.	0%	0.0%	2%	**100%**
When I become aware that I am really upset because I am losing, I criticise myself for reacting this way.	0%	**100%**	**99%**	**98%**
When I become aware that I am not focussing on my own performance, I blame myself for being distracted.	**100%**	0.3%	**98%**	2%
When I become aware that I am thinking about a past performance, I criticise myself for not being focused on my current performance.	0%	**100%**	**99%**	**100%**
When I become aware that I am angry at myself for making a mistake, I criticise myself for having this reaction.	0%	0.3%	1%	**100%**
Refocusing subscale				
When I become aware that I am tense, I am able to quickly bring my attention back to what I should focus on.	0%	0%	**98%**	**99%**
When I become aware that I am thinking about how tired I am, I quickly bring my attention back to what I should focus on.	**100%**	**100%**	2%	**100%**
When I become aware that I am not focussing on my own performance, I am able to quickly refocus my attention on things that help me to perform well.	0%	0.2%	**100%**	1%
When I become aware that some of my muscles are sore, I quickly refocus on what I have to do.	0%	**100%**	**100%**	**100%**
When I become aware that I am really excited because I am winning, I stay focused on what I have to do.	0%	0%	0%	**100%**

Note. selected item indicated in **bold** the genetic algorithm (625 runs to find short measures of each length) are presented in Table 1. The resulting items were very consistent across the 625 runs. The least consistent case was selecting 3 items for the awareness subscale. Even in this case, the top three items were selected at least 97% of the time, leaving other items selected only 3% of the time. This indicates the model achieved a very stable solution. The items selected for each short measure appear to also demonstrate face validity and coverage of the construct. For example, “I pay attention to the type of emotions I am feeling” was consistently selected for the Awareness subscale and appears to cover the essential components of the construct. To assess how well these short measures perform on quantitative measures of reliability and validity, we tested each measure in the undergraduate testing sample and the golfer, with results presented in Figure 2 and raw data in supplementary materials.

Reliability and Validity

Reliability. In the undergraduate sample, omegas for the 5-, 4-, 3-item subscales showed acceptable reliability ($\omega \geq 0.7$) for all subscales. For the 2-item subscales, we found acceptable reliability for refocusing, but poor reliability for the other two subscales ($\omega_{awareness} = 0.55$, $\omega_{non\text{-}judgment} = 0.57$). Cronbach’s alpha largely mirrored this pattern. In the golf sample, omega for the non-judgment subscale was borderline using 2-items ($\omega_{non\text{-}judgment} = .68$), but acceptable for all other subscales and all longer measures.

Test-retest reliability. In the undergraduate sample, correlations between mindfulness subscales at the start of the semester and those three months later were all strong ($.43 \leq r \leq .61$). While correlations were somewhat stronger when using 5-items per subscale ($.57 \leq r \leq .61$), correlations were acceptable even when using 1-item per subscale ($.43 \leq r \leq .47$).

Content validity. In the undergraduate sample, correlations between the full and short measures were stronger than .75 for all subscales. Correlations were almost perfect when

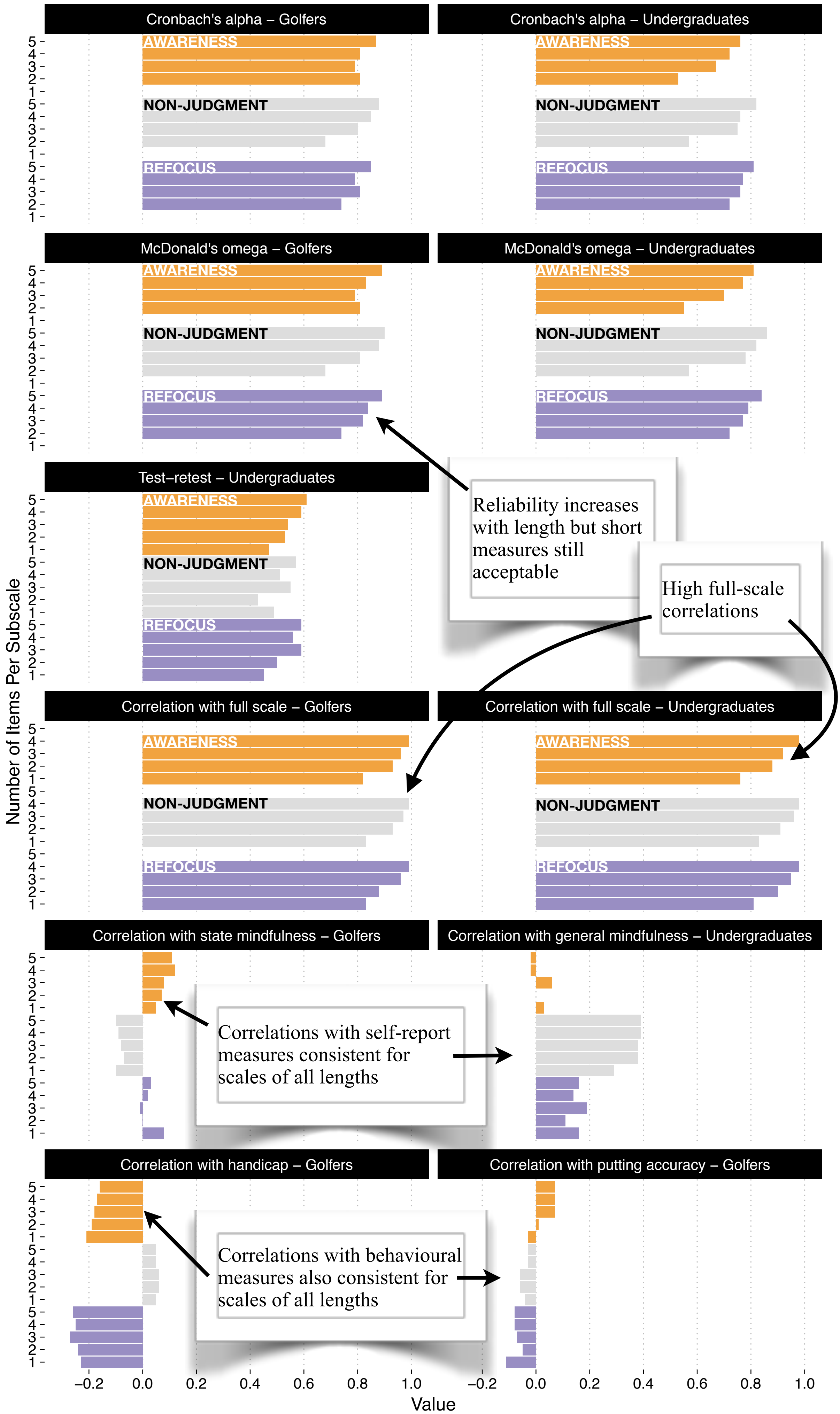

Figure 2. Reliability and validity metrics for full measure (5-items per subscale) and shortened measures

using 4 items per subscale (r = .98), and were modestly weaker when using 3 items (.93 ≤ r ≤ .96), 2 items (.88 ≤ r ≤ .91), or one item per subscale (.76 ≤ r ≤ .83). In the golf sample, correlations between the full and short measures were stronger than .80 for all subscales. Correlations were almost perfect when using 4 or 3 items per subscale (.96 ≤ r ≤ .99), and were modestly weaker when using 2 items (.89 ≤ r ≤ .94) or one item per subscale (.82 ≤ r ≤ .85).

Factorial validity. The factor structure was generally maintained for all shorter measures (see Table 2). In the undergraduate sample, there was a close fit when using 2-items per subscale. The fit was less good, but still acceptable, when using 3-, 4-, or 5-items. In the golf sample, most indices described a close fit with the three-factor structure.

Criterion-related validity: self-report. In the undergraduate sample, mindfulness in sport was associated with trait mindfulness regardless of the length of the sport measure. Trait mindfulness was associated with a tendency to avoid self-judgment in sport (.29 ≤ r ≤ .39) and an ability to refocus in sporting settings (.11 ≤ r ≤ .19). In the golf sample, correlations between trait mindfulness (via the Mindfulness Inventory for Sport) and state mindfulness (via the Mindful Attention Awareness Scale) were similar regardless of the length of the measure. State mindfulness showed weak positive correlations with a tendency to be aware (.16 ≤ r ≤ .20) and weak negative correlations with a tendency to avoid self-judgement (-.14 ≤ r ≤ -.21).

Criterion-related validity: behaviour. Performance, as measured by error on ten putts, was not associated with mindfulness subscales of any length (-.11 ≤ r ≤ .07; all p > 0.38). This was possibly due to the short task (so no need to ‘refocus’) with no pressure (so no need for ‘awareness’) and no ability to compare against a standard, being a novel task with no competition (so small probability of ‘self-judgment’). However, golfers with better (i.e., lower) handicaps tended to be better able to refocus their attention (-.20 ≤ r ≤ -.23; all p <

Table 2

Confirmatory Factor Analysis Fit Indices for Short Measures in Undergraduate and Golf Samples

	Maximum Likelihood							Robust Maximum Likelihood	
Items per subscale	Chi Squared	*p*	TLI	CFI	RMSEA	SRMR	AIC	Chi Squared	*p*
Undergraduate sample									
5	172.23	0.00	0.90	0.91	0.07	0.07	9020.69	153.12	0.00
4	95.19	0.00	0.91	0.93	0.06	0.07	7416.84	86.63	0.00
3	39.04	0.03	0.95	0.97	0.05	0.06	5620.96	37.38	0.04
2	8.91	0.18	0.96	0.98	0.05	0.04	3880.07	9.17	0.16
Golf sample									
5	93.35	0.30	0.99	0.99	0.02	0.06	5295.34	81.67	0.64
4	53.18	0.39	0.99	1.00	0.02	0.06	4400.26	46.61	0.65
3	27.53	0.28	0.99	0.99	0.04	0.06	3324.06	25.50	0.38
2	2.38	0.88	1.05	1.00	0.00	0.03	2252.60	2.14	0.91

0.02). Lower handicaps were associated with higher awareness ($-.12 \leq r \leq -.16$), but correlations were only significant for one and two item versions (p = .03 and .049 respectively; others ps between .052 and .083). Handicap did not appear to strongly relate to a tendency for self-judgment ($.04 \leq r \leq .05$; all $p > 0.5$). As seen by the narrow range between the highest and lowest correlation, there were no meaningful differences in these associations for subscales of different lengths. The shorter versions of the Mindfulness Inventory for Sport appeared equivalent to longer versions for both types of criterion validity.

Discussion

The purpose of this study was to use the genetic algorithm procedure to abbreviate the Mindfulness Inventory for Sport and assess the utility of the resulting scales. Table 3 summarises the trade-offs that occur when using the measures shortened by the genetic algorithm. Reliability was proportional to the length of the measure, as would be expected (Tavakol & Dennick, 2011), though drop-off in reliability was particular steep between the 9- and 6-item measure. The 9- and 6-item versions may balance the needs of many researchers. The 9-item version of the measure demonstrated acceptable reliability, acceptable factor fit, equivalent criterion validity and 40% lower response burden than the 15-item measure. The 6-item version demonstrated poorer reliability, but better factor fit, equivalent criterion validity, and 60% lower response burden compared with the 15-item measure. Short measures demonstrated very high correlations with the long version of the questionnaire (rs between .88 and .97). All things considered, we believe the 9-item version fulfils the needs for most researchers who want to balance brevity with common standards of reliability and validity.

Reliability and Validity of Shortened Measures

The psychometric properties of our shortened measures replicate those of the initial development paper. The reliability metrics for the 9-item measure were as good as those

Table 3

Questionnaire Options with Benefits and Drawbacks

	Benefits	Drawbacks
15-item version	• Highest reliability	• Highest questionnaire burden • Poor factor fit in some samples
12-item version	• Near perfect correlation with full scale (*r*s = .98-.99) • Good reliability • Acceptable factor fit • 20% lower response burden than full measure • Equivalent criterion validity with full measure	• Poorer factor fit than 6-item scale
9-item version	• Near perfect correlation with full scale (*r*s = .93-.97) • Acceptable reliability • Acceptable factor fit • 40% lower questionnaire burden than full measure • Equivalent criterion validity with full measure	• Poorer factor fit than 6-item scale
6-item version	• High correlation with full scale (*r*s =.88-.93) • Strongest factor fit across samples • 60% lower response burden than full measure • Equivalent criterion validity with full measure	• Lowest reliability
3-item version	• Good correlation with full scale (*r*s = 75-.83) • 80% lower response burden than full measure • Equivalent criterion validity with full measure	• Cannot calculate reliability

found in the original validation paper (Thienot et al., 2014). Thienot and colleagues found acceptable internal consistency in both their samples of undergraduates ($.77 \leq \alpha \leq .78$) and elite athletes ($.70 \leq \alpha \leq .79$). While the 6-item measure showed poorer internal reliability, the fit indices were as good, if not better, than those presented in the initial validation study ($.89 \leq \text{CFI} \leq .92$ and $.88 \leq \text{TLI} \leq .90$ from Thienot et al., 2014). In our undergraduate sample the full measure showed similar, borderline fit indices (TLI = .90; CFI = .91; RMSEA = .07, SRMR = .07). By contrast, the fit was much closer for the 9-item (TLI = .95; CFI = .97; RMSEA = .05, SRMR = .06, AIC = 5620.96) and 6-item versions (TLI = .96; CFI = .98; RMSEA = .05, SRMR = .04, AIC = 3880.07). Both 6- and 9-item versions demonstrated close fit in the golf sample. These data indicate that the factorial validity for the short measures is at least as strong as that of the full scale. Our data suggest that, in some samples, shorter measures may be preferable, suggesting that longer measures may assess latent variables outside the core hypothesised constructs.

Neither the criterion validity nor the test-retest reliability appeared to be influenced by the length of the questionnaire. Correlations with other measures of mindfulness were weak. The same was true for behavioural measures of performance. This was a consistent pattern for all short measures, including that with one item per subscale. These weak correlations are similar to those demonstrated in previous studies, which found weak correlations between non-judgment, refocusing and other trait mindfulness measures (Thienot et al., 2014). While these correlations were weak, we could not find evidence that the validity of the Mindfulness Inventory for Sport was compromised by shortening the measure.

As outlined in Figure 2, we found no meaningful associations between any mindfulness subscale and putting performance or state mindfulness. We found only weak associations between golf handicap and the mindfulness constructs of awareness and refocusing. Non-judgement demonstrated a negligible association with handicap. These weak correlations raise doubts about the predictive utility of self-reported mindfulness. These data are consistent with previous studies that found weak positive correlations of mindfulness with performance (r = .17, Blecharz et al., 2014; r = .33, Röthlin et al., 2016; r = .19, Sarnell, 2012). All three of these studies used general mindfulness measures rather than those specifically assessing mindfulness in sporting contexts. They also used different measures of performance (soccer skills, Blecharz et al., 2014; self-report, Röthlin et al., 2016; coach report, Sarnell, 2012). Nevertheless, the consistency of these associations suggest a small percentage of the variance in performance may be explained by an athlete's dispositional mindfulness. Our findings suggest that these associations may be better explained by an athlete's tendency to be aware of their thoughts and refocus, rather than their tendency to avoid self-judgement. Teasing apart these components allows researchers to understand why mindfulness may be helping athletes, allowing for more targeted interventions.

Comparisons with Other Genetic Algorithm Implementations

Our findings are consistent with previous implementations of genetic algorithms for questionnaire abbreviation (Basarkod et al., 2018; Eisenbarth et al., 2015; Sahdra et al., 2016; Sandy et al., 2014). Those implementations have sometimes shortened long, multi-factor questionnaires while maintaining explained variance (e.g., *r*s between .90 and .97; Sahdra et al., 2016). Some have also abbreviated unidimensional questionnaires that were already quite short (e.g., 12-items to 5-items with $r = .98$; Basarkod et al., 2017). In all of these examples, genetic algorithms have produced shorter versions of questionnaires while maintaining high full-scale correlations. Reliability metrics are usually compromised to some degree, but other implementations have also found long and short versions perform comparably on criterion validity (Basarkod et al., 2018; Eisenbarth et al., 2015; Sahdra et al., 2016; Sandy et al., 2014). Our findings add to the evidence for using machine learning as an efficient tool for abbreviating questionnaires. Genetic algorithms appear to reliably produce short questionnaires that meet standards of reliability and validity. These algorithms still require some researcher input to tune parameters (e.g., the item-cost parameter). However, compared with traditional approaches to measure abbreviation (Marsh et al., 2005), these approaches require less subjectivity and judgment.

Other implementations of genetic algorithms have selected an item-cost parameter to find one shorter version of the questionnaire that fits the goals of the researchers (Basarkod et al., 2018; Eisenbarth et al., 2015; Sahdra et al., 2016). In contrast, our implementation was the first to offer end-users flexibility to choose their own level of abbreviation. Rather than determining scale length by tuning the item-cost parameter ourselves, we have provided future researchers with a transparent description of the costs and benefits of using questionnaires of various lengths. These shorter measures appear to be reliable and valid approaches to measuring mindfulness components in athletes. For example, the 9-item measure was 40% shorter than the full measures and demonstrated strong reliability and validity. It may allow for more flexible and accurate data collection while reducing cost and questionnaire burden (Basarkod et al., 2017).

Limitations and Future Directions

Our study only assessed reliability and validity in a narrow section of the athletic population: Australian undergraduate students and golfers. As a result, reliability and validity is yet to be assessed in different athletic populations and different cultures. They are also yet to be validated with children, and the items included in our measures may not be developmentally appropriate. For children, new items with simpler language might be required for reliable, valid measurement. For all other populations, researchers may benefit from checking whether the psychometric properties hold before making strong claims based on the use of these shorter measures.

The Mindfulness Inventory for Sport assesses distinct components of mindfulness, and these components vary in their association with unidimensional measures of mindfulness. Many previous studies have used scales that treat mindfulness as a unidimensional trait, although separating the construct into these components may have theoretical and practical benefits for future research. For example, if refocusing is the best predictor of performance and refocusing is amenable to change, then interventions may be more effective if they target refocusing, rather than awareness or non-judgementality. Mindfulness and acceptance interventions vary in the degree to which they address these components, and the focus of the intervention is sometimes poorly reported (Noetel et al., 2017). It is not yet clear which of these processes is the most important and which is the most amenable to change, but we hope that these shorter measures make it easier to explore those questions. Shorter measures likely make it more feasible to assess mindfulness before, during, and after an intervention. Doing so would allow the research community to identify whether mindfulness does mediate the relationship between interventions and performance outcomes, and if it does, which processes of change are most important.

More broadly, this paper describes an option for the research community to develop robust versions of short measures. Genetic algorithms are not, however, completely autonomous. As mentioned earlier, researchers are still involved in determining the ultimate length of the questionnaire, or in setting the item-cost parameter. These algorithms are not a shortcut for abbreviating questionnaires for people with no psychometric or subject-matter expertise. Instead, they are a tool for guiding researcher decision-making in a way that is both cognitively and computationally efficient (Yarkoni, 2010). If researchers do not want to rely exclusively on the algorithm, they can use it in conjunction with other methods of questionnaire abbreviation (e.g., Marsh et al., 2005).

Conclusions

More than half of high-performance athletes fill out questionnaires every day (Taylor et al., 2012) and questionnaire length is one of the biggest barriers to completion (Saw et al., 2015). Genetic algorithms can identify robust shorter version of questionnaires while maintaining sound psychometric qualities. We demonstrated that these methods are useful in sport and exercise psychology by creating multiple short versions of the Mindfulness Inventory for Sport. While our data found weak associations between performance and some components of mindfulness, we hope the measures encourage more research into the construct. We also offer researchers a new tool for shortening questionnaires, facilitating flexible, efficient, affordable, and accurate data collection.

Conflicts of Interest

None

References

Aherne, C., Moran, A. P., & Lonsdale, C. (2011). The effect of mindfulness training on athletes' flow: An initial investigation. *Sport Psychologist*, *25*(2), 177–189. doi: 10.1123/tsp.25.2.177

Baer, R. A., Smith, G. T., Hopkins, J., Krietemeyer, J., & Toney, L. (2006). Using self-report assessment methods to explore facets of mindfulness. *Assessment*, *13*(1), 27–45. doi: 10.1177/1073191105283504

Basarkod, G., Sahdra, B., & Ciarrochi, J. (2018). Body Image-Acceptance and Action Questionnaire-5: An abbreviation using genetic algorithms. *Behavior Therapy*, *49*(3), 388–402. doi: 10.1016/j.beth.2017.09.006

Beilock, S. L., Jellison, W. A., Rydell, R. J., McConnell, A. R., & Carr, T. H. (2006). On the causal mechanisms of stereotype threat: can skills that don't rely heavily on working memory still be threatened? *Personality & Social Psychology Bulletin*, *32*(8), 1059–1071. doi: 10.1177/0146167206288489

Birrer, D., Rothlin, P., & Morgan, G. (2012). Mindfulness to enhance athletic performance: Theoretical considerations and possible impact mechanisms. *Mindfulness*, *3*(3), 235–246. doi: 10.1007/s12671-012-0109-2

Blanton, H., & Jaccard, J. (2006). Arbitrary metrics in psychology. *The American Psychologist*, *61*(1), 27–41. doi: 10.1037/0003-066X.61.1.27

Blecharz, J., Luszczynska, A., Scholz, U., Schwarzer, R., Siekanska, M., & Cieslak, R. (2014). Predicting performance and performance satisfaction: Mindfulness and beliefs about the ability to deal with social barriers in sport. *Anxiety, Stress, and Coping*, *27*(3), 270–287. doi: 10.1080/10615806.2013.839989

Brown, K. W., & Ryan, R. M. (2003). The benefits of being present: Mindfulness and its role in psychological well-being. *Journal of Personality and Social Psychology*, *84*(4), 822–

848. doi: 10.1037/0022-3514.84.4.822

Bühlmayer, L., Birrer, D., Röthlin, P., Faude, O., & Donath, L. (2017). Effects of mindfulness practice on performance-relevant parameters and performance outcomes in sports: A meta-analytical review. *Sports Medicine*, *47*(11), 2309–2321. doi: 10.1007/s40279-017-0752-9

Cathcart, S., McGregor, M., & Groundwater, E. (2014). Mindfulness and flow in elite athletes. *Journal of Clinical Sport Psychology*, *8*(2), 119–141. doi: 10.1123/jcsp.2014-0018

Chiesa, A. (2012). The Difficulty of Defining Mindfulness: Current Thought and Critical Issues. *Mindfulness*, *4*(3), 255–268. doi: 10.1007/s12671-012-0123-4

Davidov, E., Schmidt, P., & Schwartz, S. H. (2008). Bringing values back in: The adequacy of the European Social Survey to measure values in 20 countries. *Public Opinion Quarterly*, *72*(3), 420–445. doi: 10.1093/poq/nfn035

Djokovic, N. (2013). *Serve to Win*. New York: Zinc Ink.

Eisenbarth, H., Lilienfeld, S. O., & Yarkoni, T. (2015). Using a genetic algorithm to abbreviate the Psychopathic Personality Inventory-Revised (PPI-R). *Psychological Assessment*, *27*(1), 194–202. doi: 10.1037/pas0000032

Fan, W., & Yan, Z. (2010). Factors affecting response rates of the web survey: A systematic review. *Computers in Human Behavior*, *26*(2), 132–139. doi: 10.1016/j.chb.2009.10.015

Fletcher, R. B., & Hattie, J. A. (2004). An examination of the psychometric properties of the physical self-description questionnaire using a polytomous item response model. *Psychology of Sport and Exercise*, *5*(4), 423–446. doi: 10.1016/S1469-0292(03)00036-0

Glidewell, L., Thomas, R., MacLennan, G., Bonetti, D., Johnston, M., Eccles, M. P., … Grimshaw, J. M. (2012). Do incentives, reminders or reduced burden improve healthcare professional response rates in postal questionnaires? Two randomised controlled trials.

BMC Health Services Research, *12*, 250. doi: 10.1186/1472-6963-12-250

Greco, L. A., Baer, R. A., & Smith, G. T. (2011). Assessing mindfulness in children and adolescents: Development and validation of the Child and Adolescent Mindfulness Measure (CAMM). *Psychological Assessment*, *23*(3), 606–614. doi: 10.1037/a0022819.supp

Green, S. (2008). Preparing a Cochrane review. In J. P. T. Higgins & S. Green (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester, UK: John Wiley & Sons.

Gross, M., Moore, Z. E., Gardner, F. L., Wolanin, A. T., Pess, R., & Marks, D. R. (2018). An empirical examination comparing the Mindfulness-Acceptance-Commitment approach and Psychological Skills Training for the mental health and sport performance of female student athletes. *International Journal of Sport and Exercise Psychology*, *16*(4), 431–451. doi: 10.1080/1612197X.2016.1250802

Gustafsson, H., Davis, P., Skoog, T., Kenttä, G., & Haberl, P. (2015). Mindfulness and its relationship with perceived stress, affect, and burnout in elite junior athletes. *Journal of Clinical Sport Psychology*, *9*(3), 263–281. doi: 10.1123/jcsp.2014-0051

Halson, S. L. (2014). Monitoring training load to understand fatigue in athletes. *Sports Medicine* , *44 Suppl 2*, S139–S147. doi: 10.1007/s40279-014-0253-z

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction*. New York, NY: Springer. Retrieved from http://thuvien.thanglong.edu.vn:8081/dspace/handle/DHTL_123456789/4053

Higgins, J. P. T., Altman, D. G., Sterne, J. A. C., & on behalf of the Cochrane Statistical Methods Group and the Cochrane Bias Methods Group. (2011). Assessing risk of bias in included studies. In J. P. T. Higgins & S. Green (Eds.), *Cochrane Handbook for Systematic Reviews of Interventions* (Vol. 5.1.1). The Cochrane Collaboration.

Holland, J. (1975). *Adaption in natural and artificial systems*. Ann Arbour, MI: The University of Michigan Press.

Honaker, J., King, G., & Blackwell, M. (2012). AMELIA-II: A Program for Missing Data (Version 1.6.2). Retrieved from https://r.iq.harvard.edu/docs/amelia/amelia.pdf

Jakobsen, J. C., Gluud, C., Wetterslev, J., & Winkel, P. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials - a practical guide with flowcharts. *BMC Medical Research Methodology*, *17*(1), 162. doi: 10.1186/s12874-017-0442-1

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning: with Applications in R*. Springer. doi: 10.1007/978-1-4614-7138-7

Jamshidian, M., Jalal, S., & Jansen, C. (2014). MissMech: An R Package for Testing Homoscedasticity, Multivariate Normality, and Missing Completely at Random (MCAR). *Journal of Statistical Software, Articles*, *56*(6), 1–31. doi: 10.18637/jss.v056.i06

John, K. S., Kumar, V. S., & Lal, K. G. (2012). The effect of music therapy and meditation on sports performance in professional shooters. *International Journal Of Research In Ayurveda And Pharmacy*, *3*(1), 133–136. Retrieved from http://www.ijrap.net/admin/php/uploads/761_pdf.pdf

Josefsson, T., Ivarsson, A., Lindwall, M., Gustafsson, H., Stenling, A., Böröy, J., … Falkevik, E. (2017). Mindfulness Mechanisms in Sports: Mediating Effects of Rumination and Emotion Regulation on Sport-Specific Coping. *Mindfulness*, *8*(5), 1354–1363. doi: 10.1007/s12671-017-0711-4

Kaplan, D., & Su, D. (2016). On matrix sampling and imputation of context questionnaires with implications for the generation of plausible values in large-scale assessments. *Journal of Educational and Behavioral Statistics*, *41*(1), 57–80. doi:

10.3102/1076998615622221

Kaufman, K. A., Glass, C. R., & Arnkoff, D. B. (2009). Evaluation of Mindful Sport Performance Enhancement (MSPE): A new approach to promote flow in athletes. *Journal of Clinical Sport Psychology*, *3*(4), 334–356. doi: 10.1123/jcsp.3.4.334

Kee, Y. H., & Wang, C. K. J. (2008). Relationships between mindfulness, flow dispositions and mental skills adoption: A cluster analytic approach. *Psychology Of Sport & Exercise*, *9*(4), 393–411. doi: 10.1016/j.psychsport.2007.07.001

Kinovea. (2016). Kinovea (Version 0.8.15). Retrieved from http://www.kinovea.org/

Marsh, H. W., Ellis, L. A., Parada, R. H., Richards, G., & Heubeck, B. G. (2005). A short version of the Self Description Questionnaire II: operationalizing criteria for short-form evaluation with new applications of confirmatory factor analyses. *Psychological Assessment*, *17*(1), 81–102. doi: 10.1037/1040-3590.17.1.81

Miller, P., Henry, D., & Votruba-Drzal, E. (2016). Strengthening Causal Inference in Developmental Research. *Child Development Perspectives*, *10*(4), 275–280. doi: 10.1111/cdep.12202

Moen, F., & Wells, A. J. (2016). Can the attention training technique reduce burnout in junior elite athletes? *International Journal of Sports Science & Coaching*, *10*(1), 53–64. Retrieved from https://www.dbpia.co.kr/Article/NODE06602978

Moorman, R. H., & Podsakoff, P. M. (1992). A meta-analytic review and empirical test of the potential confounding effects of social desirability response sets in organizational behaviour research. *Journal of Occupational and Organizational Psychology*, *65*(2), 131–149. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2044-8325.1992.tb00490.x

Moritz, S. E., Feltz, D. L., Fahrbach, K. R., & Mack, D. E. (2000). The relation of self-efficacy measures to sport performance: A meta-analytic review. *Research Quarterly for*

Exercise and Sport, *71*(3), 280–294. doi: 10.1080/02701367.2000.10608908

Noetel, M., Ciarrochi, J., Van Zanden, B., & Lonsdale, C. (2017). Mindfulness and acceptance approaches to sporting performance enhancement: a systematic review. *International Review of Sport and Exercise Psychology*, 1–37. doi: 10.1080/1750984X.2017.1387803

Ojaghi, A., Gholizade, H., & Mirheidari, L. (2013). The effect of mindfulness techniques training on anxiety and sport performance among table tennis players. *Life Science Journal*, *10*(S1), 225–230. Retrieved from http://www.lifesciencesite.com/lsj/life1001s/037_15426life1001s_225_230.pdf

Parker, P. D., Jerrim, J., Chmielewski, A. K., & Marsh, H. W. (2017). Predicting university entry using machine-based models and solutions. In I. Schoon & R. Silbereisen (Eds.), *Educational opportunities, motivation and attainment in times of social change*. London: UCL IOE Press.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference* (2nd edition). Cambridge University Press. Retrieved from https://www.amazon.com/Causality-Reasoning-Inference-Judea-Pearl/dp/052189560X

Pineau, T. R., Glass, C. R., Kaufman, K. A., & Bernal, D. R. (2014). Self- and team-efficacy beliefs of rowers and their relation to mindfulness and flow. *Journal of Clinical Sport Psychology*, *8*(2), 142–158. doi: 10.1123/jcsp.2014-0019

Podsakoff, P. M., MacKenzie, S. B., & Podsakoff, N. P. (2012). Sources of method bias in social science research and recommendations on how to control it. *Annual Review of Psychology*, *63*, 539–569. doi: 10.1146/annurev-psych-120710-100452

R Core Team. (2018). R: A language and environment for statistical computing (Version 3.5.1). Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Rolstad, S., Adler, J., & Rydén, A. (2011). Response burden and questionnaire length: is shorter better? A review and meta-analysis. *Value in Health: The Journal of the International Society for Pharmacoeconomics and Outcomes Research*, *14*(8), 1101–1108. doi: 10.1016/j.jval.2011.06.003

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, Articles*, *48*(2), 1–36. doi: 10.18637/jss.v048.i02

Röthlin, P., Horvath, S., Birrer, D., & Holtforth, M. (2016). Mindfulness promotes the ability to deliver performance in highly demanding situations. *Mindfulness*, *7*(3), 727–733. doi: 10.1007/s12671-016-0512-1

Sahdra, B. K., Ciarrochi, J., Parker, P., & Scrucca, L. (2016). Using Genetic Algorithms in a Large Nationally Representative American Sample to Abbreviate the Multidimensional Experiential Avoidance Questionnaire. *Frontiers in Psychology*, *7*, 189. doi: 10.3389/fpsyg.2016.00189

Sandy, C. J., Gosling, S. D., & Koelkebeck, T. (2014). Psychometric Comparison of Automated Versus Rational Methods of Scale Abbreviation. *Journal of Individual Differences*, *35*(4), 221–235. doi: 10.1027/1614-0001/a000144

Sappington, R., & Longshore, K. (2015). Systematically reviewing the efficacy of mindfulness-based interventions for enhanced athletic performance. *Journal of Clinical Sport Psychology*, *9*(3), 232–262. doi: 10.1123/jcsp.2014-0017

Sarnell, L. B. (2012). Motivation, mindfulness, performance and commitment in young female athletes. *Dissertation Abstracts International: Section B: The Sciences And Engineering*, *75*(5-B(E)), No Pagination Specified. Retrieved from http://gradworks.umi.com/35/78/3578597.html

Saw, A. E., Main, L. C., & Gastin, P. B. (2015). Monitoring athletes through self-report: Factors influencing implementation. *Journal of Sports Science & Medicine*, *14*(1), 137–

146. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/25729301

Science & Motion. (2016). SAM PuttLab (Version 5). Floersheim, Germany: Science & Motion GmbH.

Scrucca, L. (2013). GA: A Package for Genetic Algorithms in R. *Journal of Statistical Software, Articles*, *53*(4), 1–37. doi: 10.18637/jss.v053.i04

Scrucca, L., & Sahdra, B. (2016). GAabbreviate: Abbreviating Questionnaires (or Other Measures) using Genetic Algorithms (Version 1.3) [R package]. Retrieved from http://CRAN.R-project.org/package=GAabbreviate

Smith, G. T., McCarthy, D. M., & Anderson, K. G. (2000). On the sins of short-form development. *Psychological Assessment*, *12*(1), 102–111. doi: 10.1037/1040-3590.12.1.102

Solberg, E. E., Berglund, K. A., Engen, O., Ekeberg, O., & Loeb, M. (1996). The effect of meditation on shooting performance. *British Journal of Sports Medicine*, *30*(4), 342–346. Retrieved from http://europepmc.org/backend/ptpmcrender.fcgi?accid=PMC1332422&blobtype=pdf

Stocker, E., Englert, C., & Seiler, R. (2018). Self-control strength and mindfulness in physical exercise performance: Does a short mindfulness induction compensate for the detrimental ego depletion effect? *Journal of Applied Sport Psychology*, 1–22. doi: 10.1080/10413200.2018.1471754

Tamminen, K. A., & Poucher, Z. A. (2018). Open science in sport and exercise psychology: Review of current approaches and considerations for qualitative inquiry. *Psychology of Sport and Exercise*, *36*, 17–28. doi: 10.1016/j.psychsport.2017.12.010

Tavakol, M., & Dennick, R. (2011). Making sense of Cronbach's alpha. *The Journal of the International Association of Medical Science Educators: JIAMSE*, *2*, 53–55. doi: 10.5116/ijme.4dfb.8dfd

Taylor, K., Chapman, D., Cronin, J., Newton, M. J., & Gill, N. (2012). Fatigue monitoring in high performance sport: a survey of current trends. *J Aust Strength Cond*, *20*(1), 12–23. Retrieved from https://www.researchgate.net/profile/Dale_Chapman/publication/270905024_Fatigue_Monitoring_in_High_Performance_Sport_A_Survey_of_Current_Trends/links/54b9af6d0cf24e50e93dc960.pdf

Thienot, E., Jackson, B., Dimmock, J., Grove, J. R., Bernier, M., & Fournier, J. F. (2014). Development and preliminary validation of the mindfulness inventory for sport. *Psychology of Sport and Exercise*, *15*(1), 72–80. doi: 10.1016/j.psychsport.2013.10.003

U.S. Department of Health and Human Services FDA Center for Drug Evaluation and Research, U.S. Department of Health and Human Services FDA Center for Biologics Evaluation and Research, & U.S. Department of Health and Human Services FDA Center for Devices and Radiological Health. (2006). Guidance for industry: patient-reported outcome measures: use in medical product development to support labeling claims: draft guidance. *Health and Quality of Life Outcomes*, *4*, 79. doi: 10.1186/1477-7525-4-79

Yarkoni, T. (2010). The abbreviation of personality, or how to measure 200 personality scales with 200 items. *Journal of Research in Personality*, *44*(2), 180–198. doi: 10.1016/j.jrp.2010.01.002

Zhang, C. Q., Chung, P. K., & Si, G. (2017). Assessing acceptance in mindfulness with direct-worded items: The development and initial validation of the Athlete Mindfulness Questionnaire. *Journal of Sport and Health Science*, *6*(3), 311–320. doi: 10.1016/j.jshs.2015.09.010

Zhang, C. Q., Chung, P. K., Si, G., & Gucciardi, D. F. (2016). Measuring decentering as a unidimensional construct: The development and initial validation of the Decentering

Scale for Sport. *Psychology of Sport and Exercise*, *24*, 147–158. doi: 10.1016/j.psychsport.2016.02.006

Zhang, C. Q., Si, G., Duan, Y., Lyu, Y., Keatley, D. A., & Chan, D. K. C. (2016). The effects of mindfulness training on beginners' skill acquisition in dart throwing: A randomized controlled trial. *Psychology of Sport and Exercise*, *22*, 279–285. doi: 10.1016/j.psychsport.2015.09.005

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's α, Revelle's β, and Mcdonald's ωH: their relations with each other and two alternative conceptualizations of reliability. *Psychometrika*, *70*(1), 123–133. doi: 10.1007/s11336-003-0974-7