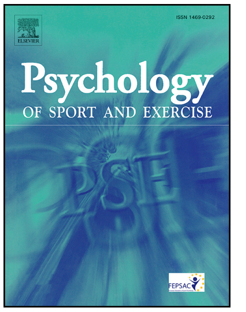
Journal Pre-proof



Reliable measurement in sport psychology: The case of performance outcome measures

Geoffrey Schweizer, Philip Furley, Nicolas Rost, Kai Barth

PII: S1469-0292(19)30756-3

DOI: https://doi.org/10.1016/j.psychsport.2020.101663

Reference: PSYSPO 101663

To appear in: *Psychology of Sport & Exercise*

Received Date: 31 October 2019

Revised Date: 20 January 2020

Accepted Date: 2 February 2020

Please cite this article as: Schweizer, G., Furley, P., Rost, N., Barth, K., Reliable measurement in sport psychology: The case of performance outcome measures, *Psychology of Sport & Exercise* (2020), doi: https://doi.org/10.1016/j.psychsport.2020.101663.

Reliable Measurement in Sport Psychology: The Case of Performance Outcome Measures

Geoffrey Schweizer[1], Philip Furley[2], Nicolas Rost[3] & Kai Barth[1]

[1]Department of Sport Psychology, Heidelberg University, Heidelberg, Germany

[2]Institute of Cognitive and Team/Racket Sport Research, German Sport University Cologne

[3]Max Planck Institute of Psychiatry, Department of Translational Research in Psychiatry, Munich, Germany; International Max Planck Research School for Translational Psychiatry, Munich, Germany

Author Note

Geoffrey Schweizer, Heidelberg University, Department of Sport Psychology, INF 720, 69120 Heidelberg, Germany, Phone: +49 (0) 6221 546033, Email: geoffrey.schweizer@issw.uni-heidelberg.de

Philip Furley, German Sport University Cologne, Institute of Cognitive and Team/Racket Sport Research, Am Sportpark Müngersdorf 6, 50933 Köln, Germany, Email: p.furley@dshs-koeln.de

Nicolas Rost, Max Planck-Institute of Psychiatry, Department of Translational Research in Psychiatry, Kraepelinstr. 2-10, 80804 München, Germany; International Max Planck Research School for Translational Psychiatry, Munich, Germany, Email: nicolas_rost@psych.mpg.de

Kai Barth, Email: kai-barth@web.de

Correspondence concerning this article should be addressed to Geoffrey Schweizer, Heidelberg University, Department of Sport Psychology, INF 720, 69120 Heidelberg, Germany, Phone: +49 (0) 6221 546033, Email: geoffrey.schweizer@issw.uni-heidelberg.de

Declarations of interest: none.

Reliable Measurement in Sport Psychology: The Case of Performance Outcome Measures

Abstract

Objectives: The present research addresses a neglected aspect within the current Zeitgeist of improving methodological standards in (sport)psychology: reliable measurement. We discuss and highlight the importance of reliable measurement from different perspectives and empirically assess reliability of three commonly used performance outcome measures in order to give guidelines to researchers on how to increase reliability of measurements of performance outcomes.

Method: In three studies we estimate 5 different reliability coefficients for three performance outcome measures based on 14 golf putts (study 1; N = 100), 14 dart throws (study 2; N = 200; 100 sports students; 100 non-sports students) and 14 free throws in basketball (study 3; N = 192; 100 non-basketball players; 92 basketball players).

Results: The highest reliability was the odd-even reliability for darts for the whole sample (.888), followed by golf putts (.714 for distance from the hole, .614 for successful putts) and free throws (.504 non-basketball players; .62 for basketball players; and .826 for whole sample).

Conclusions: Based on theoretical considerations and our empirical findings we give practical guidelines to improve reliability for performance outcome measures in sport psychology.

183 words

Keywords: Classical Test Theory, replicability, research quality, golf putts, darts, basketball

Reliable Measurement in Sport Psychology: The Case of Performance Outcome Measures

"For my money, the #1 neglected topic in statistics is measurement" (Gelman, 2015).

In the past ten years, there has been a controversial discussion regarding the quality of psychological research (e.g., Nelson, Simmons, & Simonsohn, 2018). Many authors have criticized what they perceive as problematic research practices, that lead to low rates of replications and weaker empirical support for psychological theories and interventions than researchers themselves may believe. Some authors even announced a "crisis of confidence" (Pashler & Wagenmakers, 2012, p. 528). Currently, the aforementioned discussion has developed from pointing out existing problems to creating viable practices for strengthening the quality of empirical research in the future (Lakens & Evers, 2014; Munafo et al., 2017; Nelson et al., 2018). For example, future research is supposed to benefit from appropriate sample sizes (Button et al., 2013; Fraley & Vazire, 2014; Schönbrodt & Perugini, 2013; Schweizer & Furley, 2016), methodological advances and new software[1] (e.g., Wagenmakers et al., 2018; Love et al., 2019), replicability projects (Camerer et al., 2018; OSC, 2015; Soto, 2019), the avoidance of p-hacking (e.g., Simmons, Nelson, & Simonsohn, 2011), the opportunity to preregister plans for studies and data analyses (e.g., Nosek, Ebersole, DeHaven, & Mellor, 2018), transparency (e.g., Nosek et al., 2015), and, maybe most importantly, a heightened awareness for the importance of these and other related issues. In the light of these changes, some authors have even gone as far as to proclaim a "renaissance" of psychological research (Nelson et al., 2018, p. 511).

However, as indicated by the quote above, with few exemptions, measurement has long been absent from the current discussion on research quality[2]. This is particularly remarkable as measurement is strongly connected to many of the issues that have been debated, as we will show below. In the present publication, we will focus on one particularly important aspect of measurement, namely reliability. Below, we will outline why we focus on reliability instead of other aspects of measurement. The main goal of the present paper is twofold: First, we aim to alert readers to the importance of reliable measurement. We illustrate why reliability is crucial for high-quality research in general, and particularly so for sport psychology, as we will argue. Second, we aim to assess reliability estimates for three performance outcome measures (golf putts, dart throws, and basketball free throws) commonly used in sport psychological research. Based on these estimates, we make recommendations for future research in sport psychology.

What is Reliability?

In Classical Test Theory (CTT), every observed value ($Y_{observed}$) consists of a true value (T_{true}) and measurement error (Y_{error}). Measurement error (Y_{error}) is defined as random in this context (e.g., Bühner, 2011; Lord & Novick, 1968; Steyer & Eid, 1993; Vaughn, Lee, & Kamata, 2012).

Equation 1

$$Y_{observed} = T_{true} + Y_{error}$$

This means that every measurement consists of the respective person's true value and a random error which is due to the measurement's imperfectness. The smaller a measurement procedure's error, the closer each single measurement will be on average to the true value. As measurement error is defined as random, it must have an expectancy value of 0. Furthermore, it is defined as having a finite variance. Measurement error is usually supposed to be normally distributed. This means that averaged over a large number of individual measurements the mean of the measurement error is supposed to be 0. In other words, larger errors are supposed to be less likely than smaller errors, and errors cancel each other out. With only enough measurements, therefore, the mean of the observed values will equal the mean of the true values. The more measurement error, the larger this number of measurements has to be (see Appendix A for more details). However, for every single measurement we cannot say how large its individual error (Y_{error}) is. In this context measurement error refers to the random error associated to every single measurement due to the measurement's imperfection. It does not refer to sampling error (i.e., differences between samples due to interindividual variation of the true values) or intraindividual variation (i.e., differences between measurement points due to intraindividual variation of the true values).

Reliability can be understood as the inverse of measurement error. In other words, the less error a measurement contains, the higher its reliability and vice versa. Formally, reliability is defined as the proportion of true variation among the entire variation:

Equation 2

$$r_{yy} = \frac{\sigma^2_{y_{true}}}{\sigma^2_{y_{true}} + \sigma^2_{y_{error}}}$$

Reliability can be estimated via several approaches (e.g., test-retest-reliability; split-half reliability; parallel-test reliability; Cronbach's alpha; see Appendix A for more information on different coefficients).

Why Does Reliability Matter?

"Measurement error adds noise to predictions, increases uncertainty in parameter estimates, and makes it more difficult to discover new phenomena or to distinguish among competing theories" (Loken & Gelman, 2017, p. 584).

The question why reliability matters could be answered in one sentence: The higher a measurement's reliability (i.e., the less measurement error contained on average in a single observation), the more precise is every single measurement on average. However, it is possible to look at this feature from several perspectives, and different researchers may be more familiar with some of these perspectives than with others. Therefore, we will explore the benefits of reliable measurement in more detail below. Importantly, the following arguments are not independent of each other, but they are all a direct consequence of the above mentioned observation.

Reliability matters for Type-1 and Type-2 errors.

Both Type-1 errors (false positives) and Type-2 errors (false negatives) can be found in psychological research. Researchers commit a Type-1 error (or false positive) when they report finding an effect when truly there is none; or when refuting a null hypothesis when in truth the null hypothesis is correct. Conversely, researchers commit a Type-2 error when they report not finding an effect when truly there is one; or when retaining a null hypothesis when in truth the null hypothesis is false (Fraley & Vazire, 2014). Reliability plays a role for both these errors (Loken & Gelman, 2017). Type-1 and Type-2 errors are often presented in the framework of Null-Hypothesis-Significance-Testing (NHST)[3], however, they are relevant for all statistical perspectives that contain binary decisions such as accept-or-reject, present-or-absent, retain-or-dismiss.

It has been understood for more than a century that measurement error attenuates effect sizes (in this case correlations; Spearman, 1904, 1910). This means that when measuring an effect with less than perfect reliability, the estimated effect size will be smaller than the real-world effect size. The lower the reliability, the smaller the estimated effect size will be compared to its real size. For example, when researchers estimate a correlation coefficient between two measures, then the estimated coefficient will be smaller than the true coefficient to the extent that one or both of the measurements are less than perfectly reliable. As larger effects are more easily detected, measurement error thus increases the likelihood of committing a Type-2 error (or false negative): When planning for power, researchers usually plan their sample size with the goal of achieving a certain power (e.g., .9) based on an expected real-world effect size (e.g., r_{expected} = .7). The less reliable their measurement, the smaller the effect size they estimate in their study gets (e.g., $r_{\text{estimated}}$ = .5). However, the likelihood of finding this smaller effect size is lower than the likelihood of finding the real-world (and larger) effect size. Thus, the researchers in this example risk committing a Type-2 error: They do not find the effect although it truly exists.

In order to obtain the same power to detect an effect of the same (true) size, researchers need smaller samples when using reliable measures than when using unreliable measures. Therefore, research using more reliable measures is, all else equal, more economical than research with less reliable measures.

Only recently have researchers suggested that in addition to deflating coefficient estimates, measurement error can also inflate coefficient estimates (Loken & Gelman, 2017). Whether measurement error deflates or inflates coefficient estimates depends primarily on sample sizes: In large samples, measurement error nearly always deflates coefficient estimates. Here, measurement error primarily leads to Type-2 errors. In small samples, measurement error can deflate and inflate estimates. Therefore, in small samples, measurement error can lead to Type-1 and Type-2 errors. When researchers are more likely to publish larger effect sizes than smaller ones (e.g., because they are more likely to be statistically significant or because they seem to be more impressive), measurement error is likely to contribute to the potentially high proportion of false-positive findings that have been diagnosed for psychological research, because it inflates effect sizes in small samples (Loken & Gelman, 2017).

Taken together, all else being equal, a research field will benefit from more reliable measurement, due to less Type-1 and less Type-2 errors. Or stated differently, the field with more reliable measures is both less likely to miss potentially important effects and it is less likely to report an effect that in reality does not exist.

Reliability matters for replications.

In past replication projects, several definitions and operationalizations of successful or unsuccessful replication attempts have been employed (Camerer et al., 2018; OSC, 2015; Soto, 2019). What exactly constitutes a successful or unsuccessful replication remains a matter of debate (for an enlightening discussion and one possible solution see Simonsohn, 2015). Broadly, one can distinguish two strategies for defining a successful replication. The first strategy considers a replication as successful when both the original study and the replication study have significant results in the same direction. The second strategy compares the effect sizes of the original study and the replication study with each other. An effect is considered to be successfully replicated when the effect sizes produced by the original and the replication study do not differ.

For the first strategy, reliability matters because it affects the studies' power. The probability of replicating a true effect depends, among other factors, on the reliability of the measure with which the to-be-replicated effect is assessed (Stanley & Spence, 2014). The lower the reliability of a measure, the lower the probability of replicating an effect, even when the effect is true. Therefore, when measurement is unreliable, unsuccessful replication attempts may both be a consequence of an effect not being true or of an effect being measured with much error. For example, Soto finds that the successful replication rate in a large scale replication project from personality psychology was substantially higher than the respective rates in large scale replication projects from other behavioral sciences (2019). He goes on to speculate that one reason for this striking discrepancy (among others, such as sample sizes and type of focal effects) might be that personality psychological research uses more standardized and thus more reliable measurements than, for example, social psychological research.

For the second strategy, reliability matters because, as explained above, measurement error may both inflate and deflate estimates of effect sizes, making it less informative to compare two effect sizes to each other. As replication attempts are becoming ever more important in science, so does the role of reliable measurement, as without reliable measurement replication attempts are at least hard to interpret or at worst futile.

Reliability matters for the impact of p-hacking.

P-hacking refers to the practice of “selectively reporting data and analyses” or, in other words, “conducting multiple analyses on the same data set and then reporting only the one(s) that obtained statistical significance” (Nelson et al., 2018, p. 513). When researchers employ p-hacking, the likelihood of obtaining a false-positive increases “dramatically” beyond the level usually assumed by researchers (Nelson et al., 2018, p. 513). Typical examples of p-hacking[4] include a) having two correlated dependent variables and selectively reporting one of them, b) adding observations to the sample and stopping once statistical significance has been reached, c) deciding whether to drop one out of several experimental conditions, d) selectively controlling for gender or for the interaction of gender with treatment, or e) combinations thereof (Simmons et al., 2011). In the simulations run by Simmons and colleagues, p-hacking could lead to a likelihood of obtaining a false-positive (i.e., finding a significant result when truly there is no effect) of up to 61% (Nelson et al., 2018; Simmons et al., 2011). As a result of p-hacking, there is supposed to be a high proportion of false and therefore non-replicable findings in the psychological literature. Nelson and colleagues (2018) suggest that p-hacking (i.e., analyzing the data until researchers find a significant result) and not publication bias (i.e., simply not publishing non-significant results) is the real answer to the decades-old question how psychologists manage to publish such a high proportion of significant results when their studies typically have rather low power. In a recent paper, Friese and Frankenbach (in press) suggest that p-hacking and publication bias interact: In their simulation study, the extent to which p-hacking distorts meta-analytic effect size estimates depends on the level of publication bias and on true effect sizes.

Although the problematic influence of p-hacking on research quality has been described in detail (Simmons et al., 2011), it remains an open question how to reduce its impact in future research. Again, reliability plays a role as p-hacking exploits random variation. For example, the strategy of successively increasing the sample size until a certain difference becomes (randomly) significant works “best” when this very difference is subject to lots of random variation. As random variation increases with measurement error, so do the opportunities to employ p-hacking. It follows then that highly reliable measurements should be less vulnerable to p-hacking. Therefore, one (rather indirect) method among others to reduce the impact of p-hacking would be to employ highly reliable measurements. Obviously, this would not entirely rule out the possibility of p-hacking, but it would at least to some extent decrease the potential for employing them.

Reliability matters for comparisons between measurements.

Whereas the abovementioned benefits of high reliability follow directly from its definition, namely a measurement with comparably low error, the reliability of measures is consequential beyond simple estimations of the parameter of interest. Reliability is particularly important for comparisons between measurements. This can be comparisons between different studies, comparisons between different measurements in a single study or comparisons of some effect on different measurements in a single study.

Suppose some researchers are interested in the question whether some treatment has a (differential) effect on two different variables. They find several studies reporting an effect on variable A, and several studies not finding an affect on variable B. They conclude that the treatment works for variable A, but nor for variable B. However, and unfortunately so, in this example variable A was assessed with a more reliable measure than variable B. Therefore, the observed difference might simply be due to measurement error. This hypothetical scenario gets worse when we assume that preferred measurement and theoretical background of researchers might be correlated. In this case, different theories might appear to be differentially supported by evidence, while the only real difference is measurement error. These observations also hold when researchers compare effects on different variables within one study[5]. For example, researchers might conclude that their treatment affects variable A (e.g., some symptom of a disease), but not variable B (e.g., an unwanted side effect). Again, this conclusion is only legitimate when both variables are measured with the same high reliability.

Finally, this also holds for all kinds of multiple regression strategies and related attempts to control for one variable when estimating associations between two or more additional variables (Westfall & Yarkoni, 2016). For example, when one predictor significantly predicts the outcome variable while controlling for another predictor, researchers often interpret this finding as indicative of incremental validity, which itself may be interpreted as signifying that both predictors measure “strongly related but conceptually distinct constructs” (Westfall & Yarkoni, 2016, e0152719). However, as Westfall and Yarkoni show, “… a simpler interpretation that is often equally consistent with the data is that both predictors are simply noisy indicators of the same construct” (2016, e0152719). Westfall and Yarkoni conclude that reliable measurement is particularly important when trying to assess incremental validity in regression models.

Why reliability matters: A summary.

Taken together, increasing reliability should lead to both less Type-1 and less Type-2 errors, a higher chance of replicating an effect given it is true as well as making replication attempts more informative in general. Additionally, when measurements are more reliable, smaller sample sizes are needed in order to safeguard against statistical errors, p-hacking, and biases. Likewise, comparing between measurements is easier when both measurements are reliable. Taken together, it seems safe to conclude that research without reliable measurement does not make much sense in general, and particularly it does not make much sense in the age of replicability.

Reliability in Sport Psychology

As we have outlined above there are good reasons to make efforts to increase reliability in science. Nevertheless, it is important to note that not every field of investigation or every measurement tool faces comparable challenges when it comes to both validity and reliability. In terms of validity it seems clear that certain psychological measures (e.g., a questionnaire measuring a person’s tendency to behave aggressively) struggle with more problems regarding validity as physical measures, for example assessing a person’s weight or body size, which has led researchers to speak of a validation crisis within the field of psychology (Schimmack, 2010; 2019).

While questionnaires can be both problematic in terms of validity and reliability, other measurement techniques have high face-validity (i.e., there is little doubt as to whether they actually measure what they claim to measure). For example, most people would probably agree that measuring a person's performance shooting basketball free throws is a valid measure of this person's ability to shoot free throws. However, it is less clear how reliable this measure is, or what has to be taken into account when reliably trying to assess perceptual-motor performance in answering different research questions in sport psychology. Hence, the present research focused on reliability of commonly used individual sport performance outcome measures.

Discussions of reliability have not been absent within sport science (Hopkins, 2000; 2017; Zhu, 2013), and reliability has also been the focus of increasing research endeavors in some subfields of sport science, for example in determining both the validity and reliability of new technologies like GPS (Global Positioning System) in assessing sport performance data (Barbero-Álvarez, Coutts, Granda, Barbero-Álvarez, & Castagna, 2010; Coutts & Duffield, 2010; Jennings, Cormack, Coutts, Boyd, & Aughey, 2010; Johnston, Watsford, Pine, Spurrs, Murphy, & Pruyn, 2012; Petersen, Pyne, Portus, & Dawson, 2009). In sport and exercise psychology, Eklund, Tenenbaum and Kamata (2012) provide an extensive overview about nearly all potentially important aspects of measurement in sport and exercise psychology, from basic concepts to specific issues, such as cognitive, motivational, emotional and behavioral measurement. These discussions and analyses have shown that reliable measurement of behavior in sports, although these measures appear high in face validity, is not a trivial topic. While measurement is an important topic at all levels of analyses within sports (e.g., biochemical measures, physiological measures, biomechanical measures, psychological questionnaires, anthropometric measures, behavioral measures, etc.), some of the most relevant measures (at least in terms of spectator interest or financial reward) are outcome measures of sports performance.

Somewhat surprisingly, we are not aware of any literature systematically analyzing the reliability of typically used sport performance outcome measures, although plenty of research uses sport performance measures as dependent variables. Maybe claims like “sport measures outcome with a finality of judgment that scientific papers would not pass” (Walsh, 2014, p. 860) have led researchers to simply assume sport outcome measures are reliable without needing to pay special attention to this. To address this shortcoming in the literature, we decided to first identify the most commonly used outcome measures of skilled perceptual-motor performance in sport psychology and subsequently calculate different reliability indices of these measures in a series of empirical studies.

A literature search identified 40 papers using golf putts as a dependent variable, 37 papers using darts, and 28 using free throws in basketball (see the reference list in the supplement for an overview). Therefore, it seems safe to argue that these are frequently employed individual sport performance outcome measures in sport psychology[6]. Reliabilities were not reported in any of these papers[7]. It is important to note that we are not pointing out or criticising these papers. We ourselves have not reported reliability coefficients in most of our papers, when employing other measures than questionnaires.

However, this is precisely our point: Whereas everybody cares about reliable measurement when reporting questionnaire data, hardly anybody does when reporting performance outcome measures. There are probably several reasons for this. One, whereas it may seem rather straightforward how to compute reliabilities for questionnaires (e.g., most examples from the methodological literature refer to questionnaires), it may seem to be more unclear how to compute reliabilities for performance outcome data. Second, there is a common perception that a measure must have been reliable when there has been a significant result for this variable

(Loken & Gelman, 2017). Therefore, it may seem unnecessary to examine its reliability. As explained above, this is problematic for several reasons (see Loken & Gelman, 2017, for more details on this misconception).

The Present Research

The main goal of the present research was to estimate reliability coefficients for three commonly used individual performance outcome measures in sport psychology, namely golf putts (study 1), darts (study 2) and free throws in basketball (study 3). Furthermore, we aimed to investigate whether these reliability coefficients are dependent upon different samples in general and upon participants' experience with the respective task in particular.

General Method

Here, we describe the rationale common to all three studies. In the section below we describe characteristics unique to each study. In all studies, participants provided informed consent before commencing the study and were thanked and debriefed before receiving some candy as compensation for participating. Participants were neither paid, nor were they incentivized dependent on their performance. In all studies, participants first performed 20 training trials, before executing 14 test trials. We only estimated reliability coefficients for the 14 test trials, not for the training trials. The training trials were intended to reduce the influence of potential short-term learning effects on the reliability estimates (Hopkins, 2000), to get participants calibrated to the performance context (Ajemian, D'Ausilio, Moorman, & Bizzi, 2010; Wunderlich, Heurer, Furley, & Memmert, 2019), and in turn decrease measurement error. We assumed potential learning and calibration effects to reach an asymptotical level after the learning trials and therefore performance to be stable for the test trials. In all studies, the setup was highly standardized, including videos demonstrating the correct execution of the required movements to all participants (regardless of their familiarity with the task). Participants were told to try to achieve optimal performance, but experimenters emphasized that we would not evaluate individual performance in order not to induce pressure (e.g., Baumeister & Showers, 1986).

For all variables, we estimated their *split-half reliability* using two different methods, one method splitting each test in a first half and a second half and the odd-even method (please see Appendix A for further elaboration on our statistical approach). When using the “first-half vs. second half” method, for every participant, we computed one mean across the first seven test trials (i.e., trials 1-7) and one mean across the second seven test trials (i.e., trials 8-14). We then computed Pearson’s correlation coefficient for the correlation between the first and the second mean. When using the odd-even method, for every participant, we computed one mean across the seven odd-numbered test trials (i.e., trials 1, 3, 5, 7, 9, 11, 13) and one mean across the seven even-numbered test trials (i.e., trials 2, 4, 6, 8, 10, 12, 14). We then computed Pearson’s correlation coefficient for the correlation between the first and the second mean. According to Classical Test Theory (CTT), the resulting correlations can be considered one estimate for the respective measures’ reliability. Reliability depends on the number of items, and split-half reliabilities thus estimate the reliability for a test of half its original length (i.e., in our case, for seven instead of 14 trials). Therefore, we used the Spearman-Brown formula to estimate the reliability of all 14 trials, based on the obtained reliability coefficients (please see Appendix A for the formula). Additionally, we computed *Cronbach’s Alpha* for all measures.

In all studies, we used performance measures (and respective instructions) that allowed for obtaining continuous measurements, which is a prerequisite for estimating reliability according to CTT (please see Appendix A for further information). Furthermore, where necessary, we aimed at minimizing the number of missing values, that is attempts that we could not measure.

We planned to total 200 participants per study (please see Appendix A for further elaboration on sample size planning). In all studies, we planned to collect data in two subsamples that differ regarding sports experience, each subsample totalling 100 participants. Below, we describe all studies in detail in order to facilitate interpretation and replication. We encourage researchers to contact us for more details. When we do not refer to a particular reliability coefficient, we always refer to the odd-even reliability, as it is usually considered superior to the "first-half vs. second half" reliability. In all studies and in all subsamples, there is no significant difference between the mean value for the odd and for the even items, and neither do standard deviations differ, which is considered a prerequisite for estimating odd-even reliability.

Study 1: Golf Putts

Participants

One hundred students of Heidelberg University participated in the study (58 men and 42 women; M_{age} = 24.9; SD_{age} = 7.9). Sixty-six of them were sports students, 34 were not. None of them reported having experience playing golf that went beyond participating in one basic course. Contrary to our plans, we did not collect data from a second subsample (for an explanation why, please see the Discussion section below).

Apparatus and Procedure

Participants were positioned 200 cm away from the hole and instructed to assume a typical putting position. They could choose between a putter for left-handers and for right-handers. In order to maximize standardization, the study was conducted in a laboratory room.

Therefore, participants did not perform on a real green but on a putting mat made of plastic, as is common in sport psychological studies (see Supplement for several examples).

Participants were instructed to aim for the hole and informed that performance would be measured as distance from the hole. This allowed us to measure performance in a continuous way. Simply counting successful putts is also a commonly used measure and we therefore also estimated reliabilities for number of successful putts. This allowed us to compare the reliabilities of two different performance outcome measures constructed from the same task. In order to obtain continuous measurement, we added up all attempts, as sums of binary variables can be treated as continuous variables (e.g., Lunney, 1970).

Results and Discussion

In the whole sample, the average distance to the hole across all 14 putts was 257 mm (*SD* = 175; *Md* = 240; *Mode* = 241). On average, 53% of all putts were successful. When looking at the mean performances for each of the 14 putts separately, no learning trend was apparent (see Figure 1). Due to failures in data recording, in total the results of 15 putts out of 1400 were not recorded. Results do not change when excluding the respective participants.

For the continuous performance outcome measure (i.e., distance from the hole), the different reliability coefficients do not differ from each other (see Table 1). Thus, results are not dependent on a particular coefficient. Reliability coefficients can not be considered acceptable for seven putts only. When estimating reliabilities for all 14 putts using the Spearman-Brown-Formula, reliability coefficients expectably get higher, but they are still lower than what is usually considered acceptable (e.g., Vaughn et al., 2012). The Spearman-Brown corrected odd-even reliability for the whole sample is .714 (CI_{95} [.602; .798]).

For the binary performance outcome measure (i.e., number of successful putts), the different reliability coefficients do not differ from each other (see Table 1). None of them reaches a level commonly considered as acceptable, with the highest estimate for all trials being .614 (CI_{95} [.475; .723]). Thus, at least descriptively, reliability estimates for the binary performance outcome measure are lower than for the continuous one. However, confidence intervals overlap. Based on our data and our sample, golf putts as conducted in the present study did not possess sufficient reliability to be employed as a performance outcome measure in a sport psychological study. Reliability estimates for distances from the hole are somewhat better, but they still do not reach levels usually considered acceptable for other psychological measurement procedures. Based on the reliability estimates for the number of putts in the current study, the Spearman-Brown formula allows to calculate what number of putts would be necessary in order to achieve a certain level of reliability (e.g., .8 or .9). We present these calculations in the section *"comparison between performance outcome measures"*.

Putting distances in sport psychological studies vary from 100 to 400 cm, with 200 cm being common (see Supplement). Therefore, we decided to use 200 cm in our study. However, we realized that not only putting performance, but also reliability in golf putts probably strongly depends on the distance to the hole (as reliability of the performance measurement depends on the true score of the performance, which probably varies with distance). At the same time, there is no standard putting distance. Therefore, we decided not to conduct a second study with another 100 participants, as initially planned, but instead to move on to another performance measure. Our next two performance measures (darts and free throws) feature standard distances. Therefore, the problem described above does not apply to them.

Study 2: Darts

Participants

Study 2 consisted of two subsamples (sample 2a and sample 2b). Both samples consisted of 100 participants, totalling 200 participants. In sample 2a, there were 50 women and 50 men (4 left-handed). All of them were students at Heidelberg University. In sample 2b, there were 44 women and 56 men (10 left-handed). All of them were sports students at Heidelberg University. Thus, in the whole sample there were 94 women and 106 men (14 left-handed), 100 non-sports students and 100 sports students.

Apparatus and Procedure

In line with the World Darts Association's standards, we placed the dart board in such a way that the centre of the bull (also called bullseye) was at 173 cm (5 ft 8 inches) above ground. The diameter of the dart board was 400 mm. Participants were positioned behind a line (the so-called oche) that was 237 cm (7 ft 9.25 inches) away from the board. Behind the dart board we placed a board made of rigid foam (size: 120 cm x 120 cm; thickness: 40 mm). This setup allowed us to measure throws that missed the dart board but got stuck in the foam board, in order to minimize missing values. When a throw did not reach the board, participants were allowed to repeat the attempt (however, this happened hardly ever, due to the size of the foam board). We utilized regular tournament darts with a length of circa 155mm and a weight of circa 18g. Tournament darts come in different variants. Our darts consisted of a steel point, a brass barrel, an aluminium shaft and a standard shape flight.

Participants were instructed to aim for the bull and informed that performance would be measured as distance from the bull. This allowed us to measure performance in a continuous way (i.e., to obtain interval scale data), which is a prerequisite for estimating reliability. The usual scoring system in darts, however, would probably not have produced continuous measurements

(Selkirk, 1976; Tibshirani, Price, & Taylor, 2011). Aiming for the bull is a common instruction in sport psychological studies (see Supplement). In line with our measurement and the instructions, we did not utilize a dartboard with radial sections and double and triple rings. Instead, we used a dartboard with concentric rings of equal width.

Results and Discussion

In the whole sample, the average distance to the bull across all 14 darts was 87 mm (*SD* = 39; *Md* = 79; *Mode* = 65; see Table 2). The sports students (sample 2b) performed better than the non-sports students (sample 2a), and the men better than the women (see Table 2). When looking at the mean performances for each of the 14 darts separately, no clear learning trend was apparent (see Figure 2).

The different reliability coefficients do not differ from each other in each sample (see Table 1). Thus, results are not dependent on a particular coefficient. Likewise, reliability coefficients do not differ between both subsamples. Neither does each reliability coefficient for the whole sample differ from the respective coefficients in both subsamples.

Reliability coefficients can already be considered acceptable for seven throws only: The odd-even reliability for the whole sample based on seven throws is .799 (CI_{95} [.743; .844]). When estimating reliabilities for all 14 throws using the Spearman-Brown-Formula, reliability coefficients are high: The Spearman-Brown corrected odd-even reliability for the whole sample is .888 (CI_{95} [.855; .914]). At least in this study, both in the subsamples and in the overall sample, dart throws seemed to capture a substantial proportion of systematic variation as opposed to random variation and therefore seemed to be able to capture variation in participants' true score rather well.

Furthermore, we investigated into the question whether reliability coefficients varied between different samples or between different groups of participants. We did so in a more exploratory manner, based on assumptions that we consider to be common when planning sport psychological experiments. As different reliability coefficients do not differ from each other, from now on we refer to the odd-even reliability, as we consider it the most appropriate (see Appendix A). In our further analyses, we distinguished between a) sports students and non-sports students, b) women and men, and c) participants who play darts and participants who do not. We distinguished between darts players and non-darts players based on a median split on participants' answers to the question "How often did you play darts during the past twelve months?". All participants who reported never to have played darts in the past twelve months were assigned to the group of non-darts players (n = 106), whereas all other participants were assigned to the group of darts players (n = 94). We conducted a median split in order to obtain groups of roughly equal sample size, although this approach has some disadvantages. The main disadvantage here is that the group of darts players did not only contain participants who played regularly, but also participants who had only played a couple of times. We address this shortcoming in study 3.

First, mean performance differs between the two groups in all three comparisons (see Table 2). That means, a) sports students were significantly better than non-sports students (t[167.58] = 4.6, p < .001, d = 0.65), b) men were significantly better than women (t[137.62] = 9.46, p < .001, d = 1.39), and c) darts players were significantly better than non-darts players (t[198] = 4.83, p < .001, d = 0.68).

As already reported above, reliability coefficients did not differ between sports students and non-sports students (i.e., subsample 2b and subsample 2a, see Table 2). Descriptively, reliability was somewhat higher for women than for men, but confidence intervals overlap (see Table 2). Finally, reliability coefficients do not differ between darts players and non-darts players (see Table 2). If at all, non-darts players have a slightly higher reliability coefficient, but again, confidence intervals overlap.

Thus, and contrary to what one might have intuitively expected, neither gender, nor studying sports nor playing darts had an impact on reliability estimates. However, at least the last finding might be due to the fact that we distinguished darts players from non-darts players based on a median split, which is not the best method to compare different levels of experience. We address this issue in study 3.

Study 3: Basketball Free Throws

Participants

As one goal of study three was to further investigate into the role of sports experience for reliability, we aimed to obtain two different subsamples. One subsample should consist of experienced basketball players, whereas the other one should consist of comparably inexperienced players. We assigned potential participants to the sample of experienced players, when they were active members of a basketball club and reported their free throw success rate to be at least 30%. We assigned potential participants to the sample of inexperienced players when they did not fulfil the inclusion criteria for the experienced sample. These rules were defined prior to data collection. Additionally, participants had to be able to hit the rim or score a basket at least ten out of 20 times during the practice trials in order to make sure that they were sufficiently skilled.

Therefore, study 3 consisted of two subsamples (sample 3a and sample 3b). Sample 3a (the inexperienced sample) consisted of 100 participants (M_{age} = 24.8; SD_{age} = 4.2). Sample 3b (the experienced sample) consisted of 92 participants (M_{age} = 25.3; SD_{age} = 7.7), totalling 192 participants. In sample 3a, there were 50 women and 50 men. All of them were sports students at Heidelberg University. In sample 3b, there were 42 women and 50 men. All of them were players in regional basketball clubs. Thus, in the whole sample there were 92 women and 100 men, 92 non-sports students and 100 sports students. As a result of the above mentioned criteria for inclusion of participants into the different subsamples, both subsamples differed considerably with regard to basketball experience. On average, participants in subsample 3a (the rather inexperienced) reported to play basketball for 28 minutes per week (*SD* = 52), whereas participants in subsample 3b (the experienced) reported to play basketball for 314 minutes per week (*SD* = 206).

Apparatus and Procedure

Participants conducted all free throws in line with the regulations of the Fédération Internationale de Basketball (FIBA, 2018a; b). Participants were positioned behind the free throw line 422.5 cm from the middle point of the basket. According to FIBA, basketballs for men are supposed to weigh 567-650 g with a circumference of 74.9-78.0 cm. Basketballs for women are supposed to weigh 510-567 g with a circumference of 72.4-73.7 cm. We used two balls: one ball for all men and one ball for all women. From time to time, we made sure that both balls were still within the limits specified by the regulations. According to FIBA, the basket ring has to be positioned at a height of 304.8 cm +/- 0.6 cm and have an inside diameter of 45.0-45.9 cm. We made sure that the baskets utilized were within these specifications.

We coded each shot as either successful (the ball went through the basket) or not (the ball did not go through the basket). There were no missing values, as all shots could be coded. We did not distinguish between different kinds of unsuccessful shots, for example, balls hitting the rim or air balls, as researchers sometimes do. The reason for this is that assessing the difference between successful shots and different kinds of misses does not produce a continuous (i.e., interval scale) measurement. In order to obtain continuous measurement, we added up all attempts, as sums of binary variables can be treated as continuous variables (e.g., Lunney, 1970). Examples for this practice can be found, for example, in intelligence tests, where each single item produces a binary datum, however, items are summed up along scales and then treated in a continuous manner.

Results and Discussion

Experienced participants performed better than inexperienced participants. In the experienced sample (3b), the average success rate was 74% (*SD* = 17). In the inexperienced sample (3a), the average success rate was 32% (*SD* = 17%). When looking at the mean performances for each of the 14 throws separately, no learning trend was apparent, neither for sample 3a nor for sample 3b (see Figure 4 and Figure 5; see Appendix A for further elaboration on this issue).

The different reliability coefficients do not differ from each other in each sample (see Table 1). Thus, results are not dependent on a particular coefficient. At least descriptively, reliability coefficients are higher for the experienced sample (sample 3b) than for the novice sample. However, confidence intervals still overlap for both groups. In both subsamples, reliability coefficients for seven shots only are low, and they are still not acceptable when correcting for all 14 shots using the Spearman-Brown formula. The Spearman-Brown corrected odd-even reliability for sample 3a is .504 (CI_{95} [.341; .637]) and for sample 3b it is .62 (CI_{95} [.475; .732]).

Interestingly, the reliability coefficients for the whole sample are substantially higher than for each of the subsamples alone (i.e., confidence intervals do not overlap). The Spearman-Brown corrected odd-even reliability for the whole sample is .826 (CI_{95} [.755; .866]). This observation cannot be explained by the sample size itself. Instead, it means that in the whole sample the ratio of systematic variation to random variation was higher than in each of the subsamples. This makes sense when considering that the whole sample included both inexperienced and experienced players (i.e., much variation between participants), whereas in each of the subsamples there were only inexperienced or only experienced players (i.e., less variation between participants). This observation illustrates a common misconception when planning sport psychological studies: Researchers are sometimes tempted to think that reliability must be higher the more experienced or the better athlete participants are. However, this is not true, as illustrated by our data.

Comparison between Performance Outcome Measures

The highest reliability estimated in our study was the odd-even reliability for darts, in the whole sample for all 14 darts (.888; see Table 1). The respective estimates for golf putts are .714 (for distance from the hole) and .614 (for successful putts). The respective estimates for free throws are .504 (for the inexperienced sample 3a), .62 (for the experienced sample 3b) and .826 (for the whole sample). As the reliability for darts is acceptable, we further calculated what number of putts and what number of free throws would have been necessary in order to achieve the same reliability as in darts (i.e., .888). The Spearman-Brown formula allows to calculate these numbers based on the existing reliability estimates (see Equation 4 in Appendix A for the Spearman-Brown formula solved for k, which is test length). In brackets, we report the number of items researchers would have needed to achieve a reliability of .8 (instead of .888), which is often considered acceptable.

Based on the procedures and samples as employed in our studies, estimations based on the Spearman-Brown formula suggest that in order to achieve the same odd-even reliability as with 14 dart throws, researchers would have needed 44 (22) golf putts (when measuring distance from the hole) or 70 (35) golf putts (when counting successful putts). Likewise, they would have needed 109 (55) free throws (in the inexperienced sample), 68 (34) free throws (in the experienced sample) and 23 (12) free throws in the whole sample. These results demonstrate that differences that may look small when expressed in reliability coefficients may have large consequences for number of items or trials: In the above example, in order to achieve the same reliability, number of trials varies from 14 to 109!

The increases in item numbers in order to achieve a reliability of .8 are still substantial, however far smaller than the ones reported above for .888. This observation illustrates that increasing the number of items at first leads to relatively high increases in reliability, however, further gains in reliability need increasingly more items (Amelang & Zielinski, 2002).

General Discussion

Discussion of Results

Our results suggest that common sport performance outcome measures exhibit reliabilities whose interpretations when computed for psychological questionnaires would range from good (for darts =.888 and free throws in the combined sample =.826) to barely acceptable for golf putts (.714) (e.g., Vaugh et al., 2012; see Table 1). Thus, our results suggest that different sport performance outcome measures may have different reliabilities. Furthermore, our results demonstrate that reliability estimates increase when the number of items increases, which is a well known property of reliabilities according to CTT. Our results also demonstrate, that depending on a measurement's reliability, vastly different numbers of items are required in order to achieve the same acceptable level of reliability. This observation has consequences for the construction of performance outcome measures as we will discuss in more detail below.

Most importantly, our results demonstrate that reliabilities of sport performance outcome measurements may strongly depend on sample characteristics: Reliabilities estimates for both subsamples in basketball were very low (.504 and .62), however for the whole sample the estimated reliability was substantially higher (.826). This observation underlines the necessity of having samples with true-score variation when researchers want to obtain reliable measurement (see also Appendix A; Vaughn et al., 2012). At the same time this observation debunks what we (anecdotally) perceive to be a common misconception in sport psychological research, namely that reliability will be higher in expert samples than in non-expert samples. This also means that researchers need to be particularly careful when conducting studies with rather homogenous expert samples, as this approach might lead to low reliabilities.

When interpreting the present results, it is important to keep in mind that all estimates reported here depend on the respective samples and operationalizations and therefore do not necessarily generalize to other situations or samples (see Appendix A). Furthermore, it is important to keep in mind that our results may paint a rosier picture of reliabilities than is actually warranted when looking at existing studies: One reason why estimates appear to be rather high in our studies is that we employed 14 trials. To the extent that past studies may have employed fewer trials, all else equal, they had less reliable measurements.

Limitations and Unintended Consequences

We based our main conclusions on CTT in general and on specific estimators (i.e., odd-even reliability and the Spearman-Brown formula) in particular. However, alternative approaches exist. They comprise a) different estimates that have existed within the framework of CTT for a long time (e.g., the Kristof or the Guttman formulas, see Bühner, 2011); b) novel estimates that have been proposed only recently (e.g., omega as an alternative to Cronbach's alpha [McNeish, 2018]; weighted kappa [Robinson & O'Donoghue, 2007] to assess agreement amongst observers in performance analysis; and special coefficients for particular research designs within sport science [Hopkins, 2017]); and c) estimates computed via structural equation modeling (SEM; e.g., Raykov, 1997). All of these approaches have advantages and disadvantages and it is impossible to say that one of them is per se superior. Just as one example, one presumed advantage of omega above Chronbach's alpha is that omega relies less on modeling assumptions (McNeish, 2018). However, the advantages of omega have been questioned and currently there is a controversial discussion regarding its merits (Raykov & Marcoulides, 2019; Savalei & Reise, 2019).

Furthermore, CTT itself has some well-known weaknesses, for example its dependency on sometimes questionable assumptions and on sample characteristics (Bühner, 2011; see also Appendix A). Item Response Theory (IRT) in turn allows for modeling the probability of a response to an item as a joint function of both this item's difficulty (the item parameter) and a person's ability (the person parameter), which is a substantial advantage over CTT (Bühner, 2011). Taken together, we consider it to be important to keep in mind that alternatives to the particular estimates that we employed and to CTT exist. We hope to stipulate a discussion on which theoretical approach and which coefficients are best suited for measurement in sport psychology. In order to foster this discussion, we make all of our raw data public, so that researchers may take these data and calculate other estimates of reliability, SEMs or parameters from IRT.

As we have mentioned above, our estimates depend both on our samples and on our exact operationalization of the different measures. For example, maybe the reliability estimate for the free throws would have been different had participants been closer to the basket. Hence, theoretically, reliability (and its estimates) always refer to a measurement, *not* to an instrument. Therefore, we cannot say that we estimated *the* reliability of darts, or *the* reliability of free throws, instead we estimated the reliabilities of our specific measurements. An unintended consequence of our study would be if from now on researchers in sport psychology would predominantly use darts as dependent variable, because “it has been proven to be reliable”. Future studies with different samples might be different in terms of reliability.

Moreover, reliability must not be confused with validity. It would be a mistake if researchers simply used certain measures because they are reliable, and did not care about validity, a concern that has been raised in psychometrics (e.g., Bühner, 2011). To test theories that relate theoretical constructs to each other (e.g., construct A influences construct B for individuals drawn from population P under conditions C), it is necessary to not only have reliable measures, but also valid measures that actually measure construct A and B and control for P and C. Validity typically refers to whether a given measure in fact measures what it claims to measure. Unfortunately, frequently used measures within psychology (e.g., Schimmack, 2019) and sport science (Fischman, 2015) might not measure what they claim to measure. Although, the present paper focused on reliability and not validity, high quality measurement in any scientific field needs to focus on both. However, high reliability is a prerequisite for validity: A measurement that is not reliable cannot be valid. Finally, we would like to emphasize that our results do not intend to undermine the credibility, quality or replicability of prior studies that have employed golf putts, darts, or free throws. Instead, they should draw attention to the importance of reliable measurement in sport psychology with the aim of securing it in the future.

Conclusions

Sport performance outcome measures may substantially differ regarding their reliability and may have different reliabilities for different samples (and not necessarily in an intuitive way). Suppose three research teams each used a different one of our measures with their respective reliabilities to answer a research question (e.g., the effects of pressure or fatigue on perceptual-motor performance). All else equal, these teams would have substantially different likelihoods of a) finding an effect, given it exists, of b) replicating an effect found in a prior study, and c) being able to make meaningful comparisons between studies, variables, and theories.

When conducting studies, we hope that researchers in sport psychology will try to construct reliable measurements, that they will assess their measurement's reliability, and that they will interpret their results in the light of these reliabilities. Reliabilities need to be *high*, and moderate reliabilities may exacerbate methodological problems. For example, Westfall and Yarkoni (2016) report that the Type-1 error rate when assessing incremental validity via regression models was highest for moderate reliabilities (at least for certain sample sizes).

Regarding conclusions, we hope that researchers will be very careful when comparing findings to each other that may stem from measurements with different reliabilities. Likewise, we hope that researchers will consider the role of (different) reliabilities when assessing replications as being successful or not. If possible, we suggest that researchers pretest their performance outcome measure and try to determine an optimal number of trials that provides sufficient reliability, but that does not induce threats to validity (such as fatigue or learning effects) and is still economically feasible (see Appendix A for more information and guidelines on these issues). Whereas increasing reliability by adding items only works to a certain extent for common psychological measurement procedures such as questionnaires, for performance outcome measures, such as discussed in this paper, it seems to be more promising (for more information see Appendix A). Furthermore, as mentioned above, in order to obtain the same power to detect an effect of the same (true) size, researchers need smaller samples when using more reliable measures than when using less reliable measures. Therefore, there is a trade-off regarding research economy[8]: On the one hand, adding items or trials to a measurement in order to make it more reliable will make the measurement less economical by increasing its duration. On the other hand, this approach will make the measurement more reliable and thus more economical because smaller sample sizes are needed. It seems to be an interesting endeavor for future research to try and formalize this trade-off depending on its various costs and benefits.

In this endeavor, experimenters should attempt to use individual performance outcome measures that allow for sufficient variation in performance that is indicative of true performance variation and not random performance fluctuation and measurement error. To this end the following guiding questions might prove helpful (see also Table 3): a) what is my precise research question and how well do the variables in my research design measure the constructs in my research question; b) what is the skill level of my participants or how experienced are participants with the (or similar) tasks being measured; c) how difficult does the task have to be (e.g. putting distance in golf); and d) how many trials are sufficient to achieve adequate reliability, while not threatening validity (e.g. motivation, calibration, learning, fatigue, etc.).

References

Ajemian, R., D'Ausilio, A., Moorman, H., & Bizzi, E. (2010). Why professional athletes need a prolonged period of warm-up and other peculiarities of human motor learning. *Journal of Motor Behavior, 42,* 381–388. doi:10.1080/00222895.2010.528262

Amelang, M, & Zielinski, W. (2002). *Psychologische Diagnostik und Intervention*. Heidelberg: Springer.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., … Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour, 2,* 6–10. doi:10.1038/s41562-017-0189-z

Barbero-Álvarez, J. C., Coutts, A., Granda, J., Barbero-Álvarez, V., & Castagna, C. (2010). The validity and reliability of a global positioning satellite system device to assess speed and repeated sprint ability (RSA) in athletes. *Journal of Science and Medicine in Sport, 13,* 232-235. doi:10.1016/j.jsams.2009.02.005

Baumeister, R. F., & Showers, C. J. (1986). A review of paradoxical performance effects: Choking under pressure in sports and mental tests. *European Journal of Social Psychology, 16,* 361-383. doi:10.1002/ejsp.2420160405

Bühner, M. (2011). *Einführung in die Test- und Fragebogenkonstruktion, 3. aktual. Auflage*. München: Pearson.

Button, K. S., Ioannidis, J. P. A., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S. J., & Munafò, M. R. (2013). Power failure: Why small sample sizes undermine the reliability of neuroscience. *Nature Reviews Neuroscience, 14,* 365-376. doi:10.1038/nrn3475

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., … Wu, H. (2018). Evaluating the replicability of social science experiments in *Nature* and *Science* between 2010 and 2015. *Nature Human Behaviour*, *2*, 637-644. doi:10.1038/s41562-018-0399-z

Charter, R. A. (1999) Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients. *Journal of Clinical and Experimental Neuropsychology, 21,* 559-566. doi:10.1076/jcen.21.4.559.889

Coutts, A. J., & Duffield, R. (2010). Validity and reliability of GPS units for measuring movement demands of team sports. *Journal of Science and Medicine in Sport, 13,* 33–135. doi:10.1016/j.jsams.2008.09.015

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. New York: Routledge.

Cumming, G. (2014). The new statistics why and how. *Psychological Science*, 25, 7-29. doi:10.1177/0956797613504966

Danner, D. (2015). *Reliabilität: Die Genauigkeit einer Messung. GESIS Survey Guidelines*. Mannheim: GESIS – Leibniz Institut für Sozialwissenschaften. doi:10.15465/sdm-sg_011

Diedenhofen, B. & Musch, J. (2015). cocor: A comprehensive solution for the statistical comparison of correlations. *PLoS ONE, 10*. e0121945. doi:10.1371/journal.pone.0121945

Diedenhofen, B., & Musch, J. (2016). cocron: A web interface and R package for the statistical comparison of Cronbach's alpha coefficients. *International Journal of Internet Science, 11*, 51–60.

Eklund, R. C., Tenenbaum, G., & Kamata, A. (2012). *Measurement in sport and exercise psychology*. Champaign, IL, Human Kinetics.

Fischman, M. G. (2015). On the continuing problem of inappropriate learning measures: Comment on Wulf et al. (2014) and Wulf et al. (2015). *Human Movement Science, 42*, 225-231. doi:10.1016/j.humov.2015.05.011

Fédération Internationale de Basketball (FIBA) (2018a). *Official Basketball rules 2018.* Fédération Internationale de Basketball: Mies, Switzerland. https://www.fiba.basketball/documents/official-basketball-rules.pdf

Fédération Internationale de Basketball (FIBA) (2018b). *Official Basketball rules: Basketball equipment.* Fédération Internationale de Basketball: Mies, Switzerland. http://www.fiba.basketball/OBR-2018-Basketball-Equipment-Yellow-Version-2.pdf

Fraley, R. C., & Vazire, S. (2014). The N-Pact Factor: Evaluating the quality of empirical journals with respect to sample size and statistical power. *PLoS One, 9*, e109019. doi: 10.1371/journal.pone.0109019

Friese, M., & Frankenbach, J. (in press). P-hacking and publication bias interact to distort meta-analytic effect size estimates. *Psychological Methods*.

Gaito, J. (1980). Measurement scales and statistics: Resurgence of an old misconception. *Psychological Bulletin, 87*, 564-567. doi:10.1037/0033-2909.87.3.564

García-Pérez, M. A. (2017). Thou shalt not bear false witness against null hypothesis significance testing. *Educational and Psychological Measurement, 77*, 631-662. doi:10.1177/0013164416668232

Gelman, A. (2015, April 28). What's the most important thing in statistics that's not in the textbooks? [Web log post]. Retrieved from https://statmodeling.stat.columbia.edu/2015/04/28/whats-important-thing-statistics-thats-not-textbooks/

Gelman, A., & Stern, H. (2006). The difference between “significant” and “not significant” is not itself statistically significant. *American Statistician, 60*, 328-331.

Hopkins, W. G. (2017). Spreadsheets for analysis of validity and reliability. *Sportscience*, *21*.

Hopkins, W. G. (2000). Measures of reliability in sports medicine and science. *Sports Medicine*, *30*, 1-15. doi:10.2165/00007256-200030010-00001

Jennings, D., Cormack, S., Coutts, A. J., Boyd, L., & Aughey, R. J. (2010). The validity and reliability of GPS units for measuring distance in team sport specific running patterns. *International Journal of Sports Physiology and Performance*, *5*, 328-341. doi:10.1123/ijspp.5.3.328

Johnston, R. J., Watsford, M. L., Pine, M. J., Spurrs, R. W., Murphy, A. J., & Pruyn, E. C. (2012). The validity and reliability of 5-Hz global positioning system units to measure team sport movement demands. *The Journal of Strength & Conditioning Research*, *26*, 758-765. doi:10.1519/JSC.0b013e318225f161

Lakens, D., & Evers, E. R. K. (2014). Sailing from the seas of chaos into the corridor of stability: Practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science, 9*, 278-292. doi:10.1177/1745691614528520

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science, 355*, 584-585. doi:10.1126/science.aal3618

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading: Addison-Weasley.

Love, J., Selker, R., Marsman, M., Jamil, T., Dropmann, D., Verhagen, A. J., Ly, A., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Wild, A., Knight, P., Rouder, J. N., Morey, R.

D., Wagenmakers, E.-J. (2019). JASP: Graphical statistical software for common statistical designs. *Journal of Statistical Software, 88*. doi:10.18637/jss.v088.i02

Lunney, G. H. (1970). Using analysis of variance with a dichotomous dependent variable: An empirical study. *Journal of Educational Measurement, 7*, 263-269.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*, 412–433. doi:10.1037/met0000144

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. L. (2019). Abandon statistical significance. *The American Statistician, 73*, 235-245. doi:10.1080/00031305.2018.1527253

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., Simonsohn, U., Wagenmakers, E.-J., Ware, J. J., Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*, 0021. doi:10.1038/s41562-016-0021

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … Yarkoni, T. (2015). Promoting an open research culture. Author guidelines for journals could help to promote transparency, openness, and reproducibility. *Science, 348*, 1422-1425. doi:10.1126/science.aab2374

Nelson, L. D., Simmons, J., & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology, 69*, 511-534. doi:10.1146/annurevpsych-122216-011836

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*, 657–660.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349,* aac4716. doi:10.1126/science.aac4716

Pashler, H., & Wagenmakers, E.-J. (2012). Editors' introduction to the special section on replicability in psychological science: A crisis of confidence? *Perspectives on Psychological Science, 7*, 528-530. doi:10.1177/1745691612465253

Petersen, C., Pyne, D., Portus, M., & Dawson, B. (2009). Validity and reliability of GPS units to monitor cricket-specific movement patterns. *International Journal of Sports Physiology and Performance, 4,* 381–393. doi:10.1123/ijspp.4.3.381

Raykov, T. (1997). Estimation of composite reliability for congeneric measures. *Applied Psychological Measurement, 21*, 173-184.

Raykov, T., & Marcoulides, G. A. (2019). Thanks coefficient alpha, we still need you! *Educational and Psychological Measurement, 79*, 200–210. doi:10.1177/0013164417725127

Robinson, G., & O'Donoghue, P. (2007). A weighted kappa statistic for reliability testing in performance analysis of sport. *International Journal of Performance Analysis in Sport, 7,* 12-19. doi:10.1080/24748668.2007.11868383

Savalei, V., & Reise, S. P. (2019). Don't forget the model in your model-based reliability coefficients: A reply to McNeish (2018). *Collabra: Psychology, 5*, 36. doi:10.1525/collabra.247

Schimmack, U. (2010). What multi-method data tell us about construct validity. *European Journal of Personality*, *24,* 241-257. doi:10.1002/per.771

Schimmack, U. (2019, under review). The validation crisis in psychology. *Meta-Psychology*. (https://replicationindex.files.wordpress.com/2019/04/validation.crisis.v3.pdf)

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? Journal of Research in Personality, 47, 609e612. http://dx.doi.org/10.1016/ j.jrp.2013.05.009

Schweizer, G., & Furley, P. (2016). Reproducible research in sport and exercise psychology: The role of sample sizes. *Psychology of Sport and Exercise, 23*, 114-122. doi:10.1016/j.psychsport.2015.11.005

Savalei, V., & Dunn, E. (2015). Is the call to abandon *p*-values the red herring of the replicability crisis? *Frontiers in Psychology, 6*, 1-4. doi:10.3389/fpsyg.2015.00245

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality, 47*, 609e612. doi:10.1016/j.jrp.2013.05.009

Selkirk, K. (1976). Re-designing the dartboard. *The Mathematical Gazette, 60*, 171-178.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology : Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science, 22*, 1359-1366. doi:10.1177/0956797611417632

Simonsohn, U. (2015). Small telescopes: Detectability and the evaluation of replication results. *Psychological Science, 26*, 559-569. doi:10.1177/0956797614567341

Soto, C. J. (2019). How replicable are links between personality traits and consequential life outcomes? The Life Outcomes Of Personality Replication Project. *Psychological Science, 30*, 711-727. doi:10.1177/0956797619831612

Spearman, C. (1904). The proof and measurement of association between two things. The *American Journal of Psychology, 15*, 72-101. doi:10.2307/1412159

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology, 3*, 271-295. doi:10.1111/j.2044-8295.1910.tb00206.x

Stanley, D. J., & Spence, J. R. (2014). Expectations for replications: Are yours realistic? *Perspectives on Psychological Science, 9*, 305-318. doi:10.1177/1745691614528518

Steyer, R., & Eid, M. (1993). *Messen und Testen*. Springer: Heidelberg.

Tibshirani, R. J., Price, A., & Taylor, J. (2011). A statistician plays darts. *Journal of the Royal Statistical Society, 174*, 213-226.

Vaughn, B. K., Lee, H.-Y., & Kamata, A. (2012). Reliability. In R. C. Eklund, G. Tenenbaum, & A. Kamata (Eds.), *Measurement in sport and exercise psychology*. Champaign, IL: Human Kinetics.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, A. J., Love, J., Selker, R., Gronau, Q. F., Šmíra, M., Epskamp, S., Matzke, D., Rouder, J. N., Morey, R. D. (2018). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*, 35-57.

Walsh, V. (2014). Is sport the brain’s biggest challenge? *Current Biology, 24*, R859-R860. doi:10.1016/j.cub.2014.08.003

Westfall, J., & Yarkoni, T. (2016). Statistically controlling for confounding constructs is harder than you think. *PLoS ONE, 11*, e0152719. doi:10.1371/journal.pone.0152719

Wunderlich, F., Heuer, H., Furley, P., & Memmert, D. (2019, iFirst). A serial-position curve in high-performance darts: The effect of visuomotor calibration on throwing accuracy. *Psychological Research*, 1-8. doi:10.1007/s00426-019-01205-2

Zhu, W. (2013). Reliability: what type, please! *Journal of Sport and Health Science, 2,* 62-64. doi:10.1016/j.jshs.2012.11.001

Footnote

[1]For example, the JASP software package includes Bayesian parameter estimation and Bayes factor hypothesis testing via a graphical user interface (see Love et al., 2019). Furthermore, free and open-source packages for specific procedures in R make it more feasible for researchers to use these procedures, to name only two examples.

[2]Our point is not that measurement is generally absent in the methodological literature in psychology, quite the contrary. However, in the context of the current debate on methodological practices (as described by Nelson et al., 2018) only few papers focus on measurement (e.g., Loken & Gelman, 2017).

[3]There is an ongoing debate in psychology whether researchers should abandon Null Hypothesis Significance Testing (NHST), and, if they do, which methods they should use instead. Some authors suggest abandoning not only NHST, but the frequentist perspective altogether by employing Bayesian methods (e.g., Wagenmakers et al., 2018). Some suggest abandoning statistical significance as a threshold, but to retain p-values and treat them as one (albeit continuous) piece of information among others (McShane, Gal, Gelman, Robert, & Tackett, 2019). Some authors retain a frequentist perspective, but suggest replacing NHST by focusing on confidence intervals (e.g., Cumming, 2012, 2014). Some authors defend the utility of NHST (e.g., García-Pérez, 2017; Savalei & Dunn, 2015), and some have even suggested improving NHST by redefining statistical significance (Benjamin et al., 2018). We would like to note that in this article, we do not take any position regarding these questions. Instead, we emphasize that reliable measurement plays a key role for all of the methods discussed above.

[4]As these examples refer to choices researchers can make, the underlying construct was initially called "researcher degrees of freedom" (Simmons et al., 2011, p. 1359). Later, Simmons and colleagues adopted the term p-hacking (Nelson et al., 2018, p. 513).

[5]Generally, one needs to be careful when comparing significant and non-significant effects to each other: When one effect is significant and the other one is not, this does not mean that the difference between them is significant. This holds both for differences between groups (Gelman & Stern, 2006) and for differences between correlations (Diedenhofen & Musch, 2015).

[6]First, we looked through the latest issues of sport psychological journals in order to identify generally used performance outcome variables. This search led us to golf putts, darts and free throws. Then, we conducted a literature search in google scholar using the key words "golf putts"; "darts"; "free throws". We combined these key words with different search terms, such as "psychology", "performance"; and "experiment". Our criterion for inclusion was that the paper reported a study in which the respective outcome measure had been used (as compared to, for example, a mathematical model of darts performance).

[7]At least in the ones we could access, we could not check the full text of nine articles due to difficulties acquiring the full text.

[8]We thank an anonymous reviewer for this idea.

Table 1

Reliability estimates for different measures

	N	split-half (odd-even)	split-half (half-half)	split-half (odd-even)	split-half (half-half)	Cronbach's alpha
Golf (distance)	100	.555 (.402; .678)	.552 (.399; .675)	.714 (.602; .798)	.711 (.598; .796)	.670 (.567; .758)
Golf (successful putts)	100	.443 (.27; .588)	.360 (.176; .52)	.614 (.475; .723)	.529 (.371; .657)	.598 (.472; .705)
Darts-I (sample 2a)	100	.797 (.712; .859)	.700 (.584; .788)	.887 (.836; .923)	.824 (.749; .878)	.855 (.81; .894)
Darts-I (sample 2b)	100	.742 (.639; .819)	.736 (.631; .815)	.852 (.787; .898)	.848 (.782; .895)	.834 (.782; .878)
Darts total	200	.799 (.743; .844)	.732 (.66; .79)	.888 (.855; .914)	.845 (.8; .881)	.863 (.833; .889)
Basketball-I (sample 3a)	100	.337 (.151; .5)	.300 (.11; .469)	.504 (.341; .637)	.462 (.292; .604)	.502 (.346; .634)
Basketball-II (sample 3b)	92	.449 (.269; .599)	.392 (.204; .552)	.62 (.475; .732)	.563 (.405; .688)	.547 (.399; .672)
Basketball total	192	.703 (.623; .768)	.677 (.592; .747)	.826 (.775; .866)	.807 (.751; .851)	.812 (.771; .849)

Note. The first two columns for the split-half reliabilities report the simple correlation between the two test halfs. The third and the fourth columns report the respective coefficients for all 14 items, computed using the Spearman-Brown formula. Cronbach's alpha also refers to all 14 items.

Numbers in brackets are 95% confidence intervals (rounded to the third decimal place).

For estimating the 95% CIs for Cronbach's alpha we used the cocron package in R (via its web interface) (Diedenhofen & Musch, 2016).

Table 2

Darts performance for different subgroups

	N	mean performance	split-half (odd-even)
Darts-I (students)	100	99 (45)	.887 (.836; .923)
Darts-II (sports students)	100	75 (28)	.852 (.787; .898)
Darts total	200	87 (39)	.888 (.855; .914)
Women	94	111 (41)	.871 (.812; .913)
Men	106	66 (22)	.742 (.642; .817)
Darts players	94	73 (30)	.85 (.78; .90)
Non-darts players	106	99 (43)	.89 (.84; .92)

Note. The third column reports the mean distance from the bulls eye averaged over all 14 throws (in mm). Numbers in brackets are standard deviations.

The fourth column reports the split-half reliability coefficient for all 14 items, computed using the Spearman-Brown formula. Numbers in brackets are confidence intervals (rounded to the third decimal place).

Table 3

Guidelines for creating reliable performance outcome measures

Initial considerations	• Select a measurement procedure based on theoretical grounds and research goals.
Pretest	• Pretest the measurement's reliability for a specific number of trials. • Use a sample that is drawn from the same population that you intend to draw your main study's sample from. • When considering sample size, think about precision, not statistical significance (see, for example, Appendix A; Charter, 1999; Schönbrodt and Perugini, 2013). • Consider different estimates of reliability: Which one is best suited for your measurement based on practical and statistical assumptions (see Appendix A for a brief overview)? • Estimate reliability.
Main study	• Based on the above estimate of reliability, calculate the number of trials that you need in order to achieve a certain level of reliability. • Conduct your main study, and estimate reliability again.
Future studies	• Take into consideration that reliability is dependent (among other factors) on samples. • Different measurements using the same instrument may therefore lead to different estimations of reliability, for example when samples differ regarding true score variation.

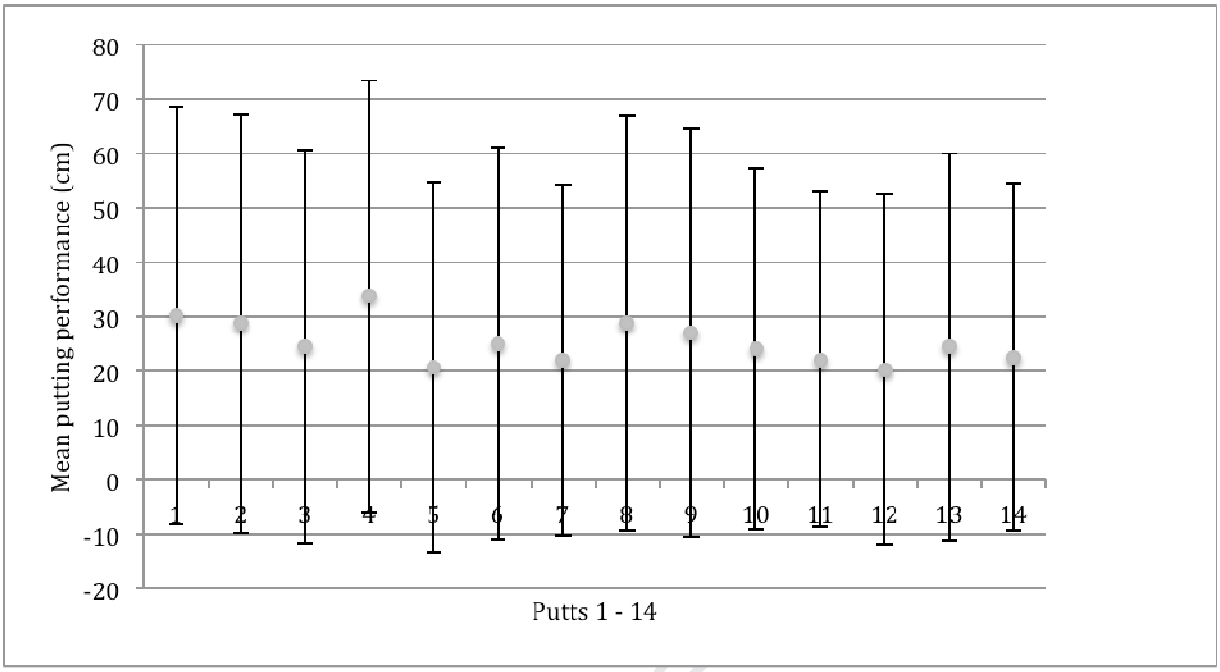


Figure 1. Mean putting performance for all 14 putts (distance from the hole in centimeters) in Study 1. Error bars are standard deviations.

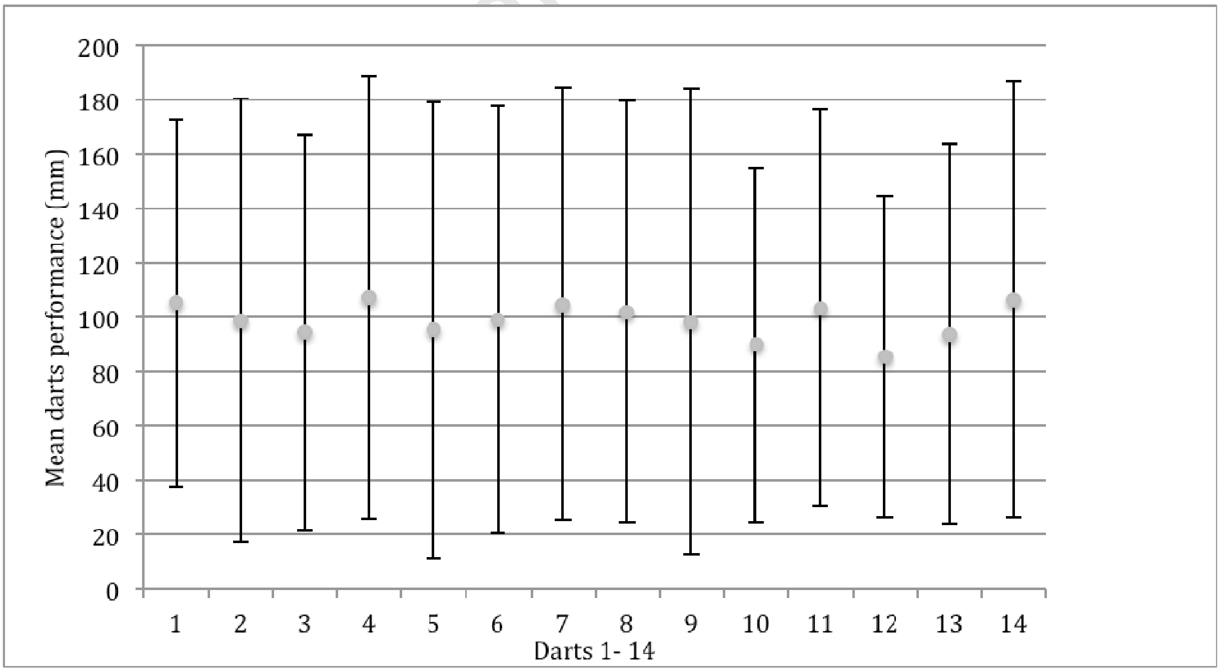


Figure 2. Mean darts performance (distance from the bull in millimeters) in Study 2, sample 2a (students). Error bars are standard deviations.

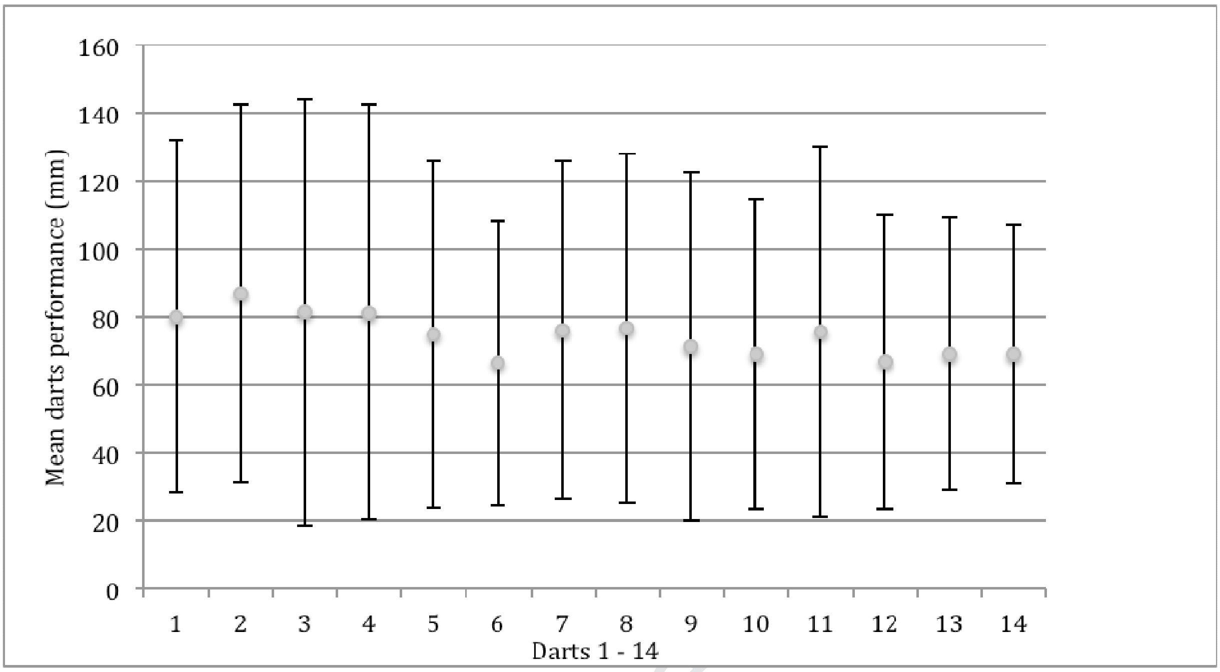


Figure 3. Mean darts performance (distance from the bull in millimeters) in Study 2, sample 2b (sports students). Error bars are standard deviations.

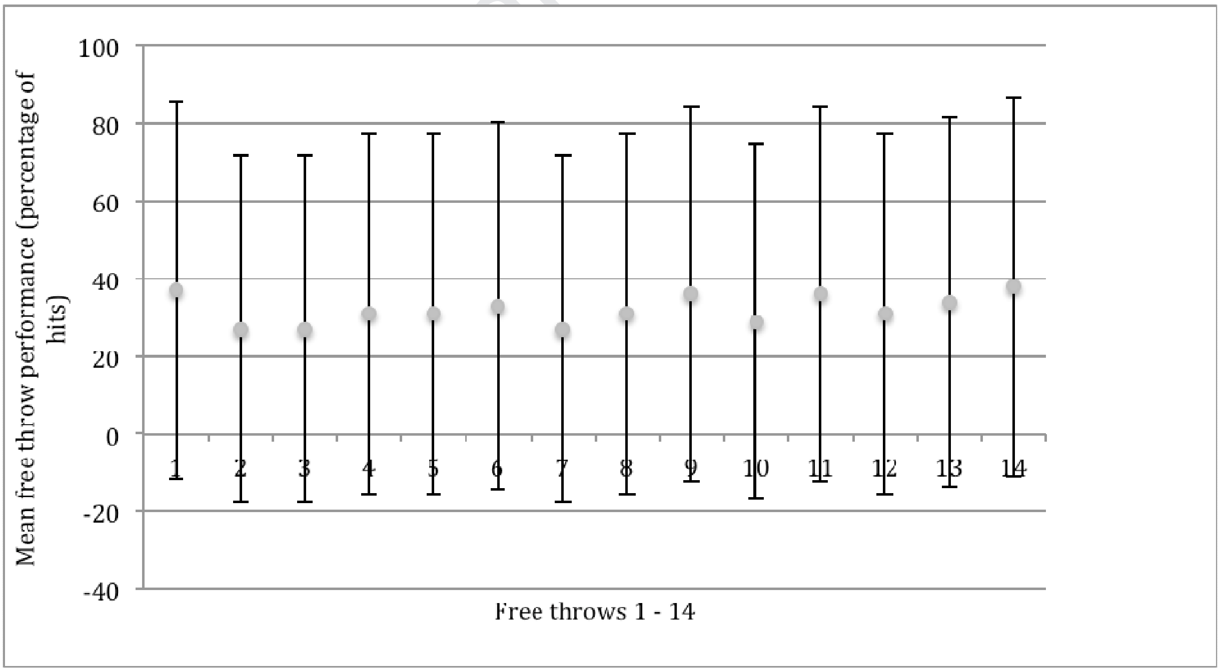


Figure 4. Mean free throw performance (percentage of hits) in Study 3, sample 3a (inexperienced). Error bars are standard deviations.

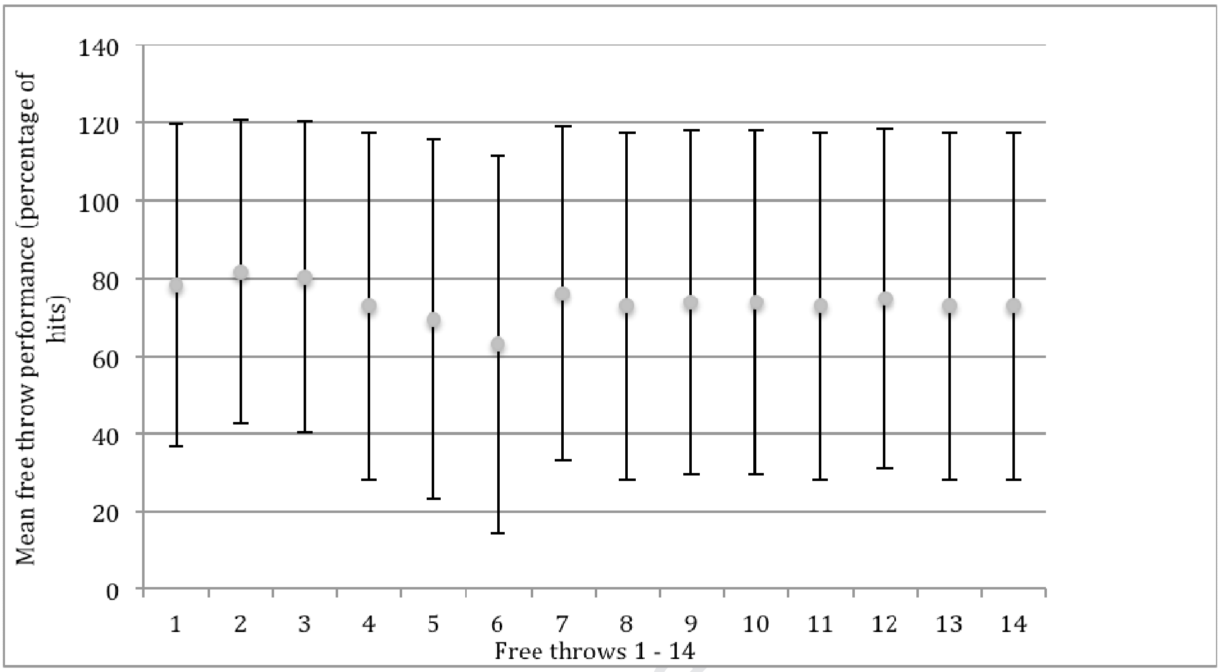


Figure 5. Mean free throw performance (percentage of hits) in Study 3, sample 3b (experienced). Error bars are standard deviations.

Appendix A: Statistical considerations

Justification of sample size planning

In the context of CTT, reliability coefficients are estimated using correlations. Therefore, their size is independent of the respective sample size (i.e., unlike p-values, correlations do not change as a mere function of sample size). Thus, the sample size in a reliability analysis does not affect the estimated reliability per se, but instead the precision of the estimate. The precision of the estimate can be captured by the confidence interval around the estimate. As correlations stabilize only at rather high numbers of participants (e.g., Schönbrodt and Perugini [2013] suggest 250 participants as a reasonable sample size for interpreting single correlation coefficients), large sample sizes have been suggested for reliability analyses (e.g., Charter, 1999). For example, Charter suggests at least 400 participants in order to conduct reliability analyses. This is particularly important for studies that aim to estimate a reliability coefficient which can be interpreted largely independent of the respective sample, for example when reporting *the* reliability of a certain questionnaire, as is often done in questionnaire construction. For this kind of analysis, samples are required that are representative of an underlying population. Still, reliability analyses are possible with fewer participants. They simply lead to somewhat less precise estimates. For example, with 250 participants, the 95% confidence interval for a correlation $r = .85$ ranges from .812 to .881, whereas for 200 participants it ranges from .806 to .884. Even for 100 participants, the respective CI still ranges from .785 to .897. In light of these considerations, and given that we had to conduct single-participant sessions, we considered 200 participants per variable to be acceptable for our approach.

The role of sample homogeneity or heterogeneity for reliability

Whereas sample size does not affect the reliability estimate itself, (i.e., larger sample sizes do not lead to larger or smaller reliability estimates), true score variation in the respective sample does (e.g., Bühner, 2011; Steyer & Eid, 1993). That means that more heterogeneous samples may lead to higher reliability estimates than more homogenous ones. This follows directly from Equation 2: When variation in the true score increases more strongly than variation due to measurement error, reliability increases. This observation has some consequences: First, one and the same measurement instrument may have different reliabilities for different kinds of samples (or populations, respectively, from which these samples are drawn). For example, an instrument assessing political attitudes may be more reliable in a moderate sample (where people have different political attitudes) than in an extremist sample (where people have rather similar attitudes) (see Danner, 2015, for this example). This transfers to applications in sport psychology: When used in a high-performance sample (where there is low variation in athletes' performance), a measurement instrument might have lower reliability than when used in a sample of more moderate performance (where there is substantial variation in athletes' performance). Second, a certain reliability that was estimated based on a representative sample of the population may not apply to a more homogenous subsample of that same population.

When one argues that larger samples are more likely to be heterogeneous, then it follows that increasing reliability may be an indirect consequence of increasing sample sizes. However, this only holds when heterogeneity increases (more technically, as described above, when due to increased heterogeneity true score variation increases more strongly than error variation).

Restrictions and assumptions underlying reliability in CTT

Importantly, Classical Test Theory can only be applied to measurements that produce interval-scale (i.e., continuous) data (Bühner, 2011; Steyer & Eid, 1993; for a disagreement with this position see Gaito, 1980). The reason for this restriction is that reliability is defined as a proportion of variances (see Equation 2), and variances can only be calculated for continuous data. Furthermore (following Equation 1), measurement error is defined as the difference between the observed value and the true score ($Y_{error} = Y_{observed} - T_{true}$), which again is only possible for continuous measurements. This restriction can be misunderstood as meaning that reliability coefficients can only be *calculated* for continuous data, however its consequences are more far reaching: Indeed, reliability according to CTT is only *defined* for continuous measurements. It follows that when one wants to estimate reliability coefficients for a certain measurement procedure, this procedure must yield continuous measurement outcomes.

The core of CTT are three definitions, sometimes also called axioms. The first definition states that every observed value ($Y_{observed}$) consists of a true value (T_{true}) and random measurement error (Y_{error}). That is, $Y_{observed} = T_{true} + Y_{error}$. The second definition states that measurement error has an expectancy value of 0 and a finite variance. The third one states that a) measurement error of a test t_1 is independent from this test's true values, b) measurement error of one test t_1 is independent from measurement error of another test t_2, and c) that measurement error of a test t_1 is independent from the true values of another test t_2.

Relatedly, CTT contains five models, that describe assumptions that are necessary for estimating reliability (Bühner, 2011; Steyer & Eid, 1993). These five models are a) the model of parallel measurement, b) the model of essentially parallel measurement, c) the model of tau-equivalent measurement, d) the model of essentially tau-equivalent measurement, and e) the model of tau-congeneric measurement. These models contain assumptions regarding the true scores and (the intercorrelations of) measurement error.

When accepting the axioms of CTT, and when the above described modelling assumptions hold, it can be shown that the different reliability coefficients are estimates of the measurement instrument's reliability. That is, their respective formulas can be converted (only if one assumes that the axioms hold) into the definition of reliability according to Equation 2 (e.g., see Steyer & Eid, 1993). However, when the modelling assumptions do not hold, estimates can either under- or overestimate a measurement's reliability (Savalei & Reise, 2019; Steyer & Eid, 1993). The exact nature of the deviation depends on the exact nature of the violation of the assumptions. In cases of extreme violations of the assumptions, reliability estimates can become entirely meaningless and unrelated to a measurement's true reliability (Steyer & Eid, 1993). Different reliability coefficients require different modelling assumptions.

Some considerations on different reliability coefficients

Whereas a measurement only has one reliability (defined by Equation 2), this reliability can be assessed or estimated via different reliability coefficients. Reliability itself (and not only the estimate) is sample dependant. That means that one and the same measurement may have different reliabilities for different samples. When reliability was estimated using a representative sample, one may assume that the same reliability holds for samples that are either a) also representative or b) drawn randomly from the same population and sufficiently large.

So, why are there several reliability coefficients? First, as mentioned above, different reliability coefficients require different modelling assumptions, and only when these are met can the respective coefficients be used in order to estimate reliability.

Second, there are some conceptual and practical considerations. Using *test-retest reliability* only makes sense when the construct to-be-measured is stable between the two measurement points and when the measurement is not (differentially) affected by memory effects or learning. When test-retest reliability is calculated although these conditions are not met, the resulting coefficient will underestimate an instrument's reliability. Using *parallel-test reliability* only makes sense when two absolutely parallel tests exist for measuring the same construct, in other words when two tests exist that measure the same construct with different items. When parallel-test reliability is calculated although these conditions are not met, the resulting coefficient will underestimate an instrument's reliability. In case that the above mentioned conditions are not met, *split-half reliabilities* can be estimated. These tend to be higher the more homogenous the measurement instrument is. Calculating split-half reliabilities requires to divide all items of the measurement procedure into two equivalent halves. Subsequently, the correlation between the two halves is calculated. This only makes sense when the two halves are indeed equivalent. In the light of these considerations, we considered split-half reliabilities, and particularly odd-even reliabilities to be the most appropriate estimators for our measurements.

The role of the number of items for reliability

Reliability itself, and not only its estimate, increases when the number of items that measure the same underlying construct increases. This is a property of reliability according to CTT (Bühner, 2011; Steyer & Eid, 1993). Intuitively, it can be understood when one considers that according to CTT, measurement errors cancel each other out, and the more measurements one has, the more they cancel each other out. Mathematically, the relationship between the length of a measurement procedure and reliability is described by the Spearman-Brown-Formula (Bühner, 2011; Steyer & Eid, 1993).

Equation 3: Spearman-Brown-Formula. r_{tt} is the reliability of the current form of the measurement procedure; k is the factor by which the length of the current measurement procedure changes (e.g., k = 0.5 means half the number of items, k = 2 means twice the number of items, k = 3 means three times the number of items, and so on); r_{ttcorr} is the reliability of the changed form of the measurement procedure.

$$r_{ttcorr} = \frac{k \cdot r_{tt}}{1 - (k-1) \cdot r_{tt}}$$

The Spearman-Brown-Formula can be used to predict how reliability will change when the number of items of an existing measurement procedure with a known reliability changes. Thereby, it can be used to predict how many items researchers need to add (or subtract) in order to achieve a certain reliability, once they already know the reliability of a measurement procedure with a given number of items. As one can deduce from the Spearman-Brown-Formula, increasing the number of items at first leads to relatively high increases in reliability, however, further gains in reliability need increasingly more items (Amelang & Zielinski, 2002).

Therefore, practically, researchers who want to obtain a reliable measure can increase the number of items. It seems even possible to predefine a certain reliability one wants to achieve (say, .90) and then to increase the number of items until this reliability is achieved. However, there are three potential problems with this approach (Amelang & Zielinski, 2002). First, the single items must all measure the same construct. This seems feasible for performance outcome measures as discussed in this paper. However, it may be problematic for other kinds of measurement procedures. For example, there may only be a limited number of items that are suitable for assessing a certain construct in a questionnaire (e.g., there may only be a limited number of items for assessing anxiety). Second, the relationship between reliability and economy is an inverse one: When researchers increase reliability by increasing the number of items, they also increase the time their measurement procedure takes. Both in applied and in research contexts, time is usually limited (and important to consider; e.g. for motivational reasons). This is the main reason why constructing short forms of widely used questionnaires has become common. Furthermore, when a measurement procedure takes more time, adverse effects such as fatigue, boredom or concentration problems become more likely to influence the measurement outcome, thus limiting the measurement's validity. Still, time constraints do not seem to be a major limiting factor for increasing the number of trials in studies employing performance outcome measures, as the time needed per trial (e.g., per putt, dart throw or free throw) is very short. Third, there may be a seemingly paradoxical relationship between reliability and validity. On the one hand, reliability is a prerequisite for validity. That means, a measurement that is not reliable cannot be valid. On the other hand, increasing reliability by increasing the number of items can decrease validity. The reason for this seemingly paradoxical observation is that adding items in order to increase reliability is often likely to make a measure more homogenous. To the extent that the construct one intends to assess is rather heterogenous, then, the measurement becomes less valid. One way to try and circumvent this problem is by having a measurement procedure with several subscales. Each subscale is rather homogenous and constructed in order to be highly reliable, whereas the heterogeneity of the construct is captured by the multitude of different subscales (e.g., intelligence or personality tests). However, this approach will make the measurement procedure less economical again.

Taken together, for the above mentioned reasons increasing reliability by adding items only works to a certain extent for common psychological measurement procedures such as questionnaires. For performance outcome measures, such as discussed in this paper, it seems to be more promising. Therefore, if possible, we suggest that researchers pretest their performance outcome measure and try to determine an optimal number of trials: A number, that provides sufficient reliability, but that does not induce threats to validity (such as fatigue or learning effects) and that is still economically feasible.

As mentioned above, it is possible to use the Spearman-Brown-Formula in order to estimate, based on a given reliability (r_{tt}) for x items, how many additional items would be necessary in order to achieve a prespecified reliability ($r_{tt\ corr}$). In order to do so, Equation 3 needs to be solved for k, which yields Equation 4.

Equation 4:

$$k = \frac{r_{ttcorr} \cdot (1 - r_{tt})}{r_{tt} \cdot (1 - r_{ttcorr})}$$

Importantly, k is not the number of items but the factor, with which the original number of items needs to be multiplied in order to achieve the prespecified reliability. That means, when x is the original number of items, k*x is the new number of items.

Initial considerations	• Select a measurement procedure based on theoretical grounds and research goals.
Pretest	• Pretest the measurement's reliability for a specific number of trials. • Use a sample that is drawn from the same population that you intend to draw your main study's sample from. • When considering sample size, think about precision, not statistical significance (see, for example, Appendix A; Charter, 1999; Schönbrodt and Perugini, 2013). • Consider different estimates of reliability: Which one is best suited for your measurement based on practical and statistical assumptions (see Appendix A for a brief overview)? • Estimate reliability.
Main study	• Based on the above estimate of reliability, calculate the number of trials that you need in order to achieve a certain level of reliability. • Conduct your main study, and estimate reliability again.
Future studies	• Take into consideration that reliability is dependent (among other factors) on samples. • Different measurements using the same instrument may therefore lead to different estimations of reliability, for example when samples differ regarding true score variation.

- Reliable measurement plays an important yet underrated role for research quality
- Performance outcome measures used in sport psychology differ regarding reliability
- Reliability of performance outcome measures depends on sample characteristics
- Reliability of performance outcome measures depends on item number
-

Author statement

Geoffrey Schweizer: Conceptualization; Formal analysis; Supervision; Writing - Original Draft; Writing - Review & Editing

Philip Furley: Conceptualization; Writing - Original Draft; Writing - Review & Editing

Nicolas Rost: Formal analysis ; Investigation; Writing - Review & Editing

Kai-Eric Barth: Formal analysis; Investigation; Writing - Review & Editing

Declarations of interest: none.