TEST RETEST STABILITY OF THE TASK AND EGO ORIENTATION QUESTIONNAIRE

Andrew M. Lane,

Alan M. Nevill, Neal Bowes,

University of Wolverhampton, United Kingdom

and

Kenneth R. Fox

University of Bristol, United Kingdom

Revision submitted: September 22nd 2004

Running Head: Measures of reproducibility

TEST RETEST STABILITY OF THE TASK AND EGO ORIENTATION QUESTIONNAIRE

Revision submitted: September 22nd 2004

Running Head: Measures of reproducibility

Abstract

Establishing stability, defined as observing minimal measurement error in a test-retest assessment, is vital to validating psychometric tools. Correlational methods such as Pearson, intraclass and kappa are a test of association or consistency, whereas stability or reproducibility (regarded here as synonymous) assesses the agreement between test-retest scores. Indices of reproducibility using the Task and Ego Orientation in Sport Questionnaire (TEOSQ; Duda & Nicholls, 1992) were investigated using correlational (Pearson, intraclass and kappa) methods, repeated measures multivariate analysis of variance, and by calculating the proportion of agreement within a referent value of ±1 as suggested by Nevill, Lane, Kilgour, Bowes, and Whyte (2001). Two hundred and thirteen soccer players completed the TEOSQ on two occasions, one week apart. Correlation analyses indicated a stronger test-retest correlation for the Ego subscale than the Task subscale. MANOVA indicated stability for Ego items but with significant increases in the four Task items. Proportion of test-retest agreement scores indicated that all Ego items reported relatively poor stability statistics with test-retest scores within a range of ± 1 ranging from 82.7-86.9%. By contrast, all Task items show test-retest difference scores ranging from 92.5-99%, although further analysis indicated that four Task subscale items increased significantly. Findings illustrate that correlational methods (Pearson, intraclass, and kappa) are influenced by the range in scores, and calculating the proportion of agreement of test-retest differences with a referent value of ± 1 could provide additional insight into the stability of the questionnaire. It is suggested that the item-by-item proportion of agreement method proposed by Nevill et al. (2001) should be used to supplement existing methods and could be especially helpful in identifying rogue items in the initial stages of psychometric questionnaire validation.

Key words: Goal orientation, measurement, test-retest reliability, and validity

Test-retest Stability of the Task And Ego Orientation Questionnaire

The inherent link between theory testing and construct validation suggests that researchers are indebted to investigate the validity and reliability of measures (Marsh, 1997; Schutz, 1994). Recent research has argued that procedures that are more stringent should be used to assess validity. Biddle, Markland, Gilbourne, Chatzisarantis, and Sparkes (2001) provided a review of controversial or problematic themes of research methods in sport and exercise psychology. Biddle et al.'s (2001) review highlights substantial developments in methods to assess validity, such as using structural equation modeling (Bentler, 1995; Schutz & Gessarolli, 1993). Schutz (1998) echoed this view, arguing that future research to assess stability and reliability of measures could also use structural equation modeling techniques.

Establishing stability is vital to validating psychometric tools (Anastasi & Urbina, 1997; Kline, 1993). Stability refers to the concept that constructs retain a degree of resistance to change over time. An aspect of stability is the extent to which test-retest scores are reproducible, regardless of environment conditions. Without reproducibility, the researcher cannot emphasize the validity of dispositional measures. Reliability is defined as the ratio of true variance to error variance (Cohen, 1960), and is typically assessed using correlation. A number of different techniques could be used to assess the reproducibility/stability of test-retest scores.

It is important that researchers should be aware of the limitation of the methods they use. Methods such as the Pearson Product Moment correlation, and more recently intra-class correlation and kappa have been used to assess test-retest stability, and it is common for researchers to treat reliability and stability or reproducibility as synonymous. Criterion values for showing acceptable test-retest stability using correlation suggest that the coefficient should be greater than $r = .80$ (Anastasi & Urbina, 1997; Kline, 1993). Recent

research has questioned using correlational methods as a measure of test-retest stability since correlation is a measure of relationship rather than agreement (Bland & Altman, 1986; Nevill, 1996, Nevill, Lane, Kilgour, Bowes, & Whyte, 2001; Wilson & Batterham, 1999). For example, a perfect correlation ($\underline{r}$ = 1.00) can be found with no agreement, when measures are unstable. Consider the following example to illustrate this point. Scores taken from three participants at one point in time of 1,2, and 3 will correlate perfectly with scores recorded at a second point in time of 3, 4, and 5. Thus, researchers should also assess the agreement between scores.

It is important to acknowledge that the intra-class correlation (ICC) will remove this systematic bias. Nevertheless, the intra-class correlation, like Pearson correlation coefficient, will still be highly dependent on the range of observations. Consider the following hypothetical data as examples. In example 1, seventy-two participants responded to a single item on a 5-point Likert scale on two separate occasions. As Table 1 indicates, participants used the full-range of responses (1-5), with 40 participants reporting the same value (along the diagonal from top left to bottom right) and 32 participants disagreeing by ± 1 only. The Pearson's and intra-class correlations between week 1 and week 2 scores were r = .88 and ICC = .93 respectively (both $p < .001$) with *kappa* = .44, $p < .001$, results that suggest acceptable reliability results (Anastasi & Urbina, 1997; Kline, 1993).

_____

Insert Table 1 about here

_____

_____

Insert Table 2 about here

_____

In example 2, participants were more homogeneous in their responses to the same item than participants in the first example. As Table 2 shows, participants recorded scores of only 2 or 3 on the same 5-point Likert scale, hence a far more restricted range of responses. As in example 1, Table 2 indicates the same number of participants (n = 40) responded identically to the item on the two occasions and the same number of participants (n = 32) differed by ± 1. However, the Pearson's and intra-class correlations between week 1 and week 2 scores were r = .11 and ICC = .20 (both $p > .05$) and *Kappa* = .11, p = .35, correlations suggesting poor stability.

In both examples, the test-retest differences are the same (40 participants having perfect agreement, 32 participants differing by a score of ±1 from week 1 to week 2) indicating the same degree of stability for responses to the item by both groups of participants. However, an examination of the correlation coefficients suggested dramatically different conclusions. This would have led researchers to supporting erroneously the stability of the item in example 1 and refuting the stability in the second example. Thus, it is argued that it is important to also use methods that are independent of the range of scores such as test-retest differences in addition to tests of association.

Recent research has seen developments in methods to investigate test-retest stability. Schutz (1998) and Marsh (1993) have suggested that researchers use structural equation techniques to assess test-retest stability. Using this approach, it is possible to investigate a) the stability of the traits which are free from errors of measurement, b) the stability of the measurement errors, and c) systematic variances associated with the items that underlie the traits. Thus, the advocates of structural equation modeling believe they can address the concerns of correlational methods suggested above.

However, one major limitation to using structural equation modeling is the difficulty in obtaining appropriate data. Structural equation modeling requires large sample sizes

(Bentler, 1995). It is suggested that there should be at least 10 participants per free parameter (Bentler, 1995; Tabachnick & Fidell, 1996), thus even with small questionnaires comprising, for example, of just 10 items, sample sizes will need to be 200+. This issue is complicated as testing for reliability using structural equation modeling requires at least three test completions, with ideally at least four test completions. It should not be surprising that research using structural equation modeling has tended to use data that are relatively easy to access. For example, Marsh (1993) used Student Evaluations over an eight-year period as raw data and thus could draw on a database of one million test completions. Similarly, Schutz (1995) used baseball performance data compiled from official records. Hence, these datasets do not require participants to volunteer data on a regular basis. It should be noted that if a researcher wishes to assess reliability and stability separately, at least three assessments are needed for any method of quantification. Researchers who wish to use only two assessments (and for practical reasons that is all we can expect in many cases) should not expect to obtain independent indicators of stability and reliability.

Attrition is a limitation when conducting test-retest research that involves individuals completing self-report measures. This can present a difficult hurdle for researchers planning to investigate stability of self-report measures, particularly in the initial stages of scale development. Thus, even though the approach to assessing stability proposed by Marsh (1993) might be the most robust, difficulties in recruiting sufficient sample sizes and retaining participants for subsequent completions might have contributed to few researchers using it. Altman and Bland (1987) critically evaluated the use of structural equation modeling to assess stability. They argued that using structural equation modeling approaches to assess stability lead to researchers using ‘unnecessarily complex statistical methods to solve simple problems’ (p. 225). They emphasized that this can lead to interpretation issues and can mislead researchers. Altman and Bland (1987) suggested that structural equation

modeling can lead to ‘attention being focused on technical statistical issues instead of on far more important considerations of the quality of the data and the practical interpretation of the analysis’ (p. 225).

There have been at least three other alternative approaches to using correlation (Schutz, 1998; Wilson & Batterham, 1999; Nevill et al., 2001). All methods require smaller sample sizes than structural equation modeling and require only two completions. The first by Schutz (1998) proposed using repeated measures multivariate analysis of variance to assess one component of stability, namely mean stability. MANOVA will account for differences in mean scores, but it is possible to have no significant differences between measurement occasions when the within-subject variation between test-retest differences is unacceptably large.

Second, Wilson and Batterham (1999) recommended an assessment based on the proportion of participants that record the same response on two separate occasions, referred to as the proportion of agreement (PA). The proportion of agreement does not require data to meet requirements of normal distribution. A key point from Wilson and Batterham’s (1999) work is that stability statistics should be calculated for each item of the questionnaire separately. Tests of agreement tend to be conducted following item analysis techniques such as factor analysis. Recent researchers have argued that assessment of each item should provide a more rigorous investigation of test-retest stability (Wilson and Batterham, 1999; Nevill et al., 2001). Calculating composite scores by summing items can mask individual item instability. Clearly, if each item is proposed to assess a theoretically stable construct, each item should demonstrate acceptable stability using a suitable criterion. If some items show poor test-retest stability scores, it would suggest that the underlying construct is unstable. Schutz (1994) argued that psychometric measures should be theory-driven, and thus item-analysis in terms of test-retest agreement should fulfill this aim.

However, a limitation of Wilson and Batterham's method is that they suggested that psychometric measures should show perfect agreement. Nevill et al. (2001) also recommended that researchers should calculate the test-retest differences for each item rather than calculate factor scores. Nevill et al. (2001) suggested that a dispositional construct utilizing a five-point scale should show that the majority of participants (90%) should record differences within a referent value ±1. They argued that some variation in test-retest difference scores was inevitable. They argued that it is important to acknowledge that completing a self-report scale requires participants to indicate their responses to a category, for instance report feeling 'not at all' (0), or 'very much so (4)'. Although there is some degree of continuity between responses, a likert scale yields only ordinal level data, i.e., not interval or ratio level data. Consequently, data should be treated using non-parametric methods.

A limitation of this approach is that the criterion for acceptability is arbitrary. The rationale for selecting a range of ±1 is based on the notion that the use of self-report to assess target constructs suggests that some variation is inevitable. It should be noted that self-report measures provide estimates of psychological constructs and cannot be relied on as objective and observable scores (see Nisbett & Ross, 1980; Nisbett & Wilson, 1977). For example, an individual might be genuinely unclear about what he/she is feeling. This assumption also forms part of the rationale for the use of correlation as it is proposed to be the *true variance* that reflects the reliability of measures, with *error variance* being attributed to random variation.

The aim of the study was to compare indices of stability using the Task and Ego in Sport Questionnaire (TEOSQ; Duda & Nicholls, 1992). The TEOSQ was chosen because achievement motivation has been one of the most frequently researched constructs in the sport psychology literature and recently has featured in vociferous debate (see Duda &

Whitehead, 1998; Harwood, Hardy, & Swain, 2000; Harwood & Hardy, 2001; Treasure, Duda, Hall, Roberts, Ames, & Mahr, 2001). Research investigating the validity of the TEOSQ has found support for the psychometric integrity of the hypothesized model other than problems related to items that are similarly worded on the Task subscale (see Chi & Duda, 1995). To date there have been very few tests of the stability in terms of test-retest differences. Of available research, Duda (1992) reported a correlation of r= .75 for the Ego subscale and r = .68 for the Task between scores taken over a three-week period. She also reported a correlation of r = .72 for the Ego factor and a correlation of r = .71 for the Task factor for scores taken over the course of a season. A limitation of these studies is that test-retest stability or reproducibility coefficients were not reported.

Given that the TEOSQ is proposed to assess a relatively stable construct, 90% or more of test-retest differences for each item should be within a reference value of ± 1. We also investigated stability using Pearson correlation, intra-class correlation, kappa and MANOVA results to provide clear comparisons with the proportion of acceptable agreement results within ± 1.

Methods

Participants

Participants were 213 soccer players from the US Midwest (116 males, 97 females: Age range 13-16 yr.). Of the 213, there were 15 participants aged 13 years, 57 aged 14 years, 57 aged 15 years, and 84 aged 16 years. Participants represented a broad range of ability levels, from recreational to national representative standard in the United States of America. Players varied in experience from players who were just beginning their competitive careers to players with a number of years of playing competitive soccer. They varied in terms of times they trained per week, number of competitive games per season, and numbers of years of soccer experience.

The sample size used in this study is commensurate with the sample size recommended (minimum N = 100) for assessing the reliability of psychometric questionnaires (Nevill et al., 2001).

Measure of Task and Ego Orientation in Sport Questionnaire

The TEOSQ (Duda & Nicholls, 1992) is an assessment of dispositional achievement goal orientations. The TEOSQ is a 13-item scale asking participants to respond to Task and Ego statements following from the stem "I feel successful in (soccer) when…". Each item is answered on a five-point scale. Task orientation is assessed by statements revolving around feelings of success derived from learning new skills, fun, trying hard, and practicing. Assessments of ego orientation are based upon responses concerning doing better than friends, scoring most points / goals, and being the best.

Procedure

On registration, parents/guardian were asked to complete an informed consent form, allowing their child (ren) to participate in the study. Parents / guardians were informed that participation was voluntarily. No child was withdrawn following signing this agreement.

The TEOSQ was administered under standardized conditions on two separate occasions (test-retest), separated by 5 days. The initial test was completed at the beginning of a 5-day soccer camp. Players completed a 15-hour course of soccer instruction. The course comprised instructions sessions involving individual ball skills, soccer specific skills (e.g., passing, shooting, heading, dribbling, turning), game related activities, with a 'World Cup' tournament concluding each day.

The camp comprised an achievement condition in which players have an opportunity to demonstrate physical competence. Task orientation conditions included practices that emphasized self-referenced improvement. As practices were not performed in isolation, competence could be judged in terms of an ego orientation goal disposition. Whenever an

individual describes his/her performance, it opens up the possibility of an evaluation with others. For example, if Player A and Player B both score eight goals in a shooting practice and encouraged to score more goals next time this will be a task oriented practice if players try to beat their own score. However, an ego involving condition exists whereby Player A might view success in relation to how many more goals he scores than Player B, regardless of his own improvement from previous attempts. The study did not control for players discussing their achievements and therefore it is likely that ego orientated individuals will seek out information about the performance of others. This suggests that practices such as improving the number of goals being scored are as much ego as task involving.

We argue that the more important indication of stability can be derived from the proportion of test-retest differences within (±1) as suggested by Nevill et al. (2001). As this is a relatively new technique, some explanation is warranted. Agreement between the test-retest measurements of the TEOSQ were quantified by calculating the differences between the responses recorded on two separate occasions for each item (Nevill et al., 2001). Clearly, these differences will be discrete (ranging from –4 to +4) and will follow a binomial rather than a normal distribution (see Nevill et al., 2000). Under such circumstances, Nevill et al. (2001) recommended adopting a non-parametric approach for assessing agreement of psychometric questionnaires, based on the methods originally proposed by Bland and Altman (1999). Briefly, Nevill et al. recommended reporting the proportion of differences within the criterion range (±1). The authors recommend that for each item to be stable, 90% or more of the participants should record differences within this criterion range (±1). Systematic bias from test to retest was assessed using the non-parametric median sign test.

Results

Insert Table 3 about here

The mean, standard deviation, minimum and maximum test-retest differences, the intra-class and product-moment correlations, repeated measures MANOVA results, the percentage of participants with differences within (±1), and the Median Sign Test results (the number of participants with differences above and below the median, 0) for each item of the TEOSQ are given in Table 3. Results show that Ego items have a wider range of test-retest differences as well as higher test-retest correlations. In contrast, most Task items have a relatively narrower range of test-retest scores and lower correlations. Further, comparing test-retest correlations having transformed correlations using Fisher $Z_r = ½ \log (1 + r) – ½ \log (1-r)$ for Ego items with those of the Task items identified that Ego items showed a significantly stronger relationship ($t = 1.87$, $p < .05$). Test-retest correlations coefficients for the composite Ego factor was r = .68, p < .01. Test-retest correlations for the equivalent Task factor was r = .61, p < .01. Repeated measures MANOVA results indicated a significant difference in TEOSQ items over time (Wilks' lambda $_{13,199} = .78$, p < .001, Partial $Eta^2 = .22$). Univariate results indicated that two Ego items significantly reduced (I am the only one who can do or play the skill and I can do better than my friends) and one increased (I'm the best). Four Task subscale items significantly increased (I learn a new skill and it makes me want to practice more; I learn something that is fun to do; I learn a new skill by trying hard; Something I learn makes me want to go and practice more).

Results demonstrate that <u>all</u> Ego items reported relatively poor stability statistics with test-retest scores within a range of ± 1 ranging from 82.7-86.9%. By contrast, all Task items show stable test-retest results with test-retest difference scores ranging from 92.5-99%. One

Task item ‘I do my very best’ showed a test-retest agreement score of 99%, which could be argued shows a meaningfully stronger degree of agreement that other Task items. Thus, calculating the proportion of agreement for each item demonstrates that Ego and Task items show different proportions of agreement. Using the < 90% proposed by Nevill et al. (2001) as a guide, results show that Ego items are relatively unstable, where Task items show stability.

An important feature of assessing stability is the detection of bias as it is possible for participants report an acceptable stability score but for all scores in one scale to change by ± 1. For example, if all participants report a test-retest increase of 1, this would show a systematic shift, but also would show acceptable stability coefficients in terms of a ± 1 criterion. A stable construct should show no systematic shift in scores. In the present study, results demonstrate that six items (see Table 3) had a systematic shift over the assessment period. Four Task items (2, 5, 10 and 12) increased significantly. In contrast, the systematic shift of the two Ego items (1 and 11) varied in direction. Item 1 declined significantly over the period of assessment, and item 11 significantly increased.

In summary, test-retest results show that Task items are relatively stable although it should be noted that four items showed a systematic positive shift in test-retest scores. In contrast, Ego items are unstable in terms of the significantly greater variation of test-retest differences.

Discussion

The present study investigated indexes used to assess test-retest stability. Recent research has suggested researchers use more rigorous methods to assess the validity and stability of their measures (Biddle et al., 2001; Schutz, 1998). Researchers are obliged to investigate the validity of their measures and that the concept of stability of dispositional constructs has been under researched. If the construct is proposed to be stable, stability is

imperative to the demonstration of validity. Test-retest questionnaire scores should be reproducible if the construct is stable. Validity is an ongoing process in which researchers challenge the notion that measures assess the construct under investigation. Researchers in sport and exercise psychology can never develop perfect measures, but they can get better ones (Schutz, 1994). In the present study, we focus on the TEOSQ.

The range of indices to show reliability and stability contained in Table 3 show the intraclass r and Pearson's r yield comparable results, but that the percent agreement relates negatively with the correlational results. In the present study, we argue that correlation methods (Pearson, intraclass and Kappa statistics) are influenced heavily by the range of responses, and high correlations can occur when the range of responses are considerable. Correlation results for TEOSQ scores show a similar trend to our hypothetical example. Ego items have the highest correlations, but also have the lower proportion of agreement scores. In contrast, Task items have significantly lower correlation (compared to Ego items) but have higher proportions of agreement, all greater than 90 ± 1.

Ego items with agreement values less than 90 ± 1 are also the ones with the highest standard deviations. This shows the importance of examining the range of scores when investigating stability. In the present study, MANOVA results indicated no significant bias in Ego items but this may be due to the nature of test-retest differences being both relatively large and random.

Although Task items were relatively stable according the 90% ± 1 criterion, it is possible that the agreement could be a product of a restricted range in scores, with the majority of participants reporting either 4 or 5 on the Likert scale on both occasions. It is also possible for items to show acceptable agreement primarily due to a restricted range of responses. For example, item 13, 'I do my very best' was found to be stable, but it should be noted that all participants reported either a 4 or 5 (maximum) on both completions. We argue

that it is important to consider the range of test-retest differences. For example, it is possible for an item to have 90% agreement within ± 1, but with the remaining 10% showing extreme outliers (e.g., ± 4). In this example, the median sign test might show no systematic bias. However, as the median sign test relies on rank differences, it would be unable to detect the effect of such extreme outliers. In this case, the use of MANOVA might be more appropriate as it takes the absolute variation in differences into account. In the present study, both MANOVA and the non-parametric median sign test indicated a significant shift in Task items.

We suggest that researchers interested in examining stability in the initial stages of test construction calculate test-retest differences for each item rather than calculating composite factor scores. Indeed a simple cross-tabulation of test-retest responses similar to tables 1 and 2 would be useful to assess the level of agreement along the diagonal and off-diagonal, to provide additional support and insight for the proposed 90% ±1 criterion. A limitation of assessing stability of factor scores is that it is not possible to identify rogue items that behave differently to the others in the scale.

The proposal that 90% of test-retest scores for the TEOSQ scale in the present study should lie within a reference value of ± 1 was based on the notion that researchers should set a criterion that has the most practical value (Altman & Bland, 1987). When investigating stability, researchers are interested in the magnitude and direction of test-retest differences (Bland & Altman, 1999). Recent research has emphasized the value of using the size of the effect rather than significance (Biddle et al., 2001; Schutz & Gessaroli, 1993). Interpretation of effect sizes has guidelines for interpretation rather than strict rules (Thomas & Nelson, 1996). The 90% of test-retest scores within ± 1 criterion is clearly an arbitrary value and there are a number of factors that could influence the acceptable criterion used for each

study. However, it should be noted that items in the same scale should show similar stability and reliability values including percent agreement.

Bland and Altman (1999) indicated that there are a number of factors that should be taken into accounting when considering whether a variable is reproducible. The first consideration is the extent to which the underlying construct is theoretically stable. A theoretically stable construct such as dispositional goal orientation (see Duda & Whitehead, 1998) should a show high percentage for test-retest agreement, with lower scores for less stable constructs such as psychological state variables. However, even with a theoretical stable construct, the number of choices available on the Likert scale will influence the percentage of test-retest agreement scores, and the greater the number of choices, the lower the percentage of agreement scores with the reference value ± 1 should be expected. In the present study, the categorical nature of a 1 to 5 Likert type scale used in the TEOSQ means that a participant can chose from one of five options, hence it is an ordinal scale, and therefore 90% of test-retest scores within ± 1 is acceptable.

An additional factor that can influence stability results is the interval over which data were measured can. Generally, stability coefficients reduce as the length of time increases (Anastasi & Urbina, 1997). This relationship is influenced by whether there were changes in the environment that might bring about changes in the target construct. A short completion of time might not bring about stable results if there are a number of factors that could change the target construct. In the present study, test-retest completions were only one week apart with minimal environmental changes, and therefore it is reasonable to assume a 90% ± 1 criterion value.

The method proposed for the assessment of stability should be used to compliment existing methods of assessment rather than replace them, but we emphasize the importance of researchers being clear on what aspect of stability/reliability each statistical test can

highlight. Researchers have tended to use correlation as a tool for multiple purposes, when clearly it assesses the degree of association or consistency between tests rather than the stability or reproducibility of test-retest scores. Correlation results cannot be solely relied upon as the range of scores heavily influences these, and this range can mask instability.

We suggest that the proportion of agreement method and traditional approaches could be used as a precursor to using structural equation modeling. Structural equation modeling can test for stability and association but is limited because it requires multiple measures and large samples. In the initial stages of test development, researchers are unlikely to invest such vast resources. Few researchers have used structural equation modeling to test for reliability and stability (Schutz, 1998). Other researchers have emphasized the point that structural equation modeling provides highly complex results and that as a simple alternative approach to stability, researchers should report the magnitude and direction of test-retest differences (Altman & Bland, 1987). Thus, although Marsh (1993) argued for at least three test-retest completions, given difficulties controlling the time before completions and factors that might influence the target, we argue two completions provide sufficient data to test stability especially in the initial stages of construct development. Logically, a truly stable measure would show evidence of stability from two completions.

In summary, we recommend that when assessing the stability of self-report questionnaires, researchers should calculate the test-retest differences and report the proportion/percentage of participants with differences within a reference value, thought to be of no practical importance. In the case of relatively stable dispositional constructs utilizing a five point scale, we recommend that a reference value of ±1 be adopted and argue that the majority of participants (90%) should record differences within this value. The percentage of participants within ±1 will indicate what is an acceptable or unacceptable test-retest variation/stability for each item. However, researchers need to assess whether there has been

a systematic shift in scores. Findings from the present study show that a wide and unacceptable range of random variation for Ego items, and Task items demonstrated a tendency to increase significantly over a short period of training.

References

Altman, D. G., & Bland, J. M. (1987). Comparing methods of measurement. *Applied Statistics, 36,* 224-225.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). New York: Prentice-Hall.

Bentler, P. M. (1995). *EQS Structural equation program manual.* Mulitvariate Software, Encino, CA.

Biddle, S. J. H., Markland, D., Gilbourne, D., Chatzisarantis, N. I. D., & Sparkes, A. C. (2001). Research methods in sport and exercise psychology: quantitative and qualitative issue. *Journal of Sports Sciences, 19,* 777-809.

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet i,* 307-310.

Bland, J.M., & Altman, D.G. (1999). Measuring agreement in methods comparison studies. *Statistical Methods in Medical Research, 8,* 135-160.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

Chi, L., & Duda, J. L. (1995). Multisample confirmatory factor analysis of the task and ego orientation is sport questionnaire. *Research Quarterly for Exercise and Sport, 66,* 91-98.

Duda, J. L. (1992). Motivation in sport settings: A goal perspective approach. In G. C. Roberts (Ed.). *Motivation in sport and exercise* (pp 57-92). Champaign, Illinois.

Duda, J. L., & Nicholls, J. G. (1992). Dimensions of goal achievement motivation in schoolwork and sport. *Journal of Educational Psychology, 84,* 290-299.

Duda, J. L., & Whitehead, J. (1998). Measurement of goal perspectives in the physical domain. In J. Duda (Ed.) *Advances in sport and exercise psychology measurement,* pp. 20-47. Morgantown, WV: Fitness Information Technology.

Harwood, C., & Hardy, 2001, L. (2001). Persistance and effort in moving achievement goal research forward: A response to Treasure and colleagues. *Journal of Sport and Exercise Psychology*, *23*, 330-345.

Harwood, C., Hardy, L., & Swain, A. (2000). Achievement goals in sport: A critique of conceptual and measurement issues. *Journal of Sport and Exercise Psychology*, *22*, 209-235.

Kline, P. (1993). *Handbook of Psychological Testing*. Routledge: London.

Marsh, H. W. (1993). Stability of individual differences in multiwave panel studies: Comparison of simplex models and one-factor model. *Journal of Educational Measurement, 30,* 157-183.

Marsh, H. W. (1997). The measurement of physical self-concept: A construct validation approach. In K. R. Fox, (Ed.) *The Physical Self*, pp. 27-59. Champaign, IL: Human Kinetics.

Nevill, A., Lane, A. M., Kilgour, L., Bowes, N., & Whyte, G. (2001). Stability of psychometric questionnaires. *Journal of Sports Sciences, 19*, 273-278.

Nevill, A.M. (1996). "Validity and measurement agreement in sports performance" [Editorial] *Journal of Sports Sciences, 14*, 199.

Nisbett, R., & Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment*. Englewood Cliffs, NJ: Prentice Hall.

Nisbett, R.E., & Wilson, T.D. (1977). Telling more than we know: Verbal reports of mental processes. *Psychological Review*, *84,* 213-279.

Schutz, R. W. (1994). Methodological issues and measurement problems in sport psychology. In S. Serpa, J. Alves, & V. Pataco, (Eds.), *International perspectives on sport and exercise psychology*. pp. 35-57. Morgantown, VA: Fitness Information Technology.

Schutz, R. W. (1995, August). *The stability of individual performance in baseball: An examination of four 5-year old periods, 1928-32, 1948-52, 1968-72, 1988-92.* Paper presented at the annual meeting of the American Statistical Association, Orlando, FL.

Schutz, R. W. (1998). Assessing the stability of psychological traits and measures. In J. Duda (Ed.) *Advances in sport and exercise psychology measurement*, pp. 394-408. Morgantown, WV: Fitness Information Technology.

Schutz, R. W., & Gessaroli, M. E. (1993). Use, misuse, and disuse of statistics in psychology research. In R. N. Singer, M. Murphy & L. K. Tennant (Eds.), *Handbook of research on sport psychology*, pp. 901-921. McMillan, NY.

Tabachnick, B. G., & Fidell, L. S. (1996). *Using multivariate statistics (*3rd *ed)*. New York, NY: Harper and Row.

Thomas, J.R. & Nelson, J. (1996). *Research methods in physical activity* (3rd ed.). Champaign, Ill.: Human Kinetics.

Treasure, D. C., Duda, J. L., Hall, H. K., Roberts, G. C., Ames, C., & Mahr, M. L. (2001). Clarifying misconceptions and misinterpretations in achievement goal research in sport: A response to Harwood, Hardy, and Swain. *Journal of Sport and Exercise Psychology*, *23*, 317-329.

Wilson and Batterham (1999). Stability of questionnaire items in sport and exercise psychology: Bootstrap limits of agreements. *Journal of Sports Sciences, 17***,** 725-734.

Author Note

Andrew Lane, Alan Nevill and Neal Bowes, School of Sport, Performing Arts, and Leisure. Kenneth R. Fox, University of Bristol.

Correspondence concerning this article should be addressed to Professor Andrew Lane School of Sport, Performing Arts, and Leisure, University of Wolverhampton, Gorway Road, Walsall, UK, WSI 3BD. E-mail: A.M.Lane2@wlv.ac.uk

We would like to acknowledge the helpful and insightful comments of the reviewers in preparing this manuscript.

Table 1

Test-retest Frequencies for Each Value on a 5-Point Likert Scale Over Time using Hypothetical Data (r = .88, p < .05; Intraclass correlation = .93, p < .05; kappa = .44, p < .001)

	Week 1					
	1	2	3	4	5	All
Week 2						
1	8	4	0	0	0	12
2	4	8	4	0	0	16
3	0	4	8	4	0	16
4	0	0	4	8	4	16
5	0	0	0	4	8	12
All	8	16	16	16	8	72

Table 2

Test-retest Frequencies for Each Value on a 5-Point Likert Scale Over Time using Hypothetical Data (r = .11, p > .05; Intra class correlation = .20, p > .05; Kappa = .11, p = .35)

	Week 1		
Week 2	2	3	All
2	20	16	36
3	16	20	36
All	36	36	72

Table 3

The Minimum and Maximum Test-Retest Differences, The intra-class and product-moment correlations, Kappa, F ratios, effect sizes, Percentage of Participants with Differences within (±1), and the Median Sign Test results (the number of participants with differences above and below the median, 0) for each Item of the TEOSQ

	Min	Max	Test 1		Test 2		intra-class	*r*	*Kappa*	$F_{1,212}$	Eta^2	% (±1)	≥ 1	0 diff	≤ -1
			M	*SD*	*M*	*SD*									
1. I am the only one who can do or play the skill. (Ego)	-4	4	2.67	1.22	2.51	1.24	.69*	.53*	.35*	5.34*	.03	82.7	45	98	70*
2. I learn a new skill and it makes me want to practice more. (Task)	-2	2	3.99	0.65	4.15	0.72	.65*	.48*	.32*	11.81*	.05	96.2	58*	127	28
3. I can do better than my friends. (Ego)	-4	4	2.94	1.03	2.74	1.05	.60*	.43*	.18*	7.07*	.03	86.9	53	84	76
4. The others can't do as well as me. (Ego)	-4	4	2.58	1.03	2.49	1.11	.60*	.43*	.25*	1.04	.01	84.9	53	94	66
5. I learn something that is fun to do. (Task)	-2	3	4.16	0.68	4.33	0.73	.60*	.43*	.26*	10.06*	.05	95.3	63*	120	30
6. Others mess up and I don't. (Ego)	-3	3	2.37	1.06	2.30	1.06	.65*	.48*	.26*	0.80	.00	84.6	54	96	63
7. I learn a new skill by trying hard. (Task)	-4	2	4.35	0.77	4.47	0.66	.46*	.30*	.26*	5.26	.02	92.5	53	122	38
8. I work really hard. (Task)	-3	2	4.43	0.65	4.47	0.63	.45*	.29*	.21*	0.52	.00	95.7	48	120	45
9. I score the most points/goals/hits. (Ego)	-4	3	2.72	1.09	2.86	1.16	.69*	.53*	.32*	3.53	.02	84.1	64	103	46
10. Something I learn makes me want to go and practice more. (Task)	-3	2	3.80	0.75	4.05	0.76	.62*	.45*	.29*	21.95*	.09	93.0	67*	118	28
11. I'm the best. (Ego)	-4	4	2.37	1.12	2.57	1.27	.75*	.60*	.29*	7.15*	.03	83.6	76*	98	39
12. A skill I learn really feels right. (Task)	-2	3	3.97	0.74	4.07	0.77	.59*	.41*	.28*	3.07	.01	92.9	59*	118	36
13. I do my very best. (Task)	-2	2	4.62	0.58	4.61	0.61	.70*	.54*	.37*	0.23	.00	99.0	30	150	33

* $p < .05$