Moving with the Times: Keeping up with Trends in Statistical Analysis and Research Design

Gershon Tenenbaum and Gerard Fogarty

University of Southern Queensland

In this chapter, we review some current controversies in statistics and measurement and suggest what they might mean for research in sport psychology. The first section of the chapter deals with the debate on the use of null hypothesis testing, a debate that has gathered considerable momentum over the past five years. Because most research in our field still relies on conventional tests of significance – after all, that is the approach still advocated in most textbooks - it is important that we review the arguments for and against significance testing and consider whether we need to change the way we do things. Our own summation of the debate is that it comes down to a question of reliability: the faith one has in one’s experimental outcomes. We suggest a number of ways in which researchers in sport psychology can improve the confidence they have in the results of single studies.

Rethinking the Null Hypotheses Significance Testing (NHST)

Recently one of our papers has returned from a review process. The response of one of the reviewers is used here to signify the concern that several leading statisticians in the behavioral and social sciences have with what is termed “Null Hypothesis Significance Testing” (NHST). The reviewer wrote:

“With three dependent variables, a repeated measures MANOVA would be the appropriate analysis for this data. At least with an experimentwise Bonferroni adjustment of ANOVA p values, significance criterion would be actually around .016, putting most results in doubt, and definitely attenuating the discussion presented of already non-significant results. MANOVA would also provide a report on sphericity compliance in multivariate data - supporting or rejecting the use of RM degrees of freedom (360 in the denominator) in what are already, at least, only marginal F values”.

It should be noted, before we discuss further the NHST, that the effect sizes obtained in the study ranged between 0.40 and 1.71, and although the procedures suggested by the reviewer would reduce experimentwise error rate, they would also reduce the power of the study. There is nothing technically wrong with the advice of the reviewer, but we wish to make the point that this sort of advice is driven by a concern for strict adherence to the NHST approach. We have to ask: Is “significance level” really an additional amendment that should be added to the 10 amendments already published in the old testament? We shall now address this concept in more depth.

According to Cohen (1994) who quotes Morrison and Henkel (1970) and earlier researchers, the NHST “has not only failed to support the advancement of psychology as a science but also has seriously impeded it” (p.997). It is mainly the 0.05 significance level on which Ho is rejected that concerns Cohen. The common neglect of “base-rates” before testing any hypothesis, advocated by Bayesian theorem, results in a substantial error as a consequence of adopting a low and arbitrary probability such as 0.05. Loftus (1996) further argues that “... reliance on NHST has channelled our field into a series of methodological cul-de-sacs, and it has been my observation over the years ... that conclusions made entirely or even primarily based on NHST are at best severely limited, and at worst highly misleading” (p.162).

To make sense out of NHST one should specify what is meant by a “difference” between two or more means of the population. The probability that the means will be identical is zero and therefore “meaningful” differences should be proposed. Thus, instead of asking whether there are differences between two or more means, the question should be “how big are the differences? Are they big enough for the investigator to care about and, if so, what pattern do they form?” (Loftus, 1996, p.163). When simply testing for mean differences, “rejecting a typical null hypothesis is like rejecting the proposition that the moon is made of green cheese... Well, yes, okay, but so what” (Loftus, 1995, p.163). The null hypothesis according
Moving with the Times

to Schmidt (1992) is always false and therefore the rate of Type I error is zero resulting only in Type II error. Thus, our science is going nowhere due to false results which rely on significance levels rather than magnitudes of effects. “Amounts” are more important than “directions” when verifying a theory. It is for this reason that regression coefficients are more stable than correlation coefficients and therefore recommended.

Loftus (1996) argues that “... investigators, journals, journal editors, reviewers, and scientific consumers often forget ... and behave as if the .05 cutoff were somehow real rather than arbitrary. Accordingly, the world of perceived psychological reality tends to become divided into “real effects” (p<=.05) and “non-effects” (p>.05) ... no wonder there is an epidemic of “conflicting” results in psychological research” (p.164). It is for this reason that meta-analytical studies end up with zero effect-size. When appropriate measures are applied and magnitudes estimated, base-rates could be determined and used for testing hypotheses. The 0.05 level of significance would no longer be the ultimate criterion for accepting or rejecting theories.

What then should we do in order to advance the domain of sport and exercise psychology? We advocate that instead of imitating other domains, we should develop statistical procedures which better account for the behaviors observed in our field. We summarise the recommendations made by Cohen (1994) and Loftus (1996) in Table 1.

Table 1
Recommendation to Improve the Statistical Procedures in the Social and Behavioral Sciences

<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Use graphical presentation</td>
<td>(1) Plot data rather than present in Tables plus F and p values</td>
</tr>
<tr>
<td>(2) Use effect-sizes to show magnitudes and</td>
<td>(2) Provide Confidence Intervals (CI) to assess the</td>
</tr>
<tr>
<td>confidence-intervals (CI) to replace p values in</td>
<td>statistical power of the results. It visually shows</td>
</tr>
<tr>
<td>NHST. The smaller the CIs the greater the power.</td>
<td>how the pattern of means reflects the population</td>
</tr>
<tr>
<td></td>
<td>means-pattern (see Loftus &amp; Masson, 1994 for review)</td>
</tr>
<tr>
<td>(3) Decide upon a “good enough” range to test</td>
<td>(3) Compute effect-sizes for single studies and overall</td>
</tr>
<tr>
<td>hypotheses. Determine differences in units such as</td>
<td>ES plus variation and CI for a set of studies.</td>
</tr>
<tr>
<td>effect-size; logits; etc. (see Serlin &amp; Lapsley,</td>
<td>Control for independent variables such as gender,</td>
</tr>
<tr>
<td>1993 for review)</td>
<td>culture, instrumentation, ego, type of</td>
</tr>
<tr>
<td></td>
<td>task/treatment, duration of interventions, etc.</td>
</tr>
<tr>
<td>(4) Challenge the results with alternative</td>
<td>(4) Set a quantitative hypothesis about the underlying</td>
</tr>
<tr>
<td>explanations (perceptual control over independent</td>
<td>pattern of means (i.e., assign weights) and</td>
</tr>
<tr>
<td>variables)</td>
<td>correlate with observed means (i.e., “planned</td>
</tr>
<tr>
<td></td>
<td>comparison”)</td>
</tr>
<tr>
<td>(5) Add likelihood ratios and Bayesian methods</td>
<td>(5) When interaction emerges, instead of focusing on</td>
</tr>
<tr>
<td>(Goodman, 1993; Greenwald, 1975)</td>
<td>differences between the dependent variable at a</td>
</tr>
<tr>
<td></td>
<td>fixed level of the independent variables (vertical</td>
</tr>
<tr>
<td></td>
<td>differences), look at differences between the</td>
</tr>
<tr>
<td></td>
<td>independent variable (horizontal differences) at a</td>
</tr>
<tr>
<td></td>
<td>fixed level of the dependent variable</td>
</tr>
<tr>
<td>(6) Rely on replication</td>
<td></td>
</tr>
</tbody>
</table>
These are all good suggestions and are echoed by others in the literature. Hammond (1996), for example, advocates the use of confidence intervals and effect sizes. He also recommends the use of replication to improve reliability. Gonzalez (1994) lists four principles to guide psychological research:

- the theoretical model should play a central role in guiding the analysis;
- the theoretical model should suggest parameters to estimate;
- the researcher should create a design that permits proper estimation of the parameters;
- intervals should be placed around parameter estimates.

Gonzalez goes on to advocate the use of a Bayesian approach wherein one has to estimate one’s prior belief in an hypothesis and then compute a posterior belief on the basis of data gathered in the study. The essence of the Bayesian approach is the moderation of one’s beliefs in the light of empirical data. Gregson (1997) argues that the problem is more serious than simply replacing a “significance” test by a confidence interval and also agrees with Gonzalez that a Bayesian approach is the preferred option.

Grayson, Pattison, and Robins (1997) made an interesting contribution to the debate when they summarised the alternatives as follows:

- continue as at present with objective tests of null hypotheses that severely limit what we can say about the results of a study;
- move towards a Bayesian approach that is intuitively appealing but where the requirement for prior knowledge poses some difficulties;
- adopt some intermediate position, such as a “commonsense approximation to Bayesian confidence intervals in the absence of prior knowledge...”.

Grayson and his colleagues stopped short of recommending any particular approach, preferring instead to urge researchers to be more flexible in their thinking about the role of statistical inference in research:

We also believe that the context of a problem may well affect the interpretive position that one might wish to adopt. In one situation, meta-analysis of existing, focused, pertinent research may be very useful; in another scientifically new, exploratory context, the null hypothesis could well be a very important speculation at which to address evidence; in another well-studied situation, a more quantitative Bayesian approach to inference about parameter values may be especially valuable. (p.70)

The important point made by Grayson et al. (1997) is that there is unlikely to ever be any resolution of the debate over preferred statistical approaches and that scientists should not adopt a passive role, waiting to see which side emerges the victor. Rather, scientists should recognise that they are in the best position to judge how data should be interpreted.

Implications for Sport Psychology

It would be pointless if we were to advise researchers in our field to abandon the NHST when it is still so widely accepted (and expected) by journal editors and reviewers. Nor would we wish to do so. The NHST is so well-entrenched that it is likely to take many years before it ceases to be the dominant paradigm. Certainly it will retain this status whilst the textbooks continue to favour the NHST position. Instead, we would urge researchers to take the not-quite-so-adventurous steps of reporting effect sizes and confidence intervals rather than relying solely on a test of the null hypothesis. This will give researchers wider scope for the interpretation of research findings. We would also argue that the real crux of the NHST debate hinges on the question of reliability: it is a debate not so much about alpha levels and confidence intervals as it is about the confidence we are prepared to place in our own experimental findings. Basically, with the NHST approach, one can make two kinds of errors: a Type I error where one has rejected the null hypothesis when it was inappropriate to
do so; or a Type II error where one failed to reject the null hypothesis when it should have been rejected. These errors will only be detected with replications that fail to support the original decisions; hence Hammond’s (1996) emphasis on replication. To improve the reliability of research outcomes, we agree with Hammond (1996) that replication is important, but we would disagree that it is the only way of improving reliability. Another way of improving reliability is by improving the measures one uses in a study and also by increasing the number of measures. We will illustrate both of these principles in research conducted in our own laboratories.

**Improving Reliability by Collecting Multiple Measures**

Kirker (1997) investigated how aggression and violence develop in basketball and ice-hockey. He assumed that aggression is most likely a result of a combination of factors, and therefore the more causal factors that are present, the greater the likelihood that an aggressive act will occur. Also the severity of the aggressive acts are believed to be a function of the number and intensity of actors present. The sequence or combination of causal variables is not easily specified, and factors may operate simultaneously. Thus multivariate causal traces should be considered.

Kirker believed that self-report and introspective measures alone have limited value in understanding sport aggression in real competitive settings. Such tools are best applied in conjunction with more ecologically valid and objective measures (i.e., naturalistic observation). The observation of behavior in a natural setting provides opportunities for researchers to better understand the complex dynamics of aggressive behavior in sport. Aggression is best studied in real time and in the context it occurs. Practical constraints, such as the need for extensive training of observers, expense, and lengthy data analysis have traditionally been the main barriers to observational research. Today, through the use of the computer and video technology, these logistic difficulties can be overcome, and observational analysis can be used in a more sophisticated manner.

In Kirker’s study, questionnaires were constructed to assess the attitudes of players and officials towards aggression, to determine relevant aspects of histories of aggression, and generally to gain some insight into the factors found to be related to aggression in sport but not directly observable and thus not able to be analysed through observational coding. With observational analysis, intentionality is essentially inferred. In this study, such inferences were made under rigorous conditions by experts through repeated replays of sport-specific behavioral typologies, incorporating hypothesised typical intention and severity of actions. This approach advances on the use of single measures such as officials’ ratings of penalised behaviors made without the aid of video replays and without supplementary data from the athletes themselves (Russell & Russell, 1984; Widmeyer & Birch, 1994). In the determination of causality, the use of observational analysis, questionnaire data, and players’ and officials’ comments on a video replay of behaviors of interest, advances on previous methodologies.

Experts have been used previously to assess the nature of aggressive-like behaviors (Bar-Eli & Tenenbaum, 1989; Teipel, Gerisch, & Busse, 1983; Widmeyer & Birch, 1984). In these previous studies, the experts used were not directly involved in the behaviors under investigation. They were using their personal experience to infer intention behind acts committed in general (Widmeyer & Birch, 1984) or by others (Bar-Eli & Tenenbaum, 1989; Teipel et al., 1983). Here, experts involved in the observed behaviors of interest were used. Furthermore, the role of the experts was expanded so that they became involved in the categorisation of behaviors (taxonomies for coding), questionnaire development, observation of behaviors, and inferences of causation.
To carry out the study the following measures were taken:

- The histories of games between the teams involved were reviewed and recorded.
- Four experts, two in each sport, were recruited to develop two taxonomies of violations and aggressive acts based on the literature and their experiences and the official game regulations. Taxonomies consisted of several dimensions and classifications of severity.
- Attitudes Toward Aggression Questionnaires and single items were provided to players and officials at training sessions prior to filming games.
- Four games, two ice-hockey and two basketball, were filmed. A CAMERA (Computer Acquisition of Multiple Ethnological Records and Analysis) video coding equipment was used for observational analysis. The CAMERA system contains PC-compatible computer software which records the sequence of distinct behavioral events occurring in real time, each with start and stop times. Complex interactions were broken down to manageable segments and sessions. For each game, two cameras were used: one directed to the play, the other to the court/rinksidet behavior of coaches and substitutes. Also, microphones were placed on officials and on the sidelines to pick-up comments from the bench and crowd.
- Classification, considerations of causation, and rating-like behaviors were recorded by the experts individually on the computer while watching the games on video. The taxonomies were used as references.
- The players exhibiting the aggressive behaviors were invited to observe their acts on video and reflect on the reasons behind these acts.
- The most severe aggressive acts were selected and referenced as “Zero” point. Up to four minutes of footage before and after each act from each game was analysed. Details of players exhibiting these behaviors and recipients of the behaviors were recorded on the computer output, along with game score, time phase, and any other relevant information.

The single analysis for each aggressive act enabled Kirker to integrate the information collected and generalise the findings across the two games in each of the two sports. When such procedures are adopted, more meaningful conclusions can be made as to how aggressive acts are developed and subsequently how we can modify or minimise their occurrence. The use of multiple measures was instrumental in achieving the aims of this particular study.

Another way of improving the reliability of research projects involves the use of measures that are ecologically valid. One of the most frustrating experiences that social and behavioural scientists undergo is the low amount of dependent variable variance accounted for by the independent variables. This can be improved by the use of more ecological paradigms in which performance is measured according to some objective criteria in addition to the self-ratings that seem to form the bases of so many studies in sport psychology. We argue that when the dependent variable is measured under conditions which completely mimic the real world, more variance of this dependent variable will be accounted for by other psychological variables. Again, this is easier to demonstrate by referring to another research project in our laboratory.

In exercise physiology, measures such as oxygen uptake and anaerobic threshold account for the majority of the variance of long-distance running times. The non-accounting variance is sometimes attributed to psychological variables. Recent psychological theories (see Tenenbaum, 1996 for review) have postulated that goal-orientation interacts with environmental conditions to influence effort and adherence in exertive-type tasks. Perceived ability, self-efficacy, self-control, and determination are also believed to be important mediators of behavioral outcomes. Two studies have examined this theory using real-life exertive conditions to measure consistency and adherence under such conditions. Calcagnini
(1996) asked non-active participants and anaerobic and aerobic athletes to squeeze the handbar of a dynamometer at 50% of their maximal squeezing strength as much as they could until a decrease of 10% of their designated value. Freeman (1997) asked his participants to run on a treadmill as much as they could for 90% of their maximal oxygen uptake. Measures of the psychological variables were taken prior to and after the completion of the tasks. The dependent variable was how much time participants could sustain in the zone of exertive tolerance? In each study, the “time in the zone of exertive tolerance” was the dependent variable while physical activity, goal-orientation, coping strategies, and determination were the predicting clusters in a hierarchical regression procedure. The results are presented in Table 2(a,b.).

Table 2
Summary of Hierarchical Regression Predicting “Time in Zone of Exertive Tolerance” in Aerobic (a) (Running) and Strength (b) (Squeezing a Dynamometer) Tasks

<table>
<thead>
<tr>
<th>Variable</th>
<th>Step</th>
<th>mult R</th>
<th>R²</th>
<th>ΔR²</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity Type</td>
<td>1</td>
<td>0.33</td>
<td>0.11</td>
<td>-</td>
</tr>
<tr>
<td>(aerobic, anaerobic, untrained)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goal orientation</td>
<td>2</td>
<td>0.55</td>
<td>0.31</td>
<td>0.20</td>
</tr>
<tr>
<td>(task, ego)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coping strategies</td>
<td>3</td>
<td>0.61</td>
<td>0.38</td>
<td>0.07</td>
</tr>
<tr>
<td>(self control, self efficacy, perceived ability)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Determination</td>
<td>4</td>
<td>0.69</td>
<td>0.48</td>
<td>0.11</td>
</tr>
<tr>
<td>(b)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Activity Type</td>
<td>1</td>
<td>0.33</td>
<td>0.11</td>
<td>-</td>
</tr>
<tr>
<td>(aerobic, anaerobic, untrained)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Goal orientation</td>
<td>2</td>
<td>0.56</td>
<td>0.31</td>
<td>0.21</td>
</tr>
<tr>
<td>(task, ego)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coping strategies</td>
<td>3</td>
<td>0.65</td>
<td>0.43</td>
<td>0.12</td>
</tr>
<tr>
<td>(self control, self efficacy, perceived ability)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Determination</td>
<td>4</td>
<td>0.76</td>
<td>0.59</td>
<td>0.16</td>
</tr>
</tbody>
</table>
As expected, the results in both studies revealed that psychological variables play an important role in determining how one can tolerate exertive conditions. Though the participants’ activity type determined 11% of the exertion tolerance variance, goal orientation added 20% and 21% additional explained variance, while coping strategies added an additional 7% and 12%, and determination 11% and 16%, respectively. A total of 48% and 59% of variance was accounted for. These values are far above those that are common in social and behavioral research. We use these studies to illustrate our point that this line of research should be encouraged and applied.

Improving Measurement Operations

There is yet another way of improving the reliability of individual studies that we can address here. It concerns the measurement process itself, a theme to which we have alluded elsewhere (Tenenbaum & Fogarty, 1997). Cohen (1994) stated that:

To work constructively with “raw” regression coefficients and confidence intervals, psychologists have to start respecting the units they work with, or develop measurement units they can respect enough so that researchers in a given field or subfield can agree to use them. In this way, there can be hope that researchers’ knowledge can be cumulative...A beginning in this direction comes from meta-analysis....But imagine how much more fruitful the typical meta-analysis would be if the research covered used the same measures for the constructs they studied. (p. 1001).

Meta-analyses would undoubtedly be more useful if the studies all used the same measures, but they would be doubly useful if the measures themselves satisfied basic measurement properties. As early as 1928, Thurstone stated that scales are not sufficient if they do not satisfy the requirement of having an “origin” or a defined “zero-point” with units of measurement that extend from the origin in a linear fashion. To achieve this, Thurstone (1928) stipulated that there should be a systematic attempt to select items that in fact do elicit a linear response from “low” to “high”. The requirement for a zero origin poses some difficulties for the classical measurement model. A score of zero tells us little because it does not indicate that the individual has zero ability, it simply indicates that the individual did not get any of the items in the test correct or, in the case of an attitude scale, did not select any option with a value above zero. Nor can we easily make interpretations about the intervals between different total scores. Classical measurement processes do not satisfy the requirement for a zero origin and equal units of measurement in the way stipulated by Thurstone.

The essential prerequisites for constructing such a measure comprise (a) a consistent definition of the domain of investigation (Thurstone, 1928), (b) selection of items which best represent the domain and share a common content classified under a single heading (Guttman, 1944), and (c) administration of the resulting scale to a sample of the relevant population in order to examine the response patterns. Andrich (1981) argued that the requirements outlined by Thurstone (1928) and Guttman (1944) which define the concept of psychological scaling, are solved by the Rasch model. In Andrich’s (1981) words, “The most important distinguishing feature of Rasch’s models is that, when they hold within some specified frame of reference, they provide explicit comparisons of person parameters which are independent of other persons to be compared and also independent of the parameters of the questions or items used to obtain the required responses. In achievement testing these parameters are the abilities of persons and the difficulty of items, while in attitude measurement they may be termed respectively attitudes and, following Thurstone, affective values. The explicit separation distinguishes these models from other psychometric models, generally called latent trait models, within which framework the Rasch models are often placed”. (p.2)
The Rasch method yields person measures and item values that are independent of each other. Both represent points on linear continuums and both rely on measurement units called logits that have a true zero point with equal units of measurement extending in either direction.

There are many benefits to using Rasch measurement, some of which were described in Tenenbaum and Fogarty (1997). In this section, we will show how Rasch analysis can be used to check whether a scale is suited to the population being studied. In this study, the Task and Ego Orientation in Sport Questionnaire (TEOSQ: Duda & Nicholls, 1992) was administered to 91 athletes participating in an aerobic task. The TEOSQ measures task and ego orientation in competitive activities and has quite a lot of supporting psychometric data gathered using the classical test model approach. That is, there is evidence that the TEOSQ does measure two independent factors and that scales developed on the basis of these two factors are reliable and relate in a meaningful way to external constructs. What else could we learn by using a Rasch approach?

One requirement for good measurement is that a test instrument should be appropriate for the population. Among other things, it should be neither too easy nor too difficult. This requirement also applies to attitude scales such as the TEOSQ where it is possible to translate “difficulty” into terms of how easy respondents find it to agree with the items. Where a likert scale is used, as is the case with the TEOSQ, easy items are ones which respondents feel inclined to rate highly. If the items in a test are too easy, they will not discriminate among the respondents. If they are too difficult, the same applies. The problem of matching a test with respondents is illustrated in Figure 1.

<table>
<thead>
<tr>
<th>Items</th>
<th>* * * * * * *</th>
<th>Ability Continuum</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persons</td>
<td>* * * * * *</td>
<td>------------------</td>
</tr>
</tbody>
</table>

Figure 1
Illustration of Mis-Match Between Tests and Persons

In this illustration, the test is too easy for the respondents: the items are all tapping the low end of the ability continuum whilst the respondents are located at the upper end. In a research situation, such a test would not be able to discriminate among the respondents. Administering the test would be a complete waste of time. One of the most basic applications of Rasch analysis enables researchers to draw maps showing where both items and respondents are located on the underlying ability continuum (or attitudinal continuum in the case of an attitude scale). Analyses of this type can be very useful. The item-respondent map for the TEOSQ ego scale is shown in Figure 2.
Figure 2
Rasch analysis of TEOSQ (Ego): Person and Item Locations in Logit Units
The map may be a little hard to read if you haven’t seen one before, so we will explain it carefully. The line down the middle represents the attitudinal continuum. People with a high ego orientation are represented by crosses towards the top left hand side of the diagram. Item locations are shown on the right hand side of the diagram. Because the TEOSQ uses a five-point Likert scale, the Rasch analysis shows four locations for each item. The locations represent the thresholds between the five categories. Thus, 3.1 on the bottom right of Figure 2 represents the amount of the latent trait (ego orientation) required before one marks the second Likert option rather than the first for item 3. Moving up the right hand side of the diagram, it can be seen that an attitudinal value of approximately –1.0 represents the threshold between category 2 and category 3 of this same item. People with lower amounts of ego orientation will select option 2 or option 1, people with higher amounts will select option 3 or higher. To select the highest option for item 3, one would need a value above 2.0 on the ego-orientation continuum. We can see from the crosses on the left hand side that only seven people marked option 5. Overall, the ability span of the items and the persons who are responding match very well. There are items that will discriminate among most respondents. This is a favourable outcome.

The map for the task scale of the TEOSQ paints a somewhat different picture. It is shown in Figure 3.
Item Estimates (Thresholds)
all on task (N = 91 L = 6 Probability Level=0.50)

---

3.0

XXX

2.0

XXXXXXXXXX

XXXXXXXXXX

1.0

XXXXXXXXXXXXXXXX

10.4

12.4

8.4

5.4

7.4

2.4

13.4

XXXXXXXXXXXXXXXX

12.3

0.0

XXX

2.3

5.3

8.3

7.3

10.2

13.3

-1.0

8.2

7.2

5.2

13.2

-2.0

5.1

---

Each X represents 1 student

Figure 3

Rasch Analysis of TEOSQ (Task): Person and Item Locations in Logits
Here, there is a slight mismatch between the items and the respondents. The mismatch can be summed up by saying that the respondents found the items too easy (i.e., they rated them “high”). There are no items that can discriminate reliably among the top 20 or so respondents in this study.

The sorts of problems described above can be detected quite easily using a Rasch analysis, and they can be corrected quite easily by choosing other tests or by adding items to cover the blank spots in the continuum. The importance of this kind of analysis, however, cannot be overestimated. Analysis of between-group differences, pretest-posttest comparisons, and correlational analyses will all prove rather futile if the test or attitude scale one uses is incapable of discriminating among the population being studied. Problems of this nature are as fundamental as whether one uses the NHST or some other approach. This brings us back to a point raised by Grayson et al. (1997) when they stressed the predominant role that the researcher - rather than the statistician - should play in the research process. In the field of sport psychology, we probably need to extend the whole debate one step further, to include the role of the consumers of research, the practitioners who are too-often left to convert esoteric research findings into practice that will lead to performance improvements. We would like to conclude this chapter by expanding the theme somewhat to include both researchers and practitioners.

The Scientist-Practitioner Approach

Though the role of the practitioner in the development of the behavioral and social sciences was debated for a long period of time, and much progress has been made since then, the scientific-practitioner dilemma is still misunderstood. Barlow, Hayes, and Nelson (1984) suggested methodologies which enable practitioners to empirically collect data on the interventions they use with their clients so that behavioral changes could be better accounted for and evidence of their effectiveness be more sound. Accordingly the scientific side of practice consists of three interrelated activities: (a) practitioners consume research findings on techniques they can apply when necessary, (b) practitioners use their own intervention using empirical methods to increase accountability, and (c) practitioners become researchers by producing new data from their own observations and measures to advance the scientific domain.

It is evident though that the prevailing experimental techniques which are based on large groups of participants, means and significance levels, comparisons and predictions in determining treatment effectiveness widen the scientific-practitioner gap (Barlow et al., 1984). About 21 years ago Cohen (1976) in a review of educational and health professionals reported that fewer than 20% of research articles have some applicability to field practitioners. Moreover, about 40% of mental health professionals believe that no research exists that is relevant to practice. Though Cohen (1981) raised the difficulty in defining research utilization, this concern remains today and we strongly believe that it is evident in the sport and exercise psychology domain. It is a common finding that observations made by clinicians were disregarded by scientists and research findings were perceived as inappropriate or trivial by practitioners (Strupp, 1968). Thus is seems that both scientists and practitioners seem to be insensitive to each other’s work (Lehrer, 1981).

To overcome the disputes between practitioners and scientists, Barlow et al., (1984) suggest to practitioners an integrated model of applied research that has the potential to narrow the gap and contribute substantially to any domain which involves human and social interactions. We believe that if these principles were appropriately applied to the sport and exercise psychology domain, better models which are field-driven would be established, and more accountability for interventions would be evident. Therefore, we shall briefly introduce
the principles of this integrated model.

The first stage involves an assessment of current interventions suggested in the literature. Such a review can lead practitioners to develop or enhance one or more of these interventions. In the next stage, a short-term study on the effectiveness of these techniques is necessary. Some initial comparisons are needed to establish alternative or modified interventions for specific problems. Then, long-term outcome studies should take place in order to test for intervention efficacy. This procedure is necessary in order to compare the findings reported by researchers and the findings in typical clinical settings (Agras & Berkowitz, 1980). At this stage, long-term outcomes and systematic field testing are substantially missing. Therefore, the extent of effectiveness and generalizability of the reported findings concerning intervention and treatment are very limited.

To solve some of the methodological problems associated with practice, practitioners are encouraged to treat large numbers of athletes or exercisers in diverse settings. Care should be given particularly to successes and failures as the series progresses and subsequently, to the reasons for these outcomes. In particular those alterations and/or additions that have been made to secure intervention success are to be accounted for. Failures, through the process of clinical replication, and their reason, should also be reported. At this stage the use of different measures is essential along with appropriate application of single case experimental design (see Chapters 8 - 10 in Barlow et al., 1984). A schematic representation of the various stages of this approach is displayed in Figure 4.

Figure 4

Time-series methodology is recommended to practitioners since variability attributed to other sources than that of treatment can be identified at the individual level and therefore more reliable rules can be generated that relate particular client or therapist characteristics to outcome. Thus, single-subject designs should be used with many individuals so that generalizability can be identified. Such a methodology, in contrast to group-comparison methodology, may offer more possibilities for applications but “it is only through the work of many practitioners that the development of rules of generalizability, based on the analysis of the individual, become practical” (Barlow et al., 1984, p.66). Thus large-scale, multicluster, clinical collaborative studies are believed to become an alternative to the classical positivistic approach.

An essential component in the empirical practitioner approach is accountability. Accountability can be achieved only when good measures are sufficiently valid and sensitive to the treatment provided to the clients. Each athlete, coach, or exerciser should be provided with realistic “measures of change”. Only through sufficient measures can the practitioners
evaluate their efforts. Though hundreds of measures were developed in recent years (see Ostrow, 1996), one should be cautious before choosing an appropriate measure or any observational procedure to evaluate or measure changes associated with (a) any problem, trait, and/or state, and (b) estimate “changes” in certain behavior/s. Many surveys in the past indicated that a substantial amount of measurement tools are impractical in the applied setting (Ford & Kendall, 1979; Wade, Baker, & Hartman, 1979). Thus, in the sport and exercise domain, we suggest that both practitioners and scientists first read the items before submitting questionnaires to their clients. Then, if necessary, submit the questionnaire to clients and question for clarity and appropriateness. Treatments can be improved substantially by awareness of measurement problems, measuring different aspects that treatment is aimed at, and enhancing accountability (Barlow et al., 1984).

Next we briefly specify the guidelines and principles suggested by Barlow et al., (1984) for collecting measures in practical settings. These guidelines, through their scientific perspective, may advance any social/behavioral field and establish better ecological theories. These guidelines are as follows:

- **State client’s problems or concerns in specific terms.** Specific terms are measureable or observed. Ask clients about their specific goals and wishes. Use Problem-Oriented Record (POR) and obtain subjective data (S), objective data (O), assessment (A), and plan (P), (SOAP). Goal-attainment scaling (i.e., for each goal establish a scale) is a recommended procedure.
- **Specify several problem behaviors.** Behaviors and interventions are complex and should be broken-down into dimensions and segments. Quantified measures should be obtained for each problem, regardless of their importance. In several cases, measures that are unrelated to the treatment can also be applied.
- **Obtain multiple measures for each problem behavior.** One measure is sometimes insufficient for diagnosis. Sometimes measures of the same trait/state are in accordance with each other. Thus, through clinical replications, the “best” measure for the “particular situation” can be determined. Inconsistency among measures can be then attributed to “method variance”. Frequency in reported asynchrony among motoric, physiological, and self-report measures is due to confound measurement methods and the content being measured (Cone, 1979).
- **Select measures that are both sensitive and meaningful.** Choose molecular measures that are very sensitive to intervention changes though lacking in construct validity (i.e., smiles, violent acts, eye contact, etc.) along with molar measures which have high construct validity but lack sensitivity (i.e., many of the introspective questionnaires in use). Molecular measures should be tested very frequently whereas molar measures should be tested on a monthly or longer basis.
- **Collect measures early in the course of treatment.** This secures a substantial baseline for late treatment accountability. At baseline, it is recommended that many measures are used to detect problems and concerns.
- **Some measures should be taken repeatedly prior, during, and following treatment to account for reasons and outcomes.** Such measures are used as feedback for the psychologist as to the effectiveness of the treatment/technique and to indicate when alteration is necessary.
- **Comparisons should be made within a specific measure only if data are collected under similar conditions, to enable valid comparisons to be made across measurement.** Irrelevant factors (time of talk, etc.) that occur during the intervention are kept constant to enable later cause-effect conclusions. It is of vital importance to insure that changes in
behaviors are due to real events rather than uncontrolled conditions, as behavior is situation-specific. Measures should be sensitive to situational changes.

- Quantitative data should be presented graphically. On the time axis, different measures should be displayed in parallel along with remarks of changes and alterations that have taken place in specific points of time. Graphs provide a convenient means of data storage that can be used later for subsequent analysis and conclusions.

- Convenient measures should be measured more frequently than inconvenient measures. Some measures that may supply additional information do not have to be collected frequently, though they may supply evidence for different aspects associated with the interventions.

- Selection of good and accurate measures is essential. Multiple-methods and measures are recommended. Though psychometrically sound measures are recommended, use of "real" measures of what clients do in their natural environment is necessary. These behaviors should be recorded by trained observers to secure accuracy. The quality of the self-monitoring data can be improved if clients are aware that their reliability can be detected. Thus, "reliability check-ups" are also important. The selection of instruments which have "functional utility" is therefore required.

Once these procedures are adopted, the domain of sport and exercise psychology may be recognised as a "target domain" in which the practitioner-scientist gap has been narrowed and ecological theories have been developed.

References


Loftus, G.F. (1996). Psychology will be a much better science when we change the way we analyze data. Current Directions in Psychological Sciences, 54, 161-170.


