Int. J. Sport Psychol., 20:93-113, 1989

COMPARISON OF TEST CONSTRUCTION STRATEGIES IN AN ATTEMPT TO DEVELOP AN ATHLETIC POTENTIAL SCALE

DAVID R. BROWN, WILLIAM P. MORGAN AND JOHN F. KIHLSTROM *, *USA*

Empirical and intuitive approaches to test construction were compared during an attempt to develop a scale that would discriminate between successful and unsuccessful athletes. Eight scales were developed using items from the Minnesota Multiphasic Personality Inventory item pool. Two scales were constructed using an empirical approach, five scales using an intuitive strategy, and one scale based on a random selection of MMPI items. Results of this investigation indicate that reliable scales were developed using the empirical and intuitive test construction methods, yet all scales were equally ineffective as valid measures of athletic potential. Post hoc analyses indicated that while overall classification rates were not significant, classification of successful athletes was made with greater precision than classification of unsuccessful athletes. Findings are interpreted in terms of characteristics inherent in the MMPI item pool and the domains assessed by the MMPI. Recommendations are made for future scale development in sport psychology.

Introduction

Few areas of inquiry in psychology have generated more debate than the area of personality. Controversy surrounds the different theoretical positions maintained by personality theorists to describe, explain or predict behavior, the effectiveness of objective personality tests developed to oper-

* This research was conducted as partial fulfillment for the Doctor of Philosophy Degree by the first author. Suggestions and editorial assistance were provided the first author by Drs. Mary Brennan, Richard Johnson, Henry Montoye and Margaret Safrit, and their help is gratefully acknowledged.

Professor Morgan is affiliated with the Department of Physical Education and Dance at the University of Wisconsin-Madison and Professor Kihlstrom is currently with the Psychology Department at the University of Arizona, Tucson.

Address for correspondence: David R. Brown, Miami University, Ohio, Department of Physical Education, Health and Sport Studies, Room 107, Phillips Hall, Oxford, Ohio 45056, U.S.A.

ationalize personality theories, and the actual test construction methods used during the development of personality scales. In the field of sport psychology, scale construction strategies have received little attention, as they have been overshadowed by personality research designed to evaluate the effectiveness of objective personality assessment *per se* (Brown, 1989).

The issue of which test construction strategy is best to employ during the development of a personality scale has been examined by Burisch (1984). After analyzing the literature in which the inductive (internal), external (empirical) and deductive (intuitive) test construction strategies were evaluated, Burisch endorsed an intuitive approach. This approach requires test developers to write or select items that possess intuitive appeal or high face validity for measuring a particular construct. According to proponents of this approach, such items can be used to develop valid and reliable personality scales in a relatively brief period of time, and the scales are comparable in effectiveness to those developed using factor analytic (internal) or group discriminative (empirical) test construction methods (Ashton & Goldberg, 1973; Burisch, 1984; Hase & Goldberg, 1967; Jackson, 1971).

It was with this literature base in mind that the current investigation, using the Minnesota Multiphasic Personality Inventory (MMPI), was conducted. The value of using the MMPI as a predictive instrument in the field of sport psychology has been limited. The MMPI has been employed with some success in discriminating between athletes and nonathletes (Booth, 1958; Slusher, 1964; Johnson & Morgan, 1981), successful and unsuccessful athletes (LaPlace, 1954; Booth, 1958; Morgan & Johnson 1977, 1978; Johnson & Morgan, 1981), and athletes participating in different sports (Blaser & Schilling, 1976; Johnson & Morgan, 1981). However, the results from these investigations, in which the MMPI was used to test athletes, do not generate confidence for using the inventory in an applied context or field setting.

The study described in this article evaluated empirical and intuitive approaches to test construction to determine if a reliable and valid short form of the MMPI could be developed to predict athletic potential (i.e., success), more effectively than had been accomplished using the full length MMPI in earlier investigations by Morgan and Johnson (1977, 1978; Johnson & Morgan, 1981). An important feature of the current investigation is that the large data base was the same as that used in the Morgan and Johnson research (Johnson & Morgan, 1981; Morgan and Johnson, 1977, 1978). The data base was created in the early 1960's and has been preserved for research purposes by the University of Wisconsin-Madison Counseling Services (Drake and Oetting, 1959). While this investigation used the same subject pool,

data base, and definition of the independent variable as employed in the previous studies (Morgan & Johnson, 1977, 1978; Johnson & Morgan, 1981), the methodology differed, in that a double cross-validation design was employed. In addition, individual items from the MMPI, rather than the global factors assessed by the MMPI validity and standard scales, were evaluated to determine their effectiveness in discriminating between successful and unsuccessful athletes.

Empirical and intuitive test construction strategies were used to construct Athletic Potential Scales. It was hypothesized that the intuitive and empirically derived Athletic Potential Scales would effectively discriminate between successful and unsuccessful athletes. In addition, it was hypothesized that the Athletic Potential Scales developed using intuitive strategies would be at least equal in effectiveness to the empirically developed scales. These hypotheses were consistent with what could be expected based on analysis of the literature comparing different test construction strategies (Burisch, 1984).

Materials and method

Subjects

This investigation was prospective in nature. The subjects were former male athletes who enrolled at the University of Wisconsin-Madison during the years 1960 through 1964, and who eventually graduated from the University of Wisconsin-Madison. The time interval 1960 through 1964 was selected because it was the last five year period in which the Minnesota Multiphasic Personality Inventory (MMPI) was routinely administered to all entering University of Wisconsin freshman students. Permission was obtained from the athletes in the study to retrieve and analyze their MMPI data.

The athletes were classified as successful or unsuccessful. Consistent with earlier investigations (Johnson & Morgan, 1981; Morgan & Johnson, 1977, 1978), successful athletes were defined as those who earned a freshman numeral and two or three varsity letters in at least one sport. Those athletes who earned only their freshman numeral during their athletic careers were considered unsuccessful.

As illustrated in Figure 1, two groups, Group A and Group B, each contained subgroups of 92 successful and 92 unsuccessful athletes. Therefore, the number of subjects in this study totalled 368.

Design

A double cross-validation design was employed in this investigation, as illustrated in Figure 2. Two empirically-derived scales (Scale A and Scale B) were constructed during Phase I. Double cross-validation of these scales with each group of subjects (Group A and Group B) occurred in Phase II. Also illustrated in Figure 2 is the Phase III development

of intuitive scales, and the development of a scale based on random item selection. These scales were cross-validated with Group A and Group B subjects in Phase IV.

Empirical Scale Development

During the Phase I development of empirical scales, the responses of the successful and unsuccessful athletes to the true-false MMPI items were analyzed using a chi square statistical analysis procedure. The development of two empirical scales required the computation of 1100 chi square statistics, as the MMPI contains 566 items, 16 of which are duplicates. Items that differentiated between the successful and unsuccessful athletes at the .05 level of significance were retained, and two empirical scales designed to measure athletic potential were constructed. One empirical scale was developed based on responses made by the athletes in Group A (Athletic Potential Scale-A or APS-A), and one developed based on the responses of the athletes in group B (Athletic Potential Scale-B or APS-B).

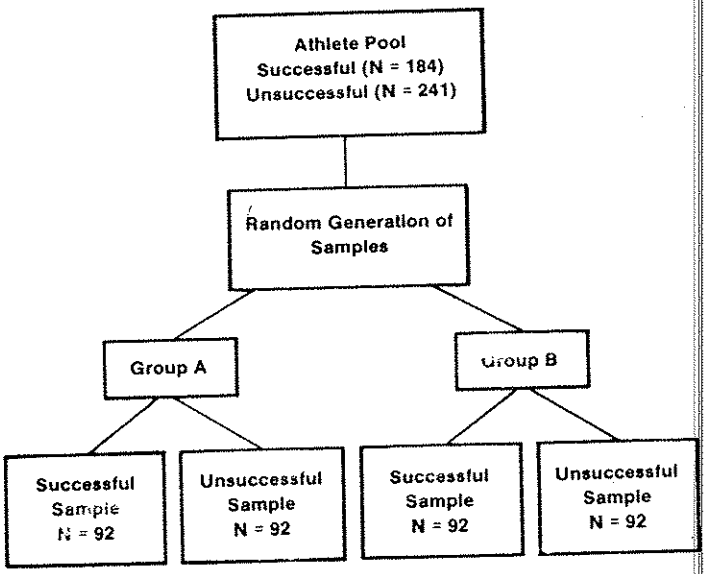

Fig. 1. - Schematic depicting allocation of subjects to Group A and Group B.

Two global or total Athletic Potential Scale scores, one for APS-A and one for APS-B, were imputed for each athlete by adding together all of the one point responses given by an athlete. The scoring key was designed to favor successful athletes, and was developed based on the results of the chi square analyses. An athlete's answer to an item was given one point if it was consistent with the answer made to the item by the majority of the athletes in the successful group.

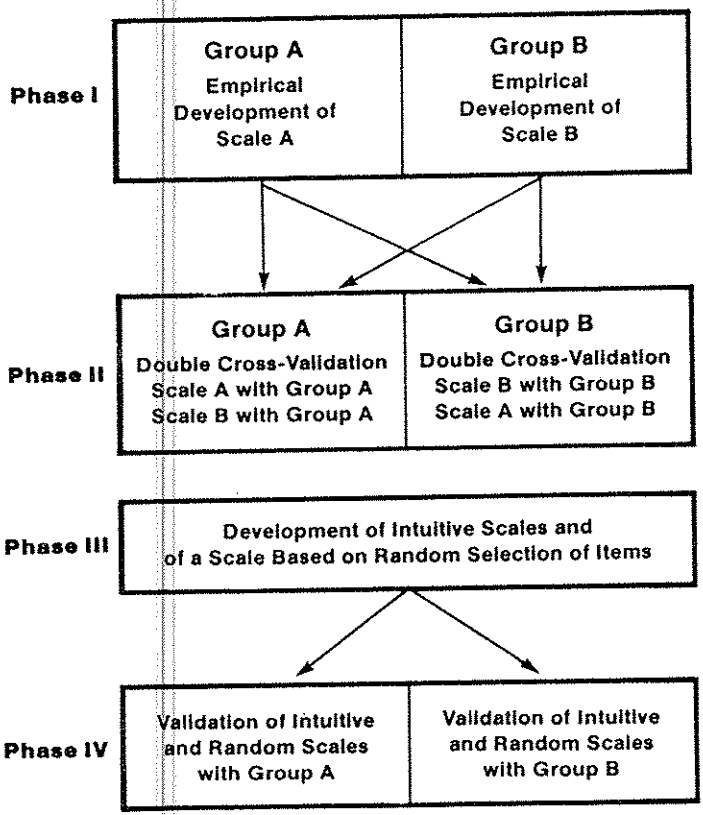

Fig. 2. - Schematic depicting the sequence of scale development and validation analyses used in this study.

Intuitive Scale Development

During Phase III of this investigation, four scales were developed intuitively by MMPI item selectors. One scale was developed by each of four groups of judges, and there were ten judges in each group. The groups consisted of undergraduate athletes (Scale UA), undergraduate nonathletes (Scale UNA), graduate students in physical education (Scale PE), and graduate students in clinical psychology (Scale CP). The forty judges were asked to read each MMPI item and to indicate whether or not an item could potentially discriminate between successful and unsuccessful athletes. The directions that were given to each judge are as follows:

«Dear Study Participant:

You are participating in an investigation in which an attempt is being made to construct a psychological inventory for measuring athletic potential. The investigation is aimed at identifying those items in an already-existing psychological inventory, the Minnesota Multiphasic Personality Inventory (MMPI), that may measure such potential.

Athletic potential is being defined as the potential for a collegiate athlete to earn varsity letters in sports such as football, basketball, hockey, baseball, track, cross country, tennis, swimming, etc. Athletes who participate in sports in college may either be skilled enough to become members of an athletic team, may be «cut» from the team and not given the opportunity to participate in athletics, or may elect not to continue participating in athletics. Those athletes who become members of a team and who are very successful can earn a varsity letter at the end of each season of athletic eligibility. For example, presently a successful athlete who participates in a sport for four years at the University of Wisconsin-Madison could earn four varsity letters. Athletes who are good enough to become members of a team but are not good enough to «start» or play a great deal while participating in the sport may not earn any varsity letters, or perhaps may earn only one letter. These athletes could be considered less successful than those who earned four letters during their four years of eligibility. Several years ago freshman athletes could not compete on a varsity team. Therefore, an athlete could earn a frosh numeral and one, two, or three varsity letters during four years of participation in a sport.

Enclosed is an MMPI test booklet and answer sheets for recording your answers to each MMPI item. As you read each MMPI item, you should decide whether you believe that the item could potentially discriminate between successful and unsuccessful athletes. In this investigation, successful athletes are being defined as those having the skill and potential to earn three or four varsity letters, or those athletes who in the past earned a frosh numeral and two or three varsity letters. Unsuccessful athletes are being defined as those who do not have the ability to earn varsity letters. As you read an MMPI item, you may decide that successful athletes would answer the item one way, while unsuccessful athletes would answer the item the opposite way. Thus, the item would discriminate between successful and unsuccessful athletes.

If you feel that an item would discriminate between successful and unsuccessful athletes, and that successful athletes would answer the item TRUE, please darken in CIRCLE A on your answer sheet. If you read an MMPI item and decide that the item would discriminate between successful and unsuccessful athletes and that successful athletes would answer the item FALSE, then please darken in CIRCLE B on your answer sheet. If you read an MMPI item and decide that there would be NO DIFFERENCE in the way that successful and unsuccessful athletes would respond to the item, then darken in CIRCLE C on your answer sheet. If you have any questions, please ask the person administering this test to provide you with clarification before you complete your answer sheets.

Thank you.

It was decided *a priori* that in order for an item to be included on a group's intuitive scale, it had to be selected as a discriminating item by at least seven of the group's ten judges. In addition, all seven or more judges selecting the item had to agree about how successful athletes would answer the item. The items included on the intuitive scales were subsequently keyed to favor successful athletes. An athlete was given one point for every true or false response made that was in agreement with the judges' belief about how a successful athlete would respond. A total scale score was computed for each athlete using each intuitive scale (UA, UNA, PE and CP) by summing the number of one point responses made by an athlete on a scale.

A fifth intuitive scale was developed from items that formed the four intuitive scales developed by the judges. An item was included on the fifth or combined scale (Scale COM) only if it was selected by at least three of the four *groups* of item selecting judges.

Scale development of randomly assigned items

In addition to the scales developed using the empirical and intuitive approaches to scale construction, an additional scale was developed by randomly selecting MMPI items to be included on the scale. Items on this scale were keyed by assigning one point to a randomly selected true or false response. This scale of randomly selected items (Scale RAN) served as a basis for comparing the effectiveness of the emprical and intuitive scales to discriminate between criterion groups above a chance level.

Data analysis-reliability

Inter-judge reliability coefficients of agreement were computed between the different groups of judges using a method developed by Ebel (1951). The coefficients were computed by analyzing each item included on each intuitive scale, taking into account the number of judges from the four groups (N = 40 possible) who selected the item as having the potential to discriminate between the criterion groups. The coefficient of agreement also took into account the true or false response selected by a judge as the response that a successful athlete would give.

The reliability of each scale was also evaluated using the alpha coefficient method. Since the data were dichotomous in nature (true-false responses to MMPI items), the coefficient alpha is equivalent to the Kuder-Richardson-20 reliability coefficient. The significance of the reliability coefficients was determined by using a formula described by McNemar (1962).

Data analysis - validity

Hull and Nie (1981) indicate that while reliability using coefficient alpha can be employed «to assess the reliability of the test, it would provide no information on the ability of the scale to distinguish between the two groups. For this type of application, discriminant function analysis is more appropriate» (p. 250). A discriminant analysis was conducted to assess the effectiveness of using the empirically, intuitively, and randomly developed athletic potential scales to classify Group-A and Group-B athletes as successful or unsuccessful.

In order to evaluate the predictive accuracy of the total scores obtained from using the athletic potential scales a base-rate value was compared to the prediction rate (i.e., hit rate) of subjects correctly classified. Base-rate expectations were maintained at 50% by using an equal number of successful and unsuccessful athletes. In other words, classifying all 184 subjects in Group A (or Group B) as successful, or all as unsuccessful, would result in a 50% correct classification rate because there were 92 successful and 92 unsuccessful athletes in Group A (and Group B). The actual percentage of athletes accurately predicted to be successful and unsuccessful using total scale scores from the empirically, intuitively and randomly developed scales, was compared to the base-rate 50% expectation. The prediction percentage gain then became Delta, where Delta equals the Prediction-Rate minus Base-Rate values.

Results

The Scales

The two empirical scales, five intuitive scales, and one random scale ranged in length from 14 to 49 items. The empirically developed Athletic Potential Scale-A (APS-A) and Athletic Potential Scale-B (APS-B) were 36 and 49 items in length respectively. The scale constructed by the 10 undergraduate athlete judges (Scale UA) contained 29 items, and the scale developed by the 10 undergraduate nonathlete judges (Scale UNA) contained 41 items. The scales constructed by the 10 graduate students in physical education (Scale PE) and the 10 graduate students in clinical psychology (Scale CP) contained 19 and 21 items respectively. The combined scale (Scale COM), contained 14 items and each item was selected by at least three of the four groups of judges. The randomly developed scale (Scale RAN) was 32 items long, which was the average length of the intuitive scales developed by the four groups of judges (UA, UNA, PE and CP) and the two empirically derived scales (APS-A and APS-B).

There was very little item overlap between the two empirical scales, APS-A and APS-B. Of the 36 items forming APS-A and the 49 items forming APS-B, only two items appeared on both scales. This suggests that the items which significantly discriminated between successful and unsuccessful athletes in Group A and Group B may have done so due to chance occurrence as a result of generating 550 chi square analyses. In addition, there was virtually no item overlap between either empirical scale compared with any of the intuitive scales. For example, the 14 items included on Scale COM are listed in Table I. Of 14 items, only MMPI item #83 appeared on one of the two empirical scales, APS-B. Thus, out of 28 opportunities, there was only one «hit». Item overlap between the other intuitive scales and the empirical scales was similar, and practically nonexistent.

Reliability

The degree of agreement between the different groups of judges was calculated using the method developed by Ebel (1951), and the inter-judge reliability coefficients of agreement are listed in Table II.

The coefficients ranged from .70 to .83, indicating that there was good objectivity among the judges. In other words, there was good agreement between the judges as to which MMPI items would discriminate between successful and unsuccessful athletes, and the true or false direction in which successful athletes were predicted to respond to an item. It is also noted that the intuitive scales were 41, 29, 21 and 19 items in length for scales UNA, UA, CP and PE respectively. Out of these items, 14 were included on the combined scales of intuitive items (i.e., Scale COM). Scale COM was constructed from items selected and keyed in the same direction by three or four groups of judges. Although descriptive in

TABLE I

Items forming the intuitive scale - combined (COM). Items were selected by 3 or 4 groups of judges.

MMPI item # and keyed 1 point response	Item selected by Intuitive Group *			
	UA	UNA	PE	CP
1) 3 (True)	X	X	X	X
2) 32 (False)	X	X	X	X
3) 35 (False)	X	X	X	
4) 41 (False)	X	X		X
5) 73 (True)	X	X	X	X
6) 83 (True)	X	X	X	X
7) 86 (False)	X	X	X	
8) 138 (False)		X	X	X
9) 142 (False)	X		X	X
10) 163 (True)	X	X		X
11) 257 (True)	X	X	X	X
12) 357 (False)	X	X	X	X
13) 487 (False)	X	X		X
14) 523 (True)	X	X	X	X

* X = an item that was selected by 7 or more judges out of the 10 in each group.

nature, this finding also seems to indicate that the four groups of different judges were making many similar judgments.

The alpha reliability coefficients that were obtained for each scale are listed in Table III. The two scales yielding the highest reliability coefficients of internal

TABLE II

Inter-judge reliability coefficients of agreement among judges, regarding the selection and keying of items contained on the four intuitive scales.

Scale constructed by the judges in group		Coefficient of agreement among all judges
Undergraduate athletes	(N = 10)	.70
Undergraduate non-athletes	(N = 10)	.83
Graduate students-physical education	(N = 10)	.81
Graduate students-psychology	(N = 10)	.77

TABLE III

Alpha reliability coefficients and corresponding z-values and probabilities for the empirical, intuitive, and random scales.

Scales (scale length)	Group A			Group B		
	r	z	p	r	z	p
Empirical						
APS-A (36 items)	.62	3.80	<.01	.43	2.68	<.01
APS-B (49 items)	.71	5.04	<.01	.78	5.56	<.01
Intuitive						
UA (29 items)	.67	3.61	<.01	.73	3.94	<.01
UNA (41 items)	.62	3.99	<.01	.72	4.60	<.01
CP (21 items)	.61	2.72	<.01	.72	3.21	<.01
PE (19 items)	.53	2.10	<.05	.61	2.45	<.01
COM (14 items)	.46	1.72	NS	.65	2.43	<.01
Random						
RAN (32 items)	.11	0.61	NS	.05	0.27	NS

consistency, taking into consideration the reliability coefficients obtained as a result of using each scale with the subjects in both Group A and Group B, are the empirical APS-B scale and the intuitive UA scale. The empirical Scale APS-A, the intuitive Scale UNA, Scales CP and PE, and Scale COM (Group B) also yielded reliability coefficients which were significant at the .05 level or higher. The reliability coefficients for the intuitive Scales UNA and CP were comparably high between the two groups of subjects, ranging from .61 to .72, while the reliabilities for APS-A (Group B) and PE (Group A) were slightly lower, .43 and .53 respectively. The reliability coefficient for Scale COM (Group B) was significant at the .01 level; however, the reliabilities obtained for Scale COM (Group A) and the scale of randomly selected items, Scale RAN (Group A and Group B), were not significant. It was concluded that the two empirical scales and four intuitive scales (UA, UNA, CP, and PE) were reliable in terms of a measure of internal consistency.

DISCRIMINANT VALIDITY

A discriminant function analysis was conducted to determine the effectiveness of each scale to discriminate between successful and unsuccessful athletes. For each scale and each subject group, the percentages of total correct classifica-

tion are listed in Table IV. These percentages were compared to a base-rate value of 50% using z-values to assess the significance between proportions (Downie and Heath, 1970).

TABLE IV

Total percent of cases correctly classified using the empirical, intuitive, and randomly developed scales, and corresponding z-values and probabilities.

Base-rate: Prior probability for each group = 50%							
Scales	Group A			Group B			Average %
	%	z	p	%	z	p	
Empirical							
APS (validated)	63	2.40	<.05	72	4.26	<.01	67.5
APS (cross-validated)	54	0.73	NS	51	0.21	NS	52.5
Intuitive							
UA	49	0.10	NS	52	0.42	NS	50.5
UNA	43	1.25	NS	51	0.21	NS	47.0
PE	51	0.10	NS	52	0.42	NS	51.5
CP	47	0.63	NS	49	0.21	NS	48.0
COM	50	0.00	NS	51	0.21	NS	50.5
Random							
RAN	54	0.84	NS	50	0.00	NS	52.0

Based on these analyses, the only scales that discriminated at significantly better than base-rate levels were the empirical scales, which were validated on the groups of subjects from which they were derived. The Athletic Potential Scale constructed using Group A subjects and validated using Group A subjects correctly classified 63% ($p<.05$) of the athletes, while the Athletic Potential Scale developed using Group B subjects correctly classified 72% ($p<.01$) of the subjects in the Group B sample.

Cross-validation of the empirical scales did not result in a classification accuracy rate that differed significantly from base-rate. Similarly, the intuitive scales did not significantly improve the ability to classify athletes correctly as successful or unsuccessful. In fact, the intuitive scales as well as the cross-validated empirical scales were no better in discriminating between criterion groups than the scale developed by randomly selecting MMPI items. A comparison of the empirically derived (cross-validated) scales, the

intuitive scales, and the scale developed by random item selection indicates that the greatest increase in total correct classification was 4% over the base-rate value (54% vs. 50%). The 54% correct classification of Group A subjects was obtained using the empirical APS developed on Group B subjects, and also as a result of using the scale of randomly selected items. It was concluded that none of the empirical or intuitive scales were valid measures of athletic potential as defined in this study.

Post hoc analyses

In light of the finding that none of the empirical or intuitive scales successfully discriminated between the criterion groups, several additional post hoc discriminant function analyses were conducted. The purpose of these analyses was to compare the effectiveness of the empirical and intuitive scales with some of the validity and standard MMPI scales. Thus, it could be determined whether the MMPI scales were more effective than the empirical and intuitive scales in discriminating between the successful and unsuccessful athletes in Group A and Group B.

The scales that were used for the post hoc analyses were the two empirically derived athletic potential scales, the scale developed by random selection of MMPI items, and Scale COM (i.e., the scale developed using items selected by three of four groups of judges). Scale COM was selected for use because none of the intuitive scales was found to be significantly better than any other in correctly classifying subjects, and because Scale COM contained items representative of all the other intuitive scales. The MMPI scales that were selected for the post hoc analyses were the Psychopathic Deviate (Pd), Depression (D), Hysteria (Hs), Social Introversion (Si), Psychopathic Deviate paired with Schizophrenia (Pd with Sc), Lie (L) and F scales. These scales were not K corrected in accordance with recommended MMPI scoring procedures. The Pd and Pd paired with Sc scales were used because earlier research had shown that elevated Pd scores ($T \geq 70$) and elevated Pd paired with Sc profile patterns were obtained by a significantly greater number of unsuccessful athletes compared to successful athletes (Johnson & Morgan, 1981). The D and Hs scales were chosen based on a theoretical rationale that it is reasonable to expect that athletes who are depressed (Scale D) or overly preoccupied with bodily concerns or injuries (Scale Hs) might be less successful than athletes who are not depressed or hypochondriacal. In the case of depression, there is some empirical evidence to support this hypothesis (Morgan, Brown, Raglin,

O'Connor & Ellickson, 1987). The F and L scales were selected for the post hoc analyses because of their value as validity checks against response distortion, and because the scales discriminated between successful and unsuccessful college oarsmen in an earlier study (Morgan & Johnson, 1978). The Si scale was chosen because eleven items from this scale qualified, based on significant X^2 values, for inclusion on the empirically derived Athletic Potential Scales (5 Si items on APS-A and 6 on APS-B).

As a result of conducting the discriminant function analyses, it was found that the validity and standard MMPI scales tended to correctly classify a higher total percentage of subjects than did the cross-validated empirical athletic potential scales, the intuitive scale-COM, or the Scale-RAN. The MMPI scale that correctly classified the most subjects was the Pd paired with the Sc scale. The use of Pd paired with Sc to classify Group A and Group B subjects resulted in an average derivation of 57% of the athletes being correctly classified. The average derivation scores for other MMPI scales listed in rank order from highest to lowest were F (56%), D (54%), L (53%), and Hs (53%). The average derivation score for the MMPI Scale Si was 51%, and this value was comparable to the values obtained from the empirical, intuitive, and randomly developed scales. Differences between any of the scales and the 50% base-rate value were not statistically significant and, therefore, these differences are also of no practical significance.

The finding that the Pd paired with Sc scale was the highest discriminating variable, however, suggests the possibility that a multivariate rather than univariate model may be preferable in research of this nature. It may also be of some possible clinical interest to note that the MMPI scales did a much better job of classifying successful athletes than unsuccessful athletes. The Hs scale scores used to discriminate subjects in Group B is representative of these findings. Scale Hs correctly classified 65% of the successful athletes in Group B, while correct classification of unsuccessful athletes resulted in a 44 percent hit rate. Out of 22 analyses (11 scales × 2 groups), 18 resulted in higher correct classification rates for successful athletes compared to the unsuccessful athletes, and the difference was significant in 8 of these 18 analyses.

Discussion

The four intuitive and the two empirical scales developed in this investigation were all internally consistent, and the alpha reliability coefficient obtained for each scale was significant at the .05 level or greater.

While the empirical and intuitive scales were internally reliable, the empirical validity of these scales was not confirmed as the discriminant function analyses indicated. The classification of subjects improved when the empirical scales were used to classify subjects in the groups from which each scale was itself derived, but these results were not replicated during efforts to cross-validate the scales. The empirical, intuitive, and randomly constructed scales all failed to improve upon a 50% base-rate classification. Consequently, the original hypothesis of this study, which stated that a reliable and valid athletic potential scale could be constructed using items from the MMPI, was only partially supported. This was in terms of the internal reliability of the scales.

Two important points need to be considered in conjunction with this finding. First, an important measurement issue is highlighted by the finding that a scale can be developed which is *reliable* but not *valid*. Second, personality studies in sport psychology have at times employed single- rather than double-cross validation designs. As example, a study directly related to this investigation used an abbreviated or short form of the MMPI to assess the personality characteristics of athletes (Booth, 1958). However, double cross-validation was not used during the development of the MMPI short form. If a double-cross validation design had not been used in this investigation, the results would have been very different. The empirical scales were effective in discriminating between criterion groups when used to classify subjects in the groups from which the scales were derived (single validation). Only after attempts were made to cross validate the empirical scales was it concluded that the scales were not effective in discriminating between successful and unsuccessful athletes.

An additional, noteworthy outcome of this investigation is that interjudge reliability coefficients of agreement suggest that the four «intuitive» groups of judges generally agreed upon athletic stereotypes and on the items from the 550 MMPI item pool that had the potential to discriminate between successful and unsuccessful athletes. In other words, judges differing in athletic experience as well as psychological training were in substantial agreement as to which items had the potential to be effective in discriminating between the criterion groups. Why, then, were the intuitive scales, developed by judges who generally agreed on item selection, no more effective than chance, or a scale developed by randomly selected items? Likewise, why were the empirically developed scales equally ineffective? While definitive answers to these questions cannot be provided, several possible reasons for the scales' ineffectiveness exist.

First, it is possible that fragmentation of the MMPI item pool simply

was not effective. That is, with regard to the construct «athletic potential», it may be that the «whole is better than its part» and that using global MMPI scale scores in a multivariate analysis might have enhanced the correct classification rate of athletes in this investigation. While it was hypothesized that an analysis of individual MMPI items would lead to the development of an effective Athletic Potential Scale, the results were less gratifying than those obtained in previous research using the *intact* MMPI validity and standard scales to discriminate between successful and unsuccessful athletes (LaPlace, 1954; Booth, 1958; Morgan & Johnson, 1977, 1978; Johnson & Morgan, 1981). This finding takes on additional importance when one considers that subjects used in some of the earlier research (i.e., Morgan & Johnson, 1977, 1978; Johnson & Morgan, 1981) were from the same pool as the subjects used in this investigation. In addition, the MMPI scales used in the post hoc phase of this study tended to yield slightly higher percentages of correctly classified subjects compared to the percentages obtained using the intuitive, random, and cross-validated empirical scales. Also, the Pd paired with Sc scales correctly classified more subjects than any single scale. This evidence lends support for the possibility that, with regard to the construct athletic potential, the intact MMPI may be more effective than any MMPI short form, and that a multivariate approach may be preferable to a univariate model when attempting to predict athletic success. These findings also provide some support to those sport psychologists who recommend the use of a multivariate and psychophysiological model to describe, explain, and predict athletic behavior (Morgan, 1973; Hatfield & Landers, 1983).

A second alternative explanation for the findings obtained in this study can be advanced. It may be that the MMPI items simply cannot predict in an effective manner who will experience athletic success. Such a possibility may be related to two issues; one issue about conducting personality assessment with «normal» subjects, and a second issue about the criteria or domains of personality that are sampled by the MMPI item pool.

Noting that the MMPI is a measure of psychopathology, Butcher and Tellegen (1978) have stated that the MMPI is often erroneously thought to be a comprehensive personality assessment instrument that is sensitive to normal-range personality traits. Rather than using the MMPI to assess subjects from the general population, Butcher and Tellegen suggest that it would be better to consider other measures, such as the California Psychological Inventory (CPI), which focus upon normal-range personality characteristics. With regard to the issue of psychopathology versus normal range personality assessment, it is emphasized that the present investiga-

tion did not address this point. Subjects having pathology or the absence of pathology were not identified as criterion groups, nor was psychopathology used to predict athletic potential.

It is also important to recognize that a certain portion of individuals in selected «normal» groups will manifest «abnormal» profiles. If the purpose of a given investigation involves some aspect of psychopathology in the «normal» population, then the MMPI might be quite useful. The question being asked should govern the selection of psychometric tools. Thus, arguments can be made against using the MMPI, a measure of psychopathology, with an athlete population, especially if it is assumed that all athletes are psychologically «normal» and mentally healthy. On the other hand, Morgan (1980, 1985) discussed research conducted with both elite and college athletes which has led to the formulation of a mental health model that predicts athletic success based on the presence or absence of positive mental health characteristics. «A mental health model has been found to be effective in predicting success in athletics, and the model specifies that psychopathology and success are inversely proportional» (Morgan, 1980, p. 62). This model is testable. For example, the value of the MMPI may be in using it as one diagnostic tool to help identify those athletes experiencing emotional or mental difficulties. Once identified, successful or unsuccessful performance could be defined, predicted and subsequently documented.

The second issue relates to item ineffectiveness and pertains to the domains of personality that are assessed by the MMPI item pool. An argument can be made that the empirical and intuitive scales developed in this study were not effective because the MMPI item pool does not tap the domain of «athletic potential». This is a viable explanation, but the question then becomes whether any other test, including tests that measure normal-range personality characteristics, can assess «athletic potential». There is currently no empirical support for such a valid and reliable measure. Again, perhaps the value of the MMPI is not as a measure of athletic potential, but rather as a measure of identifying problematic athletes who may subsequently experience performance difficulties. To date, studies in sport psychology that have used the MMPI to predict athletic behavior have been descriptive and correlational in nature, and have not predicted performance based on the absence or presence of psychopathology (e.g., depressive symptomatology).

Yet a third possible explanation for the findings in this study can be advanced. The series of post hoc discriminant function analyses indicate that the failure of the scales to significantly discriminate above base-rate is primarily due to the ineffectiveness of the MMPI items to correctly clas-

sify unsuccessful athletes. This was especially true when the MMPI validity and standard scales were used. While the total, combined correct classification percentages obscure this finding, there was a discrepancy between a higher percentage of correctly classified successful athletes and a low percentage of correctly classified unsuccessful athletes. The inability to correctly classify unsuccessful athletes largely contributed to the many non-significant overall results obtained in this study. This suggests that using MMPI data to predict athletic success may lead to relatively accurate predictions. The prediction of failure, on the other hand, may be more difficult and, at least in terms of using MMPI data, could result in the identification of several athletes who will be «false negatives». That is, such athletes will be athletes who succeed despite a prediction that they will not. The dangers of this situation should be readily apparent and the findings from this investigation argue against using personality data for making «cuts» with regard to player selection.

A fourth explanation which could possibly account for the findings of this investigation is that the definitions of successful and unsuccessful athletes were not precise enough to achieve discrimination between the two groups using items from the MMPI. That is, college athletes who earned a freshman numeral and two or three varsity letters, and those athletes who only earned a freshman numeral, may be more alike than unalike. However, if the criterion groups were homogeneous in nature, it seems reasonable to expect that using selected MMPI scales to discriminate between groups would result in equivalent findings for both groups. As noted previously, even though the total correct classification rate was not significantly improved above base-rate, some of the MMPI scales correctly classified successful athletes significantly better than unsuccessful athletes.

Recommendations for future study

Four possible explanations for the findings that resulted from this investigation have been advanced. While the explanations are speculative, the implication of each is the same with regard to the future use of item selectors versus item writers for the construction of personality tests. The capabilities of *item selectors* to develop effective personality scales, even where there is a high degree of agreement between selectors, may be hindered by the limitations of the relationship between a specific item pool and the defined criterion of interest, rather than by the decision making capabilities of the item selectors. Obviously the potential of *item writers*

to develop effective personality scales would not be restricted to any particular item pool, but rather by the writers' own creative limitations. It is therefore recommended that, if item selectors are used to choose items for inclusion on personality scales, they also be afforded an opportunity to write items. This should enhance the potential effectiveness of any new instrument being developed, using the intuitive method of test construction.

While there has been a call for the construction of sport specific personality inventories (Kroll, 1976; Martens, 1975; Rushall, 1970; Singer, Harris, Kroll, Martens and Schrest, 1977), little attention has been focused on the actual development of new scales (Brown, 1989). Sport psychologists who are constructing new measures will be in an ideal position to further evaluate different test construction strategies. For example, one question is whether the intuitive test construction strategy can be used to develop a new sport/athlete specific personality scale that is any more effective than currenctly existing objective personality measures. This would require item writers or item selectors to spend a few hours developing items for inclusion on the new personality scale. Perhaps items could be selected from a «normal range» personality inventory to avoid the controversy surrounding the use of an item pool which measures psychopathology. On the other hand, perhaps psychopathology may be of central importance to the scale developer. Whether an existing item pool is used or not, there is evidence than item writers are capable of developing valid and reliable tests with only a few hours of work invested (Burisch, 1984). It is recommended that, for future scale development in sport psychology, the intuitive approach be evaluated to determine if it can be used effectively in exercise science and sport contexts. In other words, can intuitively derived scales be made in a short period of time that are at least as effective as empirically and internally derived scales?

The results of this investigation indicate quite clearly that differences in personality, at least as measured by a short form of the MMPI, cannot predict athletic success or failure. After many years of research, perhaps the safest conclusion that can be made about personality assessment in sport psychology is the following. Subtle individual differences in personality, or differences in personality between athletes that fall within a normal or average range, should not be used for predicting athletic success or failure in terms of player selection. The findings of this investigation suggest that it may be especially difficult to accurately predict failure among athletes when relying on such differences. For this reason alone, tests of psychopathology may remain attractive for use with athletes on occasion, in that they hold the potential as diagnostic tools to distinguish between ex-

tremes: athletes experiencing mental health problems or emotional disturbance from those who do not. The current state-of-the-art with regard to personality assessment in sport contexts suggests that the ability to make such determinations may ultimately lead to the most accurate predictions of sport performance.

RÉSUMÉ

L'approche empirique et intuitive à la construction d'un test ont fait l'objet d'une comparaison pendant la construction d'une échelle pour distinguer entre les athlètes gagnants et les non-gagnants.

On a développé 8 échelles utilisant l'item du Minnesota Multiphasic Personality Inventory (MMPI). Deux échelles ont été réalisées par l'approche empirique, tandis que cinq échelles ont été utilisées par l'emploi de l'approche intuitive et une se basant sur une sélection random des items du MMPI. Les résultats indiquent les bons niveaux de confiance pour les échelles réalisées par l'approche empirique et intuitive, tandis que toute les échelles se révélées inefficaces pour déterminer le potentiel athlétique. Aussi, les analyses ont indiqué que la classification des athlètes gagnants était plus précise que la classification des athlètes non-gagnants. Les résultats sont interprétés en termes de caractéristiques concernant le pool de item du MMPI et de zones examinées par le MMPI. On donne des indications sur le développement à venir des échelles en matière de psychologie du sport.

RESUMEN

Durante el trabajo de construcción de una escala para discriminar entre atletas vencedores y no-vencedores, han sido comparados el acercamiento empírico y aquel intuitivo para la construcción de un test. Se han desarrollado 8 escalas utilizando item del Minnesota Multiphasic Personality Inventory (MMPI). Dos escalas han sido construidas empleando el acercamiento empírico, cinco de ellas sirviéndose del acercamiento intuitivo y una escala basándose en una selección random de los item del MMPI. Los resultados han puesto en evidencia buenos niveles de confiabilidad para las escalas construidas utilizando el acercamiento empírico y aquel intuitivo, mientras que todas las escalas se han revelado ineficaces para medir el potencial atlético. Además, los análisis han puesto en claro que la diversificación de atletas vencedores estaba hecha con mayor precisión de la clasificación de atletas no-vencedores. Se interpretan los resultados en términos de características inherentes al pool de item del MMPI y de áreas evaluadas por el MMPI. Se ponen indicaciones referentes al desarrollo futuro de escalas en sicología del deporte.

ZUSAMMENFASSUNG

Der offensichtliche und der empirische Ansatz zur Konstruktion eines Tests ist mit der Konstruktion einer Skala verglichen worden, die während der Auswertung von gewinnenden und nicht gewinnended Athleten hergestellt wurde.

Es wurden acht Skalen angefertigt, indem man das Item des Minnesota Multiphasic Personality Inventory (MMPI) benutzte. Zwei Skalen wurden mit dem empirischen Ansatz hergestellt, fünf Skalen mit dem offensichtlichen Ansatz und eine Skala mit der randomisierten Auswählung der Items des MMPI. Die Ergebnisse haben eine gute Zuverlässigkeit der Skalen bewiesen, die den empirischen und den offensichtlichen Ansatz benützt haben, während alle Skalen sich gegenüber der Bewertung der athletischen Leistungsfähigkeit als unwirksam gezeigt haben. Die Studie hat außerdem gezeigt, daß die Einstufung der gewinnenden Athleten präziser war als die Einstufung der nicht gewinnenden Athleten. Die Ergebnisse wurden in Verbindung mit dem Pool der Items des MMPIs und in Verbindung mit den Gebieten, mit denen sich das MMPI befaßt, ausgelegt. Es werden Hinweise gegeben, wie man in Zukunft Skalen der Sportpsychologie weiterentfalten könnte.

RIASSUNTO

Durante il lavoro di costruzione di una scala per effettuare una discriminazione tra atleti vincenti e non-vincenti sono stati confrontati gli approcci empirico e intuitivo alla costruzione di un test.

Sono state sviluppate 8 scale utilizzando item del Minnesota Multiphasic Personality Inventory (MMPI). Due scale sono state costruite utilizzando l'approccio empirico mentre cinque scale sono state costruite servendosi dell'approccio intuitivo e una scala è stata costruita basandosi su una selezione random degli item dell'MMPI.

Dai risultati emergono buoni livelli di affidabilità per quanto concerne le scale costruite utilizzando l'approccio empirico e quello intuitivo, mentre tutte le scale si sono rivelate inefficaci come misure del potenziale atletico.

Le analisi hanno inoltre evidenziato che la classificazione di atleti vincenti era fatta con maggior precisione rispetto alla classificazione di atleti non-vincenti.

I risultati vanno interpretati in termini di caratteristiche inerenti al pool di item dell'MMPI e di aree valutate dall'MMPI e si raccomanda che in futuro vengano sviluppate le costruzioni di scale applicate alla psicologia dello sport.

REFERENCES

Ashton, S.G., & Goldberg, L.R. (1973) In response to Jackson's challenge: the comparative validity of personality scales constructed by the external (empirical) strategy and scales developed intuitively by experts, novices and laymen. *Journal of Research in Personality, 7,* 1-20.

Blaser, P., & Schilling, G. (1976) Personality tests in sport. *International Journal of Sport Psychology, 7,* 22-35.

Booth, E.G., Jr. (1958) Personality traits of athletes as measured by the MMPI. *Research Quarterly, 29,* 127-138.

Brown, D.R. (1989) Test development in sport psychology. Manuscript submitted for publication.

Burisch, M. (1984) Approaches to personality inventory construction: A comparison of merits. *American Psychologist, 39,* 214-227.

Butcher, J.N., & Tellegen, A. (1978) Common methodological problems in MMPI research. *Journal of Consulting and Clinical Psychology, 46,* 620-628.
Downie, N.M., & Heath R.W. (1970) *Basic statistical methods* (3rd ed.). New York: Harper and Row Publishers.
Drake, L.E., & Oetting E.R. (1959) *An MMPI codebook for counselors.* Minneapolis: University of Minnesota Press.
Ebel, R.L. (1951) Estimation of the reliability of ratings. *Psychometrika, 16,* 407-424.
Hase, H.D., & Goldberg, L.R. (1967) Comparative validity of different strategies of constructing personality inventory scales. *Psychological Bulletin, 67,* 231-245.
Hatfield, B.D., & Landers D.M. (1983) Psychophysiology - A new direction for sport psychology. *Journal of Sport Psychology, 5,* 243-249.
Hull, C.H., & Nie, N.H. (1981) *SPSS update 7-9.* New York: McGraw-Hill Book Company.
Jackson, D.N. (1971) The dynamics of structured personality tests. *Psychological Review, 78,* 229-248.
Johnson, R.W., & Morgan, W.P. (1981) Personality characteristics of college athletes in different sports. *Scandinavian Journal of Sports Science, 3,* 41-49.
Kroll, W. (1976) Reaction to Morgan's paper: psychological consequences of vigorous physical activity and sport. In M.G. Scott (Ed.), *The Academy Papers.* Iowa City: American Academy of Physical Education.
LaPlace, J.P. (1954) Personality and its relationship to success in professional baseball. *Research Quarterly, 25,* 213-319.
Martens, R. (1975) The paradigmatic crisis in American sport personality. *Sportwissenschaft, 5,*9-24.
McNemar, Q. (1962) *Psychological statistics* (3rd ed.), New York: Wiley.
Morgan, W.P. (1973) Efficacy of psychobiological inquiry in the exercise and sport sciences. *Quest, 20,* 39-47.
Morgan, W.P. (1980) The trait psychology controversy. *Research Quarterly of Exercise and Sport, 51,* 50-76.
Morgan, W.P., & Johnson, R.W. (1977) Psychologic characterization of the elite wrestler: A mental health model. *Medicine and Science in Sports, 9,* 55-56 (Abstract).
Morgan, W.P., & Johnson, R.W. (1978) Personality characteristics of successful and unsuccessful oarsmen. *International Journal of Sport Psychology, 9,* 119-133.
Morgan, W.P., Brown, D.R., Raglin, J.S., O'Connor, P.J., & Ellickson, K.A. (1987) Psychological monitoring of overtraining and staleness. *British Journal of Sports Medicine, 21,* 107-114.
Rushall, B.S. (1970) An evaluation of the relationship between personality and physical performance categories. In G.S. Kenyon (Ed.), *Contemporary psychology of sport.* Chicago: Athletic Institute.
Singer, R.N., Harris, D., Kroll W., Martens, R. & Schrest, L. (1977) Psychological testing of athletes. *Journal of Physical Education and Recreation, 48,* 30-32.
Slusher, H.S. (1964) Personality and intelligence characteristics of selected high school athletes and nonathletes. *Research Quarterly, 35,* 539-545.