Procedures for Assessing the Validities of Tests Using the "Known-Groups" Method

John Hattie and Ray W. Cooksey
University of New England

If a test is "valid," one criterion could be that test scores must discriminate across groups that are theoretically known to differ. A procedure is outlined to assess the discrimination across groups that uses only information from means. The method can be applied to many published tests, it provides information that relates to the construct validity of the test, and it presents a way to identify how a new sample can be related to previous studies.

Validity is an elusive concept, and statistical methods related to the validity of tests are noted by their paucity. Validity has been defined in a variety of ways, for example, in terms of truthfulness (Mehrens & Lehmann, 1975), in terms of methods of investigating interpretations (Cronbach, 1971), and in terms of the appropriateness of inferences from test scores (American Psychological Association, 1974). Validity refers not to a measuring instrument but to the purpose for which the instrument is used.

According to the APA Standards for Educational and Psychological Tests (APA, 1974), a test manual

> can provide evidence that will enable the user to evaluate the appropriateness of the item content, to determine whether the test is an acceptable measure of a specified construct, and to decide whether the test has provided useful predictive validities in situations similar to his own (p. 31).

Thus, there are three major purposes of an instrument: (1) to sample a domain of content, (2) to measure some psychological trait, and (3) to determine relationships with other variables. Related to these three purposes are three types of validity: content, construct, and criterion validity.

In an early and seminal paper on the validity of psychological tests, Cronbach and Meehl (1955) discussed various methods for experimentally investigating validity, particularly construct validity. One of their methods was the known-groups method. "If our understanding of a construct leads us to expect two groups to differ on the test, this expectation may be tested directly" (Cronbach & Meehl, 1955, p. 287). Thus if a test is "valid," one criterion could be that test scores should discriminate across groups that theoretically are expected to be different on the trait measured. For example, a test of self-actualization should be able to discriminate between groups of counselors and psychiatric patients, or between persons before and after they have been to encounter groups. If this is so, there is evidence of the usefulness of the test as a decision-making instrument and evidence that it can be generalized on a meaningful psychological trait across different samples of people.

The known-groups method has been used only by a relatively small number of researchers (Ho-

APPLIED PSYCHOLOGICAL MEASUREMENT
Vol. 8, No. 3, Summer 1984, pp. 295–305

0146-6216/84/030295-11$1.80

gan, 1975a, 1975b; Pettegrew & Wolf, 1981; Rest, 1974, 1976, 1977; Rhoads & Landy, 1973; Smith & Apfeldorf, 1975). Perhaps the most comprehensive use of the method has been by Rest and his coworkers (Rest, 1974, 1976, 1977; Rest, Cooper, Coder, Masanz, & Anderson, 1974) in validating the Defining Issues Test (DIT). In the DIT a dilemma is presented such as the "Heinz and the Drug" dilemma used extensively by Kohlberg: Heinz's wife is dying of cancer and a chemist has a drug that might save her, but the chemist is charging an exorbitant price for the drug and Heinz cannot raise the money. Should he steal the drug in an attempt to save his wife? Following the dilemma, 12 statements are presented that express various considerations or questions in making a decision. The task is to decide which considerations or questions are crucially important and which are not.

Rest argued that more educated persons should generally have higher scores on the DIT than less educated subjects. From 66 studies, Rest found that there were indeed differences:

Junior high, $N = 1{,}322$, Mean = 21.9, $SD = 8.5$;
Senior high, $N = 581$, Mean = 31.8, $SD = 13.5$;
College, $N = 2{,}479$, Mean = 42.3, $SD = 13.2$;
Graduates, $N = 183$, Mean = 53.3, $SD = 10.9$; and
Adults, $N = 1{,}149$, Mean = 40.0, $SD = 16.7$.

From this and other comparisons, Rest found that education and IQ had the most consistent relations to the DIT. The relations were in the expected directions; thus Rest concluded that this evidence proffered much support for the validity of the DIT.

This example (and other examples referenced) have all assumed a priori differences between groups. Clearly, if the theory on which the test is based enables in advance the prediction of which groups will be differentiated by the test, then that seems to offer more support for construct validity than showing that after the fact, groups can be found that have different scores on the test.

Typically, the procedure in analyzing differences between known groups has been to collect raw data from a number of groups. Given raw data from a number of groups there are few problems in assessing the discrimination across groups. Yet it is rare to have access to samples of sufficient size from differing groups. Recently, there has been a surge of interest in aggregating data across many samples (e.g., meta analysis, Glass, 1977). These methods, however, are most appropriate as a means of aggregating data from studies that use different measures where all the means and variances are arbitrary and diverse. This problem does not exist when collating data from a single dependent measure, that is, from one test.

It is possible a priori to form meaningful groups from a large collection of samples and ask whether the means differ between these groups. For example, many samples of summary statistics from a test of self-actualization could be meaningfully grouped into subsets such as counselors, self-actualizers, meditators, criminals, pre- and post-encounter groups, fakers, students, and disturbed individuals. Differences could be hypothesized between these groups and the significance of these differences could be assessed using ANOVA (and planned comparisons). From summary statistics such as sample size, means, and standard deviations, analysis of variance statistics can be computed (Burrill, 1971; Gordon, 1973; Huck & Malgady, 1978).

However, there are problems with the use of a priori groups. Scott (1968) claimed that it is misleading to infer the extent of validity from the significance level of a statistical test between two means. Such tests are substantially affected by the size of groups and by the way in which the samples are selected. Very large mean differences can be obtained for instruments that have little predictive value simply by an opportunistic selection of the known groups. Scott recommended that the known-groups method should be based on representative samples of the population to which the instrument will be applied.

The purpose of this paper is to describe a procedure based on the known-groups method that attempts to avoid many of the problems of the a priori method.

A Suggested Procedure for Establishing Validity Using the Known-Groups Method

There are four steps in the suggested procedure:

1. From a review of literature, a large number of means are gathered on the test in question (or on the subtests within the test). These sets should come from a variety of studies, and the samples should reflect the kinds of groups to which the test intends to be generalized. With recent advances in computer searching of the literature (see Geahigan & Geahigan, 1982), it should not be too difficult to obtain a large set of means.
2. Distances between the group means are calculated using a Euclidean distance metric. Multidimensional scaling can then be used to represent the distances spatially. From this scaling, it is possible to assess the dimensionality of the samples and, if the test is discriminating among groups, it should be possible to identify the dimensions in a meaningful way. Certainly, it may be possible to look at the groups at the polar extremes of the dimension(s) to determine whether the dimensions are scaling the groups in the desired direction.

   Most likely, there will be much overlap between various "clearly" identified groups (simply because the groups cannot be uniquely classified). A box-plot (Tukey, 1977) of the coordinates of these groups can be used to assess whether they are reasonably homogeneous.

   The researcher could stop at this point having obtained substantial evidence as to the adequacy of the test to make meaningful distinctions between groups to which the test purports to relate. Yet there are two profitable further steps.
3. Independent of the above steps, meaningful and easily identifiable a priori group labels can be presented as pairwise stimuli to persons knowledgeable of the domain tapped by the test. The labeling of the groups may not always be easy. In such cases some consensus among various judges may be needed. The prime concern at this step, however, is to use meaningful groups that are expected a priori to fall at various points on the dimension purportedly measured by the test. All possible pairings of the groups can be presented and each person asked to first identify the group expected to score higher, on the average, on the test. Second, they are asked to indicate the degree of preference for the so-identified "higher" group. The indications of degree, with appropriate sign, then can be scaled using an individual differences model (cf. Coxon, 1982; Davison, 1983; Schiffman, Reynolds, & Young, 1981).

   From an inspection of the dimension(s), it is possible to determine, at a minimum, a rank ordering of the groups. This rank order can be correlated with the rank order from Step 2. Of course, there is much more information in the scaling solution and it is possible, for example, to correlate the mean coordinates from Step 2 with the coordinates from the scaling of this step. If this correlation is very high (and at minimum, significantly different from zero), this would suggest that the test can meaningfully scale groups in an expected manner.
4. If the scaling solution of Step 2 indicates that the test has dependable interpretations, it is possible to devise methods so that users of the test can determine where their new group should be placed in relation to other groups. If all the group means are available, it is possible to directly place a new group into the scaling solution described above. If the means are not available, it is possible to cluster analyze the coordinate values from the dimension(s) into a number of groups. For each cluster the means of the (sub)tests can be calculated. Then the squared distance between the means of each cluster (preferably standardized using total-group standard deviations so that all subtests are expressed in the same units) and the means of the new group can be determined. The new sample can then be located in that cluster where the squared distance is minimum.

Thus from the first three steps, the test user can determine how a test discriminates across many groups and whether the resulting pattern of discrimination is meaningful. From the fourth step,

the user can determine where a new sample can be located relative to other groups. These procedures directly relate to the validity of a test in terms of whether a test can make appropriate inferences across many groups, and the procedures allow indications of the underlying dimension(s) of the test.

An Example: The Personal Orientation Inventory

The Personal Orientation Inventory (POI; Shostrom, 1974) purports to measure aspects of mental well-being or self-actualization. It consists of 150 pairs of alternative value judgments that are scored on 2 major and 10 subsidiary scales. These scales measure inner directedness, time competence, self-actualizing, existentiality, feeling reactivity, spontaneity, self-regard, self-acceptance, nature of man, synergy, acceptance of aggression, and capacity for intimate contact.

Step 1

From a review of the literature, 107 samples were collected from both published and unpublished studies (full references and all data are available upon request from the authors). These samples came from a cross section of samples and are based on 11,001 persons.

Step 2

The Euclidean distances between the means on the 12 POI scales across the 107 groups were input into the ALSCAL program (Young & Lewyckyj, 1979). A classical nonmetric scaling solution was used. There was one matrix (107 × 107), the measurement level was specified as ordinal, and a simple Euclidean model was used. (If there was only one and not multiple scales in the test, then a one-dimensional scaling solution would perfectly reconstruct the distances.)

There are various ways to assess dimensionality. The most commonly used methods are to report values for stress and R-squared. The stress value is the square root of the normalized residual sum of squares. The R-squared indicates the proportion of variance of the disparities that is accounted for by the multidimensional scaling model. While Kruskal prefers stress (see Kruskal & Wish, 1978), Young prefers R-squared (see Schiffman et al., 1981; Young & Lewyckyj, 1979). From these measures, a one-dimensional solution was chosen as providing the best fit.

The stress value was .11, which falls below Kruskal and Wish's (1978, p. 54) suggested criterion. The R-squared value was very large—.97. Thus, a large amount of variance was accounted for (in all the disparities) by one dimension.

Table 1 presents the 107 samples ranked according to the stimulus coordinates. It is reasonably clear that there is a pattern in the ordering of the groups. Those that would be expected to be more self-actualized are ranked higher than less self-actualized groups. The positive end is identified by groups of persons that have been classified as high genuine counselors and by those who have been to encounter groups, advanced therapy, or who have had appropriate training in the behaviors measured by the POI. At the other end are those asked to fake the test or to appear "self-actualized," and disturbed persons. In between these extremes are groups of students, adults, those disposed to go to encounter groups, and prisoners.

From the order of the samples, it seems that it is more justifiable to divide students into university (average rank = 55, $N = 18$), college (average rank = 70, $N = 11$), and secondary school students (average rank = 86, $N = 7$). That there is a progression from secondary school through college, university, and adults (average rank = 44, $N = 17$) suggests that age may be a factor in self-actualization. When a two-dimensional solution was plotted, there were indications that the second dimension related to age. The pattern, however, was far from clear.

There were some interesting placements for some samples. For example, a group of persons from Alcoholics Anonymous is placed very high. This is probably because the program used in the treatment very much aims at increasing self-respect, self-awareness, and inner directedness.

From the 107 samples, it is possible to identify 10 groups:

adults;
disturbed persons (hospitalized psychiatric patients, practicing alcoholics, and neurotic persons);
those predisposed to attend encounter groups (includes those who have indicated a willingness to go to encounter groups);
postencounter groups (those given the POI after they have attended an encounter group);
university students;
college students;
secondary school students;
criminals;
fake "good" (those asked to try to fake the POI to appear as "self-actualized" people); and
those who have received training (this includes groups that have been given training related to self-actualization, such as counselors, and those given training in POI-related activities).

Figure 1 presents box-plots of the coordinates for these 10 groups. The center verticals are means, the length of the box is two standard deviations, and the length from the end of the box to the vertical dash (|) is one standard deviation.

Although there is much overlap between the groups, it is clear that many groups are dissimilar from each other. For example, those who had received training in self-actualized activities scored higher than criminals, college students, disturbed persons, and those who attempted to fake the test. College and secondary students are not too dissimilar, but both these groups are different from university students.

Most of the groups are reasonably homogeneous but some are not, such as those who had received training, postencounter groups, and disturbed persons. In the received-training group, one of the two groups that received training in POI-related skills had a mean coordinate much lower than the group mean ($z = -1.2$). The group of postencounter persons can be divided into two much more homogeneous groups depending on whether the participants were university (Mean = 1.54, $SD = .52$, $N = 5$) or college-secondary students (Mean = .06, $SD = .36$, $N = 4$). This is a good illustration of the problems of a priori grouping: different conclusions could have resulted depending on whether these groups were combined or separated.

Figure 1
Box Plots of the Coordinates from the Ten Groups

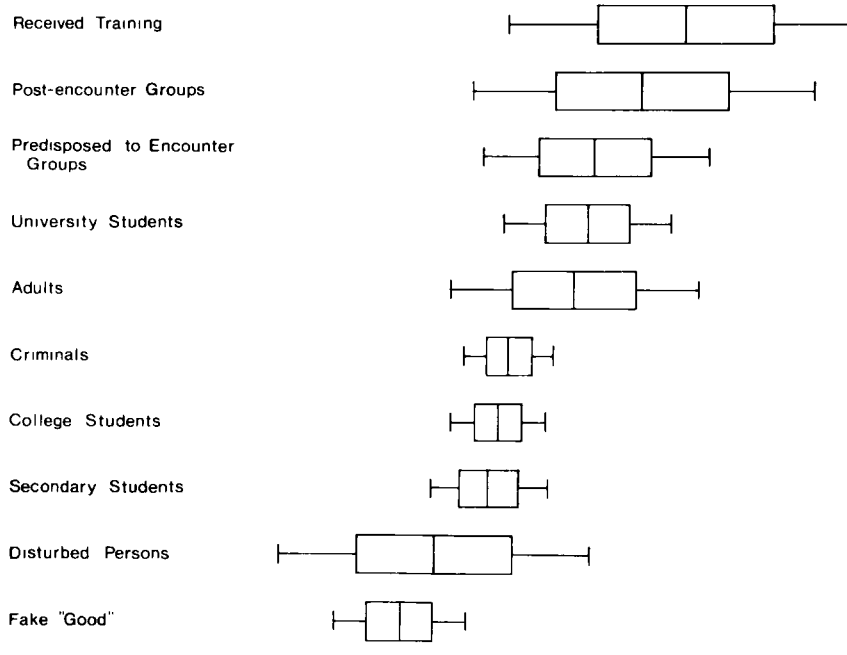

Table 1
Rank Order of MDS Coordinates For the 107 Samples

Order	Coordinate	Sample Size	Sample	Order	Coordinate	Sample Size	Sample
1	2.58	8	High Genuine	34	.44	86	University Students, etc. 2nd Administration
2	2.54	39	Advanced Therapy	35	.43	112	Navy pre-encounter
3	2.16	18	6 mth post-encounter	36	.43	227	College post-encounter
4	1.96	13	After POI training	37	.41	20	University students
5	1.81	7	Post-encounter	38	.39	40	University students
6	1.57	18	Post-encounter	39	.34	66	Parents
7	1.50	9	Post-meditation	40	.32	35	Low neurotic
8	1.43	26	Alcoholic Anonymous	41	.32	86	University students, 1st administration
9	1.41	33	Post-encounter	42	.31	14	University students
10	1.37	13	Executive	43	.28	60	Secondary post-encounter
11	1.29	15	Regular Meditators	44	.27	29	After POI training
12	1.26	29	Self-actualized	45	.27	84	Drug addicts
13	1.22	7	Pre-encounter	46	.17	39	Encounter types, male
14	1.10	54	Demonstrators	47	.16	86	University students, 1st administration
15	1.09	40	Encounter types, females	48	.11	53	University students
16	1.07	7	University Students, 2nd Administration	49	.08	59	Upper middle management
17	1.03	64	Childless parents	50	.07	18	Pre-encounter group
18	1.02	16	Educational consultants	51	.06	55	Telephone counselors, beginning of course
19	.91	86	University Students, 2nd Administration	52	.02	13	University students
20	.81	7	University Students, 1st Administration	53	.00	158	Non-demonstrators
21	.80	8	Low genuine	54	-.05	66	Nurse supervisors
22	.77	41	University Students	55	-.09	17	University students
23	.77	112	Navy post-encounter	56	-.10	132	University students
24	.72	23	Counselors	57	-.10	66	College pre-encounter
25	.70	33	University pre-encounter	58	-.12	22	Beginning meditation
26	.70	31	Phone counselors	59	-.12	56	Adults
27	.68	20	Merchant entrepreneurs	60	-.18	221	College Post-encounter
28	.65	120	Teachers	61	-.21	83	Supermarket Managers
29	.59	55	Telephone counselors, end of course	62	-.23	150	College students
30	.59	11	University Students	63	-.23	37	Beginning therapy
31	.53	158	Adults	64	-.28	150	Felons
32	.45	62	Peace Corp	65	-.29	56	Secondary students
33	.45	74	Trainee	66	-.31	399	Secondary post-encounter
				67	-.33	50	Secondary students

Table 1 (Continued)
Rank Order of MDS Coordinates For the 107 Samples

Order	Coordinate	Sample Size	Sample	Order	Coordinate	Sample Size	Sample
68	-.33	29	University students	89	-.84	816	College males
69	-.34	9	Pre-meditation	90	-.87	1254	College males
70	-.36	136	Adults	91	-.90	216	Secondary females
71	-.39	86	Supermarket Managers	92	-.96	25	Disturbed patients 4 months after treatment
72	-.44	41	College students				
73	-.46	64	Student nurses	93	-.99	11	Disturbed patients 1 year after treatment
74	-.47	278	College pre-encounter	94	-.99	412	Secondary students
75	-.49	41	Entering nurses	95	-1.08	38	High neurotic
76	-.49	100	Male prisoners	96	-1.09	196	Secondary males
77	-.57	96	Supermarket Managers	97	-1.27	20	Fake POI as "good"
78	-.62	408	College females	98	-1.28	70	Alcoholic treatment
79	-.62	24	Practising alcoholics	99	-1.31	29	Disturbed - immediate post treatment
80	-.62	19	University students				
81	-.65	76	Supermarket Managers	100	-1.46	17	Fake POI as "well adjusted"
82	-.69	761	Secondary students	101	-1.57	86	Fake POI as "good"
83	-.69	20	Craftsman entrepreneur	102	-1.65	185	Hospitalized Psychiatric
84	-.74	792	College females	103	-1.85	19	Fake POI as "good"
85	-.76	84	Navy criminals	104	-2.15	11	Fake POI as "good"
86	-.76	197	Secondary students	105	-2.19	12	Disturbed patients
87	-.77	34	Non-self actualized	106	-2.24	29	Disturbed patients
88	-.82	146	Male alcoholics	107	-2.76	12	Disturbed patients

Another example of the problems of a priori grouping relates to drug addicts. There was one sample of drug addicts and they were classified as disturbed persons, but the coordinate of the drug addicts was .27, which is much higher than the mean coordinate ($z = +1.91$) for disturbed persons. It seems that perhaps they may not be best classified as disturbed.

Figure 1 does illustrate that, overall, the POI can make distinctions between groups expected to differ in self-actualization. The above analyses of sample means indicate that a user of the POI can have much confidence in using it to make dependable decisions regarding the construct or self-actualization as it relates to specific groups of interest.

Step 3

Independently of the above steps, a group of 26 staff and advanced students in education were presented with all possible pairings of the 10 groups (adults, disturbed persons, predisposed to encounter groups, postencounter groups, university students, college students, secondary students, fake "good," criminals, and received training). They were asked to indicate their preference for one group over the other in terms of self-actualization, and also to indicate their view of the differences between the groups on a scale from 1 (hardly any difference) to 99 (completely different).

A nonmetric individual differences model was used to scale the 26 10 × 10 matrices. The scaled coordinates from the preferences of the 26 persons are presented in Table 2. The correlation between the rank ordering from this sample and the rank order of the groups' mean coordinates from Step 2 was .79 ($r = .86$ for the actual coordinates). Only two groups are misplaced in order by three positions. The 26 persons classified adults as higher and criminals lower than the test results. Generally, the correspondence is moderately strong and gives the user much confidence in the POI as a discriminator on the trait of self-actualization.

Step 4

The coordinate values from the one-dimension

Table 2
MDS Coordinates from the Preferences of 25 Persons

	Mean Coordinates	Rank Order from Scaling 107 Groups	Rank Order from Scaling 25 Persons
Received training	1.18	1	1
Post-ecounter group	.82	2	4
Predisposed to encounter group	.51	3	5
University students	.83	4	3
Adults	.88	5	2
Criminals	-1.12	6	9
College students	.47	7	6
Secondary students	-.79	8	7
Distrubed persons	-1.75	9	10
Fake "good"	-1.04	10	8

solution from Step 2 were then clustered into four groups using a modification of the ISODATA procedure (Ball, 1970; Blashfield & Aldenderfer, 1978; Cooksey, 1982). ISODATA yields successive nonhierarchical partitions of a sample into from 1 to 10 mutually exclusive clusters. The modification entailed using increase in eta-squared in a scree-type test as a criterion for establishing the most likely number of clusters. Four clusters seemed to best represent the composition of the sample (eta-squared = .89; a five cluster solution would only have added a trivial 3.8% additional explanation of differences in the coordinate values).

The first cluster contained 12 samples with a cluster mean of 1.85 (Groups 1 to 12 in Table 1); the second cluster had 34 samples with a mean of .64 (Groups 13 to 46), the third had 46 samples and a mean of −.47 (Groups 47 to 92), and the fourth had 15 samples and a mean of −1.40 (Groups 93 to 107).

To further investigate these results from the ISODATA clustering, it is prudent to work back to the individual POI scales to see how well they differentiated the four clusters of samples. Thus, the means on the 12 POI scales for the four clusters were computed. These means are presented in Table 3. The total group is based on 11,001 persons. Clearly, the clusters differ in terms of their degree

Table 3
Means of the 12 POI Scales Clustered Into Four Groups and the Total Sample Means and Standard Deviations

POI Scale	1 (n=12)	2 (n=34)	3 (n=46)	4 (n=16)	Total (n=107) $\bar{X}$	s.d
Time Competence	96.90	87.59	79.08	71.72	82.65	10.28
Inner Directedness	18.69	17.43	15.32	14.56	16.24	3.07
Self-Actualizing Values	21.76	20.38	18.84	17.41	19.44	2.98
Existentiality	25.28	21.64	19.03	15.96	20.09	4.15
Feeling Reactivity	17.98	16.11	14.57	13.13	15.22	2.91
Spontaneity	14.26	12.76	11.22	9.74	11.83	2.52
Self Regard	13.02	12.58	11.11	10.57	11.71	2.37
Self Acceptance	18.82	16.66	14.87	13.59	15.68	3.30
Nature of Man	12.81	12.12	11.30	11.11	11.70	2.07
Synergy	7.71	7.13	6.52	6.09	6.78	1.45
Acceptance of Aggression	18.29	16.47	15.35	13.89	15.81	3.12
Capacity for Intimate Contact	21.13	18.68	16.81	14.75	17.58	3.51

of self-actualization. There were significant differences between the four means on each scale, but this should not be too surprising as the clustering procedure aims to partition samples into groups so as to maximize group differences.

To ascertain where a new sample can be placed along the dimension is easy if the means from all the groups are available. The user merely reruns the scaling procedure adding in the new group. If all the means are not available, then the procedure for including a new group is not so straightforward. In this latter case, the new sample can be placed along the self-actualizing dimension by calculating, for each cluster separately, the squared Euclidean distance between the means from the new sample and the means in the cluster (standardized using the information for each scale in the Total column of Table 3). The minimum squared distance over the four clusters indicates in which of the clusters the new sample is most likely to be located.

For example, Osborne and Steeves (1982) presented means for a group of counselors who had completed a counseling practicum. These means were (in the same order as in Table 3) 101.9, 20.0, 23.1, 21.0, 18.8, 15.2, 14.4, 19.3, 12.6, 7.9, 18.7, 22.9. The squared distances were 2.6 for Group 1, 8.8 for Group 2, 23.3 for Group 3, and 40.1 for Group 4. Since the minimum squared distance is 2.6, this sample is closest to Group 1 and can be classified as a very self-actualized group. At the other end is a sample of nonmethadone-treated addicts (Cryns, 1974). The means were 74.4, 12.6, 16.6, 16.0, 14.0, 10.5, 7.5, 10.4, 9.9, 5.7, 12.9, 16.1. The squared distance values were 41.69, 21.07, 7.80, and 4.00, for the four clusters respectively. The minimum squared distance is from the last cluster, thus this sample is very low on the dimension of self-actualization.

Another group consisted of 36 YMCA administrators before and then after going to an encounter group (Reddy, 1973). Before attending the encounter groups, the administrators were classified into Group 2 (means: 88.1, 17.8, 20.4, 20.5, 16.0, 12.9, 12.4, 17.4, 12.2, 7.6, 16.8, 18.9, and squared distances of 4.08, .27, 4.71, 13.96), whereas after attending the encounter group, they were in the most self-actualized group (means: 94.8, 18.5, 21.1, 18.0, 14.1, 12.6, 18.1, 12.5, 8.1, 18.3, 21.1, and squared distances of .27, 3.56, 14.08, 29.11). Interestingly, six months later the administrators remained in the top group.

Conclusions

One of the most important characteristics of a test, if not the most important, is its validity. The quality of interpreting meaningful group differences very much depends upon evidence of the validity of the test.

One of the problems has been that there are very few empirical methods that assess the validity of a test, or more correctly, the validities of a test. The two most commonly used empirical methods are factor analysis (cf. Hattie, 1981 for an example of factor analysis and the POI), or multitrait-multimethod analysis (Campbell & Fiske, 1959; Watkins & Hattie, 1981). Another method first suggested by Cronbach and Meehl (1955) and termed the known-groups method has not been used very extensively.

As initially conceived and used, the known-groups method involved comparing groups that were theoretically expected to be different on the construct measured. It has been pointed out that there are problems with a priori groupings, such as using samples based on a small number of persons or using unrepresentative samples, and there are difficulties classifying samples into meaningful and homogeneous groups.

This paper has suggested an alternative method for assessing the validity of a test. This involves locating a large sample of means and then scaling the distances between the sets of means. The resulting dimension(s) should be interpretable and should be meaningful in terms of the construct(s) the test purports to measure.

A group of experts can be asked to scale various groups, and the rank order of the scaled coordinates can then be compared to the order from the scaling of the actual test means. Provided that the groups presented to the "experts" are reasonably homogeneous and clearly distinct from each other, there should be a close correspondence between these

orderings if the test is to be considered valid. This procedure is very similar to the procedure for naming factors that was presented in Hattie (1981). Finally, it was suggested that the information from the above procedures can be used to assign new groups to some point on the dimension(s) of the test.

These methods were illustrated using the POI. The means from 107 samples were found to scale along one dimension. Inspection of how the groups were ordered along the dimension indicated that groups expected to be more self-actualized were at the higher end of the dimension and low self-actualized groups were at the lower end of the dimension. There appears to be much evidence that the POI can discriminate between groups in a meaningful way. The ordering of groups along the dimension was very similar to the ordering from a group of experts. Further, after clustering the coordinates into four clusters, it was demonstrated how new groups could be assigned to one of these four clusters.

Overall, the results suggest that the POI appears to be reasonably valid. The POI seems to reflect an underlying construct of self-actualization and can be used to meaningfully discriminate between various groups of persons.

The method has been used here to demonstrate only one aspect of validity, and other procedures such as multitrait-multimethod analyses and factor analysis can provide additional information. Given the paucity of methods to assess validity, the method of discrimination between groups using multidimensional scaling and clustering is offered as an additional procedure for the test user to marshall evidence in support of claims of validity.

References

American Psychological Association. (1974). *Standards for Educational and Psychological Tests and Manuals*. Washington DC: Author.

Ball, G. H. (1970). *Classification Analysis*. Menlo Park CA: Stanford Research Institute.

Blashfield, R. K., & Aldenderfer, M. S. (1978). Computer programs for performing iterative partitioning cluster analysis. *Applied Psychological Measurement, 2,* 533–541.

Burrill, D. F. (1971). *Analysis of variance generalized: Parameters other than the mean*. Paper presented to Canadian Educational Research Association, Newfoundland.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Cooksey, R. (1982). *A modified version of the ISODATA program*. Unpublished manuscript, University of New England, Centre for Behavioural Studies in Education, Armidale, Australia.

Coxon, A. P. M. (1982). *The user's guide to multidimensional scaling*. Exeter NH: Heinemann.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational Measurement* (2nd. ed., pp. 443–507). Washington DC: American Council on Education.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52,* 281–302.

Cryns, A. G. (1974). Personality characteristics of heroin addicts in a methadone treatment program: An exploratory study. *The International Journal of Addictions, 9,* 255–266.

Davison, M. L. (1983). *Multidimensional scaling*. New York NY: Wiley.

Geahigan, C., & Geahigan, P. (1982). Using computers to search the educational literature: A primer. *Contemporary Education Review, 1,* 179–193.

Glass, G. V. (1977). Integrating findings: The meta-analysis of research. In L. S. Shulman (Ed.), *Review of Research in Education* (Vol. 5, pp. 351–379). Itasca IL: Peacock Publications.

Gordon, L. U. (1973). One-way analysis of variance using means and standard deviations. *Educational and Psychological Measurement, 33,* 77–88.

Hattie, J. A. (1981). A four-stage factor analysis approach for studying behavioral domains. *Applied Psychological Measurement, 5,* 77–88.

Hogan, H. W. (1975a). Test of the validity of the Wilson-Patterson conservatism scale. *Perceptual and Motor Skills, 40,* 795–801.

Hogan, H. W. (1975b). Validity of a symbolic measure of authoritarianism. *Psychological Reports, 37,* 539–543.

Huck, S. W., & Malgady, R. G. (1978). Two-way analysis of variance using means and standard deviations. *Educational and Psychological Measurement, 38,* 235–237.

Kruskal, J. B., & Wish, M. (1978). *Multidimensional Scaling*. Beverly Hills CA: Sage.

Mehrens, W. A., & Lehmann, I. J. (1975). *Measurement and Evaluation in Educational Psychology* (2nd Ed.). New York: Holt, Rinehart & Winston.

Osborne, J. W., & Steeves, L. (1982). Counseling practicum as a facilitator of self-actualization. *The Alberta Journal of Educational Research, 28,* 248–256.

Pettegrew, L. S., & Wolf, G. E. (1981). *Validating*

Measures of Teacher Stress. Nashville TN: George Peabody College for Teachers. (ERIC Document Reproduction Service No. ED 213 743)

Reddy, W. B. (1973). The impact of sensitivity training on self-actualization: A one-year follow-up. *Small Group Behavior, 4,* 407–413.

Rest, J. (1974). *Manual for the Defining Issues Test: An Objective Test of Moral Judgment Development*. Minneapolis MN: University of Minnesota.

Rest, J. (1976). *Moral Judgment Related to Sample Characteristics*. (NIME Report No. 24988). Minneapolis MN: University of Minnesota, College of Education, Department of Educational Psychology.

Rest, J. (1977). *Development in Judging Moral Issues—A Summary of Research Using the Defining Issues Test*. (Technical Report No. 3). Minneapolis MN: University of Minnesota, College of Education, Department of Educational Psychology.

Rest, J. R., Cooper, D., Coder, R., Masanz, J., & Anderson, D. (1974). Judging the important issues in moral dilemmas—An objective measure of development. *Developmental Psychology, 10,* 491–501.

Rhoads, R. F., & Landy, F. J. (1973). Measurement of attitudes of industrial workgroups towards psychology and testing. *Journal of Applied Psychology, 58,* 197–201.

Scott, W. A. (1968). Attitude measurement. In G. Lindzey & E. Aronson (Eds.), *The Handbook of Social Psychology* (Vol. 2, pp. 204–273). Reading MA: Addison-Wesley.

Schiffman, S. S., Reynolds, M. L., & Young, F. W. (1981). *Introduction to Multidimensional Scaling: Theory, Methods and Applications*. New York: Academic Press.

Shostrom, E. L. (1974). *Manual for the Personal Orientation Inventory*. San Diego CA: Educational and Industrial Testing Service.

Smith, W. J., & Apfeldorf, M. (1975). Scales which measure behavioral reactions to illness during hospitalization and attitudes towards hospitals. *Psychological Reports, 36,* 719–724.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading MA: Addison-Wesley.

Watkins, D., & Hattie, J. A. (1981). An investigation of the constructs validity of three recently developed personality instruments: An application of confirmatory multimethod factor analysis. *Australian Journal of Psychology, 33,* 227–284.

Young, F. W., & Lewyckyj, R. (1979). *ALSCAL-4 User's Guide*. Chapel Hill NC: Data Analysis and Theory Associates.

Acknowledgment

Preparation of this article was supported by a grant to the first author from the Australian Research Grants Committee.

Author's Address

Send requests for reprints or further information to John Hattie, Centre for Behavioural Studies, University of New England, Armidale, N.S.W., Australia, 2351.