Dawes, R. M., & Meehl, P. E. (1966). Mixed group validation. *Psychological Bulletin*, *66*, 63-67.



# MIXED GROUP VALIDATION:

## A METHOD FOR DETERMINING THE VALIDITY OF DIAGNOSTIC SIGNS WITHOUT USING CRITERION GROUPS[1]

ROBYN MASON DAWES
*Veterans Administration Hospital, Ann Arbor, and University of Michigan*

AND

PAUL E. MEEHL
*University of Minnesota*

To determine whether a sign *x* distinguishes between *X*-type individuals and *Y*-type individuals (*Y* = not-*X*), it is necessary to obtain estimates of $p(x/X)$ (the probability an *X*-type shows sign *x*) and $p(x/Y)$. The traditional method for determining these probabilities, criterion-group validation, involves observing the incidence of *x* in one group consisting entirely of *X*-types and in a second group consisting entirely of *Y*-types. Here, equations are developed for obtaining these probabilities by observing the incidence of *x* in any two groups having different, known base rates of *X*-types. The validation procedure making use of these equations is termed mixed group validation, and it is pointed out that criterion-group validation is a special (limiting) case of mixed group validation. Advantages of mixed group validation are discussed.

A medical practitioner discovers a sign that he believes will distinguish between patients who will survive lockjaw and patients who will not. A psychologist discovers a sign that he believes will distinguish between schizophrenic and nonschizophrenic adults. An educator discovers a sign that he believes will distinguish between students who will be successful in graduate school and those who will not.

In all the above examples, the diagnostician is concerned with a universe of people that can be exhaustively divided into two types (survivors and nonsurvivors, schizophrenics and nonschizophrenics, successful students and unsuccessful students); let these types be labelled *X* and *Y* (*Y* = not-*X*). In all the examples, the diagnostician has discovered some sign, label it *x*, that he thinks may distinguish between *X*-type individuals and *Y*-type individuals. "Type" is used here in a very general sense—almost synonymous with "class" or "category." A stricter meaning of "type" (e.g., Cattell, 1957, pp. 364-381) is not required in the present context.

The problem of validating the sign *x* is the problem of finding some method for determining $p(x/X)$ and $p(x/Y)$—the probability that an *X*-type

[1] The statements in this paper are those of the authors, and do not necessarily reflect the views of the Veterans Administration. The second author's work was supported in part by grants from the Ford Foundation and the National Institute of Mental Health, United States Public Health Service (Research Grant M-4465).

individual shows sign $x$ and the probability that a $Y$-type individual shows sign $x$. This paper presents what we believe to be a new method for determining these probabilities.

The standard method for determining $p(x/X)$ and $p(x/Y)$ is *criterion-group validation*. The diagnostician obtains a group of $X$-type people and a group of $Y$-type people; the incidence of sign $x$ among the $X$-type people yields $p(x/X)$; the incidence of sign $x$ among $Y$-type people yields $p(x/Y)$. Obtaining such (pure criterion) groups is full of difficulties, which will be discussed later (not the least of which is that, in many cases of interest, the diagnostician would have to wait many years in order to determine who is an $X$-type and who is a $Y$-type). What we wish to point out is that it is not necessary to have such groups in order to obtain $p(x/X)$ and $p(x/Y)$; the desirability of having such groups is discussed later.

Consider a *mixed group* consisting of a set containing people of both type $X$ and type $Y$. Consider, moreover, that we know the proportion of $X$-type individuals in the group, *even though we do not know specifically who is an* X-*type and who is a* Y-*type*. Labelling our group as Number 1, we may write this proportion as $p_1(X)$; similarly, $p_1(Y) = 1 - p_1(X)$ is the proportion of type $Y$ individuals in the group. Let $p_1(X)$ be termed the *base rate* of $X$ in Group 1.

Now consider $p_1(x)$, the proportion of sign $x$ in Group 1. Term this the *sign rate* of $x$ in Group 1.

Since $X$-types and $Y$-types are mutually exclusive and exhaustive, $p_1(x) = p_1(xX) + p_1(xY)$; that is, the sign rate in Population 1 is equal to the proportion of people in Population 1 who are type $X$ and who show sign $x$ plus the proportion in Population 1 who are type $Y$ and who show sign $x$.

But, $p_1(xX) = p(x/X)\,p_1(X)$; that is, the proportion of people who are type $X$ and who show sign $x$ is equal to the proportion of people who are of type $X$ multiplied by the probability that a type $X$ individual shows sign $x$. It is assumed that this conditional probability is independent of the population (See Section 3).

Similarly, $p_1(xY) = p(\mathrm{x}/Y)\,p_1(Y)$; and substituting in the original equation we obtain:

$$p_1(x) = p(x/X)\,p_1(X) + p(x/Y)\,p_1(Y). \qquad [1]$$

We have assumed that we know $p_1(X)$; then, since $p_1(x)$ may be observed empirically, the only unknowns in the above equations are $p(x/X)$ and $p(x/Y)$.

Consider, now, a second mixed group with a *different* base rate $p_2(X)$ and a sign rate $p_2(x)$. As above:

$$p_2(x) = p(x/X)\,p_2(X) + p(x/Y)\,p_2(Y). \qquad [2]$$

Equation 2 has the same unknowns as Equation 1. Moreover, the two

equations are independent. For if the vector $[p_1(x), p_1(X), p_1(Y)]$ were a multiple of the vector $[p_2(x), p_2(X), p_2(Y)]$, the condition that both $p_1(X) + p_1(Y)$ and $p_2(X) + p_2(Y)$ must equal 1 would be violated. Therefore, the two equations may be combined and solved for $p(x/X)$. The solution is:

$$p(x/X) = \frac{[p_1(x)p_2(Y) - p_2(x)p_1(Y)]}{[p_1(X)p_2(Y) - p_2(X)p_1(Y)]} \quad . \qquad [3]$$

Equation 3 yields the desired information necessary for validating the sign $x$. $p(x/Y)$ may be trivially obtained from Equation 1 or 2 once $p(x/X)$ is obtained from Equation 3.

We conclude that it is not necessary to observe the incidence of $x$ in two pure criterion groups to determine $p(x/X)$; the incidence of $x$ in any two (distinct) groups with known and different base rates yields the same information. We shall refer to validation making use of Equation 3 as *mixed group validation*.

One special case should be mentioned here, because it is counterintuitive (at least for some researchers). A brief glance at Equation 3 shows that if $p_1(x)$ were to be equal to $p_1(X)$ and $p_2(x)$ were to be equal to $p_2(X)$, then $p(x/X)$would be 1. In other words, if the sign rate matched the base rate in the two populations, then the sign would be perfectly valid. Again, this is true of *any* two populations. (Since perfect signs are never found in the field of psychology, the above conclusion might be restated as: no matter what the sign, its rate cannot match the base rate in any two populations.)

Another special case is that in which the sign rate does not differ in two populations with differing base rates. In this case, the sign is totally useless. That is, if $p_1(X) \neq p_2(X)$ yet $p_1(x) = p_2(x)$, then $p(x/X) = p(x/Y) = [p_1(x) = p_2(x)]$. This conclusion may be confirmed from Equation 3 by substituting $[1 - p_1(Y)]$ for $p_1(X)$ and $[1 - p_2(Y)]$ for $p_2(X)$ in the denominator of the fraction; the denominator then reduces to $p_2(Y) - p_1(Y)$. The numerator in the case where $p_1(x) = p_2(x)$ is equal to $p_1(x)\,[p_2(Y) - p_1(Y)]$. Hence, so long as $p_2(Y) \neq p_1(Y)$, the fraction reduces to $p_1(x)$. However, if $p(x/X) = p_1(x)$, and $p_1(X) \neq 0,1$, the sign $x$ is by definition independent of the type $X$; $p(x/X)$ will equal $p(x/Y)$.

Finally, it must be pointed out that criterion-group validation is just that special case of mixed group validation in which $p_1(X) = p_2(Y) = 1$ and $p_2(X) = p_1(Y) = 0$; then Equation 3 reduces to $p(x/X) = p_1(x)$.

*The Bootstraps Phenomenon*

One particular property of mixed group validation should be mentioned before discussing its practical advantages: the base rates in mixed groups may

be determined from a sign with little validity, while a sign would have to be perfect if it were to be used for establishing pure criterion groups.

Since the sign validated from two mixed groups may be more valid than the sign used to obtain these groups, we refer to the resulting validation as a "bootstraps" operation. Suppose, for example, we had a sign $x'$ of poor validity; say, $p(x'/X) = .60$ and $p(x'/Y) = .40$. If we were to take a population in which $p(X) = p(Y) = .50$, and if we were to divide this population into those who showed sign $x'$ and those who did not, we would then have two subpopulations, one of which would have a base rate of $X$ of .60, the other of which would have a base rate of $X$ of .40. If, then, another sign $x$ showed a sign rate of .60 in the first subpopulation and a sign rate of .40 in the second subpopulation, we could conclude that *sign* x *has perfect validity*. Yet, it was *validated on* a sign with little validity. Of course, most signs $x$ will not be perfectly valid, but their degree of validity, as determined from Equation 3, is entirely independent of the degree of validity of sign $x'$, when sign $x'$ is used as a criterion for determining the base rates in two mixed groups.

Cronbach and Meehl (1955), in their article on construct validity in psychology, have pointed out the bootstraps phenomenon. They argued that, in practice, a psychological test may be validated on a test of lesser validity. The conclusions presented here give a precise mathematical explication of how this phenomenon works, at least in one class of validation contexts. Moreover, the fact that one may determine the validity of signs from mixed groups argues for the essential merit of the construct-validity approach.

### *The Assumption that* p(x/X) *Does Not Vary from Population to Population*

This assumption is undoubtedly an approximation to the state of nature, and sometimes a rather poor one. However, the assumption's validity is not germane to the distinction between mixed group validation and criterion-group validation. If $p(x/X)$ varies greatly from population to population, the attempt to obtain a single value for it from criterion-group validation is just as suspect as the attempt to obtain a single value for it from mixed group validation. If $p(x/X)$ varies slightly, we see no reason why its value as determined from mixed group validation should be any further from its mean (or median, or modal) value than should its value as determined from criterion-group validation. (The problem of sampling error will be discussed in the last section of this paper; the reader should be forewarned, however, that we are not experts on distribution theory, and we will not derive precise estimates of the variance of $p(x/X)$ using the mixed group method—or the criterion-group method, for that matter.)

*Advantages of Mixed Group Validation*

The most obvious advantage of mixed group validation is that it allows the immediate estimate of sign validity in investigations in which the category to which an individual belongs cannot be determined until some time after the observation of whether or not the person displays the sign whose validity is of interest. Typical of such studies are those that attempt to discover signs of mortality associated with a given disease or (physical or mental) syndrome. If the investigator uses criterion-group validation, he must wait after observing which people show the sign until every person in his sample can be classified. If he uses mixed group validation, he can determine the sign validity if he is able to find two groups with known base rates among his subjects.

Consider, for example, the problem of the psychologist who is attempting to discover signs associated with future suicide. He can only classify a given subject as suicidal or nonsuicidal after the subject is dead, because one cannot be sure a subject will not commit suicide until he has died from some other cause (or is physically incapacitated to the point he cannot commit suicide, even though he may wish to). If the psychologist is to use criterion-group validation in determining the validity of the signs he is studying, he must wait until every last subject is dead. Aside from the problem that it is highly likely that some of his subjects will outlive him, the psychologist is faced with all sorts of demands for results now. If the psychologist feels compelled to use criterion-group validation, he is likely to give up, or not begin the project in the first place. On the other hand, he can use mixed group validation, because there is extensive sociological literature on the differential suicide rates among various social groups, and he can make use of this literature in determining the base rates of suicide among various subpopulations in his sample. He can then validate any sign he is investigating by using Equation 3.

It is generally recognized that studies involving classification based on future events are becoming extremely important, especially in medicine. Medicine has advanced to the point that one of its major concerns is the prevention, as opposed to the cure, of disease (see DeBakey, 1964). Prevention can be effective only if it is known who is likely to suffer some disease *in the future* and who is not. The validation of present signs for predicting future illness is therefore a pressing concern. The mathematics of this paper demonstrate that such studies need not take the form of "follow-up studies," in which people with various signs (e.g., heavy smoking) are classified on the basis of their health over a number of years after the sign is observed. Rather, the extensive statistical literature on prevalence rates may be used to determine sign validity here and now—by Equation 3.

Another, but not so obvious, advantage of mixed group validation is that it allows us to determine the validity of signs associated with categories that are "open concepts." An example will best illustrate this point. Suppose a psychologist has an "open concept" of schizophrenia. That is, he doesn't have a precise idea of what he means by the term, but he feels the term has meaning; he doesn't have an operational definition of "schizophrenia," but he does not wish to abandon the concept, which is loosely, approximately specified by a set of observable indicators of unknown relative weight ("validities"). Without a precise operational definition, he cannot say with certainty whether each given subject he studies is, or is not, a schizophrenic. Hence, he cannot use criterion groups to determine the validity of a sign he wishes to study. He can, however, use mixed group validation if he is willing to estimate the base rates of schizophrenia in two different populations. It might be asked how he is able to estimate base rates when he cannot state precisely who is and who is not schizophrenic; and an answer to this question would involve a long argument on the virtues of open concepts—particularly for "beginning" sciences—as opposed to pure operational concepts. This argument is not presented here since it is not central to the method of mixed group validation. (But see Meehl, 1965.)

*The Problem of Sampling Error*

Equation 1 is just as valid whether one is dealing with a universe or a sample from that universe. Hence, any sample deviation of $p_1(x)$ from the universe $p_1(x)$ is due to the fact the sample values of $p_1(X)$ or $p(x/X)$ or $p(x/Y)$ differ from the universe values of $p_1(X)$, $p(x/X)$, or $p(x/Y)$.

It would be difficult to determine how the sample values of $p(x/X)$ and $p(x/Y)$ differ from the universe values for these probabilities because these probabilities are what is solved for in Equation 3. On the other hand, it can be easily seen that the sample value of $p_1(X)$ may differ from the universe value. One is sampling from a mixed universe with proportion $p_1(X)$ of people who are type $X$, and one happens to get proportion $\hat{p}_1(X)$ in the sample.

To determine how such a discrepancy would affect an investigator's estimate, we must begin by assuming that $p(x/X)$ and $p(x/Y)$ are constant from population to population. (As pointed out previously, this assumption is also made in criterion-group validation.) Suppose, further, that the proportion of people of type $X$ in Population 1 in the investigator's sample is $\hat{p}_1(X)$ and the proportion in Population 2 is $\hat{p}_2(X)$; whereas, the investigator supposes these proportions are $p_1(X)$ and $p_2(X)$ (the universe values). The *actual value* of $p(x/X)$ is equal to

$$\frac{[p_1(x)\hat{p}_2(Y) - p_2(x)\hat{p}_1(Y)]}{[\hat{p}_1(X)\hat{p}_2(Y) - \hat{p}_2(X)\hat{p}_1(Y)]}$$

But the value the investigator estimates, label it $\overset{*}{p}_1(X)$, is

$$\frac{[p_1(x)p_2(Y) - p_2(x)p_1(Y)]}{[p_1(X)p_2(Y) - p_2(X)p_1(Y)]}$$

The difference between true and estimated values simplifies to

$$p(x/X) - \overset{*}{p}(x/X) = \frac{[p_1(x) - p_2(x)][\hat{p}_1(Y)p_2(Y) - \hat{p}_2(Y)p_1(Y)]}{[p_2(Y) - p_1(Y)][\hat{p}_2(Y) - \hat{p}_1(Y)]} \quad . \qquad [4]$$

Various examples can be worked out from this formula by the investigator who is interested in knowing how mistaken he might be in his estimate of $p(x/X)$; of course, he will have to start with some idea of how much his sample values, $\hat{p}_1(X)$ and $\hat{p}_2(X)$, might differ from the universe values.

We have not developed any analytic expression for the error of estimate more general than Equation 4. This equation may, however, serve as the basis for one further conclusion: In general, the validity of a sign $x$ will be underestimated if and only if the investigator overestimates the differences between his two groups, and it will be overestimated if and only if the investigator underestimates.

This conclusion is supported as follows: Without loss of generality, we may suppose $x$ is associated with $X$ (not $Y$) and that Population 1 has a higher base rate of $X$ than Population 2. Then, the factor $[p_2(Y) - p_1(Y)]$ will be positive, and unless the investigator's sample proportions go in the opposite direction from the universe proportions, $[\hat{p}_2(Y) - \hat{p}_1(Y)]$, and $[p_1(x) - p_2(x)]$ as well, will be positive. Therefore, it follows from Equation 4 that $p(x/X) - \overset{*}{p}(x/X)$ will be positive, that is, the investigator will *underestimate* the validity of sign $x$, if and only if the factor $[\hat{p}_1(Y)\,p_2(Y) - \hat{p}_2(Y)\,p_1(Y)]$ is positive. *In general*, this latter condition will be true when $\hat{p}_1(Y) > p_1(Y)$ and $\hat{p}_2(Y) < p_2(Y)$, which is the case when there are fewer $X$'s in Sample 1 than the investigator believes and more in Sample 2—that is, when he has overestimated the difference between Groups 1 and 2.

The limiting case occurs when the investigator believes he has pure criterion groups, but actually has mixed groups; in this case, the investigator will *always* underestimate the validity of his sign—provided it has any validity at all.

Another consideration relevant to the problem of sampling error is that the denominator of Equation 3 may be written as $p_2(Y) - p_1(Y)$ or, equivalently, as $p_1(X) - p_2(X)$. The error in estimating this difference will increase as these probabilities approach .50, because the variance of a difference between two independent quantities is equal to the sum of their variances, and the variance of a probability increases monotonically as the probability approaches .50. Therefore, the more the base rates approach .50 (i.e., the greater the "mixture" in the mixed groups), the larger the sample the investigator should use.

*Conclusion*

In the above discussion, the word "sign" has been used as a generic term. A "sign" can refer to anything from a particular symptom to a summary score on a heterogeneous personality inventory. Moreover, all signs have been treated as if they are either present or absent; the problem of *determining* (what constitutes) a sign has been ignored. If, for example, we suspect that subjects' scores on a given personality inventory might be useful signs, we must determine optimal "cutting points" on this inventory. While a variant of the method of mixed group validation can probably solve this cutting-point problem, it has not yet been developed. (But see Meehl, 1965). Finally, treatment of the sampling problem has been cursory.

The purpose of this paper was to present the basic logic of the method of mixed group validation, not to explore all of its many ramifications.

## REFERENCES

CATTELL, R. B. *Personality and motivation structure and measurement*. Yonkers-on-Hudson: World Book, 1957.

CRONBACH, L. J., & MEEHL, P. E. Construct validity in psychological tests. *Psychological Bulletin*, 1955, *52*, 281-302.

DEBAKEY, M. E. (Chm.) *Report to the President: A national program to conquer heart disease, cancer and stroke*. Washington, D.C.: Presidents' Commission on Heart Disease, Cancer and Stroke, 1964.

MEEHL, P. E. *Detecting latent clinical taxa by fallible quantitative indicators lacking an accepted criterion*. Research Report PR-65-2, 1965, University of Minnesota, Department of Psychiatry.

Pdf by LJY, September 2019