MATCHING AND PREDICTION ON THE PRINCIPLE OF BIOLOGICAL CLASSIFICATION

WILLIAM A. BELSON
Research Techniques Unit,
London School of Economics

In this article Dr Belson describes a technique for matching population samples. This depends upon the combination of empirically developed predictors to give the best available predictive, or matching, composite. The underlying principle is quite distinct from that inherent in the multiple correlation method.

Matching criteria have often been selected on the basis of either custom or hunch, and their application usually involves discarding the test results of those who do not fit into the matching pattern. There are, however, better ways of selecting the matching criteria and more efficient ways of applying them.

The all-important point in matching is the use or development of *relevant* matching criteria. A simple test of the relevance of a matching criterion is whether or not it is associated with or *predictive* of whatever is being studied. Here are some examples: (i) Suppose that the whole purpose of a particular matching operation is to study the relative effects of two different stimuli upon 'activity X'. If the matching criteria used are correlated with 'activity X', they qualify as relevant. If not, they serve no useful purpose. (ii) Similarly, if consumer or listening panels are to be made representative of the public (i.e. matched to them), the matching criteria used will be relevant only if they are associated with the measure or measures which the panel is meant to provide. (iii) The same applies to the controls used in quota surveys.

Relevance is, however, a matter of degree and the best matching criteria are those which, taken together, have the highest available matching power (granted, of course, adequate reliability). *This can only be achieved, with any approach to certainty, by empirical methods.* These involve establishing the predictive power of each of a wide range of possible matching variables and then developing from them the best possible predictive composite.

In earlier work[1] I had established the relative predictive power of the proposed matching variables by ordinary correlation methods. After this, the selection of the best composite of these had been effected by means of the Wherry-Doolittle technique.[2] This procedure was a considerable departure from 'matching by hunch', but it had certain methodological weaknesses and it was tedious in the extreme. The method which is described here was subsequently developed to simplify the prodecure and to increase the efficiency of the selective process.

One of its features is that it makes no use of the more formalised correlation procedures.

Establishing the Relative Predictive Power of the Proposed Matching Variables

The selection of a matching composite can often be built into the enquiry itself. Suppose that groups I and II are being matched to study the effects of two different stimuli on 'activity X'. Along with the measurement of 'activity X' (in groups I and II), respondents would be asked questions about a wide range of things put forward as possible predictors of activity X: e.g. age, marital status, membership of specific societies, whether or not respondent has ever been out of the country, number of brothers or sisters as a child, income level, experiences of a specific kind, family composition. There might be anything from 20–100 of these, but the answers to as many as possible should be reduced to simply 'Yes' or 'No'. This opens the way for assessing the power of each proposed matching variable to predict the criterion (i.e. activity X). Such assessments, however, are based upon the records of the group which is to be matched to the other. Thus if Group I is to be matched to the characteristics of Group II, the predictive or matching variables should be developed from Group I records.

When one is selecting predictors from as many variables as this procedure involves testing, it is not practical to work through full-blooded correlational procedures. That takes far too long. But quite apart from this, it is necessary to distinguish between a theoretical index of the association between two variables, and the degree to which a response to a particular question is *in effect* a predictor of response to another question. In the first place, correlation coefficients usually involve assumptions about distributions and, as abstracted indices, they give no warning about peculiarities at some particular point in the original distribution. This is specially relevant in psychological work where characteristics tend not to be distributed on the 'normal' pattern. Secondly, many (potential) predictors consist of a simple 'yes/no' response to a question. Granted certain assumptions, a tetrachoric correlation may indicate (fairly accurately) that there exists a high degree of association between, on the one hand, a particular criterion response and, on the other hand, a variable which the yes/no question was meant to tap. Yet, from the *predictive* point of view, the information available allows of no more accurate estimate than that indicated by the answer 'yes' or 'no'. An extreme case of this is where the 'yes' response cuts off no more than (say) 10% of the sample. From this, it is possible to get a high tetrachoric coefficient and a good indication of the criterion scores of those in the 'yes' group. But for that great majority answering 'no', one can tell very little indeed about the criterion score. The same general argument applies, of course, where the number of response categories available for predictive work exceeds two but is still limited.

Because of this, my strong inclination is to work directly from the

information which one *has* and *can use*. This inclination, plus the need for speedy and simple computation, has led me to the following procedure in estimating the predictive power of the many proposed matching variables.

The first stage in the estimation of predictive power is of a fairly standard kind. Respondent record cards (punched cards) are split into two groups: the higher and the lower scoring cards in terms of the criterion. Let us call this criterion 'activity X'. This process is most unlikely to give an exact 50/50 split, but the split should be as close to this as possible. The high and the low scoring cards are then machined on each of the proposed matching variables and the predictive power of each of them is worked out as follows. Suppose, by way of illustration, that one variable being tested for predictive power is 'whether or not respondent says he had any further education since ceasing full-time school'. A second variable might be 'whether or not he usually reads newspaper P'. (There would, of course, be a great many more variables to be tested.) Suppose that the results from machining these two variables are as in Table I.

TABLE I

Testing for predictive power

Criterion Score (Activity X)	Further Education		Read Paper P		All Cards
	Yes	Not Yes*	Yes	Not Yes	
Higher Scores	200	400	24	576	600
Lower Scores	100	400	1	499	500
ALL	300	800	25	1075	1100
Expected number in lower group if no association ..	136	364	11	489	
Expected minus actual number in 'lower' group ..	+36	−36	+10	−10	

* This includes cases not replying either 'yes' or 'no'.

Take 'further education' first. If there were no association at all between further education and the criterion (activity X) then we should expect the number of *Yes's* in the high and in the low scoring cells to be approximately proportional to the total number of cards in the high and low scoring groups (i.e. 600 : 500). Similarly for the number of *No's*. If they do not mirror this proportion, then we can conclude that they have some degree of predictive power. Thus, if there were no association between further education and the criterion score, we should expect 136 of the 300 people saying 'Yes' to be in the low scoring group [i.e. $(300 \times 500)/1100$]. But there are in fact 100 in that cell. The difference $(136 - 100 = 36)$ is a direct measure of the predictive power of this variable. [I use the word 'direct' in the sense that the difference (36) gives the actual number of cases 'displaced' or 'found out of position'

by using the variable concerned—whereas a correlation coefficient is both an abstraction and an index and hence is an 'indirect' measure.] It can of course be exactly reproduced from the records of those *not* claiming further education, and it can also be thought of as the variable's displacement or discrimination power. The same procedure indicates that the readership of newspaper P had a predictive power of 10.

Ordinarily, up to five 'Yes-No' variables can be tested in the one column, and a careful look along ten to fifteen columns will quickly show which dozen or so variables warrant doing the calculations just described. The best predictor is simply the one with the highest displacement or predictive power.

In view of the reasons given for preferring this method to standard correlational procedures, it is interesting to compare the results when different computational methods are used. The ϕ-coefficient based upon readership of paper P is of approximately the same size as that based on further education, whereas the tetrachoric coefficient based on paper P is much the higher. Hence neither coefficient would have been of much help in establishing the relative predictive power of these two proposed matching variables, and one of them would have been misleading. The point of the matter is that *too few people* read paper P for response about this to be of practical value in prediction. We can accurately place the few people who say 'yes, P', but we can tell practically nothing about the great majority who say 'no'. [The word 'power' (as in 'predictive power') raises problems. There is a sense in which the correlation coefficient reflects potential or maximum power (i.e. granted normal distribution and an infinite number of usable cutting points), whereas what I am seeking is a measure of its *effective* or actual power, its distribution being what it is.]

There is one further departure from standard procedure which can properly be made. Where a variable is such that people can be separated into more than two groups on the basis of it (e.g. age), it may be profitable to work out the predictive power for each of those groups

TABLE II

The process which leads to the grouping together of those at the extremities of the one scale (e.g. age)

Criterion Score (Activity X)	Age					
	21–30	31–40	41–50	51–60	61–65	All
Higher Scores	119	219	174	122	32	666
Lower Scores	151	142	149	129	41	612
ALL	270	361	323	251	73	1278
Expected number in lower group if no association ..	129	173	155	121	35	
Expected minus actual number in 'lower' group ..	−22	+31	+6	−9	−6	

and then to combine all those with positive predictive power to form a single group. In the example given in Table II, this would mean combining the 21–30's with the 51–65's (all negative), the other group being those between 31–50 years. Note that the combined positive and the combined negative predictive scores are equal.

Of course this technique is sufficient only to put the individual variables in *order* of predictive power. It tells us which is the *best* of them, but it does *not* tell us which of the others we should take in combination with it. Some of the better predictors will be correlated with each other and the variable with the second best predictive power may add very little to the predictive power already available through the first.

Selecting the Composite of Predictors on the Principle of Biological Classification

The development of the whole composite is achieved by selection on the principle of 'biological classification', a term suggested by Mr L. T. Wilkins. It is illustrated diagrammatically in Fig. 1.

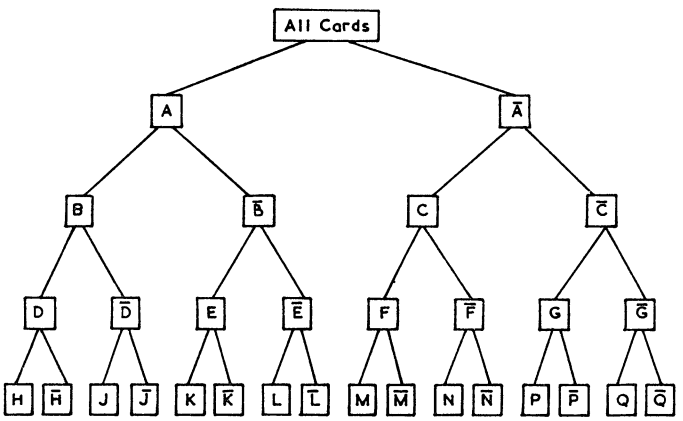

FIG. 1. Progressive splitting on the biological pattern. (It does not follow that every split will be in terms of a new variable; in practice it is quite feasible for group G to be split on variable D, or for group C to be split on variable B.

The first-order predictor is selected after an analysis of all the cards, using the method just described. Thus it will have been selected as the most powerful single predictor among the many different variables tested. This leads directly to the splitting of the full pack of cards into two parts—those punched 'Yes' to question A and those not punched 'Yes' to question A. This is in principle what might have happened after the selection of the first predictor by the Wherry-Doolittle method. From here on, however, the similarity disappears. Since the A and the non-A group are different in at least one telling way, it does not follow that the second-order predictor for group A will be the same as the second-order predictor for group non-A (as, for instance, when group A are females and group non-A are males). Accordingly, the search for a second-order predictor within group A is made quite separately from the search in group non-A. In each case the technique used is precisely

the same as that used in selecting the first predictor. Thus, group A cards are sorted into higher and lower scoring groups (with as near as possible a 50/50 split) and each of them is machined on all possible matching variables. The one chosen is that with the highest predictive power. This is shown as variable B in Fig. 1, and its selection allows the breaking of the A pack of cards into B and non-B. In group non-A, on the other hand, the best predictor is shown as C and this serves to break group non-A into groups C and non-C.

In just the same way, the search for the third predictor is made quite separately in each of the different sub-groups (i.e. separately for groups B, non-B, C, and non-C). This procedure is carried on until either (i) the sub-groups become too small for meaningful analysis; or (ii) the best available (next) predictor displaces too few people to warrant its use, i.e. it has very low predictive power for the sub-group concerned. [It is only rarely an advantage to express the predictive power of a variable in percentage form (e.g. a displacement of 72 in a group of 1100 might be expressed as 7200/1100%). To do this only obscures the relative effectiveness of the first split compared with the second, third, etc., and it may hide the fact that a further split would displace too few actual cases to make it worth while.]

An Example of a Predictive Composite. Fig. 2 is an example of a predictive composite developed through selection on the pattern of biological classification (included by the kind permission of the British Broadcasting Corporation). It was developed for use in a study of the effects of exposure to television upon the degree to which individuals participate with others in the home in their various household activities, i.e. the degree of 'joint activity' (a BBC enquiry into the effects of television on aspects of family life and sociability). It shows also something of the wide range of predictors which are in practice usable, and also the odd order in which various of them emerge.

Reasons for Not Using the Wherry-Doolittle Method

There are two reasons for preferring the present method to the Wherry-Doolittle technique, the first theoretical and the second practical.

In the Wherry-Doolittle method, the predictor with the highest correlation with the criterion is selected as the first in the predictive battery. The second is selected after mathematical consideration of the other variables in terms of their correlations both with the criterion and with each other. It turns out to be that one which adds more to the correlation of the first predictor with the criterion than does any other of those available for selection. The procedure is continued until the addition of one more variable adds nothing to the multiple correlation achieved by those already selected. But a difficulty arises at each selection stage and it can be illustrated by reference to the selection of the second predictor: the second is selected on the basis of correlations derived from the *whole* sample; but it is well known that the predictive

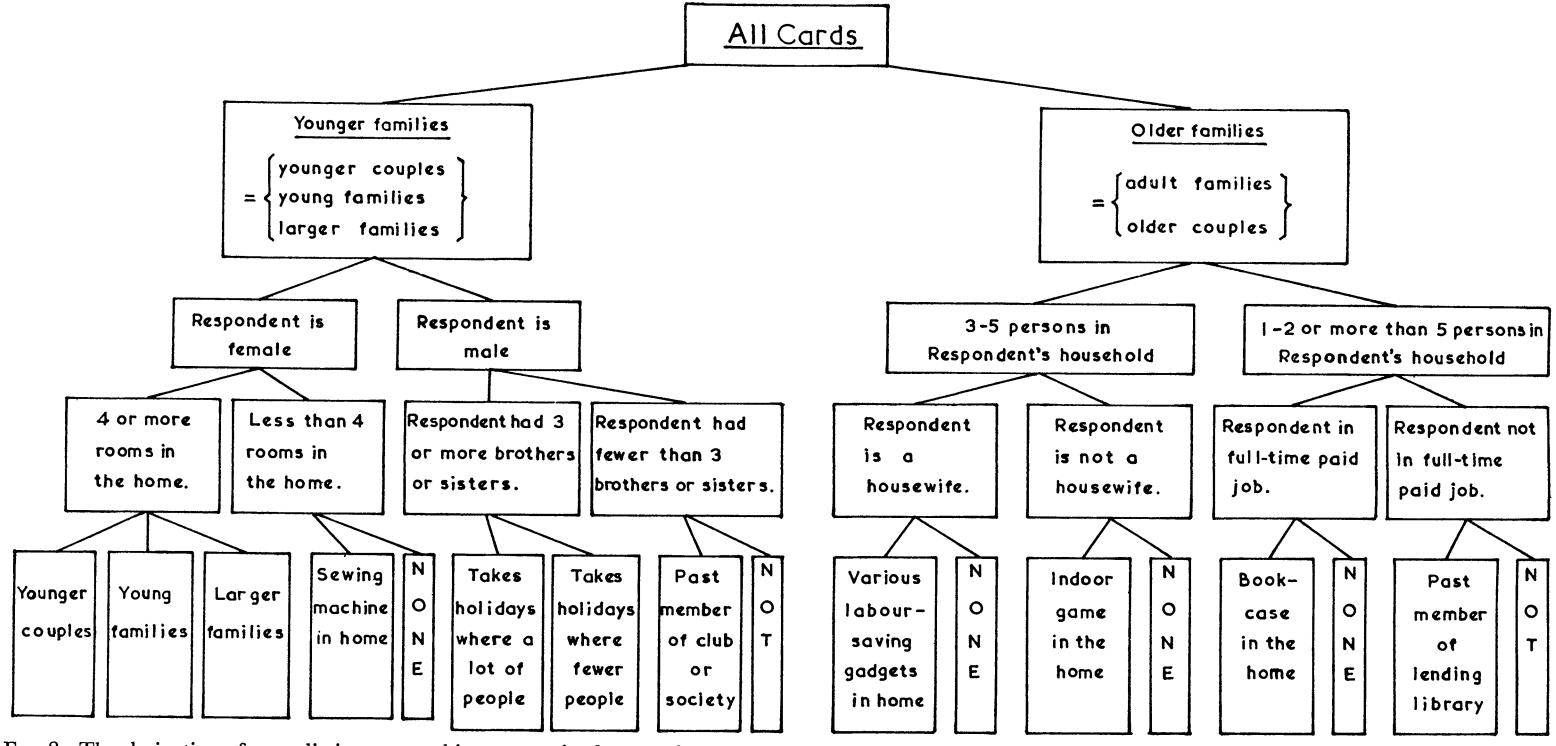

FIG. 2. The derivation of a predictive or matching composite for a study concerning 'Joint Activity' in the home. (To save overcrowding in the bottom row direct negatives of certain classes have been indicated by 'NOT' or 'NONE'.)

power of a variable can differ in different sections of the one sample. (Thus amongst old people, the predictors of, say, the amount of time spent at home are quite likely to be sharply different from those for young people; other examples are not hard to find.) Accordingly it is quite likely that for at least one section of the sample the second predictor will be ineffective—it will, as it were, be carrying dead weight. In practice, there is frequent evidence that this is happening and it militates against the maximisation of the available predictive resources by the Wherry-Doolittle method. Indeed, it seems to be responsible in part for the fact that the multiple correlation, between the composite of predictors on the one hand and the criterion on the other, starts to decline after the inclusion of more than about four or five predictors in the composite.

The method which I have described here as an alternative requires that the selective process be carried out separately for each of the increasingly homogeneous sub-samples. This does not entirely eliminate the difficulty, but it reduces it considerably.

However, the less technical difference between the two methods is perhaps the more important. The Wherry-Doolittle method involves a great amount of computation, first in the development of a matrix of correlations, and then (to a greater extent) in the selection process. In turn, the heavy computation serves to put a limit on the number of items which it is expedient to try out as possible matching variables. This can be specially limiting because it works against the trying out of the less orthodox of the possible predictors.

Using the Matching Composite to Achieve Matching

Once the matching variables are established, the two packs of cards which are to be matched (i.e. Packs I and II) are machined into the various sub-groups dictated by the matching composite. This is done quite separately for each of the two packs of cards (I and II). Thus, Pack I would first be split into A and non-A, then A would be split into B and non-B, whereas non-A would be split into C and non-C. Continuing in this way there would, in Fig. 1, be sixteen matching sub-groups. If Pack I is to be matched to Pack II, then the final step is to ensure that each of the sixteen sub-groups in Pack I has the same number of cards in it as the equivalent sub-group in Pack II.

It is important to avoid any throwing-out of cards and this can be accomplished in several different ways. If it does not matter *which* pack is matched to the other, the simplest method is to weight-up the number of cards in the smaller of the equivalent sub-groups (whether this be in Pack I or Pack II), doing this afresh for each of the sixteen equivalent sub-groups. If it is necessary that Pack I be matched to Pack II, then if any of the Pack II sub-groups is smaller than the equivalent sub-group in Pack I, it will probably be expedient to double the number of cards in all Pack II sub-groups, so that the Pack I sub-groups can in all cases be weighted *up* to them. Weighting-up is by the

replication of systematically selected cards in the sub-group concerned.

Another way of achieving the matching is by simple multiplication. Here, too, suppose Pack I is to be matched to Pack II. The average criterion score (i.e. for activity X) for each sub-group in Pack I is calculated. For each sub-group, this is multiplied by the number of cards in the equivalent sub-group in Pack II. The sum of these sub-products is divided by the total number of cards in Pack II and the result is recorded as the adjusted or weighted Pack I score. This and the unweighted Pack II score can then be regarded as averages for two closely matched groups. An example of adjustment by the multiplication method is given in Table III.

TABLE III

The multiplication method of applying the matching composite*

Sub-Group	Number in Pack I Sub-Group (f_I)	Average Criterion Score for Pack I Sub Group (M_I)	Number in Pack II Sub-Group (f_{II})	Product $f_{II} \cdot M_I$
1	4	13·50	10	135·0
2	20	14·30	38	543·3
3	52	19·30	52	1002·0
4	35	24·10	30	723·0
5	22	24·32	14	340·5
6	54	10·36	61	632·0
7	81	17·50	134	2345·0
8	44	19·41	67	1300·5
9	24	22·67	21	476·1
10	6	28·50	7	199·5
	$\Sigma f_I = 342$		$\Sigma f_{II} = 434$	$\Sigma f_{II} M_I = 7696{\cdot}9$

Unadjusted Pack I Score = 18·32 (working not shown)

Adjusted Pack I Score $= \frac{7696{\cdot}9}{434} = 17{\cdot}80$

* Table III is reproduced from my work on the effects of television upon viewers' interests.

Applications of the Biological System of Matching

Empirical matching is essentially *prediction*. We are asking what would have been the score or the response of a particular group had certain of their characteristics or certain of the conditions surrounding them been the same as those for another group of people. Seen in this way, empirical matching is of central importance and can have many different applications. I have described it so far in relation to the matching of two groups. But it might have been described equally well in relation to four other operations. (i) One is the matching of panel members to the public to secure an estimate of what the panel would have registered on a particular issue, had that panel been representative of the public in terms of relevant characteristics (see above for the implications of 'relevance'). This can be thought of either as a correction process or as the setting-up of controls for the selection of the panel. (ii) The same approach can be used in adjusting

the average obtained from groups undergoing intensive study or being tested under controlled conditions. Put in another way, what is involved here is the matching of volunteer groups to the public from which they were drawn. (iii) Closely allied to this is the setting up of a limited number of controls for quota survey work. (iv) Finally, the method has been used fairly extensively in isolating the effects of television.

COMMENTS ON THE METHOD

For the sake of a straight-forward presentation, I have held over the discussion of certain technical details until now. Most of the following points are the result of talking about the technique with colleagues.

1. Whilst the technique described here is shorter and more accurate than that involving multiple correlation, it is still a lengthy and demanding process, particularly as each different variable under study is logically bound to have its own predictive composite. At the same time, this is not so much a consequence of the method, as of the ordinarily rigorous demands of effective matching.

2. In evaluating the technique, some may deem it an advantage that it is statistically unsophisticated and that this happens to go along with increased efficiency. It may also be to the point that with the exception of the multiplication process (see above), it is independent of any specific assumptions about distributions. At the same time, its unsophisticated character has put the method out of line with standard practice and ready formulae, and obviously a lot of thought must be given to the question of what constitutes an adequate measure of either standard error or the significance of residual differences.

3. If we turn to detail, there is a sense in which the predictive power of a composite of the kind described here 'capitalises' on errors in the sample from which it is derived. This would mean that its predictive power would shrink somewhat in going from the original sample to an entirely fresh one. If the sole purpose of the matching were to adjust the score of the sample from which the composite was derived (and this is the usual aim of matching) the problem of shrinkage would be less likely to arise. But were the aim to develop (say) a control for recruiting panel members, it *could* matter. What should be kept in mind, however, is that this method turns not so much upon the absolute predictive power of the composite as upon the empirical selection of the best available predictors.

4. One obvious question about the development of a predictive composite is just where, and on what grounds, a halt is called to the addition of further predictors. As yet there is no cut and dried answer to this question, though it should not be difficult to find an approximate solution in the form of a simple rule. My own practice has been to stop the process after the fourth or fifth splitting stage. There are two grounds for this: The first is that one or two of the sub-groups of the sample to which the matching is to be done may have no cases in them at all, so

that the cases in the equivalent sub-groups in the other sample (that from which the matching composite was developed) have to be discarded. Along with this there are, of course, theoretical arguments which work against further reduction of the numbers in the various sub-groups. The second reason for stopping at the fourth or fifth splitting stage is that to go further becomes arduous and markedly increases the chance of there being errors in processing. *Against* stopping at this stage is the fact that the total number of cases being displaced (by all the predictors taken together) is still appreciable and may account for more than 5 % of the total sample.

5. Another problem of central importance is the adequacy of the matching. It is unlikely that the matching will ever be complete: accordingly the aim of this method is simply to get nearer to completeness than is ordinarily possible through 'matching by hunch' or through multiple correlation methods. At the same time, methods do exist for testing the hypothesis that the matching has eliminated all the unwanted or extraneous differences between the groups,[1] and in various uses of empirical matching these methods have indicated a high degree of effectiveness.[1, 3]

6. The origination of the many proposed matching variables is a crucial part of the operation, but yet depends upon the originating capacity of the research team. I think that good origination can only come out of a growing experience with the sort of thing that works, along with a considerable amount of willingness to try the unusual. Apart from this, the main safety in this vital phase lies in the trying out of a *lot* of possible predictors.

There is one concluding point to be made. The technique which I have described is by no means complete and obviously there are still loose ends to be tied up. What I hope, however, is that the tying up of loose ends will not be allowed to take the simplicity or the commonsense out of the method. Let me be more specific. The method as I have described it is, it is true, a movement towards a more empirical way of doing things; but it is just as much a movement *away from* a sophistication which is too often either baffling or misleading.

REFERENCES

[1] Belson, W. A. (1956). 'A technique for studying the effects of a television broadcast', *Applied Statistics*, **5**, 195.

[2] Garrett, H. E. (1947). *Statistics in Psychology and Education*, pp. 435-448. Longmans Green, New York.

[3] Belson, W. A. (1957). 'The effects of television upon the interests and the initiative of adult viewers', *Impact*, No. **21**.