Costa J Tchaoussoglou
Cebuco
Amsterdam, Netherlands
Joop L van Vliet
RTV-Combinatie
Amsterdam, Netherlands

4.5 Frequency scales and their use

INTRODUCTION

Although the main subject of this paper is the nature and usage of frequency scales we should first look at some more general problems in measuring readership. There is a difference between getting data and getting the data you want and need. So you have to be careful about the way in which you ask for certain kinds of information. It is not enough simply to ask for the things you would like to know. A seemingly easy question, for example may cause our respondent trouble. It is possible that he does not understand the question as we meant him to understand it, or that we have not given him enough alternative answers or that the choice of possibilities causes embarrassment or . . . .

Sometimes straight questions do not work and we have to go for our information in a roundabout way. Some questions will lead to 'don't know' answers when asked in a personal interview and will be answered in full when asked in another way. If we have the same question for each of a great number of publications this may be much more annoying or even irritating in a personal interview than in self completion.

Related to this problem is that almost identical questions may lead to different results, and furthermore that questions and questionnaires can deteriorate over time. What was acceptable ten years ago may be wrong now or perhaps insufficient. This can be due to a change in our requirements, or a change in the attitude of respondents, or both. Therefore every survey that pretends to be more than a run of the mill data collection should use pilot testing to check the questionnaire.

However, reliable data is not equal to reliable information. The raw data will usually undergo a set of transformations and we should bear in mind that a set of logical operations is not always a logical set of operations.

TESTING A QUESTIONNAIRE

Growing criticism

The NOP questionnaire for 1979 (NOP is the Dutch NRS) was not very different from the 1972 questionnaire. That there were of course adaptations and improvements, nobody can deny, but there were also doubts and unfulfilled wishes and the number of voices clamouring for a complete overhaul grew.

The board of the NOP foundation decided therefore to set up a committee of researchers and media experts to find out if and how future NOP surveys should be different. This TOEKNOP committee (TOEKomstig NOP = future NOP) deemed it necessary – among other sensible things – to do pilot research on the questionnaire. The total reach (TR) and Average Issue Readership (AIR) measures needed to be considered, because they form the basis for media selection. In theory, TR and AIR should approximate to reality as well as possible or – if that target could not be reached – have the same amount of error for each of the media under consideration. Alas, we have no idea what reality looks like, so how could we choose between questionnaires?

An important part of this last question is answered by William A Belson in his still outstanding book *Studies in Readership* (1962). Belson describes a sequence of two interviews with the same respondent. In the first interview the 'normal' reading behaviour questions are asked, for each of the publications. In the second interview one *goes for the truth* for four selected titles. This is done by trying to connect the act of reading with the circumstances during the time of reading. Let us give an example:

A respondent has claimed having read *Time* during the last week. When asked about the place of reading, the exact time of day and other circumstances, his first answer 'last week' is proved wrong. The actual reading took place in the living room while the respondent waited for his wife who was dressing to go to the theatre. The day of this visit, however, was more than a week earlier, as could be easily confirmed.

Conclusions from this test

That this testing procedure seems to ask a lot of the respondent, may be true, but it works. Respondents are willing to help if they get a proper explanation of the hows and whys of the procedure. They may be surprised by the weakness of their memory but they must not be angered by even very slight implications of having lied before.

From the example we can draw some conclusions:
(a) memory is a tricky thing – but didn't we know that before?
(b) memory, however, can be aided when we strive to

put certain actions into their context of place and time; we think there is a parallel between the three conditions for the classical Greek drama – unity of action, time and place – and looking for a coherent picture of past reality. **(c)** we may assume that some of our questionnaires are less likely to 'help' our respondents in finding the correct answers than others. Accordingly it can be proved that the answers on some of the questionnaires are 'better' than the answers on other questionnaires.

We say 'better' because we still have to prove that our 'better' questionnaire is really better, or at least have to prove that our 'better' one is likely to be better in reality.

What we try to find out with the double interview is whether the respondent answers consistently or not – in other words 'are the results stable or not?' It goes without saying that unstable results must be unreliable results too. We have not yet found a way to prove the positive counterpart but we think we have at least partially justified our third conclusion.

Troubles

It was our intention to include in this paper some of the results of our pilot experiment with intensive interviewing. We soon found out, however, that there were more problems between heaven and earth than we could imagine. Our adapted Belson interview worked smoothly in the small-scale pre-pilot, but then we encountered another major problem: interviewer bias. We did not expect to find interviewer bias in such a small (10 interviews) sample, especially as we used the already tested NOP – 79 questionnaire and interviewers who were familiar with it. As it is necessary to have the same questions for each of the publications any deviation of the exact wording of the questions may cause bias. It turned out that the frequent repetition of the same phrases causes irritation with both parties, respondent and interviewer. Therefore, the interviewer changes to short descriptions of his own devising, or the respondent starts to answer as soon as the first words of the question have been spoken. We deemed it necessary first to correct our test versions before starting with the definite pilot study. We expect this study to go into the field in the first week of April – so we have no results yet. There are, however, lots of related problems and the frequency scale with all its features is one of the most relevant.

READING FREQUENCY – HOW TO MEASURE IT?

How many scale points?

An excellent review concerning the number of scale points is from Eli P Cox III and appeared in the *Journal of Marketing Research*, November 1980. Cox cites Miller and writes: "Miller notes a remarkable similarity of findings. The mean across the experiments for the maximum number of stimulus categories that could be utilised successfully by subjects was approximately 6.5 (2.6 bits), with one standard deviation including from four to ten categories." There were, however, some limitations to Miller's conclusions, one of them that perception rather than memory was being assessed. Further on, Cox/Miller mention that "the accuracy of human judgement may be expanded when mnemonic processes are employed. Coupling human information processing with memory greatly increases channel capacities." Although we certainly may not take for granted that all these findings apply to frequency scales, we think that the recommendations for the applied researcher given by Cox may be followed without danger.

Recommendations

We quote again: "Certain things do appear clear. First, scales with two or three response alternatives are generally inadequate in that they are incapable of transmitting very much information and they tend to frustrate and stifle respondents. Second, the marginal returns from using more than nine response alternatives are minimal and efforts for improving the measurement instrument should be directed toward more productive areas. Third an odd rather than an even number of response alternatives is preferable under circumstances in which the respondent can legitimately adopt a neutral position."

It seems that the protagonist of the six-point frequency scale has fortified arguments against his 12-point scale adversary.

How do we use our frequency data?

If we go along with the fairly common opinion that AIR is a more trustworthy base for media planning purposes than, eg information about the reading frequency, what do we need the latter for? After all it is a rather 'soft' measure and certainly less stable than AIR. The Dutch situation is that reading frequency is asked for but does not appear on the tape for media planning purposes. Formerly the frequency scores were directly used for the calculation of the reading probability. Now we use a segmentation analysis in which it turns out to be the most important variable.

The reason why the reading frequency cannot be used directly is the discrepancy between AIR as measured and CIR (Calculated Issue Readership) based on the formula:

$$\text{CIR} = \sum_i f_i \cdot \frac{i}{12}$$

where i is an integer between 1 and 12 and f_i the number of readers in the frequency group.

This discrepancy stems from the fact that the number of people in a frequency group who read the last issue usually differs from the theoretical number, and we cannot simply say that six out of 12 does not mean .5 but .43.

This is, however, exactly what we do when using reading probabilities, where we can usually choose between .01 and all the intermediate two-digit figures up to .99.

In reality the number of reading probabilities is limited. Assuming an observation period of a year we have probabilities of m/12 for monthlies, w/52 for weeklies and d/309 for dailies, with m, w and d integers. Another thing is that if we only use reading frequencies some of the frequency classes are almost empty and furthermore we have the problem of the readers with a smaller than one out of . . . probability.

If the reading frequency scale extends over the last 12 issues the category 0 out of 12 does not exist for monthlies, but for weeklies and the dailies we have a residual group with a greater than zero probability.

Binomial versus hypergeometric

To calculate the cumulated audiences of two to 12 issues the binomial distribution is usually used. In reality, however, the hypergeometric distribution should be used. The following example shows why.

Let us suppose a group of 300 readers with a reading probability of .5. We have in some way obtained the exact information about their reading behaviour with regard to the last four issues and can tabulate the results (**Table 1**).

By the binomial approximation we obtain the following cumulated audiences: 150, 225, 262.5 and 381.25 for one, two, three and four issues respectively. The reality however is easily confirmed from the table and reads: 150, 250, 300 and 300. By using the hypergeometric distribution we obtain the same results. The binomial distribution falls short of reality.

TABLE 1

Group no	*Issue read or seen* 1	2	3	4	*Total*
A			50	50	50
B		50	50		50
C		50		50	50
D	50			50	50
E	50		50		50
F	50	50			50
Total	150	150	150	150	300

We must, however, pay attention to the fact that the hypergeometric distribution can be used only if the number of connected intervals of the measuring period (ie the number of points on the frequency scale) is not exceeded by the number of intervals during the planning period. If that should be the case we must stretch our scale and we have also to take into account the (not quite) 0 out of 12 group. Perhaps comparing our frequency scale with the results of panel registration (with regard to the relative occupation of the different frequency classes), could give us some clue for solving the problems.

Favourite classes

As mentioned before, some of the frequency categories are almost empty. If we look at the distribution patterns we find irregularities that cause at least some doubt about the trustworthiness of the scores. It is like the reports of the smaller weather stations on daily temperatures in tenths of degrees. One should expect an even distribution over the ten digits nought to nine, but the last digit is in about 25% of the cases a zero and in another 20% a five, with two and four especially unpopular. Therefore we should use a smoothing procedure to correct for this obviously faulty way of reporting.

The same applies to our reading frequencies. If we may use corrected reading probabilities, why shouldn't we use corrected reading frequencies? By shifting people from a higher class to a lower or vice versa we can obtain such a filling of the different categories that the formula:

$$\mathrm{AIR} = \Sigma f_i^+ \cdot {}^i/12$$

gives the right answer (f_i^+ = is the corrected number of persons in class i).

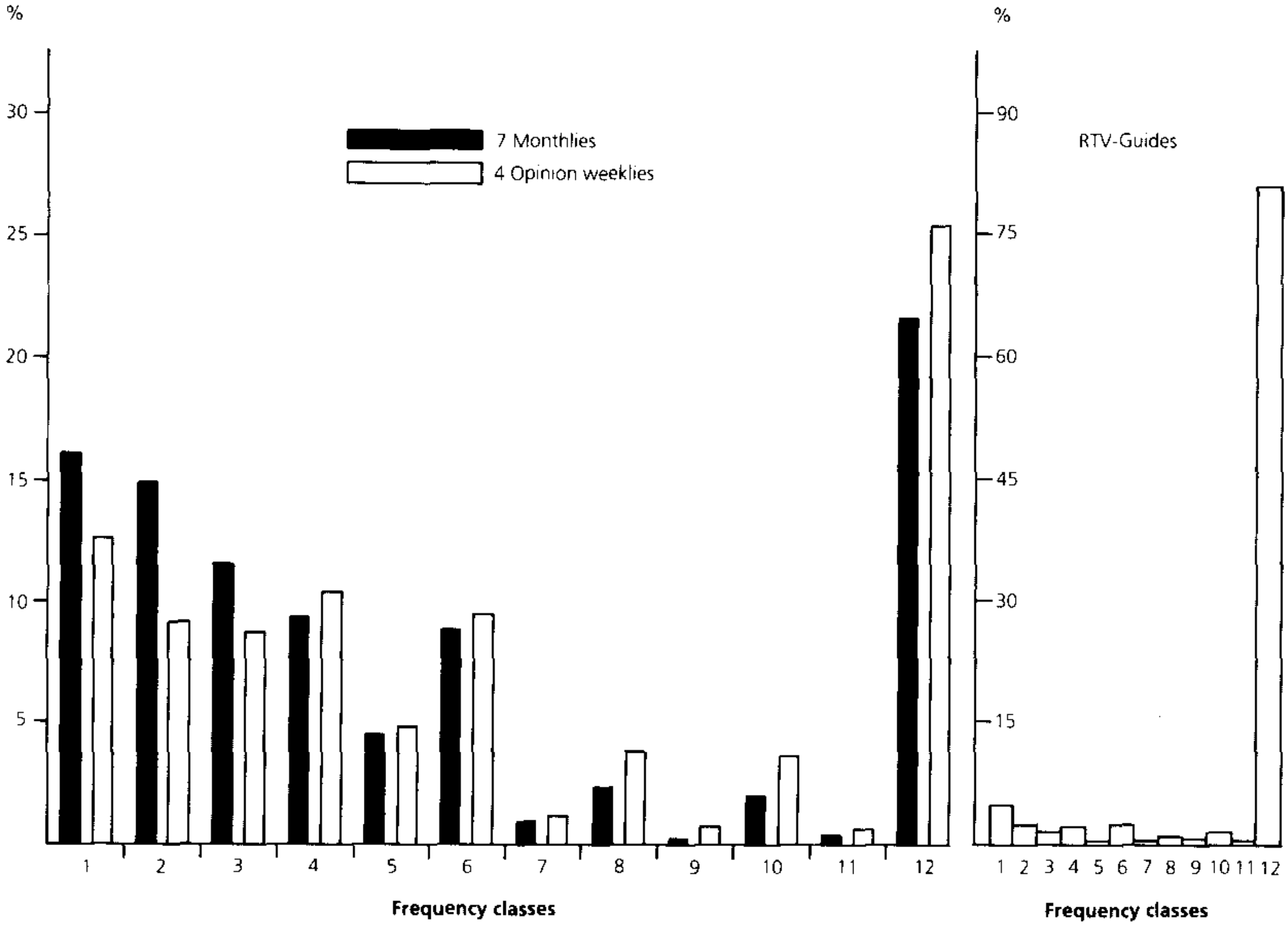
Problems remain, however. We can encounter the situation that the proportion four out of 12 readers that read the last issue exceeds the proportion of five out of 12, or even six out of 12 readers. For the monthly *Avenue* we found that the group four out of 12 with a theoretical reading probability of .33 had a reading probability of .37, whereas the six out of 12 group with .5 theoretically registered only .44, and the five out of 12 group (theoretically .42) reached .52. If the five and six groups were changed the result would have been much better.

To illustrate both these peculiarities **Figures 1** and **2** give results for combined groups of media.

As can be seen, taking together the frequency classes two by two will solve the problem of irregularities in occupation of the classes almost completely. It remains to be seen, however, if – as we would still suggest – asking the reading frequency in a shorter scale version will give

4.5 Frequency scales and their use

FIGURE 1
Relative frequency distribution



better results. Perhaps we should develop a non-numeric (whether verbal or pictorial) scale that bears no connotations with favourite digits for the respondent.

The big problem remains that respondents in most of the cases will overestimate their reading frequency. To give an illustration of this phenomenon the 12 out of 12 group for the monthlies scores an average reading probability of .83 with .72 as the lowest and .90 as the highest figure. In the case of the .72 reading probability even a shift to the category of nine out of 12 would give an over-rating. So we must ask ourselves whether shifts over this range are justified. We should bear in mind that it is not reasonable to put all the 12 out of 12 readers into the nine out of 12 group. There will be a not negligible fraction that really reads 12 out of 12 issues and these readers will have claimed to be last-month readers. If, however, we move only a fraction of this 12 out of 12 group to the lower end of the scale we must go much further down than nine. This move may be logical, but is it reasonable?

Overlap between media

Media planning programmes calculate the overlap between media as if the readership of the different media were independent, ie p(ab) = p(a).p(b). How should this be handled, since it can be shown that there is an amount of interdependence? We give examples of the audience figures calculated according to the formula above and

FIGURE 2
Reading probability and frequency class (measured values vs theoretical values)

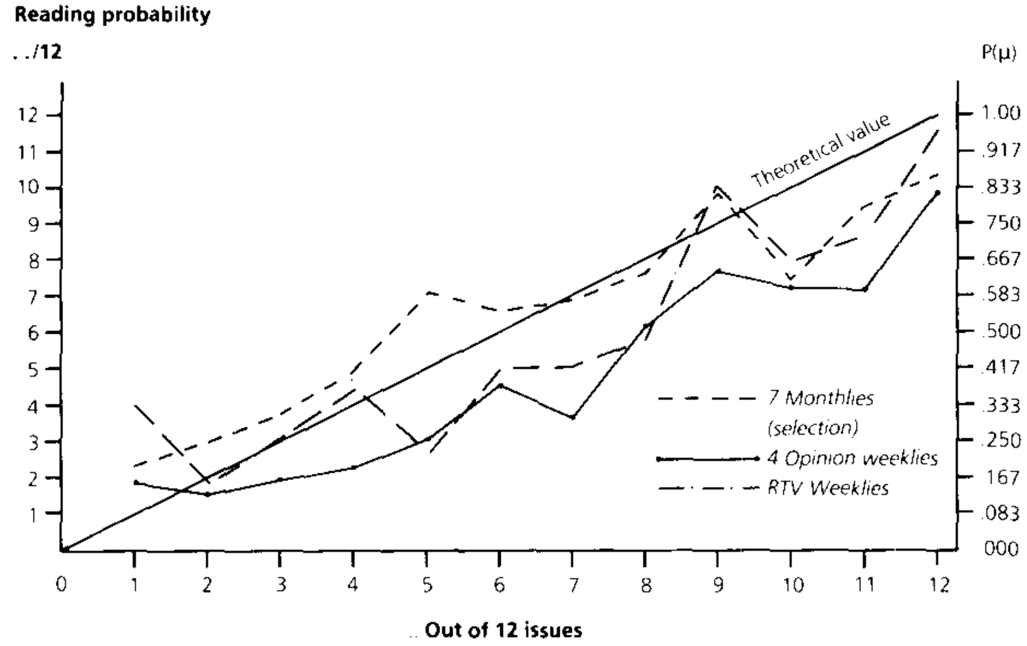


TABLE 2
Combined audiences (calculated versus direct measures)

	Formula P(ab . . . z) = P(a).P(b) . . . P(z)		Observed	
	abs × 1000	%	abs × 1000	%
Media group				
4 DHZ-Woonbladen	1677.4	15.9	1539.3	14.6
2 Handwerkbladen	1033.1	9.8	1000.4	9.5
4 Jongeren maandbladen	1412.5	13.4	1381.7	13.1
7 Vrouwenweekbl	6129.6	58.2	5685.3	54.0
9 RTV-bladen	7990.5	75.9	8064.1	76.6
PAN/NWR/TEV	3771.8	35.8	3492.2	33.2
4 Land avondbl	1038.8	9.9	1037.4	9.9

measured directly (**Table 2**).

In five of the cases the calculated audience exceeds the observed audience. For the four dailies the figures are almost the same and only for the RTV guides is the calculated audience smaller than the observed one.

If we look at the differences between the highest favoured group (the seven women's weeklies and the RTV guides) then we may conclude that the assumption of independence gives the women's weeklies a relative gain of 8.8%. This may seem only a slight gain but translated into media planning terms it can be decisive sometimes.

We can solve the problem by using the directly measured values if we speak of combinations with one issue of the titles under consideration. We have to develop a correcting method if we look at combinations of cumulated audiences. We may, however, assume that for cumulated audiences the error is smaller than for single issue combinations, for it is obvious that if there is independence we shall find it first among casual readers.

CONCLUSIONS

The problems in media research are manifold and complex. We could only investigate a small proportion of these problems and discuss probable solutions. We cannot give definite answers but we feel it acceptable to offer some recommendations.

First, we would advise you to be very critical of your questionnaire, so critical that if you have the slightest doubt you do a pilot test.

Second, we wish to emphasise that the usual binomial approximation will under-rate cumulated audiences and therefore should be replaced by the hypergeometric formula that gives exact results at least between the range determined by the number of intervals researched.

Third, we think it necessary to pool our efforts to develop a frequency scale that gives the respondent enough opportunities without disturbing him by its complexity. This scale must give smooth frequency distributions or we must develop a method to transform it into a smooth distribution.

Last, but not least, we think it advisable to do much more in the field of exchanging ideas and – if possible - solutions.