AUTOMATIC SEGMENTATION FOR REACH/FREQUENCY ESTIMATION OF NEWSPAPER SECTIONS AND INTERNET PAPERS

Kristian Arnaa, Taylor Nelson Sofres Gallup A/S
Peter S. Mortensen, The Aarhus School of Business

Synopsis

This paper will present a new way of estimating reach and frequency without asking the frequency question or conducting double interviewing. Instead, the sample is segmented automatically by a CHAID-analysis, maximising the differences in reading probabilities among the segments. Typically, many segments are created, individualising the reading probabilities more than when using frequency groups.

Two examples are presented: First, an experiment in which heavy users of the Internet are sampled on the Internet itself. The readers of each “Internet paper” are segmented by variables on their use of the Internet including types of sites visited, on their use of other media, on attitudes and on demography. Next, the Gallup Denmark’s new section readership measuring, being a separate telephone survey calculating reading probabilities of the sections to the readers of the paper in the national readership survey. Finally, the question is raised: Is the frequency question needed at all in readership measuring

1. Introduction Describing the Complexity and Limitations of Existing Systems

Up until a few years ago, the need for media information was mainly limited to readership for dailies and major magazines, ratings for the large TV-stations, and listening for the major radio stations. In addition to this, the need for extracting these data from one single source had not yet manifested itself, as media planning was conducted mainly with regard to solely one of the three media groups. In most countries, three overall surveys were conducted; the NRS covered readership, a diary or interview based survey covered radio listening, and either a people meter or a diary panel would be covering the TV-ratings.

In recent years, the development in this area has evolved rapidly due to improved computer facilities which permit the inclusion of a number of background variables (TGI, brands, life-style segmentation, etc) and an increased number of media groups (professional magazines, the Yellow Pages, out-door, etc). In addition to this, a need has arisen for collecting all these data in a common database rendering it possible to gather information across media groups, i.e. conducting multi-media planning.

In a number of countries, the development described above has led to the expansion of nationally syndicated surveys with regard to both the information bulk and the number of interviews. In Denmark, all information about dailies, magazines, local papers, out-door, the Yellow Pages, cinema, attitudes, interests, activities, consumption, ownership, and brands are united in a common database, the Index Danmark/Gallup, which is the official NRS in Denmark.

At the same time, we have come to the point where it is no longer justifiable to add any more questions to the existing interview. This would be of no consequence, if this development was coming to a halt impeding further needs for the inclusion of additional media groups and novel information. This is however not the case.

In Denmark as well as in the rest of the world, the last few years have seen the birth of a new media on the market: the Internet. The Internet is the fastest growing media in Denmark and naturally the need has arisen to document the use of this media among both advertisers and advertising agencies. Note in this connection that a number of traditional media (particularly dailies) have chosen to do an Internet version of their media.

In addition to the Internet, the market has expressed a wish for a higher degree of documentation of the already existing printed media in recent years. This need has risen partly due to the highly detailed documentation, which the TV-media is able to produce. In TV, ratings are made both for individual programmes and commercials whereas the printed media only establishes the number of persons who have been in a sensory position for ‘a random/average’ edition of the media in question. In this context, it is worth noticing that the figures in the NRS tend to be much more reliable due to the relatively larger sample and smaller bias.

The disproportion between the two media groups has given rise to speculation as to whether, in the case of dailies, it is possible to probe any deeper, providing figures being more reliable for the various subsections of dailies, e.g. readership on weekdays, readership of sections and supplements, and visits to the dailies’ Internet-sites. For these parts of the dailies, not only ratings are required, but figures which may be applied for reach/frequency-calculations as well as for various more qualitative media information.

In Denmark, it is commonly accepted that the Index Danmark/Gallup is not able to include all new media/media groups as single-source. Consequently, it is necessary to gather this information via external surveys and subsequently link these data to the Index Danmark/Gallup by means of one or more statistical techniques.

This general acceptance of the fact that all data do not necessarily have to come from the same set of respondents offers the possibility of constructing separate systems tailor-made to measure the media or problem which one wants to shed light on. By applying fusion, ascription or similar techniques, information is lost or is unreliable and the results are only to some degree correct. On the other hand, thorough questioning using a method, which has been adapted to that particular media group provides a substantially higher degree of quality in the basic material.

This also applies to measuring of the Internet which in itself provides a very complex problem and requires a substantial period of interview time. This is due to the fact that it is important to establish whether the Internet has been used, which sites have been visited (both main sites, sub sites and pages), how much time was spent there, when the visit was paid, what was actually read, and finally how often visits like these are made. These are all things, which necessitate a longer interview even if one wishes to measure a relatively small number of sites. Among other things, a definition of ‘visit’ and ‘reading’ is required, and it will be necessary during the interview to allow for the respondent to take the necessary time to remember when the visit was made.

As for measuring sections as described in (Arnaa;Randrup,1997), a similar problem is encountered, even when aiming to measure solely a small number of the existing bulk of published sections and supplements. Furthermore, this requires special questioning techniques in order to disclose reading over the course of time for the measurement of accumulated coverage.

Regardless of which technique is applied for transferring information from an external survey to the NRS, a number of common variables in the two surveys are required. Consequently, there is a demand for research investigating for each type of information, which variables are best suited to discriminate in order to ensure that these variables are included in both surveys.

As mentioned before, a distinct problem is securing information about reading over the course of time. In readership measuring, the most commonly applied method is letting the respondents allocate themselves to frequency groups and subsequently assign reading probability to each group. Nevertheless, getting respondents to make valid estimates of the frequency of their reading of sections and supplements and their visits to Internet sites is very difficult. As for sections/supplements, the problem is the number of sections per daily during the course of a week, and the sections’ various rates of publication. As for Internet sites, the problem lies in limiting the period plus the fact that a site may be visited several times per day. However, it will still be possible to filter never-readers (have not read the daily/do not have Internet access). Moreover, posing a more general frequency question to the respondents who pass through the filter, will be rendered possible. e.g.

- How often do you read media xxx? (every issue, quite often, sometimes, seldom, never)
- Did you read media xxx/visit site xxx within the last 3 months?

2. Segmentation Using CHAID

Organising respondents in a number of groups may solve the problem of getting information for calculating valid reach/frequency-estimates. The ratings for each group (the number of persons exposed in the group divided by the number of respondents) can be calculated and used as reading probabilities for the respondents. Such a procedure would be identical to the way frequency groups are used in some readership surveys, the ratings of the frequency groups being used as approximations for individual reading probabilities.

The organisation of groups has to be made in such a way that there is maximal variation in the ratings among the groups formed. The organisation will be individual for each media in the readership survey – and the groups of each media need not look alike. There are a number of statistical techniques available for organising a sample of respondents in a number of groups, called segments:

- **The Cluster Analysis** automatically organises the respondents in a number of segments in a way, so they vary as much as possible in relation to various selected variables. We only need a segmentation which maximises the variation in relation to one variable (read/not read). Thus, the Cluster analysis is not very appropriate for our purpose, but could be applied if no other technique were available.
- **The Discriminant Analysis** sets up functions from a number of variables. These functions disclose which of the outcomes of a classification variable, respondents are expected to belong to. This might look like what we need, the classification variable being read/not read. Using this technique would, however, only provide us with two groups, i.e. one group consisting of expected readers and one consisting of expected non-readers. Moreover, this technique entails a number of statistical assumptions and is not able to handle dependencies between the variables used in the functions.

The AID-analysis (Automatic Interaction Detection) is a non-parametric technique, where the maximal dependency between the classification variable (read/not read) and one of the other variables is used to allocate the respondents. This division is repeated for a number of levels in such a way that the potential segmentations of respondents are analysed for each outcome of the former level. The dependencies of the outcomes of the former level normally differ, so AID-analyses often result in different ways of segmenting the various outcomes.

The Cluster analysis and the AID-analysis are exploratory techniques and this is precisely what we need. We are not particularly interested in studying which variables are efficient and inefficient in segmenting the respondents according to their level of rating. We are primarily interested in organising the respondents in segments in such a way that the variation between the groups in relation to their ratings is sufficiently large. The maximal variation would be achieved with two groups with reading probabilities of 0% and 100% respectively, but we are not seeking a segmentation like that. The reason for this is that the segmentation is to be used as an approximation to individual reading probabilities, and they are mostly distributed in the full interval from 0% to 100%. Typically, the respondents are unevenly distributed with the largest proportion of respondents with zero reading probability for most printed media (never readers from the filter question) opposed to the largest but one which is close to 100% (loyal readers). A typical distribution is illustrated in Figure 1 for 'Hemmets Veckotidning', a Swedish magazine[1].

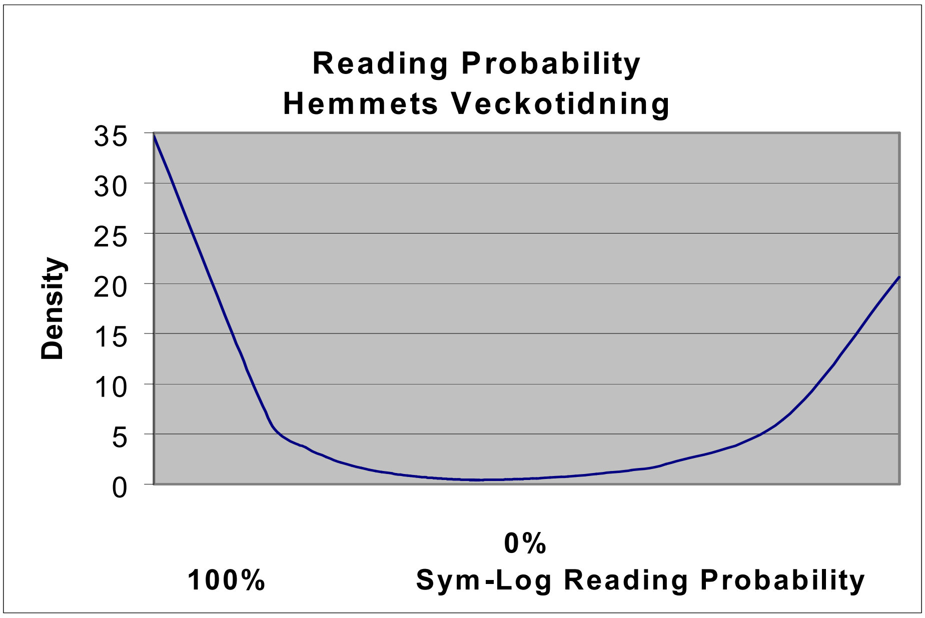

Figure 1: Distribution of individual reading probabilities

When searching for variables for allocating the respondents in relation to reading/non-reading of a printed media, one needs a measure to relate which of the variables in question is the best for segmenting the respondents. The measure chosen is values of the χ^2-distribution. The value can be calculated from the cross tabulation of 'read/not read' with the variable in question. The χ^2-value is calculated as the weighted squared deviations between the observed number and the number expected with independency. An illustration will be given by the cross tabulation of 'Visited the site of the Politiken[2]' with 'Subscriber - Politiken':

[1] A Beta distribution has been smoothed to a 12-weeks panel.

[2] The Politiken is one of the major Danish papers.

Politiken_www	Subscribing	Not subscribing	Total
Visited site	369 (**57.8%**)	692 (**17.7%**)	1061 (**23.3%**)
Not visited site	269 (42.2%)	3222 (82.3%)	3491 (76.7%)
Total	638 (100%)	3914 (100%)	4552 (100%)

Table 1: Cross tabulation: Visited site and being a subscriber

We wish to measure the effectiveness of the variable of subscribing in relation to segmenting the visitors of the Politiken site. From the column of 'Total', one may observe that 23.3% of the web-users have visited this site. Organising the Web-users into those subscribing and those not subscribing changes the rating to 57.8% and 17.7%, respectively. If subscribing had no effect, the ratings would still be 23.3% in both groups. This would mean an expected number of visitors among the subscribers of 0.233 * 638 = 148.7 instead of 369. The χ^2-value is calculated as the weighted squared deviation between these numbers for all four outcomes:

$$\chi^2 = \frac{(148.7-369)^2}{148.7} + \frac{(489.3-269)^2}{489.3} + \frac{(3001.7-3222)^2}{3001.7} + \frac{(912.3-692)^2}{912.3} = 495$$

χ^2-values are calculated for all potential variables in every new attempt to allocate a group of respondents. The variable with the largest χ^2-value or the smallest p-value[3] is accepted as the next variable for organising the respondents in question.

New segments are formed from all outcomes of the former level until certain criteria are met. In a CHAID-analysis three kinds of criteria can be set up:

- The p-value exceeds a given level of significance, typically set to α=5% or α=1%.
- The number of levels reaches a given maximal number of levels, set to 6-8 in the examples in this paper.
- The sample size of all potential new segments will be below a given minimal sample size, set to 25 or 50 in the examples in this paper.

The last rule on the sample size is that it is important to prevent the segments from becoming too small when the reading probabilities of each segment are ascribed to the same segments in the National Readership Survey. An analysis of this is presented in Section 4. Another way of meeting this problem is to merge segments with the same level of reading probability. The effect of this is analysed in Section 3.

The full segmentation may be illustrated in a tree diagram. The root and the first two levels are illustrated in Figure 2 for the Web-site of the Politiken ('Politiken_www'). The segmentation variables are 'Politiken: Computer', i.e. read the computer section in the Politiken last week and 'Politiken', i.e. read the Politiken (the newspaper) yesterday. In the outcome to the left of the Figure, one can see that only 12% of the respondents who neither read Politiken yesterday nor read Politiken Computer this week, visited the Web-site of the Politiken.

[3] The p-value is defined as P ($\chi^2 > \chi^2$-value). The p-value has to be used when there are 3^+ outcomes of the segmentation variable.

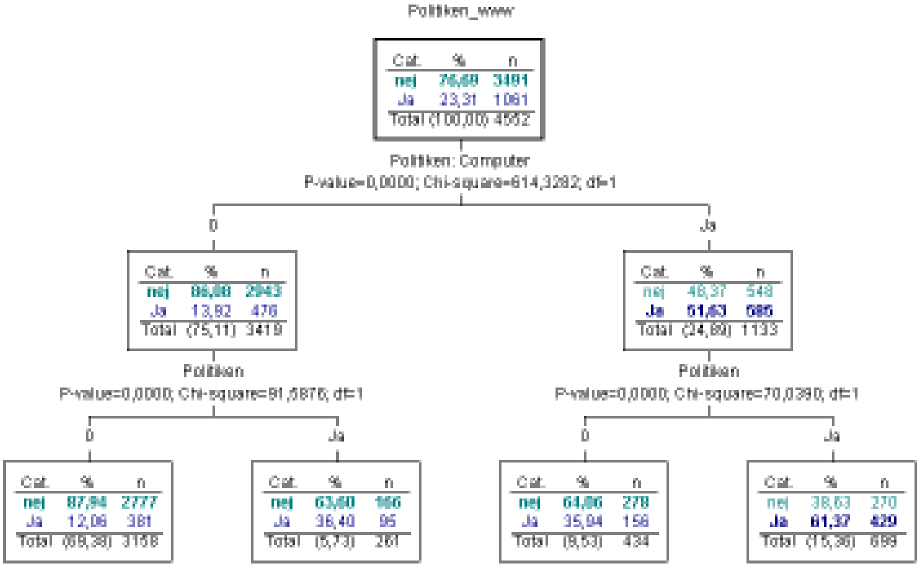

Figure 2: Tree diagram from CHAID, first two levels.

New, more advanced procedures for doing Automatic Interaction Detection have improved the usefulness of AID. The original AID-procedure assumes that all variables are recoded to only two outcomes and the procedure is purely hierarchical. The automatic part of the procedure is limited to the selection of variables to be used for allocating the respondents. Newer procedures include more automatic elements:

- accepting more than two outcomes in any variable and even using quantitative variables like income and age directly.
- automatically selecting if and how the outcomes of variables are to be classified: number and contents of outcomes.
- being non-hierarchical or recursive, which means that for every new level it is investigated whether changes have to be made on the former levels to further maximise the variation of the reading probability. These changes could concern both the variables chosen and their classification.

The automatic classification of outcomes differs from categorical to quantitative variables. With categorical variables the procedure checks whether some outcomes may be merged, e.g. geographic groups. With quantitative variables the procedure calculates how the outcomes are best split in intervals, so that the variation is maximised. A procedure including all these facilities is called CHAID - CH referring to the use of the χ^2-distribution as a measure of discrimination. Software for doing CHAID-analyses exist, see e.g. AnswerTree from SPSS.

Two examples of the use of CHAID-segmentation in readership measuring studies will be presented in the next sections: An experiment on estimating exposures to Web-sites from printed media and the estimation of reach/frequency for newspaper sections in the Danish National Readership Survey.

3. Web-sites For Printed Media.

A large number of printed media have launched their own Web-site and most of the remaining media are expected to follow. In Denmark, we have Web-sites from most of the national and local newspapers as well as from a number of magazines, dependent of the theme of the magazine. New sites from printed media turn up regularly. The same growth of sites for printed media may be observed in all Western countries.

The background for this development is that the Internet has turned into a new kind of media, which may be called an electronic media, but also bears resemblance to a printed media via the way visitors actively select which sites and pages they are to be exposed to. The explosive growth in the use and the number of users means that new types of visitors are in front of the screen. It is no longer just younger well-educated men. Of course, there is still a considerable demographic bias, illustrated in Table 2, but the bias is levelling fast.

	Web-sample	NRS
Women	24%	47%
Age: >50	2.5%	31%
Social democrats	15%	(25%)

Table 2: Demographics, a Web-sample and the NRS.

Other things have changed. The Internet is now being used less for surfing aimlessly around and more for deliberately visiting specific sites using bookmarks and links. In this way, visitors become customers of the sites. Also, the financing of the sites is gradually changing from free services and voluntary work to advertising and subscription.

Many sites on the Internet are thus looking more and more like other media, and they have to be handled accordingly by their publishers, as well as by advertisers and media planners. This means that we need to conduct surveys or create panels in order to estimate ratings and reach/frequency for the commercial sites, including the sites from printed media. This has to be conducted in such a way that reach/frequency for the sites of the printed media may be combined with the similar figures for the printed media themselves.

As of yet, there is no national survey for Internet sites in Denmark, but an experimental survey was conducted in two weeks of January 1999. The survey had multiple objectives, including the objective of surveying both printed media and Web-sites among Internet-users. Another objective was to test the cheapest way of collecting data directly from the Internet: would the answers be valid; would the sample be representative for Internet-users.

The data was collected from 21 selected Web-sites, of which 6 were from printed media. A number of randomly selected visitors were exposed to a banner inviting them to complete a questionnaire – to improve the site and to participate in a lottery and win cash. 4552 replies were submitted in the two weeks. The sites were not able to state the total number exposed, but it is assessed to be considerably higher. This means that the response rate of the study is rather low. Thus, the ratings and other figures from the survey may be used as estimates for all 1.4 million Danish Internet-users only with great reservations.

Ratings were measured in this way:

- For printed media as 'reading any issue for the first time in the recent publication period', i.e. FRR-reading.
- For Web-sites the period was set to a week: 'Visited site during last week'.

A large number of other questions were posed, including questions regarding demographics and attitudes. These questions can be used for CHAID-segmentation and for an ascription with the NRS.

Ratings have been calculated for the printed media and their Web-sites. These numbers are compared with ratings for the printed media in the Danish NRS, Index Danmark/Gallup. In Table 3, ratings are given for the 6 papers and magazines included in the data collection and 10 papers/magazines not included (papers: weekday issues).

Paper/magazine		Web sample		IndexDK
Title	Type	Website	Print	
Aktuelt	small national	5.6	3.7	3.1
Berlingske Tidende	national	18.0	20.0	12.1
Børsen	financial	13.5	8.2	4.3
Ekstra Bladet	tabloid	17.8	17.8	12.7
Fyens Stiftstidende	local	6.9	6.9	4.5
Information	small national	3.1	3.2	2.2
JyllandsPosten	national	19.4	24.1	17.5
Politiken	national	23.3	21.1	12.2
Århus Stiftstidende	local	3.4	4.5	4.4
Alt for Damerne	women weekly	8.2	8.3	9.0
Den Blå Avis	priv ad-mag.	14.2	8.8	13.2
Chili	youth monthly	11.8	7.1	5.8
ComputerWorld	PC-weekly	12.5	12.5	3.5
Ingeniøren	prof magazine	5.3	6.6	3.9
Se & Hør	gossip weekly	9.2	16.7	24.7
Tjeck	youth monthly	4.0	5.9	3.7

Table 3: Ratings for printed media and their Web-sites.

The ratings for each printed media and its Web-site are not equal among the Internet-users and are also different from the national ratings. The next question, relevant for reach/frequency calculations, is then what proportion of readers is visiting the site of the paper/magazine read - and vice versa.

Title	Type	Readers visiting	Visitors reading	Duplication
Aktuelt	small national	49.1	32.7	1.8
Berlingske Tidende	national	47.0	52.0	9.4
Børsen	financial	56.0	34.1	4.6
Ekstra Bladet	tabloid	38.9	38.8	6.9
Fyens Stiftstidende	local	63.1	62.7	4.3
Information	small national	36.6	37.3	1.2
JyllandsPosten	national	42.4	52.8	10.2
Politiken	national	54.6	49.4	11.5
Århus Stiftstidende	local	35.1	47.1	1.6
Alt for Damerne	women weekly	12.6	12.9	1.1
Den Blå Avis	priv ad-mag.	42.5	26.4	3.8
Chili	youth monthly	44.8	27.1	3.2
ComputerWorld	PC-weekly	41.9	41.8	5.2
Ingeniøren	prof magazine	43.9	54.8	2.9
Se & Hør	gossip weekly	23.0	41.7	3.8
Tjeck	youth monthly	29.7	44.2	1.8

Table 4: Proportion readers visiting site and vice versa.

Table 4 reports these proportions. The proportion of FRR-readers having visited the site within the last week vary from 13% to 63% and a close to similar variation is observed for the proportion of visitors to Web-sites reading the printed media of the Web-site within the last publication period. This is a rather large difference in the overlap between Web-sites and printed media for just 16 papers/magazines. Another measure of this overlap would be the duplication defined as the proportion being exposed to an issue/period of two media – in this case a printed media and its Web-site, see the right column in Table 4. This variation in the amount of overlap makes it very important to get valid estimates of the duplications or some other measure of accumulation to ensure the validity of the reach/frequency calculations, involving printed media and their Web-sites.

The duplications and the ratings may be used as input to model based calculations of reach/frequency, e.g. the Beta Binomial Distribution. In Table 4, the duplications between the printed media and their Web-sites were calculated on the basis of the Internet survey. In the same way, it is possible to calculate the duplication between different printed media and between different Web-sites.

However, the duplication between reading the same printed media in two different publication periods or the duplication between visiting the same Web-site in two different weeks cannot be calculated from this Internet survey. We could have considered asking a frequency question like ‘number of exposures out of 4 issues/periods’. This would have been in accordance with the Danish national readership survey for printed media, but would probably not give valid answers for Web-sites. Finally, it would increase the number of questions considerably, having no filter question in the survey.

One could have asked questions about exposure to the last two (publication) periods, but it was not expected to give valid answers, see an experiment by Copeland;Mennenga(1988). Being expensive and time-consuming, breaking the anonymity and giving more problems with the response rate, interviewing the same respondents twice like in the former SMRB-surveys was not an attractive alternative either.

Instead, an experiment with CHAID-segmentation was conducted. Segments of respondents were formed for each of the printed media having a Web-site and for the Web-sites themselves with ‘read/not read’ and ‘visited/not visited’ as classification variables. Ratings per segment would be accepted as reading probabilities for each segment and used to estimate reach/frequency, if it was possible to form segments for each media with a nice variation in the ratings - from 0% to 100%. There would then be two ways of making use of the reading probabilities for reach/frequency-calculations:

- as approximations for individual reading probabilities, assigned to each respondent like in the traditional way of using frequency groups

or

- as the basis of calculation of duplications within each media, printed, or Web-based, for use with the in between duplications as input to model based calculations of reach/frequency.

A very large number of variables were introduced in the CHAID-analyses to make sure that all possible segmentations were tried for each media. In this way, both the most efficient segmentation for each media and the most efficient variables for segmentation were uncovered. The variables introduced may be classified in the following categories:

- Exposure to another media from the same publisher (Weekday, Sunday, IT-section, Web-site) and subscription.
- Extent of visits to different kinds of Web-sites. The 70 sites in the survey were divided into 16 types of Web-sites using a factor analysis.
- Functions performed on the Web: surfing, sending e-mails, chatting, shopping etc.
- Reading of other printed media, either reading specific media or the amount of reading of types of printed media (papers, weeklies, monthlies, IT-sections).
- Exposure to major radio and TV-channels.
- Demographics (sex, age, housing, education, occupation, income, geographics).
- Attitudes: General (values of life) and specific (shopping on the Internet, banners, etc).

After conducting the CHAID-analyses for all the printed media and their Web-sites included in the survey, one can see that almost all variables are used at least once to split a group of respondents. However, some variables are used in almost all analyses, while others are only used once. Every single analysis includes variables on other media from the same publisher and also some site-types, but not the same ones for all media. In fact, all 16 site-types are used once or more. Variables on functions performed on the Web are used, but to a very limited extent. A few variables on exposure to other media are also included in most of the analyses in such a way that most printed and electronic media are used a couple of times. Variables on attitudes and demographics are included very little, but a summary of all analyses show that most of these variables are used, often just once. The conclusion is that all seven types of variables do contribute to the segmentation of the readers and visitors, yet in a varying degree. The question is, whether the analyses can be just as efficient without some of the types of variables or without some specific variables. This problem is of special interest if the reading probabilities are to be ascribed to respondents in the national readership survey.

The number of segments in the CHAID-analyses varies a great deal, in fact from 13 to 60 segments when only counting segments with different ratings. This raises the question of whether a different number of segments will influence the level of reach/frequency, so that e.g. the net reach calculated from 60 segments could not be compared with a net reach calculated from 13 segments. This is an even more serious problem if the reading probabilities for the Web-sites are to be ascribed to the national readership survey. The Danish national readership survey includes 7 frequency groups, so the CHAID-based segments were combined to 7 - no matter the starting number - and used to test the effect of having a different number of segments. Intervals of reading probabilities were fixed for each of the 7 segments. A symmetric approach was used bringing the levels close to the Index Danmark/Gallup – see the first column of Table 5. The duplication was selected as the prime measure to be used, when evaluating the effect of changing the number of segments to 7 on reach/frequency. Also, the net reach was used. The calculation for the Web-site 'Politiken_www' is illustrated in Table 5. The original 54 segments have been combined to form 7 like this:

Segment	Sample	% sample	Exposed	Rating	Dupllication
1.0	70	1.54	70	100.0	1.538
0.9 - 0.999	101	2.22	97	96.0	2.048
0.625 - 0.9	520	11.42	408	78.5	7.033
0.375-0.625	479	10.52	233	48.6	2.490
0.100-0.375	1009	22.17	212	21.0	0.979
0.010-0.100	806	17.71	39	4.84	0.041
0-0.010	1567	34.42	2	0.13	0.000
Total	4552	100.0	1061	23.3	14.13
			54 segments:		14.47

Table 5: Segmentation of Politiken_www by CHAID

The effect for the Politiken_www is that the duplication is reduced from 14.47% to 14.13%, i.e. 2.4%, when the number of segments is reduced from 54 to 7. This is in accordance with the effect of all media: the duplications are reduced for all papers and Web-sites when reducing the segments to 7. The average reduction is 1.7% with a maximum of 2.9%. The effect on net reach has been calculated for the media with maximal reduction. The net reach increased by 2.7% in a plan with 12 issues when calculated in the Beta Binomial Distribution. From these figures, one may conclude that the effect of reducing 13-60 segments to 7 segments is modest. This means that the CHAID-segmentation is a very robust way of creating groups of respondents with different reading probability. This also means that it seems possible to exclude some of the less used variables in the CHAID-analyses if they are not included in the national readership survey. As a test of that, the variables on attitudes and functions performed on the Internet were excluded and none of the new duplications were below the level of the 7 segments.

The CHAID-analyses' ability to give robust estimates of duplication and net reach does not guarantee that the level of the derived reach/frequency-calculations is valid and unbiased, i.e. being at the correct level. When evaluating the validity, we can once more turn to duplication as a measure for accumulation. Research and experiments have shown that estimates of duplication from frequency groups in most cases turn out to be smaller than those of a panel or from double interviewing, see e.g. Richard;Frankel(1983) and McGlathery;Eadie(1995). The same trend must be expected to be true for duplications calculated from CHAID-based groups. Thus, the recommendation would be to use the reading probabilities from the original segments instead of the ones from the 7 combined segments. In this way, the level of duplication and reach/frequency must be expected to be closer to the true level.

The validity of the CHAID-based duplications can be further examined using data from the Web-survey. An alternative duplication for the Web-sites can be calculated from a kind of double questioning. Two pieces of information from the survey permits this:

- The site, from which the questionnaire is answered (being registered automatically).
- The site, from which the questionnaire is answered, will be among the sites, for which a question on 'visited last week' is asked.

This duplication will report the proportion of respondents having visited the Web-site in the week prior to answering of the questionnaire, which was conducted on the web-site itself. This is not identical with the CHAID-based duplication, neither concerning the respondents in the sample, nor the time period. Also, this duplication can only be calculated for the 6 papers and magazines which were part of the Web-sample.

Table 6 compares the three estimates of the duplication for the 6 Web-sites, based on CHAID-segments, merged CHAID-segments, and double questioning:

			Duplication		
Web-site	Type	Rating	Double question	CHAID 7 segments	CHAID 13-60 segments
Politiken	Nat. Paper	23.3	18.0	14.13	14.47
Fyns Stiftstid.	Local paper	6.90	5.64	4.79	4.81
Børsen	Fin. Paper	13.5	9.35	8.64	8.77
Ingeniøren	Prof. Weekly	5.29	2.65	2.58	2.65
Chili	Youth monthly	11.8	6.36	6.17	6.26
Tjeck	Youth monthly	3.98	2.37	2.33	2.36

Table 6: Rating and duplication for web-sites from printed media.

The trend is the same as in the cited experiments. Duplications based on groups are smaller than those based on double questioning for all 6 Web-sites. The differences are very modest for the three magazines, while the differences for the papers are significant. However, it seems more probable that these differences are caused by effects of prestige and local patriotism and by confusion with the paper itself than by an increased bias in the estimation of duplication for newspapers compared with magazines caused by the CHAID-segmentation. Further insight into this issue may be obtained from analyses for the 15 Web-sites in the sample not published by a printed media and from new experiments with a more valid double question or with double interviews.

All in all, this experiment indicates that it is possible to use CHAID-segmentation for reach/frequency-calculations for media plans including Web-sites of printed media.

4. Section reading as an integrated part of the Index Danmark/Gallup

In recent years, the Danish dailies have increased the number of sections and supplements. As a rule, these sections are quite thematic and are quite often aimed at special target groups. A typical weekday daily contains 4-5 sections of which 2-3 are published every day and 2-3 vary from day to day. The Saturday issue is more or less identical to the weekday issues, whereas the Sunday issue is quite different containing up to 10 different sections. Actual magazine-like supplements most often appear in connection with the Saturday or the Sunday issues.

In Index Danmark/Gallup, it has been our objective from the start to incorporate all essential sections in the same way as other media. From the beginning, it was determined that incorporating section reading as single source would not be possible as any extension of the duration of the interview would have to be marginal.

As reading probabilities provide the foundation for media planning for any kind of media allocation of reading probabilities to all respondents not only to the main newspaper, but to the individual section as well, constitutes the main core of the problem of reporting section reading.

Consequently, Gallup has been conducting a number of tests of various methods which are able to combine data from different sources/respondents. During 1998, it was agreed to perform a major experiment transferring data by means of CHAID from special surveys to the Index Danmark/Gallup.

4.1 Method

First we had to analyse which variables best discriminated whether a person would read a section or not. This was analysed by investigating a selection of sections which subsequently would provide the basis for a common choice of variables to be included in the test.

The test showed that the essential media related variables were; whether the newspaper was the respondent's primary newspaper, reading time for the entire newspaper and source-of-copy. Furthermore, a number of demographic and geographic variables proved to have good discriminating properties as well.

As these demographic and geographic criteria already existed in both Index Danmark/Gallup and in the special survey, these variables caused no problems. The questions regarding primary newspaper, reading time, and source-of-copy were added to the Index Danmark/Gallup for the three major national dailies. In addition to this, a special survey containing the same information in addition to the regular questions (filter, period and frequency) about reading of the dailies in question was established. In case a daily had been read the previous day, reading of the various sections which the daily in questions had included, would be inquired into.

The special survey applied the same selection procedure and universe as the Index Danmark/Gallup. This was necessary in order to obtain directly comparable results. All in all, approximately 30,000 interviews were conducted in this special survey which covered a total of 55 sections, of which 9 appeared on all 6 weekdays, 28 were issued only on Sundays, and 18 were distributed once a week on a weekday.

For each section, a database was created consisting of the respondents who had had the opportunity to read the section in question (the respondents who were interviewed the subsequent day who had read the main newspaper). As an example, 4,618 interviews were conducted on a Saturday and of these, 685 had read the ‘Berlingske Tidende’ (major national newspaper) the previous day (Friday). Thus, these 685 respondents formed the database for the reading of ‘Guide’, a section on news in film, music, museums, restaurants, etc which is issued as part of the Berlingske Tidende on Fridays.

For this section, a CHAID analysis was conducted, investigating which variables and combinations best classified the respondents according to their reading or non-reading of the section.

It was determined as stopping rules for the analysis that a maximum depth of 6 levels was desired and that no segment should contain less than 25 respondents. Below, the result is shown in figure 3.

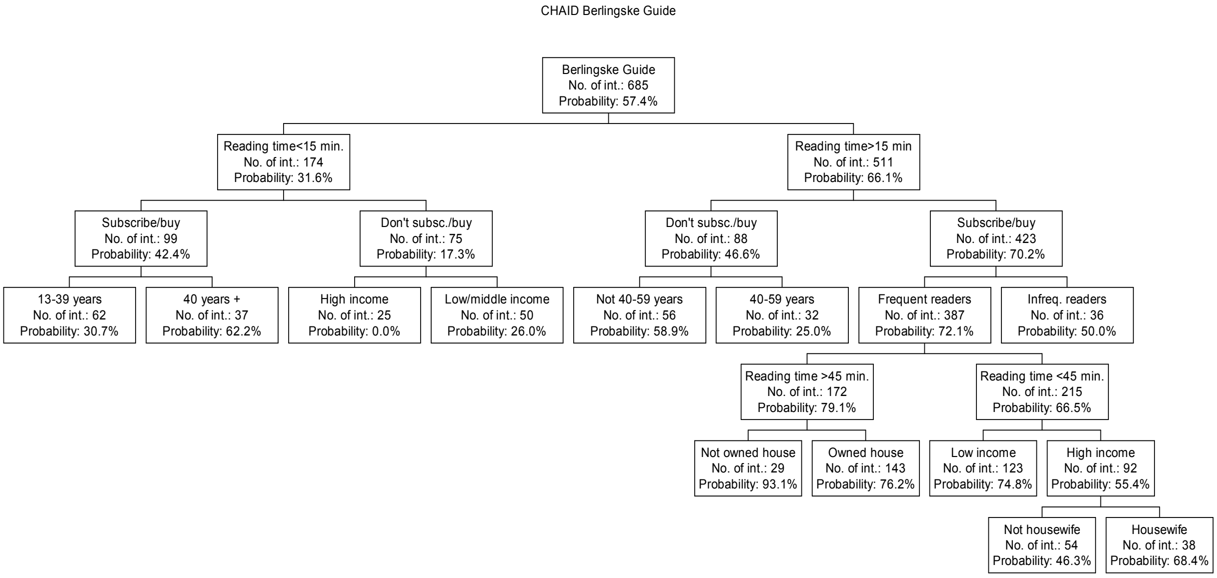
CHAID Berlingske Guide
Berlingske Guide No. of int.: 685 Probability: 57.4%
Reading time<15 min. No. of int.: 174 Probability: 31.6%
Reading time>15 min No. of int.: 511 Probability: 66.1%
Subscribe/buy No. of int.: 99 Probability: 42.4%
Don't subsc./buy No. of int.: 75 Probability: 17.3%
Don't subsc./buy No. of int.: 88 Probability: 46.6%
Subscribe/buy No. of int.: 423 Probability: 70.2%
13-39 years No. of int.: 62 Probability: 30.7%
40 years + No. of int.: 37 Probability: 62.2%
High income No. of int.: 25 Probability: 0.0%
Low/middle income No. of int.: 50 Probability: 26.0%
Not 40-59 years No. of int.: 56 Probability: 58.9%
40-59 years No. of int.: 32 Probability: 25.0%
Frequent readers No. of int.: 387 Probability: 72.1%
Infreq. readers No. of int.: 36 Probability: 50.0%
Reading time >45 min. No. of int.: 172 Probability: 79.1%
Reading time <45 min. No. of int.: 215 Probability: 66.5%
Not owned house No. of int.: 29 Probability: 93.1%
Owned house No. of int.: 143 Probability: 76.2%
Low income No. of int.: 123 Probability: 74.8%
High income No. of int.: 92 Probability: 55.4%
Not housewife No. of int.: 54 Probability: 46.3%
Housewife No. of int.: 38 Probability: 68.4%

Figure 3: CHAID analysis for the Berlingske Guide

Thus, the final result is 12 segments all of which have a reading probability between 0.0 and 93.1% assuming that they have read the main paper. Below, the same segments are displayed in table 7.

It appears from the table that the general reading probability for Guide of 57.4% covers a wide spectrum of different reading probabilities. The highest probability, 93.1%, was achieved in the group whose reading time was more than 45 minutes, who subscribe to/buy the newspaper, who are regular readers, and who live in a habitation which they do not own. All in all, 29 respondents of the 685 met these criteria. The lowest probability (0.0%) was found among the group whose reading time was less than 15 minutes, who do not subscribe/buy the newspaper, and have a high income. A total of 25 respondents out of 685 met these criteria.

Corresponding analyses were conducted for the remaining 54 sections. The results is that for the 55 sections, there is a number of segments ranging from 8 to 72 showing the reading probabilities for the section assuming that the main newspaper has been read. For each section, the respondents have been allocated to one segment only. Consequently, all segments for the individual section are mutually independent and add up to a total of 100%.

Seg-ment	1. level	2. level	3. level	4. level	5. level	6. level	Reading probabil.
1	Reading time < 15 min.	Subscribe/ buy	13-39 years				30,7%
2	Reading time < 15 min.	Subscribe/ buy	40 years +				62.2%
3	Reading time < 15 min.	Don't subscribe/buy	High income				0.0%
4	Reading time < 15 min.	Don't subscribe/ buy	Low/middle income				26.0%
5	Reading time >15 min.	Don't subscribe/ buy	Not 40-59 years				58.9%
6	Reading time >15 min.	Don't subscribe/ buy	40-59 years				25.0%
7	Reading time >15 min.	Subscribe/ buy	Frequent readers	Reading time > 45 min.	Not owned house		93.1%
8	Reading time >15 min.	Subscribe/ buy	Frequent readers	Reading time > 45 min.	Owned house		76.2%
9	Reading time >15 min.	Subscribe/ buy	Frequent readers	Reading time < 45 min.	Low income		74.8%
10	Reading time >15 min.	Subscribe/ buy	Frequent readers	Reading time < 45 min.	High/middle income	Not housewife	46.3%
11	Reading time >15 min.	Subscribe/ buy	Frequent readers	Reading time < 45 min.	High/middle income	Housewife	68.4%
12	Reading time >15 min.	Subscribe/ buy	Infrequent readers				50.0%

Table 7: 12 segments for reading the Berlingske Guide

The next step was to regenerate the same segments in the Index Danmark/Gallup. This was rendered possible by the fact that the analyses had been conducted using variables which were also found in the Index Danmark/Gallup including the new media data (primary newspaper, reading time, and source-of-copy).

Once a segment has been established in the Index Danmark/Gallup, this segment is subdivided into 7 groups according to the respondents' reply to the frequency question (reading probability for the main newspaper). The individual respondent is subsequently allocated a reading probability for the section as a multiple of the segment's reading probability for the section and the respondent's reading probability for the main newspaper based on the frequency question. This is shown for the section 'Guide' in Table 8 below.

Segment	Reading probability	No issues at all	Almost no issues	1 of 4 issues	2 of 4 issues	3 of 4 issues	Almost every issue	Every issue
Reading probability		2.0%	10.0%	26.0%	47.0%	72.0%	81.0%	92.0%
1	30.7%	0.6%	3.1%	8.0%	14.4%	22.1%	24.9%	28.2%
2	62.2%	1.2%	6.2%	16.2%	29.2%	44.8%	50.4%	57.2%
3	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
4	26.0%	0.5%	2.6%	6.8%	12.2%	18.7%	21.1%	23.9%
5	58.9%	1.2%	5.9%	15.3%	27.7%	42.4%	47.7%	54.2%
6	25.0%	0.5%	2.5%	6.5%	11.8%	18.0%	20.3%	23.0%
7	93.1%					67.0%	75.4%	85.7%
8	76.2%					54.9%	61.7%	70.1%
9	74.8%					53.9%	60.6%	68.8%
10	46.3%					33.3%	37.5%	42.6%
11	68.4%					49.2%	55.4%	62.9%
12	50.0%	1.0%	5.0%	13.0%	23.5%			

Table 8 – total reading probability for the Berlingske Guide

As an example, according to the figure above, a person whose reading time is more than 15 minutes who doesn't subscribe to or buy the media in question and who is in the age group of 40-59 years (segment 6) has a reading probability for Guide of 25% assuming that he or she reads the Berlingske Tidende. If that person at the same time replies in the affirmative to reading half of all Berlingske Tidende issues (reading probability of 47%) his reading probability for Guide is 0.25*0.47=11.8%. The blank cells are non-existing possibilities as the frequency question contributes to the formation of the segment.

Similar calculations were conducted for all respondents for all sections. Subsequently, the readership for the individual sections was calculated in a way similar to the one normally applied for calculating readership for the main newspaper. As a result hereof, the majority of the sections were given readership figures within the statistical confidence limits compared to the special survey. Comparisons were made to the share of readers of the main newspaper who had also read the section.

For the sections where discrepancies were found (some were given readership which was too high, others too low) the cause of the discrepancy was determined. In all cases the problem was that the segment's part of the universe of the special survey was significantly different from that of Index Danmark/Gallup, or that there was a marked difference from segment to segment as to the shares of regular and sporadic readers of the main newspaper in the segments. All in all, just 4 sections out of 55 were outside the accepted limits.

The reading probabilities for the 51 sections were entered into the Index Danmark/Gallup for tests among all users, and enjoyed a fine reception. Media, agencies and advertisers alike used the figures and generally agreed that this was a viable method for the incorporation of readership of newspaper sections into the Index Danmark/Gallup.

As a result hereof, the special survey has been expanded by 4 local newspapers bringing the total up to 112 sections to be reported on. The special survey has become continual and consequently, readership of sections is being reported on a quarterly basis (6 months running) similar to other readership.

At present, it is being looked into whether it would be advantageous to include any more joint questions targeted at the themes of the sections, e.g. whether the respondent is interested in the theatre, cars, or computers.

5. The Future

The CHAID method offers a number of applications within readership measuring in addition to the ones for which it is already being applied. Consequently, it seems natural to explore the areas to which this method may be applied in order to improve the quality of the information which is already being gathered today as well as establishing where it may be applied to provide new kinds of information than the ones collected today.

5.1 Application for improving the quality of existing data

The most obvious place to apply CHAID in order to improve already existing data is in connection with the allocation of reading probabilities for all types of printed media.

At present, the reading probability for a media is calculated solely based on the allocation into 7 frequency groups. By including CHAID in this process, it becomes possible to operate using a much larger number of segments than 7 and possibly also with segments which are spread more between 0-100%, i.e. the groups will become more homogeneous in relation to their reading of the media. Naturally, filter questions still have to be posed, and the ones who have not read the media in question within the filter period will be given a reading probability of 0.

Similarly, CHAID may contribute to reduce the occurrence of errors caused by respondents who misinterpret the frequency question as these groups may be further subdivided into smaller groups with different reading probability. It is vital that a number of relevant conditions for the allocation of the material with regard to the number of levels as well as minimum number of respondents in each group are laid down.

Below, the result from a test of the Berlingske Tidende, one of Denmark's major, national dailies is shown. The period figure has been established as the dependent variable. The Index Danmark/Gallup 2. half year of 1998 consisting of 9,811 interviews has been applied. For the CHAID-analysis in itself, it has been determined that each segment should contain a minimum of 50 respondents, and that the maximum depth should be set at 8 levels.

The overall filter is that respondents must have answered in the affirmative to the filter question (read within the last 3 months). The variables used are demographic, geographic, and frequency questions, interests, activities, attitudes as well as reading of other media (recency- + filter question).

This analysis resulted in 36 segments with reading probabilities varying from 1% to 100%. The deepest level was 8. The overall variables applied are listed here according to topic:

- The frequency question
- The recency question for the Politiken (rival, national daily), the Sunday edition of the Berlingske and the Weekendavisen (other intellectual Sunday paper from the same publisher).
- Filter question for the B.T. (tabloid from the same publisher), Ide-Nyt, Ud & Se, Samvirke, Motor, Falck-Bladet (all 5 members' journals/free omni-distributed magazines), Hjemmet and ComputerWorld (2 priced weeklies)
- Position in the household
- Attitude to doing without one's daily paper
- How interested one is in reading the news in the paper
- How often one makes home-baked bread, cakes etc
- How often one listens to pop/rock music

The two segments with the smallest reading probabilities are defined as follows:

1. Has not read the Sunday issue of the Berlingske within the last week. The Greater Copenhagen Area/the HT (Copenhagen City and Regional transport) area, reads 1 out of 4/1 out of 2 issues of the Berlingske Tidende, is not very interested in the news, has read BT within the last three months.
2. Has not read the Sunday edition of the Berlingske within the last week, Lolland/Falster/Bornholm/Jutland, has not read the Politiken yesterday, reads no issues at all/almost no issues/1 of 4 issues of the Berlingske Tidende, has read Ud & Se within the last year, has not read Computer World within the last year.

The first segment consists of 52 respondents and has a reading probability of 0%. The second segment consists of 224 respondents and likewise has a reading probability of 0%.

The segment with the largest reading probability is defined as follows:

1. Has read the Sunday issue of the Berlingske Tidende within the last week, very interested in reading the news in the papers, reads all issues of the Berlingske Tidende, doesn't listen to pop/rock music more than once a week, has read the Weekendavisen within the last week.

This segment consists of 123 respondents and has a reading probability of 100%.

It appears, that newspaper reading helps divide this material into reading-wise homogeneous groups. One of the reasons why the frequency question hasn't consequently been included and divided in all its replies, is caused by the fact that the middle groups contain relatively few replies, and consequently, these do not meet the requirement for a minimum of 50 respondents in each segment.

Table 9 compares the frequency group method with the test. The information provided by the table lists probability intervals, number of segments, and the number of respondents in each segment.

	Frequency question		Test	
	Number of		Number of	
Probability	Segments	Respondents	Segments	Respondents
1.00	0	0	1	123
0.900-0.999	0	0	3	251
0.800-0.899	1	744	2	143
0.700-0.799	0	0	4	254
0.600-0.699	1	145	2	126
0.500-0.599	0	0	0	0
0.400-0.499	1	88	2	211
0.300-0.399	1	236	2	112
0.200-0.299	0	0	5	333
0.100-0.199	1	391	4	286
0.001-0.099	2	1238	7	660
0	1	6969	1	7312
Total	8	9811	36	9811
Duplication		**7.154**		**7.131**

Table 9: Frequency groups and CHAID segmentation – Berlingske Guide.

5.2 Application for collecting new information

In the years to come, CHAID may well become one of the methods, which will be applied more and more frequently for combining information from multiple databases. As shown above, the system already works for measuring readership of sections, and the first test attempts of measuring the Internet have been conducted successfully.

Furthermore, Out-door is an obvious way of combining information from an external survey with the Index Danmark/Gallup. By registering a group of persons' (a panel's) exposure to selected out-door sites, this group may be subdivided into segments each having a different exposure probability to the site in question, provided the group is of an adequate size. Subsequently, these data may be regenerated in the Index Danmark/Gallup based on the common variables.

Similarly, radio is a media group which could benefit from transferring information from a diary survey to the NRS. One relevant piece of information for transferring could be the probability of listening to a particular radio on Mondays between 10 AM and 10.15 AM. In this case, the discriminative variables could be e.g. whether the respondent is employed, residence, age and whether the respondent is interested in various musical genres.

Also, TV data are relevant for CHAID surveys. Most people meter systems offer the possibility for analysing the individual panel member's use of various genres and channels. One possible piece of information for transferring is the probability of watching the news programme on TV2. Also, it would be relevant to include this piece of information into the NRS.

Generally, all media groups which offer the possibility for working with probabilities and recency measurement like e.g. household media, Direct Mail, and the yellow Pages are suitable for this type of data interconnecting.

References:

Arnaa,K;Randrup,R.(1997): Newspaper sections - a challenge to media planning. 8th readership research symposium, Vancouver.

Copeland,T;Mennenga,J.(1988): A reach and frequency experiments. Proceedings, Newspaper Research Council.

McGlathery,D.G;Eadie,W.(1995): The impact of measurement techniques on magazine audience level. 7th readership research symposium, Berlin.

Richard,A;Frankel,M.R.(1983): A comparison of reach and frequency estimates. 2nd readership research symposium, Montreal.

Weinblatt,L.(1994): The multimedia wearable passive meter – an update. Proceedings, ARF/ESOMAR-symposium, Paris.