CALIBRATED FUSION EVALUATION: THE MRI-COMSCORE CASE

David Napior and Jay Mattlin, Mediamark Research Inc.
Bob Ivins, comScore

Abstract

The paper provides a technical overview of an MRI-comScore data integration experiment. A discussion of pre-fusion cross-calibration of common Internet variables comprises the first part. Attention then turns to the fusion methodology. Following this is discussion of the fusion evaluation and the important contribution of pre-fusion calibration to the evaluation process.

1. Background

MRI and comScore have been exploring integration of MRI National Study data with cS/MediaMetrix Internet Study data. MRI offers a single source, multimedia audience and consumer targeting solution with a comprehensive, currency-level print component and limited data on Internet usage. It is based on an annual national area probability sample of about 26,000 individuals. ComScore /MediaMetrix provides a comprehensive Internet audience measurement service with limited consumer targeting or multimedia functionality. It is based on a large panel (about 60,000 strong) whose click stream activities are tracked passively either at home, at work, or both.

The goal of the R&D effort was, in essence, to develop a data integration mechanism linking the unique information in one dataset to the unique information in the other dataset. The bridge between the two data sets was, of necessity, a relatively small number of variables common to both datasets. Of particular interest was the linkage of cS Internet data to MRI print and consumer targeting data.

In this paper we report on preliminary findings from the MRI-cS data fusion experiment. We begin with a discussion of an important calibration issue that arose in the early stages of the project. We proceed to a discussion of fusion methodology. We conclude with a lengthier discussion of the evaluation findings and the important contribution of cross-calibration.

2. Measurement

The MRI experimental data were extracted from the MRI Fall 2004 National Study. The variables in the MRI dataset include:

- Last-month visitation to 28 web sites;
- Recent readership of 54 Magazines and 3 National Newspapers;
- Recent purchase of 64 consumer products;
- Age-sex demographics.

The MRI sample consists of approximately 16,000 Internet users. In the illustrative example presented in this paper, we use a subset of the total experimental dataset: 20 print variables, 21 consumer target group variables and the full set of 28 Internet variables.

The cS experimental data were compiled from one month of click stream data, transformed to binary indicators of Internet visitation in the last month. The cS variables were:

- Last-month visitation to 28 Web sites (same 28 as MRI);
- Age-sex breakouts of Internet users (balanced to the MRI Internet user enumeration model).

The cS sample consists of approximately 60,000 Internet-using panelists.

When common media variables are used as linking agents in data fusions, it is necessary and reasonable to assume that measures on both sides tap a common construct. But commonality of underlying construct does not imply equivalence of scale of measurement. In the interest of reducing the likelihood of incommensurability issues, cS transformed their detailed, click-by click accounting of Internet behavior to an MRI-like binary indicator of visitation to a given Internet site in the last 30 days. Figure 1 below presents a side-by-side comparison of one-month ratings for 28 web sites in common to MRI and cS.

Figure 1.

	Internet Ratings		
	cS	MRI	Gap
Yahoo	71	57	14
MSN	62	27	35
AOL	53	25	28
Google	43	51	-8
Ask(Jeeves)	27	6	21
Weather	23	26	-3
Lycos	19	3	16
MSNBC	18	8	10
CNN	17	11	6
CNET	16	2	14
iVillage	11	1	10
ESPN	11	12	-1
USAtoday	7	4	3
CBSsportsline	5	2	3
NYTimes	5	3	2
CBS	5	4	1
AltaVista	4	3	1
MTV	4	4	0
FOXnews	4	6	-2
ABC	3	5	-2
Excite	3	3	0
NBC	2	4	-2
PBS	2	4	-2
FOX	2	3	-1
Zdnet	2	2	0
Netscape	1	8	-7
WSJ	1	1	0
UPN	1	1	0

Correlation: 0.87 .

The correlation of .87 between paired-up user ratings over 28 data points is fairly good evidence that the variables tap a common construct. However, there are sizable inter-study gaps at the upper end of the ratings scale. Gaps of this magnitude between create significant problems for comparability of measurements of association in which the variables are involved. (Nunnally and Bernstein, 1994).

To address this issue, we developed a cross-calibration procedure grounded in psychometric theory and practice. The technique relies on the following assumptions:

- The MRI and cS Internet data tap a common latent trait: the personal probability of an individual respondent to visit a particular web site in any one month period;
- The latent personal probability of a visit to a web site can be inferred from a binary indicator of recent visit to the site in question as well as visits to other sites.
- Estimates of personal probabilities can produced by latent trait modeling techniques (Agresti, 2002,p278; Anderson & Vermunt, 2000).
- Latent personal probabilities are commensurate at an ordinal level of measurement, not higher.

Given the above assumptions, our crossover calibration proceeds, variable by variable, as follows:

- Estimate cS latent personal probabilities using latent trait modeling (Agresti, 2002, p278);
- Rank order cS respondents in ascending personal probability order;
- Record value of personal probability cumulative rank fractile score of last respondent to have a zero value on the original binary cS variable; call this the cS breakpoint;
- Estimate MRI latent personal probabilities using latent trait modeling;
- Rank order respondents in ascending personal probability order;
- Create a new MRI derived binary variable by dichotomizing the MRI ranked personal probabilities at the cS breakpoint.

The original interest in crossover calibration came out of our desire to optimize the placement of the MRI and cS respondents in a common Internet ratings space. More recently, the calibration methodology has been proposed as an essential component of the fusion evaluation methodology.

4. Fusion Methodology

The fusion mechanism uses a nearest-available-neighbor algorithm to produce solutions satisfying the statistical retention properties of constrained statistical matching (Agresti, J, 2004). Developed by MRI, the system is fast, scaleable and easily adaptable to diverse matching schemes.

5. Findings

Our entry level criteria for going forward with the fusion was a direct comparison of a fusion based on demographics-only *vs* a fusion based on the same demographics and common Internet variables. If we didn't observe a marked improvement as we moved from the baseline fusion to a fusion built around one of our core competitive assets, we knew we were in trouble. The test data set for this entry level, go/no-go fusion consisted of two common demographic variables (age and sex), 28 common Internet variables, 20 MRI readership variables and 21 MRI target group indicators (e.g. alcoholic beverages, pet food, lottery, etc).

Figure 2 presents a side-by-side comparison of the accuracy of the Demographics Only model to the accuracy of the Demographics/Internet model. Four summary statistics are used to describe the difference between the fused cS and actual MRI target group Internet ratings: mean difference, mean absolute difference, mean % relative difference and mean % absolute relative difference. In all cases, differences are calculated as MRI less cS. One of the values of the mean absolute difference and mean % absolute difference is they provide an upper bound on absolute and relative regression of the fused target group rating to the mean. In each case, the assessment of accuracy is a function of the difference between the actual MRI target group Internet ratings and the fused cS target group Internet ratings.

Figure 2. Summary statistics for target group Internet ratings

Target Group Ratings: (MRI vs fused-cS)	Demographic Fusion	Internet Fusion
Mean Difference	-2.88	-3.84
Mean absolute difference	5.17	5.77
Mean % relative difference	-42.55	-55.28
Mean % absolute relative difference	67.10	72.58
Mean MRI Internet Rating = 10.2		
Mean cS Internet Rating = 15.0		

We found the results in Figure 2 very puzzling. The fusion that incorporated Internet variables produced target group Internet ratings further from the MRI target Internet ratings than the fusion based on demographics alone. We attributed this to the distorting impact of the substantial differences between MRI and fused cS Internet ratings on the common variables. We researched the possibility of a summary level adjustment, but soon discovered that this was futile (Nunnally & Bernstein, 1994). It then occurred to us that it might be reasonable to substitute the crossover calibrated data for the observed data. The results are presented in Figure 3.

Figure 3. Summary statistics for target group Internet ratings (Calibrated)

Target Group Ratings: (MRI vs fused-cS)	Demographic Fusion	Internet Fusion
Mean Difference	1.23	0.27
Mean absolute difference	1.40	0.80
Mean % relative difference	11.41	2.78
Mean % absolute relative difference	15.23	9.39

Figure 3 fits our expectations well. Exploratory analysis prior to the fusion indicated the MRI target group Internet indices were predominantly greater than 100. As a result, we expected the mean absolute differences between MRI and cS target group ratings to be closer to zero for the demographics fusion than for the Demographics/Internet fusion. The results are totally consistent with this *a priori* reasoning.

Please note that the calibrated Internet variables were not used as link variables in the cS-MRI fusion. The calibrated data were used only in the evaluation step. This observation may alleviate concerns about circularity.

Figure 4 presents the calibrated findings for target group print/Internet duplications.

Figure 4. Summary statistics for target group print/Internet duplications (Calibrated)

MRI vs fused-cS	Demographic Fusion	Internet Fusion
Mean Difference	0.239	0.061
Mean absolute difference	0.380	0.142

Once again the results are consistent with our expectations. The Demographics/Internet fusion does much better than the Demographics-only fusion.

Next we assess the accuracy of the fusion for Internet sites not measured by MRI. This evaluation is arguably the most important of all. The vast majority of Internet sites measured by cS are not measured by MRI. There is no target group information for these Internet sites other than what the fusion can provide.

To estimate the accuracy of the fused cS target group Internet ratings for sites unmeasured by MRI, we dropped 14 of the 28 variables from the link set and reran the fusion, using only the remaining 14 variables as link variables. The results of this fusion are presented in Figures 5 and 6.

Figure 5. Summary statistics for target group Internet ratings (Calibrated)

Target Group Ratings: (MRI vs fused-cS)	Demographic Fusion	Internet (Unique/ Unlinked)	Internet (Common/ Linked)
Mean Difference	1.23	0.80	0.50
Mean absolute difference	1.40	1.10	0.80
Mean % relative difference	11.41	4.60	2.40
Mean % absolute relative difference	15.23	11.20	9.80

Figure 6. Summary statistics for target group print/Internet duplications (Calibrated)

MRI vs fused-cS	Demographic Fusion	Internet (Unique/Unlinked)	Internet (Common/Linked)Fusion
Mean Difference	0.239	.107	0.078
Mean absolute difference	0.380	.222	0.181

Once again, the figures are directionally as expected. The fusion based on 14 pseudo-unique (unlinked) Internet variables is less accurate than the fusion based 14 common/linked Internet variables. Both are more accurate than the fusion based on demographic variables only. The split-half test we have performed underestimates the power of the 28 common variables to predict target Internet ratings and duplications for sites unmeasured by MRI. Jackknife procedures will be used to develop much better estimates of the impact of the full set of available common variables.

Summary & Recommendations

In summary, we have presented an illustrative example of the application of crossover calibration to Data Fusion Evaluation. The uncalibrated evaluation was an impenetrable mess of triply confounded distortions of the latent evaluative information. The cross-calibrated evaluation looked sensible and conformed to our extremely liberal expectations. Nonetheless, the fact remains that the logic behind the fusion is based on analogies drawn to psychometric theory and practice. The appropriateness of these analogies is subject to dispute.

Recommendations for further research include:

- Expand findings on accuracy of fusion for Internet sites not measured by MRI;
- Present findings on Internet/print interactions for brother/sister Print/Internet offerings;
- Apply predictive modeling methodology to introduce unmeasured unique variables into the link set.
- Present findings of analyses of variance of fusion summary statistics by design factors; we have observed a strong interaction effect of media-rating-levels by target-group-incidence levels for all summary statistics; proper treatment of this interaction in the evaluation analyses might influence the estimated accuracy of the fusion.
- Explore use of continuous cross-calibrated personal probability fusion metrics to improve stability of models and ameliorate problems of restricted range of correlations associated with binary audience measures;
- Develop fusion-based Internet prototyping mechanism.

BIBLIOGRAPHY

Agresti, A. (2002) *Categorical data Analysis*. Wiley Inter-Science: Princeton University Press: Hoboken, NJ.

Agresti, J. (2004) D-fuse Technical documentation. Unpublished document. Mediamark Research, Inc. New York, NY.

Anderson, C.J. and Vermunt, J.K. (2000) Log-multiplicative association models as latent variables for nominal and/or ordinal data. *Sociological Methodology*. Vol.30 .

Bazaraa, M.S., Jarvis, J.J., and Sherali, H.D., 1997, *Linear Programming and Network Flows*. John Wiley & Sons. New York.

Brown, M.(1999) Effective print measurement: audiences… and more. Ipsos-RSL. Harrow.

Frankel, M., Julian Baim, Joseph Agresti and Michal Galin (2001). Toward a mathematical theory of cross-survey inerence. 10th Worldwide Readership Research Symposium. Venice.

Ho, D.E., Kosuke Imai, Gary King, and Elizabeth Stuart (2004). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. Unpublished manuscript. Harvard University (GKing.Harvard.Edu/matchit), Cambridge, MA 02138.

Kadane, J.B. (1975) Statistical problems of merged data files. OTA Paper No. 6, Office of Tax Analysis, U.S. Treasury Department, Washington, D.C. 20220 .

Lazarsfeld, P.F.(1968) Latent structure analysis. Houghton Mifflin Company. New York.

Lord, F. and Novick, M. (1968) *Statistical Theories of Mental Test Scores.* Addison-Wesley.

Moriarity , C. and Scheuren, F. (2001) Statistical matching: a paradigm for assessing the uncertainty in the procedure. *Journal of Official Statistics.* Vol. 17, No.3.

Nunnally, J.C. and Bernstein, I.H. (1994), *Psychometric Theory*, 3rd Edition. Mc-Graw-Hill , Inc., New York.

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores. Psychometrika Monograph No. 17.*

Samejima, F. (1974). Normal ogive model on the continuous response level in the multidimensional latent space. Psychometrika 39.

Soong, R. (2003) Fusion on the fly for multimedia applications. *11th Worldwide Readership Research Symposium*. Cambridge, MA USA.

Steiger, J.H. (1980) Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, Vol. 87, No. 2.

Rubin, D. (1979) Using multivariate matched sampling methods and regression adjustment to control bias in observational studies. *Journal of the American Statistical Association* 74 .

von Davier, M. and von Davier, A. (2004) A Unified Approach to Scale Linking. Research Report RR-04-09. Educational Testing Service. Princeton, NJ.