

Evaluating Merge/Purge Systems: Part Three

*By Jim Wheaton and Cynthia Baughan Wheaton
Principals, Wheaton Group*

Original version of an article appeared in the September 1987 issue of "Direct Magazine"

[Note: Despite dramatic increases in raw computing power and a proliferation of end-user software tools since the publication of this series of six articles, virtually all of the content remains highly relevant. The occasional obsolete point is highlighted.]

Statement of Purpose

In a series of six articles, we explain a number of the key concepts that mailers should understand about merge/purge, as well as reviewing (in the first article) a methodology that could be helpful in evaluating the effectiveness of either present or prospective merge/purge systems. While our comments are primarily addressed to mailers, merge/purge vendors can benefit by measuring themselves against the criteria that we have identified as important.

Our objective is to describe new and specific tools that can be used to evaluate and improve the performance of the merge/purge process. Through commentary and examples, we will attempt to translate into layman's terms the technical jargon that baffles many mailers. In the process, practical applications should become apparent.

This month's article, Part Three, focuses on the four types of unduplication errors.

Types of Unduplication Errors

Merge/purge systems must minimize four different kinds of unduplication errors, each of which incurs its own kind of cost and requires its own technique for detection. The four kinds are:

- Rented List Overkill
- Suppression List Overkill
- Rented List Underkill
- Suppression List Underkill

All systems commit these four types of errors. They are described below, along with hypothetical examples to illustrate the annual savings from small performance differences, assuming:

- A direct response marketer of moderate size mails about 10 million catalogs a year to rented names and space ad inquirers.

- 13.33 million gross names a year undergo merge/purge processing. This includes suppression list input and allows for a duplication factor of 26% to 29%, resulting in the 10 million names available to mail.
- Catalogs to rented names include a special wrap complete with “introductory offer” text describing a \$2 discount on the first order.
- Because the house list receives a catalog without the introductory discount, it must be input into the merge/purge as a suppression file.
- The “in the mail” cost per catalog is 50 cents, including 7 cents for list rental and 1 cent for the wrap.
- The overall catalog response rate for rented and space ad inquiry names is 2.0%.

Rented List Overkill

Rented list overkill occurs when separate individuals (or households) are incorrectly identified as duplicates.

- The failure to mail these legitimate prospects shrinks the pool of potential respondents from promising rental lists. Response rates are likely to decline as the direct response marketer substitutes marginal lists in an attempt to maintain targeted mail quantities.
- The rental cost per name mailed will increase for all lists requiring payment on a gross name basis. This is because the total cost is fixed, and the cost per name increases whenever a mailable name is overkilled.

In cases where a net name arrangement is in effect, but where the contract stipulates a maximum allowable duplication percentage for which a refund will be given, the cost-per-name mailed will increase to the extent that this percentage is exceeded (e.g., 15% as in the typical 85% net name agreement). Likewise, this is because at that point the total cost becomes fixed.

- Rented list overkill can be identified by visually examining the merge/purge duplicates listing for incorrect matches. This is a relatively straightforward process.

Example:

Assume Service Bureau A has 0.56 of a percent less rented list overkill than Service Bureau B, on a gross input basis (see Exhibit 1, Line A):

- Gross names of $13,333,000 \times 0.56\% \times$ response rate of $2.0\% = 1,493$ additional new customers per year generated using Service Bureau A.

Suppression List Overkill

In this case, legitimate prospect names are incorrectly identified as being current house list customers or undesirable purchase candidates, such as those with a history of bad debt, and are suppressed.

- As with rented list overkill, the failure to mail these legitimate prospects shrinks the pool of potential respondents, and is likely to increase the rental cost per name mailed.
- Suppression list overkill can usually be identified by examining the duplicates listing for incorrect matches. This process can be complicated, however, by the report formats of some vendors, where only the outside list name and not the corresponding suppression name is printed out in the listing. In such cases, the duplicates listing must be compared directly with each suppression list.

Example:

Assume Service Bureau A has 0.14 of a percentage point less suppression list overkill, on a gross input basis, than Service Bureau B (see Exhibit 1, Line B):

- Gross names of $13,333,000 \times 0.14\% \times$ response rate of $2.0\% = 373$ additional new customers per year generated using Service Bureau A.
- Total benefit due to Service Bureau A's superior performance in both categories over overkill is: $1,493 + 373 = 1,866$ additional new customers per year.

To generate these additional 1,866 new customers under Service Bureau B, marginal lists would have to be included in the mailing. The response rates from these lists would probably be lower, resulting in a higher cost per customer.

Exhibit 1						
One Year Merge/Purge Overkill Performance Comparison: Service Bureau A vs. Service Bureau B						
	Bureau A		Bureau B		Overkill Diff. A vs. B	
	Qty. (000s)	%	Qty. (000s)	%	Qty. (000s)	%
Gross names to M/P	13,333	—	13,333	—	—	—
A. Rented lists	21	0.16	96	0.72	75	0.56
B. Suppression lists	6	0.04	24	0.18	18	0.14
Total	27	0.20	120	0.90	93	0.70

Rented List Underkill

Rented list underkill occurs when legitimate duplicates from rented lists are not eliminated.

- Mailing costs are increased by sending multiple catalogs or packages to the same individual. Unfortunately, it is not unusual for one person to receive as many as four catalogs from a single company.
- There is evidence that response can increase somewhat with multiple mailings to a single household, especially households with multiple members. The increase, however, obviously pales in comparison with the higher mail costs. It is difficult to make money when mailing costs for a household are doubled, tripled, and even quadrupled!
- Rented list underkill can be identified by visually examining the final output listing for name and address records that logically appear to be duplicates. This is a relatively straightforward process. It is also critical, because underkill appears to be more of a problem than overkill, at least among the more sophisticated systems analyzed in our study.

Example:

Assume Service Bureau C has a rental list unduplication rate (exclusive of overkill) 2.33% higher than Service Bureau D, on a gross input basis (see Exhibit 2, Line A):

- Gross names of 13,333,000 x 2.33% x cost per catalog of \$0.50 = \$155,329 saved per year by eliminating multiple mailings to the same individual.

Suppression List Underkill

In this case, current house customers and/or undesirable purchase candidates are not identified on rented lists. This can have several effects:

- Customers will get multiple copies of the same promotion, whenever the house file and outside lists are scheduled to receive the same package.
- Customers may become annoyed or confused whenever outside lists are scheduled to receive a special package. This can be very damaging to companies that attempt to confer prestige on their products by offering membership to an exclusive club, society or program.
- Undesirable prospects will be mailed, such as bad credit risks and individuals who do not wish to receive third-class mail.

The costs of suppression list underkill are both direct and indirect:

- In-the-mail costs increase when current customers and undesirable prospects receive unscheduled contacts. Costs will also rise to the extent that current customers take advantage of special discounts targeted to first-time buyers.
- Customer service costs increase when current customers register their annoyance and/or confusion at membership invitations.
- Bad debt increases to the extent that poor credit risks respond to the offer.
- The direct marketing industry is hurt by promoting those who have made an effort to notify the mailer that they do not want to be contacted.

Suppression list underkill is the most difficult of the four errors to identify. Unlike rented list underkill, examination of cleaned output will not uncover the existence of a single outside name that should have “hit” against a suppression file. Only by painstakingly searching cleaned output listings for suppression names can the occurrence of this problem be identified and its magnitude estimated.

Example:

Assume Service Bureau C has a rental list unduplication rate (exclusive of overkill) against the customer file 0.58 of a percentage point higher than Service Bureau D, on a gross input basis (see Exhibit 2, Line B):

- Gross names of 13,333,000 x 0.58% x cost per catalog of \$0.50 = \$38,666 per year less in duplicate mail.
- Gross names of 13,333,000 x 0.58% x response rate of 2.0% x discount of \$2.00 = \$3,093 per year less in current customers taking the \$2.00 discount targeted for first time buyers, assuming none of these respondents pass up the discount.
- Total annual savings due to Service Bureau C's superior performance in both categories of underkill is: $\$155,329 + \$38,666 + \$3,093 = \$197,088$.

Exhibit 2						
One Year Merge/Purge Underkill Performance Comparison: Service Bureau C vs. Service Bureau D						
	Bureau C		Bureau D		Underkill Diff. C vs. D	
	Qty. (000s)	%	Qty. (000s)	%	Qty. (000s)	%
Gross names to M/P	13,333	—	13,333	—	—	—
A. Rented lists ¹	2,550	19.1	2,240	16.8	310	2.33
B. Suppression lists ¹	637	4.8	560	4.2	77	0.58
Total eliminated	3,187	23.9	2,800	21.0	387	2.90
Net names available	10,146	76.1	10,533	79.0		
¹ Correct matches. Excludes Overkill.						

The advantage from using Service Bureau C, then, is \$14.78 per thousand gross names input into merge/purge (\$197,088/13,333). But what if Service Bureau C's price were a dollar or two higher? It is likely that most direct response marketers would look no further, and choose Service Bureau D instead!

- The degree of difference in both underkill and overkill performance used in these examples is not overstated. In fact, this level was seen within the four top finalists in our study.

- The differences, and potential benefits, are much more extreme within all merge/purge systems currently on the market.
- The potential benefit will also be dramatically higher for large mailers, who drop much higher quantities of catalogs than 10 million a year.

Careful testing is required to determine system performance in light of these four types of unduplication errors. Small advantages of one system over another directly impact response rates and sales per thousand mailed by improving the quality of net names available for mailing. As shown, the improvement frequently is sufficient to overcome merge/purge cost differentials.

Because merge/purge is both art and science, even the best system will commit the four types of errors to some degree. The direct response marketer’s goal should be to find the best balance between the four, based upon his or her business priorities.

- Among the four top finalists in our study, it appears that state-of-the-art underkill performance cannot be achieved without a small increase in overkill.
- This overkill increase, however, is very minor. In fact, it would likely be much less than experienced with mediocre software that does not offer any underkill advantage.

Here is an illustration, based upon our study, which quantifies the difference in net name quality between a very good and a very bad system. Please note that the “Net Names” counts came from vendor output.

The overkill and underkill counts, on the other hand, were derived from a visual review of the output. These two numbers obviously could not be generated by the software, because software that could recognize the existence of overkill and underkill would not have made them in the first place:

	Vendor E	Vendor F
Net names	39,900	41,200
Plus Overkill	100	4,900
Less Underkill	(1,400)	(7,500)
Adjusted net output	38,600	38,600

- Vendor E identified 1,300 fewer names for mailing than Vendor F (39,000 vs. 41,200).
- Vendor E incorrectly handled only 1,500 names; that is, it had overkill of 100 names and underkill of 1,400 names.

- Vendor F incorrectly handled 12,400 names, with overkill of 4,900 and underkill of 7,500 names.
- In total, Vendor E incorrectly handled 10,900 fewer names than Vendor F. Therefore, Vendor E's output is of substantially higher quality.

Short of examining every record that enters a merge/purge, overkill and underkill must be estimated based upon samplings of ZIPs or SCFs. Therefore, the adjusted net output of two systems will be identical only to the extent that the sampling procedure has accurately reflected the entire universe of records; and can be employed as a test for such.

Tradeoff Between Underkill and Overkill

It is impossible for any vendor to perform a merge/purge perfectly. Because one can never be entirely sure what is a duplicate and what is not, the direct response marketer is faced with critical decisions when working with the vendor to set up the matching logic. There are two basic directions in which to set unduplication parameters:

First, you can "tighten" the parameters; that is, incorporate stricter duplicate requirements, which would "tilt" toward underkill.

- A simple way of visualizing this is by thinking in terms of a match code. If the match code is 16 characters long, a "tight" requirement would be that all 16 must match for two records to be considered duplicates. Thus, fewer duplicates would be identified and more names mailed.
- This is appropriate if you want to maximize the number of names available to mail, even though you will send some individuals more than one mailing piece.

Or, you could "loosen" the parameters; that is, incorporate less strict duplicate requirements. In this case, you would "tilt" toward overkill.

- Again, by thinking in terms of a 16-character match code, a "loose" requirement would be that 12 of the 16 characters must match for two records to be considered duplicates.
- Obviously, you would identify many more duplicates with a loose parameter than with a tight one. This is appropriate if you want to minimize the number of names available to mail, even if you eliminate some unique names in the process.

If you give your vendor guidance as to which of these two directions is appropriate for your business or a particular mailing, you will set the stage for the definition of the specific parameters that are appropriate for your merge/purge.

These parameter decisions may sound simple, but here are two examples that demonstrate just how nebulous things can get. Ask yourself which names should be considered duplicates:

- This first pair contains small but multiple differences in the last name, street numeric, and street name:

Record #1	Record #2
Sol Kros <u>i</u> ns <u>ch</u> wige	Sol Kros <u>e</u> nz <u>w</u> ieg
50 <u>0</u> R <u>u</u> dol <u>p</u> h <u>e</u> Road	50 R <u>a</u> ndol <u>p</u> h Road
Boston, MA 02180	Boston, MA 02180

The last names are different by four characters. The street numerics are off by one number, and the street name by two letters. Could there be this many input errors, or are these two different individuals?

- The second pair contains multiple phonetic spellings in the last name and street:

Record #1	Record #2
Richard Baughan	Richard Bawn
5 Boutelle Street	5 Bootel Street
Leominster, MA 01453	Leominster, MA 01453

One of the authors can attest to the number of times in which such a name is written phonetically (Bawn), rather than according to the correct spelling (Baughan).

But, are these records different because they represent different individuals, or because a telemarketing representative was a little lazy and did not ask for the actual spelling of the names? Although our personal experience with telemarketing representatives has been quite good, that doesn't mean that such problems cannot or do not occur.

As discussed in an earlier article in the series, you should know as much as possible about the source of the names on the lists you rent. That can be helpful in deciding on the appropriateness of certain parameters.

There really is no right or wrong answer short of tracking down every individual and verifying the data on each record.

Jim Wheaton and Cynthia Baughan Wheaton are Principals at Wheaton Group, and can be reached at 919-969-8859 or jim.wheaton@wheatongroup.com. The firm specializes in direct marketing consulting and data mining, data quality assessment and assurance, and the delivery of cost-effective data warehouses and marts. Jim is also a Co-Founder of Data University www.datauniversity.org.