Secrets of Test Design & Analysis

© 2006

A "how-to" for direct & interactive database marketers

By

Perry D. Drake
Vice President, Drake Direct
Associate Professor, New York University

Rhonda Knehans-Drake
President, Drake Direct
Associate Professor, New York University

Presenter Background and Experience

Unfortunately, Perry injured himself training for a marathon several weeks ago and cannot be with us today. He is on crutches and resting per his doctors request but wishes he could be with us today.

Presenter Background and Experience
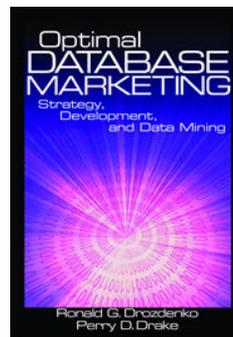
Perry D. Drake

- Current Responsibilities
    - Vice president & General Manager, Drake Direct, New York, NY
    - Associate Professor, New York University, Center for Direct and Interactive Marketing, New York, NY. Currently teaching: Database Mining, Advanced Database Mining, Statistics for Direct Marketers, Data Analytics & Reporting, Testing Best Practices and Web Mining. Recipient of the *"Outstanding Master's Faculty Award."*
- Prior Responsibilities
    - Director, Marketing Services, The Reader's Digest Association, Pleasantville, NY
    - Associate Director, Magazine Marketing, The Reader's Digest Association, Pleasantville, NY
    - Statistician, Quantitative Analysis, The Reader's Digest Association, Pleasantville, NY
    - Biostatistician, CIBA-Geigy Pharmaceuticals, Summit, NJ
- Education
    - Master of Science, Applied Statistics, University of Iowa
    - Bachelor of Science, Business Economics, University of Missouri

Presenter Background and Experience

<u>Perry D. Drake (Cont.)</u>

- Publications and Presentations
    - Over the years, Perry has presented at many industry conferences including the DMA, Circulation Days, and NCDM. He has also written numerous articles for the various trade journals including DM News, Direct, Target and Inside Direct Mail. To view his presentations and articles written visit his firm's web site at <u>www.DrakeDirect.com</u>.

    - In addition, Perry is the author of the book "Optimal Database Marketing" by Sage Publications. This book delves into database marketing concepts and practices such as the process to evaluate database needs and then select a database vendor, segmenting the customer file, response modeling, lifetime value analysis, and test design and analysis. It is a "how-to" book geared solely for the marketer wanting to better understand the practices and principals of database marketing.

Presenter Background and Experience

Rhonda Knehans-Drake

- Current Responsibilities
    - President & Founder, Drake Direct, New York, NY
    - Associate Professor, New York University, Center for Direct and Interactive Marketing, New York, NY.
- Prior Responsibilities
    - Director, Young Families Division, The Reader's Digest Association, Pleasantville, NY
    - Account Director, Information Resources Incorporated, Darien, CT
    - Senior Analyst, Columbia House Music and Video Club, New York, NY
- Education
    - Master of Science, Applied Statistics, University of Iowa
    - Bachelor of Science, Business Economics, University of Missouri
- Publications and Presentations
    - Over the years, Rhonda has presented at many industry conferences including the DMA, Circulation Days, and NCDM. She has also written numerous articles for the various trade journals including DM News, Direct, Target and Inside Direct Mail. To view her presentations and articles written visit her firm's web site at www.DrakeDirect.com.

Don't Leave Early

- We will raffle off a copy of "Optimal Database Marketing" at the end of the session.

- Free **take-away bags** to be handed out to all including:
    - Two additional articles on test design and analysis
    - Free point-and-click test planning software "The Plan-alyzer 4.0"
    - Pocket Calculator
    - Laminated "wallet sized" sample size estimator card
    - And much, much more

- A chance to enter a drawing for a free i-Pod nano by filling out a quick industry survey on database marketing practices.
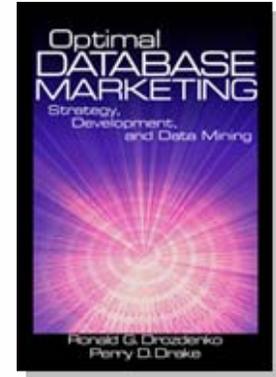
Prior survey results were published in the DMA Statistical Fact Book.
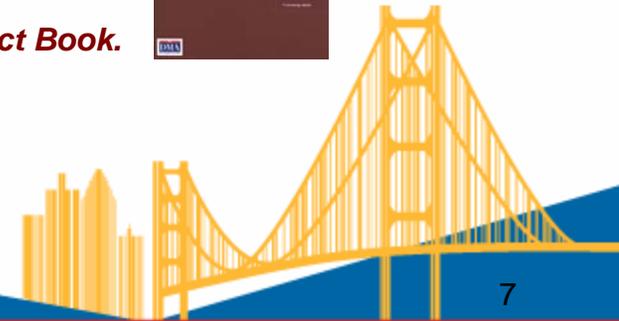
Introduction – Why We Test

- Testing is the foundation upon which one builds and grows a direct marketing firm.
- With a database, names can be selected for certain treatments and comparisons on the customer's reaction to these treatments made.
- Based on these results, in conjunction with marketing cost and revenue figures, the most profitable decision can be made.
- Without knowledge of proper test planning and analysis, one therefore is not in the strongest position to help their company grow.

Introduction – Why We Test

- The ability to easily test new marketing concepts, products or lists and read results is what sets direct marketers apart from other marketers.
- By testing:
    - we produce better front end response and better back end performance
    - we can determine the best creative approach
    - we can find the best offer
    - we find profitable mailing lists and other media

Session Objectives

During this session we will discuss six very important rules of test design and analysis to help ensure you get the most out of your testing program and make the best decisions possible.

In particular, we will discuss:

1. Why you must always back test package changes at least once
2. Why you should never fish for your confidence level
3. How to properly evaluate your test response rates
4. Why a full factorial test design is never warranted and the problems it can cause
5. The most appropriate way to assess such metrics as dollars spent
6. How to determine the amount of error you can tolerate in your test results

Six Rules of Test Design & Analysis

<u>Rule 1: Back Test Package Changes to the Control</u>

When changing to a new promotional format for the first time one should always back test (re-test) the old promotional format to validate the lift in response you had forecasted.

Without a back test you will not be able to determine if the cause for a campaign that is under forecast is due to the promotion not holding up, a list selection error or a general downturn in the business or any combination of these three.

To help illustrate this point, consider the following example….
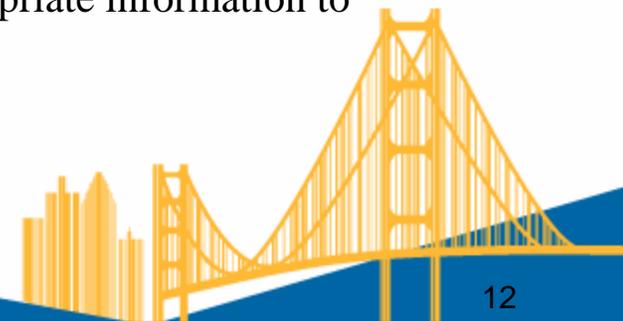
Six Rules of Test Design & Analysis

Last year you conducted a new format test and received a 1.25% response rate versus a 1.00% for the control format (a +25% lift). For your next big mailing of 1,500,000 names you decide to roll-out with the new format since it was a winner. You forecast a 1.25% for this mailing. The results of the roll-out are as follows:

	Number of Customers Mailed	Number of Customers who Responded	Response Rate
New Format	1,500,000	14,250	0.95%

The resulting response rate actually came in lower than your old format did last time (0.95% vs. 1.00%).

Your boss is not happy with the results of this mailing and has questioned your decision to change to the new format. He wants to you switch back to the old format.

Question: Is this the correct decision to make? Do you have appropriate information to answer this question?

Six Rules of Test Design & Analysis

Now, suppose I told you we had conducted a reverse test of the old format within your big mailing with results as shown below.

	Number of Customers Mailed	Number of Customers who Responded	Response Rate
New Format	1,400,000	13,300	0.95%
Reverse Test of Old Format	100,000	760	0.76%

Question: Can you now defend your decision to your boss? Did the format hold up?

Question: What are our next steps to determine what the problem is?

Six Rules of Test Design & Analysis

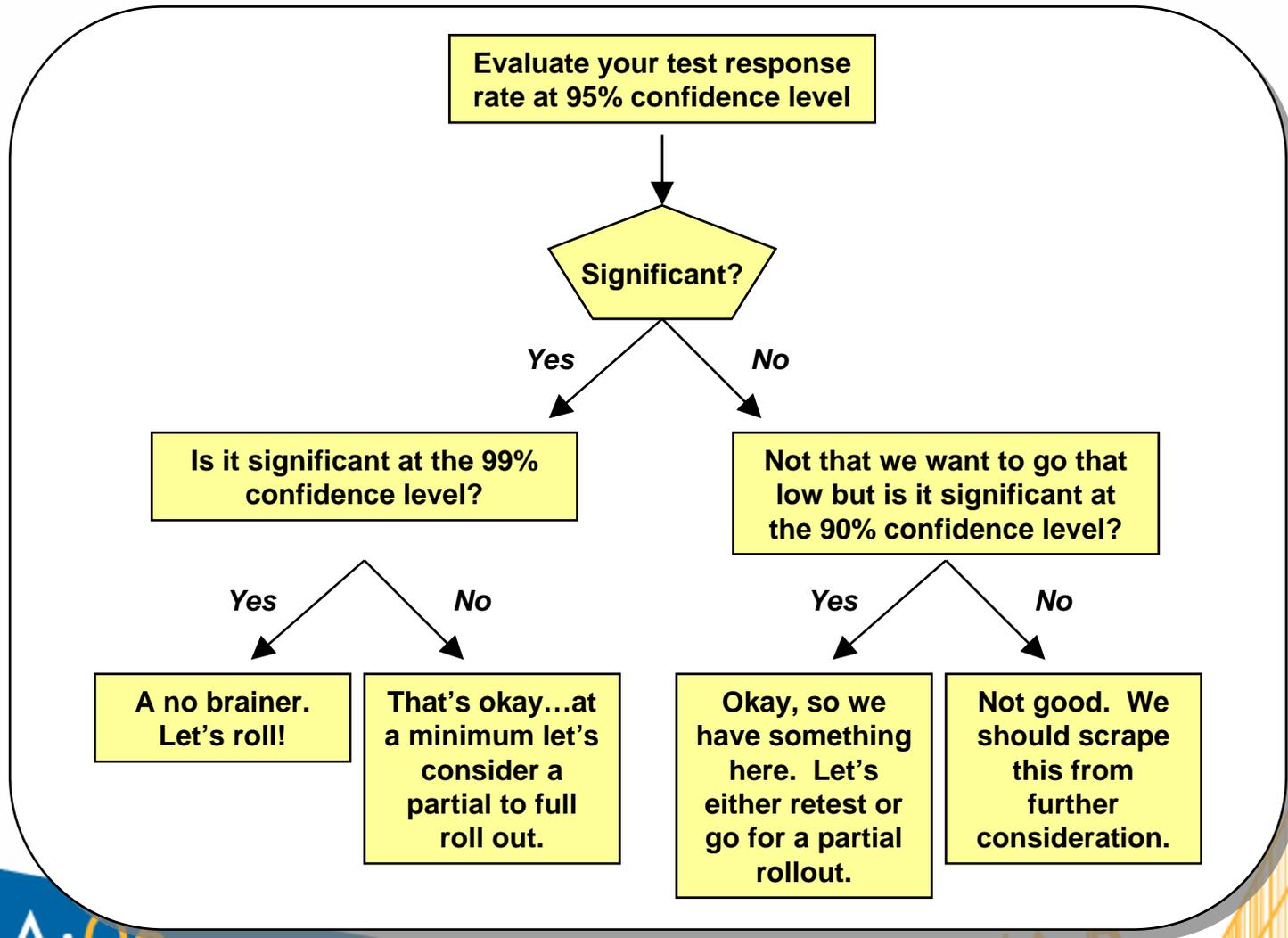**Rule 2: Put a Stake in the Ground Regarding Your Level of Confidence**

No direct marketer should ever consider evaluating their test results with a confidence level lower than 90%. To do so assumes way to much risk.

And, fishing for a confidence level that yields significance should never be practiced.

The rules that any good direct marketer should follow regarding significance are shown on the next slide.

Six Rules of Test Design & Analysis

```
┌─────────────────────────────┐
│ Evaluate your test response │
│ rate at 95% confidence level│
└─────────────────────────────┘
              │
              ▼
         ╱ Significant? ╲
        Yes          No
        ╱              ╲
```

Evaluate your test response rate at 95% confidence level

Significant?

Yes — *No*

Is it significant at the 99% confidence level?

Not that we want to go that low but is it significant at the 90% confidence level?

Yes — *No*

Yes — *No*

A no brainer. Let's roll!

That's okay…at a minimum let's consider a partial to full roll out.

Okay, so we have something here. Let's either retest or go for a partial rollout.

Not good. We should scrape this from further consideration.

Six Rules of Test Design & Analysis

As previously stated, you should never consider a confidence level less than 90%. To read a test with an 80% confidence level, for example is just too risky.

Although 80% confidence sounds quite high, on the flip side this implies there is a 20% chance or a **1 in 5 chance** you are misreading a test.

Wow! Is that really business smart?

Six Rules of Test Design & Analysis

Rule 3: Evaluate Your Tests Considering the Best and Worst Case Scenarios

One should never evaluate their test results based on a hypothesis test alone:

- *A hypothesis test will only tell you an answer to the question "Did my test beat the control, yes or no?"*

- *Just knowing the answer to this question is never enough. What you really need to know is did the test beat the control and at a minimum by how much.*

- *A confidence interval on the other hand will give you likely best and worst case scenarios for rollout regarding your test response rate.*

- *A confidence interval will tell you everything you need to know to properly assess your test panels.*

Let's consider an example….

Six Rules of Test Design & Analysis

Below are results of a new test.

Test Panel	Names	RR	Sign @ 95%
Control	75,000	1.05%	--
Test	75,000	1.15%	No

By considering only significance based on a hypothesis test alone, one would determine the test response rate is no different than the control package. And, as a result possibly pass on the test package.

Six Rules of Test Design & Analysis

When we consider a confidence interval wrapped around our test response rate, we notice the following.

Test Panel	Names	RR	Sign @ 95%	Confidence Interval	
				LB	UB
Control	75,000	1.05%	--	--	--
Test	75,000	1.15%	No	1.04%	1.36%

Question: Would you really consider this test a loser and why or why not?

Six Rules of Test Design & Analysis

You must consider the best and worst case scenarios when evaluating any test:

- *To say a test won is not enough. The question is, did it win by enough?*
- *To say a test lost is not enough. What you must ask is how close was it and how does that compare to the upside potential?*

Six Rules of Test Design & Analysis

Rule 4: Only Test for Meaningful Package Element Interactions

Generally it is unnecessary to test every possible package element combination in your test plan. For example, the Marketing Manager may be interested in testing the following changes to the control package:

- Price/Rate increase
- Addition of a premium
- Color change to outer envelope
- New format

Testing every possible combination of price, premium, outer envelope color, and format yields a total of 16 test panels. Testing all 16 is called a "full factorial test design."

When should a direct marketer consider a full factorial test design?

Six Rules of Test Design & Analysis

The only reason a marketer would test a full factorial test design is if it was truly believed interactions will occur between all four elements with respect to response.

In this example, the only possible interaction to be concerned with would be one between price and premium.

In other words, if you believe that a price increase may yield a less negative impact in a package that offers a premium and other bells and whistles than a bare bones package with no premium offer, then test that interaction panel.

Six Rules of Test Design & Analysis

Assuming you are only interested in assessing a possible interaction between price and premium, the test series would appear as shown below:

Test Panel	Description
1	Control test package
2	Price test package - as control package with $2 price increase
3	Premium test package - as control but with premium for order added
4	Format test package - as control but with a new format
5	New OE - as control but with a new OE
6	Price and premium test package - as test panel #3 with $2 price increase

Question: Based on this test series, how would you determine if the addition of a premium to the control package offset any or all of the negative effect that a price increase might have on response?

Six Rules of Test Design & Analysis

Keep in mind the main drawback of a full factorial test design is that it causes you to place mail quantity in panels that are not needed and takes away mail quantity from panels where more mail quantity is needed.

Consider the following example:

Suppose you are planning to test 5 segments, 2 formats and 2 offers all within a budget of 25,000 names total. A full factorial design will yield 20 test cells with 1,250 names per cell as shown below:

List	Format 1		Format 2	
	Offer 1	Offer2	Offer 1	Offer 2
List A	1,250	1,250	1,250	1,250
List B	1,250	1,250	1,250	1,250
List C	1,250	1,250	1,250	1,250
List D	1,250	1,250	1,250	1,250
List E	1,250	1,250	1,250	1,250

Six Rules of Test Design & Analysis

With 1,250 names per cell, the test results will have so much error variance associated with them that they will be difficult to interpret. Your decisions of which packages are working and which are not will not be sound.

Confidence bounds around various test response rates based on sample sizes of 1,250 names will be quite wide as shown below. Based on ranges that are this wide it will be difficult for you, the marketer, to make a sound decision regarding roll-out.

Response Rates of Test on 1,250 Names	95% Confidence Bounds Around the Test Response Rate
0.5% Response Rate	.11% to .89%
1.0% Response Rate	.45% to 1.55%
1.5% Response Rate	0.83% to 2.17%
2.0% Response Rate	1.22% to 2.78%
3.0% Response Rate	2.05% to 3.95%
4.0% Response Rate	2.91% to 5.09%
5.0% Response Rate	3.79% to 6.21%
6.0% Response Rate	4.68% to 7.32%

Six Rules of Test Design & Analysis

If there is no reason to assume interactions between the lists, the offers and the formats then the test series can look as follows where each test has almost three times the mail quantity. With this test plan you will be able to read your lists, offers and formats with more reliability.

List	Format 1		Format 2	
	Offer 1	Offer2	Offer 1	Offer 2
List A	3,571	3,571	3,571	
List B	3,571			
List C	3,571			
List D	3,571			
List E	3,571			

- You will read the strength of the lists using format 1 and offer 1.
- You will read the strength of the offers using only list A and format 1.
- You will read the strength of the formats using only list A and offer 1.

Six Rules of Test Design & Analysis

Rule 5: Never Assess Dollars Spent based on Mail Quantity

Many times, you will also want to assess average dollars spent for the control versus your test as opposed to a response rate. You will be interested in doing so for catalog tests, credit card spend activation tests, etc.

Assessing such metrics can be tricky. To do an accurate comparison of average spend for your test and control panels you must calculate the average based on only those that responded and not based on all names promoted.

Consider the following example.

Six Rules of Test Design & Analysis

MasterCard tested a "Spend 2 times and get a $25 statement credit" in order to stimulate spend for new card members that have not used their new card since activation. The results of this test are below:

	Control (no incentive)	Test (spend incentive)
Accounts	10,000	10,000
% Spend	20.0%	22.0%
# Spend	2,000	2,200
Total $ Spend	$800,000	$924,000
Average Spend per Account	$80.00	$92.40
Average Spend per Active Account	$400.00 ($800,000/2,000)	$420.00 ($924,000/2,200)

Question: When determining if average spend was significantly higher for the test versus the control which metric do you use?

Question: How do you calculate a worst case and best case P&L to determine if you should switch to the new spend program?
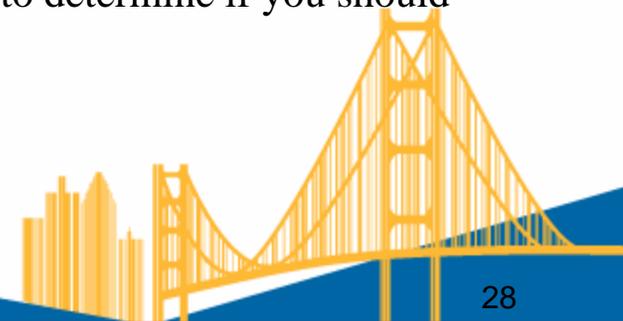
Six Rules of Test Design & Analysis

Keep in mind that to assess average dollars spent based on all names promoted is risky. This dataset may be highly skewed given there will be a lot of zeros. As such it may not assume the necessary statistical properties required for proper evaluation.

Six Rules of Test Design & Analysis

Rule 6: Determine the Amount of Error you Can Tolerate

The biggest mistake made when planning marketing tests is not properly determining the amount of error you can tolerate. This is not difficult but is a must.

There are two mistakes that can be made when you do not properly determine how much error you can tolerate in your test response rates:

1. *Test more names than are necessary given the level of precision you need in order to make a decision -- resulting in wasted testing dollars.*

2. *Test too few names and therefore have more error associated with your test than you can tolerate – placing yourself in a position where you cannot properly read your test and make a rollout decision.*

Many blindly assume you are okay as long as you can read a 10% difference as significant. After all, this is what we were taught in college.

Unfortunately, in the real world this is usually not a tight enough read. Be careful!

So how do you determine the amount of error you can tolerate?

Six Rules of Test Design & Analysis

Consider the following example:

The Marketing Director is testing the addition of a "Buck Slip" to the current control format which will highlight special product features.

- Assume the control format is known to yield a response rate of 1.80%
- Also assume the Marketing Director determined that three-quarters of an additional responses per 1,000 names mailed are required to cover the additional costs of the "Buck Slip".
- This implies an increase in response for the test of 0.075% to break-even with the control package.
- Therefore, the Marketing Director will require the test to yield a response rate of 1.875% or the test will be less profitable.

Question: So, how much error can the marketing director tolerate for this test?

Six Rules of Test Design & Analysis

In this case the marketing director can tolerate an error equal to three-quarters of a percentage point or 0.075% (1.875% minus 1.80%).

The marketing director wants to ensure he tests enough names so that he will be able to detect a 1.80% and a 1.875% as statistically significantly different. In other words, he wants to test enough names so that the confidence intervals associated with each response rate do not overlap. *And this will be the case as long as the error does not exceed .075%.*

Based on these assumptions, *"The Plan-alyzer"* yields sample sizes of close to 250,000 names per panel. WOW!

Six Rules of Test Design & Analysis

Question: What if you just don't have that many names available to test.

You will *"What-if"* by using the sample size tool backwards and determine how much higher the test response rate must be in order for you to detect it as a statistically significant winner based on your sample size constraints. And, if this required test response rate is too high you will need to think twice about conducting the test.

For example, if you only have 120,000 names total for testing *"The Plan-alyzer"* tells us the test will need to yield at least a 1.95% response rate or higher before you will be able to detect it as a statistical winner.

Does that seem realistic to you?

If so then go for it. It not then….

Plan-alyzer V4.0

Courtesy of Drake Direct
Database Marketing Consulting Firm

**Sample Size Estimation for Percentages –
Difference Between Two Estimates**

Return to previous screen

Recalculate

Input the required information and press "SUBMIT"

.018 — Estimated response rate for the control panel. Input in decimal format (e.g., input 4.00% as .04).

.0195 — Estimated response rate for the test panel. (Note: The difference between the estimated response rates for control and test should equal the "minimum difference" of concern.)

.95 — Confidence level. Input in decimal format (e.g., input 95% as .95).

The control and test panel sample sizes required for the conditions listed above are: 62,824 Samples of these sizes will guarantee a -.15 percentage point difference in response rates will be considered significant (i.e. not zero), with 95% confidence assuming the actual test results are close to the estimated response rates above.

Six Rules of Test Design & Analysis

<u>Question</u>: What if you believe the test will do better than break-even and yield a response rate closer to 2.00% based on past experience.

Well, that is good news! In this case we can tolerate more error in our test results and therefore put fewer names in the mail.

The error we can now tolerate is 0.20% (2.00% minus 1.80%). Based on these assumptions, *"The Plan-alyzer"* yields more realistic sample sizes of 36,000 names per panel. GOOD NEWS!

Questions?

<u>Perry & Rhonda's contact information is:</u>

Drake Direct
225 East 46th Street, Penthouse D
New York, NY 10017
www.DrakeDirect.com

Phone: 212-759-1225
Fax: 212-759-9756
Perry's Email: Perry@DrakeDirect.com
Rhonda's Email: Rhonda@DrakeDirect.com

DMA·06
Direct Marketing Association
CONFERENCE & EXHIBITION

And now for the drawing….

- *And, don't forget, we have free "take-away" bags in the front of the room for everyone.*

- *In addition, do not forget to fill out our "database marketing practices" survey for which we will enter you in a chance to win an i-Pod nano.*