White Paper 105

Understanding Penalty Analysis

Dave Plaehn, Senior Mathematician
Gregory Stucky, CRO

What is Penalty Analysis?

Penalty (mean-drop) analysis is used by market researchers and product developers to gain an understanding of the product attributes that most affect liking, purchase interest or any other product-related measure. Product attributes used in penalty analysis are measured with "Just -about-right" (JAR) scales. These are categorical scales in which some points represent "too little" of a particular attribute, some points represent "too much," and one point represents "Just -about-right" Penalty analysis measures the change in product liking (or any other measure) due to that product having "too much" or "too little" of the attribute of interest.

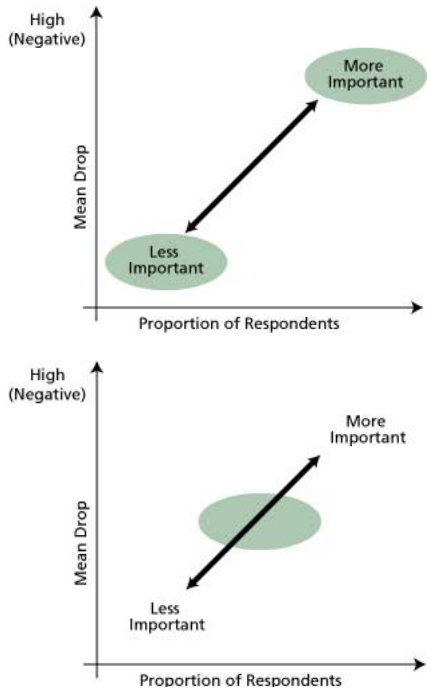


Figure 1. Penalty Scatterplots

Penalty analysis is but one of several methods used throughout the marketing research industry to reach conclusions related to the effects of a JAR variable on a different product measure. Most of these methods require substantial mathematical and statistical knowledge to implement correctly and draw appropriate inferences. At InsightsNow, we have been using, studying and writing about penalty analysis and related methods for several years; we have even developed and implemented some of our own methods. When it is implemented correctly, basic penalty analysis is a functional method that all researchers can use. This paper presents two case studies exemplifying these methods and presents recommendations for best practices.

Case Study 1: Grand Mean or JAR Mean Penalties

One of the most common methods for determining the degree to which a hedonic score is affected by a product having "too little" or "too much" of a particular attribute is to subtract the mean hedonic score across all respondents (i.e., the grand mean) from the hedonic mean of those respondents who rated the product as having "too little" of that attribute for instance (i.e., the group mean). Because the grand mean is often larger than the group mean, these values, also called mean drops, are often negative and interpreted as "penalties" due to the product having "too much" or "too little" of the attribute of interest. These "penalties" are frequently plotted against associated proportions of respondents as represented by Figure 1.

High negative penalties that are associated with large proportions of respondents (upper right quadrant) are assigned greater importance than low negative or positive penalties associated with small numbers of respondents (lower left quadrant). Unfortunately, penalties have a tendency not to organize themselves into neat quadrants on these plots. Rather, we often see penalties occupying a narrow band of space near the center (as illustrated by the bottom panel in Figure 1) instead of forming two or more distinct groups as we would prefer they do.

Grand Mean

How then do we determine the penalties on which we should focus our efforts? One method is to calculate and rank order so-called weighted penalties. A weighted penalty, as traditionally calculated, is the product of the calculated penalty and the proportion of associated respondents. If the calculated penalty uses the grand mean as its reference point, however, these weighted penalties tend to underestimate reality. The grand mean of a product measure is influenced by all respondents. This includes respondents who rated the product as "Just -about-right," along with respondents who rated it on either side of the JAR-point. As the proportion of respondents rating the product on one or the other side of the JAR-point becomes larger, the grand mean is influenced to a greater degree by these respondents. In other words, a larger proportion of respondents are "double-counted" because they are represented not only in the group mean but in the grand mean as well.

Consider the following numeric example as an illustration. Two chocolate chip cookies were rated on a JAR scale as being either "Too soft," "Too hard" or "Just-about-right." The same group of respondents rated the cookies on a 9-point overall liking scale. The following results were obtained:

	Cookie A			Cookie B		
	Too Soft	JAR	Too Hard	Too Soft	JAR	Too Hard
Proportions	0.20	.75	0.05	0.40	.55	0.05
Groups & JAR Means	5.25	6.0	5.75	5.25	6.0	5.75
Grand Means	5.84			5.69		

In the above example, 20% of respondents rated cookie "A" as "Too soft" on the JAR scale, and the mean overall liking score for those respondents was 5.25. The grand liking mean in this example is (0.2 x 5.25) + (0.75 x 6) + (0.05 x 5.75) = 5.84, and the "penalty" or mean drop associated with cookie "A" being too soft is 5.25 – 5.84 = -0.59, or a little more than half a liking scale point. The weighted penalty is -0.59 x 0.2 = -0.12.

Cookie "B" has the same mean liking associated with each of the three groups. However, instead of 20% of respondents rating the cookie as "Too soft," twice as many respondents felt this way, while the same proportion of respondents rated the cookie as "Too hard." The grand liking mean for this cookie is (0.4 x 5.25) + (0.55 x 6) + (0.05 x 5.75) = 5.69. This grand mean is smaller than the grand mean associated with cookie "A", as would be expected from a case where there were more respondents associated with a low mean and fewer respondents associated with a high mean. As a result, the penalty 5.25 – 5.69 = -0.44 is smaller in magnitude than the one associated with cookie "A", and the weighted penalty (-0.18) is only marginally larger. While one might be tempted to conclude from these results that the softness of cookie "B" is affecting liking only slightly more than is the softness cookie "A", an examination of the raw data appears to tell us that there is a much greater difference.

The problem here lies in the way the penalties, especially the weighted penalties, were calculated. Twice as many respondents were "double-counted" in the group mean and the grand mean for cookie "B" than for cookie "A". One method to correct for this is to normalize the weighted penalties on the proportion of respondents rating the products as "Just-about-right." Because this proportion decreases as the proportions of respondents rating the product on one side or another of this point increase, dividing by this proportion adjusts weighted penalties, and in effect takes into account "double-counting" respondents in both group and grand means.

Returning to the above numeric example, the normalized weighted penalty for cookie "A" is (-0.59 x 0.2) / 0.75 = -0.16 and the normalized weighted penalty for cookie "B" is (-0.44 x 0.4) / 0.55 = -0.32. The penalty for cookie "B" being too soft is now twice as large in magnitude as the same penalty for cookie "A", which makes sense given that twice as many respondents rated the cookie "B" as "Too soft."

JAR Mean

An alternative approach to calculating both penalties and normalized weighted penalties is to change the point of reference from the grand mean to the mean for only the group of respondents rating the product as "Just -about-right." The latter is called the JAR mean. Penalty analysis itself is premised on the idea that the maximum hedonic score will occur at the "Just -about-right" point. It therefore makes better sense to use the JAR mean, which is not affected by proportions of respondents rating a product on either side of the "Just-about-right" point, as the point of reference (i.e., the liking level to which we seek to optimize). The two cookies in the example have the same JAR mean and group means. Both cookies attained liking means of 6 among respondents who thought the cookie textures were "Just-about-right." If the JAR mean is used as the point of reference, the texture of both cookies is optimized to the same point, while if the grand mean is used, we would seek to optimize the first cookie to 5.84 and the second cookie to 5.69. It doesn't seem quite right to hold the

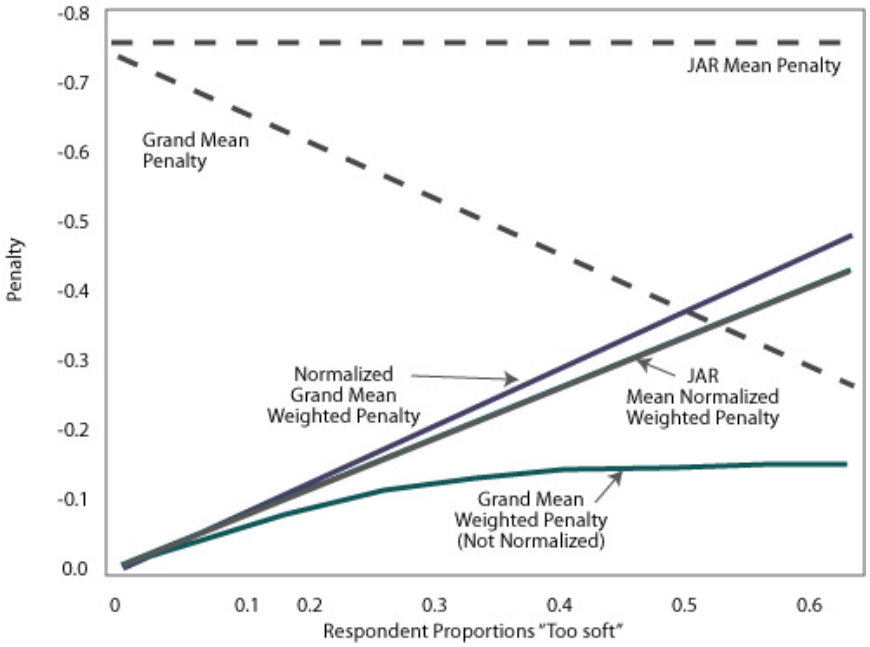
either cookie to a lower standard when both attained a higher JAR mean.

Additionally, when the JAR mean is used as the point of reference, there is no longer a need to correct for "double counting" respondents when calculating weighted penalties. As a result, no normalization is required. Instead, the JAR mean weighted penalty may accurately be calculated as the product of the penalty and the associated proportion of respondents.

From the above examples, the JAR mean penalties associated with the cookies being too soft are the same as one another: -0.75. The JAR mean weighted penalties are -0.15 and -0.3 for cookies "A" and "B" respectively. These are roughly equivalent to the normalized grand mean weighted penalties, and follow the same ratios. We recommend using the JAR Mean as the point of reference when calculating penalties.

Extending the example to more proportions of respondents, we can see a number of relationships emerge. Figure 2 shows how the various penalties change as the portion of respondents associated with "too soft" varies between 1% and 60%. The two group means and the JAR mean remain static at 5.25 ("too soft"), 6 ("Just-about-right") and 5.75 ("too hard"), and the portion of respondents associated with "too hard" remains static at 5%.

Because the two group means and JAR mean are all held constant in this example, the JAR mean penalty is constant, regardless of respondent proportions. The JAR mean weighted penalty increases linearly with respondent proportions as one would expect, given the former relationship. The grand mean penalty decreases in magnitude linearly with increasing respondent proportions as a result of the concomitant decrease in the grand mean. From the perspective of product optimization, this decrease makes little sense. All other things being equal, a penalty ought to increase in magnitude as the proportion of associated respondents increases. The non-normalized



grand mean weighted penalty appears to correct for this in some ranges of respondent proportions, but it has a curvilinear relationship with the proportion of respondents: it reaches a maximum magnitude and then begins to decrease as respondent proportions increase. Further, it is substantially smaller in magnitude than either the JAR mean weighted penalty or the normalized grand mean weighted penalty. The normalized grand mean weighted penalty has about the same magnitude as the JAR mean weighted penalty, indicating that these two will lead to about the same interpretations.

Selecting the Best

Which then is the better penalty to use: the JAR mean weighted penalty or the normalized grand mean weighted penalty? An answer to this question can be provided through modeling. Overall liking (or any other product-related measure) can be modeled from JAR variable responses, where:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

In this equation, which resembles any linear regression function, β_0 is mean overall liking across all respondents, is the intercept, which is the same as the JAR mean, and the βx's, which are often called utilities, are the same as the JAR mean weighted penalties or the normalized grand mean weighted penalties. The individual βx's may vary by small amounts depending on how they are calculated (i.e., from the JAR mean weighted penalties or the normalized grand mean weighted penalties), but the sum of the βx's for a single JAR variable will always be the same no matter how they are calculated.

Reducing this function further, the x's are the proportions of respondents (e.g., associated with "too soft" and "too hard") and the β's are the same as either the JAR mean penalties or the grand mean penalties divided by the proportion of respondents rating the product as "Just -about-right" for this attribute. Because the sum of these utilities, no matter how they are calculated, is the deviance between the intercept and mean overall liking, the JAR mean weighted penalties and normalized grand mean weighted penalties can both accurately be said to be the aggregate change in overall liking across all respondents due to the product being on one side or another of the "Just-about-right" point. However, calculating these utilities from the JAR mean weighted penalties is both mathematically more elegant and is more intuitive from the standpoint of modeling.

Given this information, we reach two

important conclusions. The first is that grand mean weighted penalties ought never to be used without normalization. Failing to normalize grand mean weighted penalties leads to underestimation of the effects of JAR variables on liking. Once this conclusion has been reached, the decision to use JAR mean penalties or normalized grand mean weighted penalties falls to personal preference. Both tell about the same story, and as has been shown, the sum of these for a given attribute is identical. Nevertheless, because of the fewer calculation steps involved and its more intuitive nature, we recommend using the JAR mean as the point of reference when calculating penalties. Rank ordering of these weighted penalties can then be used to determine relative importance.

Case Study 2: Significance Testing

As we have seen from the first case study, one way to assign relative importance to penalties is to calculate and rank order weighted penalties. But what do we do in cases of equal or nearly equal weighted penalties, and where do we draw cutoff lines? One method is to use probability theory inherent in statistical testing and declare some weighted penalties to be significantly different from zero at a certain confidence level and others not to be. Those weighted penalties that are significantly different from zero are said to exert an effect on overall liking (or another product measure), while those that are not statistically significant are said not to exert such an effect. In this case study, we examine several methods that may be used to apply statistical testing to penalties and make recommendations about best practices.

The regression equation from the first case study demonstrates that any product measure can be modeled from a JAR variable using an ordinary least squares (OLS) function. In OLS regression, the individual errors sum to zero – which is why the intercept is the same as the JAR mean – but, not all respondents are going to be perfect predictors of the regression line due to intra-respondent variation. As a result, there will be statistical error in the model. This error can be used to form statistical tests within the context of the regression equation. The most common method for doing this is to construct a t-test on the individual β's to determine if they are significantly different from zero. It should be noted though that the error in these models applies only to the β's (i.e., the JAR mean penalties) and not to the utilities (i.e., βx's, or JAR mean weighted penalties). Applying the results of statistical testing on a JAR mean penalty to its associated weighted penalty could lead to erroneous results. The JAR mean penalty for cookie "A" in the numeric example from the first case study was -0.75, and that might be significantly different from zero. But, the JAR mean weighted penalty was only -0.15, and that probably is not significantly different from zero. Further, because the utilities in these models, not the individual β's, are what represent the aggregate change in the reference variable across all respondents, it makes better sense to test the utilities or weighted penalties than it does to test the β 's or JAR mean penalties.

In order to test the weighted penalties, we need some method to estimate the statistical error associated with them. The overall model error cannot be used for the reasons described above. There are however, several other methods by which error around a value may be estimated and these center on forming a distribution around that value. Jackknifing and bootstrap-related methods are useful ways to do this*. Once a distribution has been formed around a value, statistical testing may take place on that value, given the context of the distribution. Again, a t-test is a common way to determine if a single value, such as a weighted penalty, is significantly different from zero. With bootstrap methods, which typically result in distributions having 1000 or more values, the generated distribution itself can be used to construct a statistical test. If the value stated in the null (e.g., zero in the case of weighted penalties) occurs at a percentile outside the range specified by a confidence interval, we conclude statistical significance. This method is called percentile bootstrap.

We recently tested over 350 weighted penalties for statistical significance using a variety of methods. In almost all cases where bootstrapping was used, the generated distributions of penalty weights were significantly skewed in the direction of the sign of the penalty. A t-test assumes a symmetric distribution. If a distribution is not symmetric, use of a t-test may not be appropriate.

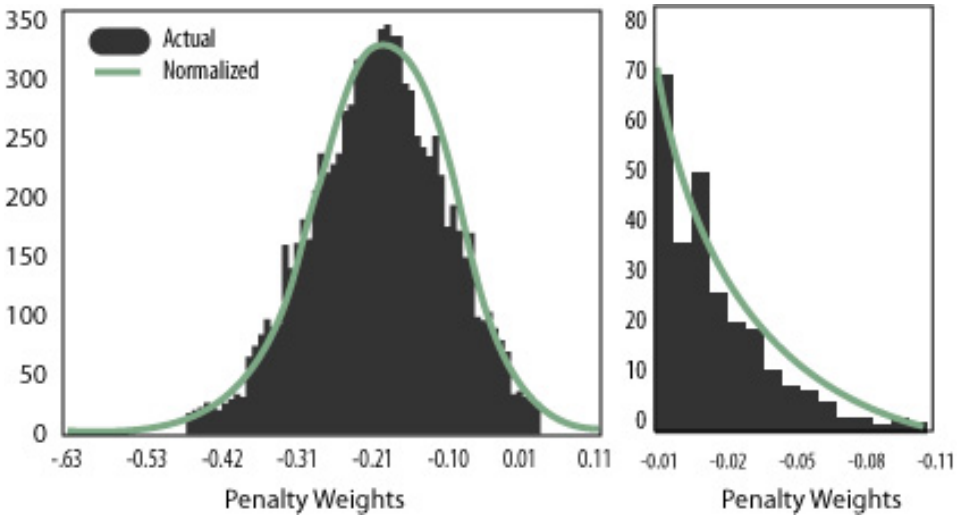
Figure 3 shows a bootstrapped distribution having 10,000 values generated around a single weighted penalty. The actual distribution in this example is skewed left relative to the normal one having the same mean and standard deviation (note the thin right tail relative to normal). The value on the x-axis associated with 2.5% under the right tail of the normal distribution (i.e., the upper limit of a normal confidence interval when testing at the 5% significance level) is greater than zero. The conclusion from this is that the confidence interval contains zero and thus is not statistically significant. The 9,751st rank-ordered value from the actual distribution, on the other hand, is less than zero. Because the confidence interval from this distribution does not contain zero, we reach the opposite conclusion. The first is likely the less trustworthy of the two given the sensitivity of t- and z-tests to violations of symmetry assumptions.

One other possibility for testing the significance of weighted penalties, given their tendency to skew, is to postulate a distribution based on the two parameters of a weighted penalty. These parameters are magnitude of the penalty, which does tend to follow a symmetric or normal distribution, and associated proportion of respondents, which follows a binomial distribution. Testing against quantiles in this postulated distribution leads to results similar to those found with the percentile bootstrap method. However, we can-

not yet be certain that the two parameters of this distribution are independent of one another, which would be a key assumption underlying its use. Because of this uncertainty, we recommend the percentile bootstrap method be used to test the statistical significance of weighted penalties. We understand that bootstrapping consumes a larger amount of both time and computing resources than does the use of a postulated distribution. The use of the postulated distribution presents a good alternative to bootstrapping if these issues are of paramount concern.

cutoff. Researchers typically ignore penalties that are associated with fewer than 20% of respondents, for instance. The rationale is that if fewer than this portion of respondents found fault with a product on a particular attribute, there is not cause to adjust the product on that attribute.

A second safeguard is a weights cutoff. While the cutoff described above refers to only one parameter in a weighted penalty, this second cutoff refers to the weighted penalty itself. We have employed a cutoff that is theoretically associated with driving statistical significance in the reference variable (e.g., overall liking). If a weighted penalty is not sufficiently large in magnitude that it would lead to a significant change in product liking, then that penalty is ignored. Determining this cutoff is more difficult than assigning and sticking to an arbitrary value. Power calculations suggest that a weights cutoff of about 0.4 is appropriate when a 9-point liking scale is used as the reference variable and samples contain about 100 respondents. Other scenarios may require different weights cutoffs. We recommend that some reasonable weights cutoff be used to determine which weighted penalties are important aside from rank ordering and significance testing by themselves.



Epilogue: Statistical vs. Real-World Significance

A consideration that must be made when doing any kind of statistical testing is whether a claim of statistical significance has any real-world implications. A weighted penalty of -0.08 may be found to be significantly different from zero if the sample size is large enough, but it may not always be reasonable to concentrate efforts on a product attribute that moves product liking less than 1/10 of a scale point.

Establishing safeguards against conflating small but significant weighted penalties with real-world importance is generally good practice. One of these is to employ a respondent proportion

Conclusion

Penalty analysis is a widely used tool for understanding how certain product attributes affect another aspect of the same product. As with any analytic tool, an understanding of the underlying basis of how penalties are constructed and tested, as well as how they ought to be interpreted is key to the proper use of the method.
We present the following guidelines: JAR mean penalties are preferred over grand mean penalties; statistical testing should properly refer to weighted penalties, not the raw penalties; and care should be taken to not over-interpret results from significance testing. Keeping these guidelines in mind and employing the recommendations contained here will lead to better use of penalty analysis as a tool for optimizing product attributes.

About InsightsNow

InsightsNow is the leading product design and development research company whose unique consumer behavior frameworks, powerful technology platform, and rapid delivery of relevant insights provide a rich, integrated, and scalable research environment that accelerates the development of products for major consumer packaged goods manufacturers. Through the company's advanced real time research and analytic solutions, such as profilesNOW and reportsNOW, InsightsNow provides faster, deeper, and more insightful results. InsightsNow's headquarters are in Corvallis, Oregon with offices throughout the United States. For more information, visit www.InsightsNow.com or call 541-757-1404.