insightsNOW

Penalty Plus

Perform complete analysis of attribute-related survey data in order to help guide product optimization with InsightsNow's unique approach to penalty testing.

From InsightsNow

The intuitive, self-contained Penalty Plus application allows more complete analysis of attribute-related survey data in order to help guide product optimization. Sophisticated, model-based approaches help you perform significance testing, and set up parameters to identify product testing study participants who fall outside of the "just about right" range.

Read on to learn more about best practices and recommendations regarding penalty analysis, so you can understand how to apply it to your next project.

What is Penalty Analysis?

InsightsNow's Penalty Plus is our approach to mean-drop analysis used to gain an understanding of the product attributes that most affect consumer liking, purchase interest or any other product-related measure. Market researchers and product developers use these insights to hone or innovate products for the maximum impact in market.

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.

Product attributes used in penalty analysis are measured with "Just-About-Right" (JAR) scales. These are categorical scales where 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 due to that product having "too much" or "too little" of the attribute of interest. When it is implemented correctly, a penalty analysis research approach is a functional method that all researchers can use.

Two Penalty Testing Approaches

First, let's look at two kinds of penalty testing: Grand Mean or JAR Mean Penalties. One of the most common methods for determining the degree to which a 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 study participants (the grand mean) from the hedonic mean of those study participants who rated the product as having "too little" of that attribute (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 study participants as represented by Figure 1.

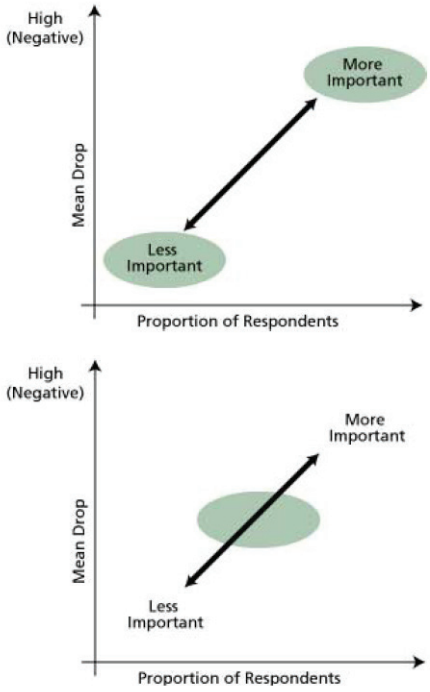

Figure 1. Penalty Scatterplots

High negative penalties that are associated with large proportions of study participants (upper right quadrant) are assigned greater importance than low negative or positive penalties associated with small numbers of study participants (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 study participants. 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 study participants. This includes study participants who rated the product as "Just-About-Right," along with study participants who rated it on either side of the JAR-point. As the proportion of study participants 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 study participants. In other words, a larger proportion of study participants are "double-counted" because they are represented not only in the group mean, but in the grand mean as well.

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 study participants 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 study participants 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).

Selecting the Best Penalty Approach

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:

$$y = \beta_0 + \beta_{1\,x1} + \beta_{2\,x2}$$

In this equation, which resembles any linear regression function, β_0 is mean overall liking across all study participants, is the intercept, which is the same as the JAR mean, and the $\beta_{x's}$, which are often called utilities, are the same as the JAR mean weighted penalties or the normalized grand mean weighted penalties. The individual $\beta_{x's}$ 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 $\beta_{x's}$ 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 study participants associated with certain product attributes and the β's are the same as either the JAR mean penalties or the grand mean penalties divided by the proportion of study participants 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 study participants 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.

We conclude then 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. Both JAR mean penalties or normalized grand mean weighted penalties tell about the same story, but 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.

Significance Testing

As we have seen, 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? 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.

Here we will examine several methods that may be used to apply statistical testing to penalties and make recommendations about best practices. The regression equation

$$y = \beta_0 + \beta_{1\,x1} + \beta_{2\,x2}$$

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 study participants 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., $\beta_{x's}$, 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. Further, because the utilities in these models, not the individual β's, are what represent the aggregate change in the reference variable across all study participants, 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. There are several methods by which error around a value may be estimated and these center on forming a distribution around that value. The bootstrap-related method is a useful way 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 1,000 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.

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 cutoff. Researchers typically ignore penalties that are associated with fewer than 20% of study participants, for instance. The rationale is that if fewer than this portion of study participants 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. If a weighted penalty is not sufficiently large in magnitude that it would lead to a significant change in, say, 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 study participants. 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.

Conclusions

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.

Interested in learning more about how InsightsNow's Penalty Plus approach can work for you?

Let's connect:
info@insightsnow.com or 541.757.1404

About InsightsNow

InsightsNow, an award-winning behavioral research firm, partners with companies across a wide array of industry verticals to accelerate marketing, branding and product development decisions for disruptive innovations achieving a cleaner, healthier, happier world. Insights are provided via custom solutions and assisted DIY tools based on proprietary behavioral frameworks to help find answers faster, improving your speed-to and success-in market.