

# O\*NET Data Collection Program: Statistical Procedures for Deviant Case Detection

National Center for O\*NET Development  
Raleigh, NC

**Overview.** This document provides a descriptive synopsis of the statistical procedures that are utilized during the deviance detection process. Deviance detection occurs in two sequential stages. The first stage involves the application of empirical procedures to identify respondent “outliers.” That is, survey cases whose responses are significantly different in comparison to other respondents for each surveyed O\*NET occupation. The outcomes of these analyses are used as input for a second stage that applies a rational review procedure conducted by trained occupational analysts. The focus of this document is on the first stage of deviance detection. It is important to note that prior to deviance detection, survey responses are subjected to various data cleaning procedures as well.

**Statistical Deviance Analysis.** The empirical procedure for deviance detection uses multiple statistical analyses to identify possible deviant respondents (see Appendix A for more details regarding the concept of “deviance”). The analysis tools used in this procedure incorporate multivariate approaches commonly implemented for outlier identification. Analyses are conducted on O\*NET task data for reasons described in Appendix B of this document. The ensuing paragraphs briefly summarize the sequence of the empirical procedure used for deviance detection for each occupational sample.

**Step 1: Minimum covariance determinant (MCD).** Task means are critical to deviance detection in that they serve as the benchmark against which to compare individual respondents. The task ratings from survey respondents are first input into a procedure to determine the “minimum covariance determinant” (MCD) of the sample. The MCD estimate is computed using a computer algorithm provided by Rousseeuw and Van Drissen (1999) and implemented in an SAS PROC IML routine.

In short, the MCD procedure (1) repeatedly samples  $(1 - h) * N$  of the total multivariate sample of size  $N$ , (2) computes the covariance matrix and multivariate means on the selected sub-sample, (3) uses these estimates to compute Mahalanobis squared distance for each observation in the total sample, (4) ranks the individual observations on their Mahalanobis squared distance values, (5) selects the lowest  $(1 - h) * N$  observations, (6) computes and records the determinant, and (7) repeats the procedure once again using the selected observations as the new sub-sample. This procedure is then repeated  $t$  times. The  $t$  determinants are then ranked and the lowest  $m$  are selected. Each of the  $m$  determinants is then used as input and the procedure iterated until the determinant reaches its minimum. The smallest determinant is then declared the MCD.



The MCD procedure has several advantages. For example, it is sensitive to situations where multiple respondents are potentially deviant. In such situations, using typical outlier analytic approaches (e.g., Euclidean distance on raw data) could skew empirical results because these approaches generally only control for biasing effects of outliers by computing means and covariance estimates over a reduced sample obtained by dropping *single observations* from the total sample. Instead, the MCD procedure assumes that outliers constitute a proportion ( $h$ ) of a total sample of size  $N$  and then finds that subset of  $N*(1 - h)$  respondents with the smallest covariance matrix. Given that the determinant is often regarded as a measure of *generalized variance*, the problem can be restated as finding the subset of  $N*(1 - h)$  observations with the minimum covariance determinant.

In addition to the advantage described above, the MCD procedure increases the efficiency and reliability of subsequent statistical analyses. In the context of deviance detection, reliability is a function of the extent to which task means are computed on a subsample of respondents belonging to the same occupation. The MCD procedure is an iterative technique for identifying the subset of respondents with the minimum covariance, with the notion that the smaller the covariance matrix, the more congruent the subset of respondent task ratings and the greater the likelihood that all selected respondents belong to the same occupation (i.e., are not “deviant”).

**Step 2: Computing Estimates of Distance.** A common approach to outlier identification is the use of various distance estimates. Two specific estimates of distance are computed: Mahalanobis squared distance ( $D^2$ ) and Robust Mahalanobis squared distance ( $RD^2$ ). These are briefly summarized below.

$D^2$  estimates are the unweighted squared distance between respondent’s profile of task ratings and a profile of task mean ratings excluding the respondent.  $D^2$  is sensitive to both shape and level profile differences. A  $D^2$  of zero between a respondent’s profile and the respondent-excluded mean profile implies that the two profiles are perfectly overlapping.

$RD^2$  estimates are the weighted sum of squared differences between a respondent’s task rating and the mean task rating across all  $p$  tasks. The weights are the individual elements of the inverse of the  $p \times p$  sample covariance matrix  $S$ . The method is termed “robust” in that both the task means and covariance matrix are computed exclusive of the respondent in question, thus preventing a deviant respondent from biasing the results. The weighted difference scores are ranked in descending order, with the most deviant respondents defined as those having the largest weighted difference from the sample corrected means. This method is similar to a  $D^2$  measure, with the exception that  $RD^2$  differentially weights each respondent–mean difference, whereas  $D^2$  treats all differences as equally important.

Procedurally,  $D^2$  estimates are computed using the total sample means and covariances for a given occupation.  $RD^2$  estimates are computed using the means and covariance matrix for the reduced (subset) sample corresponding to the MCD (from step 1).



**Step 3: Plotting Distance Estimates.** Outlier identification is facilitated by comparing the  $RD^2$  with  $D^2$  estimates using Q-Q plots (or a distance-distance plot). Under the assumption of multivariate normality, both distance measures follow a chi-square distribution. Thus, a chi-square cut-off ( $p < .0001$ ) is used to partition the plot into four separate sectors, with each sector representing the correspondence of outlier identification across the two distance measures. These sectors are described below.

- *Sector 1* contains all sample respondents that fall below the chi-square cut-off on both the regular and robust  $D^2$  values
- *Sector 2* contains respondents that are declared by the  $D^2$  measure as outliers (exceeding the chi-square cut-off) but not by the  $RD^2$  measure
- *Sector 3* includes respondents declared as outliers by both  $D^2$  measures
- *Sector 4* contains sample observations declared as outliers by the robust  $D^2$  but not by the regular  $D^2$ , these respondents are sometimes referred to as “hidden outliers” because they are masked by regular distance computation

For all subsequent analyses, outliers are defined as those respondents in Sectors 2, 3 and 4. Respondents in Sector 1 are tentatively considered being in the “inlier” group (i.e., most likely *not* deviant respondents).

**Step 4: “Corr\_plus” Deviance Test.** This regression-based procedure provides the primary test for respondent deviance (see Appendix C for statistical details of this procedure).

From Step 3 above, respondents falling into Sectors 2-4 are considered candidates for “deviant” designation. However, this initial designation is biased for O\*NET purposes because it contains both respondents whose task ratings are unexpectedly high as well as those whose ratings are unexpectedly low. Therefore, it is necessary to screen outliers so as to retain only those whose deviance is toward the low end of the scale (i.e., toward task ratings of “not relevant”). Deviant respondents can be expected to exhibit a greater-than-average number of 0 ratings (0-importance ratings correspond to marking tasks as “not relevant”).

Operationally, “corr\_plus” estimates are defined as  $Cov(T_jT)/Var(T)$  under the assumption of 0 origin, where  $T_j$  is the vector of task ratings for respondent  $j$  and  $T$  is the vector of “inlier” group (i.e., non-deviant incumbents) task means. In the context of regression analysis,  $Cov(T_jT)/Var(T) = \beta_{T_j,T}$ , which implies that a respondent’s task ratings can be considered to be regressed on the mean task ratings of the inlier group from Step 3.

$\beta_{T_j,T}$  bears a direct monotonic relation to occupation membership. That is to say, the higher the value of  $\beta_{T_j,T}$ , the greater the likelihood that respondent  $j$  belongs to the surveyed occupation. The reason for this designation is that the regression line is hinged at the 0 intercept. As the



vector of respondent ratings comes to contain higher importance ratings across the occupation's tasks, the slope increases to accommodate the upward pull of these ratings. For the situation where a respondent's rating profile is congruent with the profile of mean ratings, then  $\beta_{Tj-T} = 1$ . As the respondent's profile rises above the benchmark profile of means,  $\beta_{Tj-T}$  moves from 1 toward its maximum in a steady upward progression.

To identify a list of deviant respondents for the second stage of deviance detection (i.e., rational review by analysts), it is necessary to establish a cut-point for classification as "deviant." Because corr\_plus estimates are equivalent to regression beta-weights, the corr\_plus estimates are standardized and a one-tailed  $t_{(.05)} = -1.65$  value is used as the cut-point. A one-tailed  $p$ -value is appropriate because only those in the lower tail of the distribution are of interest.

This list of potential deviant respondents is then subjected to a rational review process conducted by occupational analysts.

## References

Rousseeuw, P. J. & Van Drissen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Techometrics*, 41, 212-223.

## Appendix A

**Overview.** The problem of central concern is the identification and evaluation of deviant survey respondents. In general, deviance connotes deviation from an expectation, a general rule, or a normative standard. In the context of the O\*NET Data Collection Program, deviance has a specific occupational interpretation. Incumbents in an occupation other than the targeted occupation can be defined as deviant with regard to their occupational membership. For survey purposes, an occupation can be conceptually equated with a population to which survey results are to be generalized. Anomalous survey cases indicate potential misclassifications into the “wrong populations” (occupations), with consequent degradation of the quality of computed sampling statistics for the target occupations.

**The Notion of Outliers.** An outlier in a data set has been defined as “...an observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data” (Barnett & Lewis, 1984, p. 4). A similar definition is offered by Grubbs (1969) who notes “An outlying observation, or ‘outlier’, is one that appears to deviate markedly from other members of the sample in which it occurs” (as quoted in Barnett & Lewis, 1984, p. 25). Use of “appears” in both definitions is central in that it implies an element of subjective post-data judgment on the part of an observer.

For O\*NET purposes in general, the terms “outlier” and “deviant” are interchangeable. The only exception is the situation in which outlier designation is based upon a “positive bias.” In other words, a respondent may be an outlier because he/she tends to rate toward the high end of the O\*NET scales (e.g., higher importance ratings for tasks). In this specific situation, an outlier respondent is not considered deviant in terms of occupational misclassification.

As the primary purpose of deviance analysis is to detect anomalous cases in order to maintain data quality, several important distinctions between occupational misclassification (i.e., deviance identification) and work analysis validity/integrity should be noted. The delineation of these differences has a significant impact on both the conceptualization of deviance and the procedures by which to test for deviance. Typically, validity in job/occupational analysis stems directly from the extent to which the collected data are indeed *relevant* to a given job, occupation, or work role. In contrast, identification of occupationally misclassified incumbents (i.e., deviants) turns the investigative scope toward relevancy of the respondent’s occupation membership status. In sum, although deviance identification for the sake of maintaining data integrity seems akin to validity-related analysis, it should not be construed as an equivalent endeavor.

One clear difference between deviance testing and validity investigation lies within the concern of population membership. In validity-related examinations of work analytic data, respondent membership is *assumed*. Thus, the focus of attention is on the job or occupation relevance of particular descriptors. On the other hand, in the context of deviance analysis, one is specifically interested in population membership (i.e., whether or not a respondent *belongs* to an occupation). The chosen descriptors, whether tasks or domain-level items, are *presumed* to be relevant to a



given occupation. A second difference between deviance and validity testing concerns the assumption of the “truthfulness” of respondents. In validity analysis, the “truth” of the respondent is always in question (e.g., does this incumbent really perform the task, or does this incumbent truly believe that numerical reasoning is important to his/her work role). This questioning of truth can even be seen in the methods commonly employed to gather validity-related evidence, such as the “carelessness index” (Green & Stutzman, 1986), “infrequency index” (Green & Veres, 1990), “false reporting scale” (Pine, 1995), and the use of “veracity items” (McCormick, 1960), all of which attempt to gauge whether or not an incumbent is responding truthfully to a work analysis instrument. Conversely, the truthfulness of a respondent is not primary in deviance testing. In deviance testing, truth is inferred from the extent to which respondents have congruent ratings relative to others within a given occupation. The significant implication of these two differences is that response consistency can now be used as a proxy for population membership, thereby allowing deviance testing to identify incumbents who “don’t belong” to a given occupation without being constrained by the validity (i.e., truthfulness) of their responses. An additional benefit of this rationale is that deviance testing is distinguished from validity examination, which is a highly contentious issue that presently lacks concise resolution in the associated literature (see Harvey & Wilson, 2000; Morgeson & Campion, 2000; Sanchez & Levine, 2000).

Importantly, validity-type examinations are not ignored within the present data cleaning processes. From a systemic perspective, the first-order data cleaning processes that occur prior to deviance analysis are analogous to validity examinations in that they serve to filter out survey respondents with seemingly dubious response sets. These examinations are critical undertakings and take place prior to the dissemination of survey results in order to at least partially bolster the validity of the collected work data. The fact that these validity-related investigations occur prior to deviance detection procedures is crucially significant as well, as the logical flow of maintaining data integrity should begin with more micro investigations (i.e., data cleaning) before proceeding to more macro analysis (i.e., deviance testing). This procedure helps to ensure that deviance results are not skewed by potentially “invalid” respondent data, thereby reducing the risk of false positives in deviance detection testing.

## References

- Barnett, V., & Lewis, T. (1984). *Outliers in statistical data* (2nd ed.). Chichester, England: Wiley.
- Green, S. B., & Stutzman, T. (1986). An evaluation of methods to select respondents to structured job-analysis questionnaires. *Personnel Psychology*, *39*, 543-564.
- Green, S. B., & Veres, J. G. III (1990). Evaluation of an index to detect inaccurate respondents to a task analysis inventory. *Journal of Business and Psychology*, *5*, 47-61.



Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, *11*, 1-21.

Harvey, R. J., & Wilson, M. A. (2000). Yes Virginia, here is an objective reality in job analysis. *Journal of Organizational Behavior*, *21*, 829-854.

McCormick, E. J. (1960). *Effect of amount of job information required on reliability of incumbents' check-list reports*. USAF Wright Air Development Division Technical Note, 60-142.

Morgeson, F. P., & Campion, M. A. (2000). Accuracy in job analysis: Toward an inference-based model. *Journal of Organizational Behavior*, *21*, 819-827.

Pine, D. E. (1995). Assessing the validity of job ratings: An empirical study of false reporting in task inventories. *Public Personnel Management*, *24*, 451-459.

Sanchez, J. I., & Levine, E. L. (2000). Accuracy or consequential validity: Which is the better standard for job analysis data? *Journal of Organizational Behavior*, *21*, 809-818.

## Appendix B

**Using Task Data for Empirical Deviance Detection.** With the scope of data collected from respondents ranging from task to domain categories (e.g., work activities, knowledge, and skills), it becomes important to consider both the ramifications and appropriateness of using either data type in deviance analysis. There are four important reasons to use task data for purposes of empirical deviance identification. These reasons are listed below and discussed in the ensuing paragraphs.

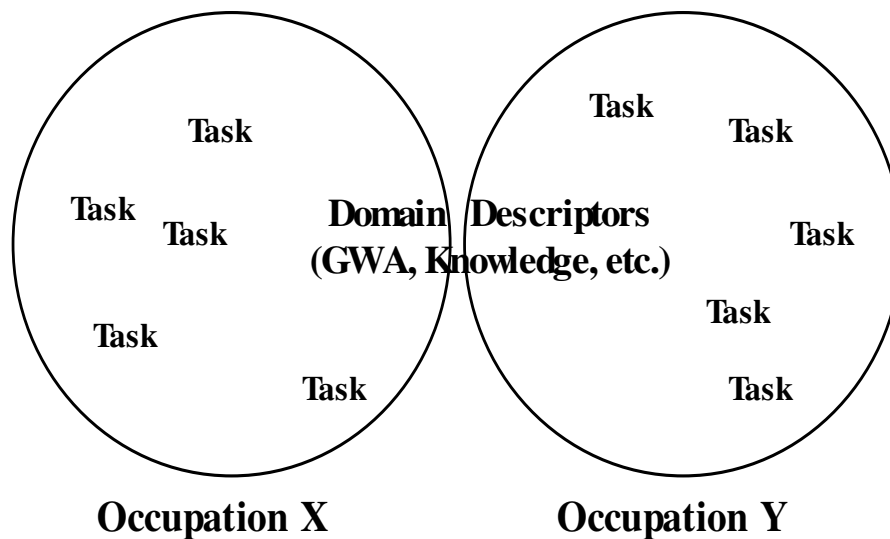
1. Historical significance
2. Level of specificity
3. Expectation of variability
4. Operational implications

The first reason speaks to the long-standing tradition in I/O psychology to partition the world of work through the use of behavioral data. This tradition is evident in McCormick's (1979) classic description of the classification of work (i.e., task > duty > position > job > occupation) that clearly begins with the simplicity of a behavioral statement. Important to the direct purpose of final deviance analysis is the definition of occupation as "a group of similar jobs found in several establishments" (Shartle, 1952, p. 26) and the fact that occupational titles usually apply to similar jobs on a nationwide basis. Moreover, it is commonly assumed that jobs under the same occupational title share "a common set of *tasks* [that are] performed or are related in terms of similar objectives, methodologies, materials, products, worker actions..." (emphasis added, U.S. Department of Labor, 1991, p. 2-1). These definitions support the notion that a job or occupation is predicated upon a common set of performed tasks/activities. Thus, the implication for empirical deviance detection is that analysis should focus on "border creating" descriptors (i.e. tasks) that serve to delineate one occupation from another. Stated another way, respondents should be distinguished, in terms of occupational classification, on the information that defines the occupation itself. As jobs and occupations are at essence creations of organizational convenience, the most defensible process of identifying deviant respondents would be one that focuses on ratings of the actual work performed within a given occupation.

Level of data specificity also plays a critical role in the use of task data for empirical deviance detection. Tasks are considered to be job- and occupation-specific (Mitchell & Wilson, 1999), whereas O\*NET domain descriptors are intentionally designed to be generic. Occupations can, and should, be distinguished by their data specificities. Additionally, tasks are more concrete than domain descriptors, rendering them "easier" to rate by respondents. Domain descriptors require individuals to rate more abstract properties of the occupation and of people, both of which involve more complex cognitive decision making processes and increased amounts of information that must be considered. Using a performance appraisal analogy, empirical deviance detection that includes domain descriptors as data input will likely have a greater chance of criterion contamination, whereas using tasks more likely leads to criterion deficiency. Criterion contamination involves the inclusion of irrelevant or invalid information, whereas criterion



deficiency stems from the omission of relevant, valid information. In terms of empirical deviance detection, deficiency in task coverage can be tolerated in that it is indeed unreasonable to assume that *all* tasks for a given occupation are subsumed under the existing task list; however, the tasks that are included are assumed to be *representative* of the occupation. On the other hand, criterion contamination in the domain descriptors, whether attributable to misclassification or to some other influence, would have a much more deleterious effect on empirical deviance detection by distorting true ratings within an occupation and thereby increasing the likelihood of error in deviant identification. Furthermore, Dierdorff and Wilson (2003) have shown that task statements are more reliably rated by incumbents, both in terms of interrater and intrarater reliability, than more generic descriptors such as generalized work activities.



**Figure 1.** *Data variability across specificity levels*

A third reason for using tasks instead of domain data for deviance testing relates to the expectation of variability across and within occupations. In order for data to adequately discriminate individuals properly classified in an occupation, they must display less variability within an occupation than variability across occupations. Figure 1 graphically depicts how tasks and domain descriptors relate to occupational boundaries. As tasks are occupation-specific and domain descriptors are intentionally generic, task statements better fulfill this requirement than do domain descriptors. Generic descriptors are designed to be applicable across a wide range of jobs and occupations (Cunningham, Drewes, & Powell, 1995) and thus, domain ratings are not as useful as task ratings for purposes of occupational identification. This is not to say that occupations will not vary in terms of mean ratings on domain items, indeed they will, but rather that survey respondents cannot be classified as belonging to a particular occupation based upon

differences in their ratings on such widely applicable items. For example, civil engineers and landscape architects are quite similar in terms of the types and levels of knowledge (see O\*NET Online®), yet are highly differentiated in the tasks that each requires. Thus, to judge deviance on rating differences from the knowledge domain would inadequately discriminate between a landscape architect and a civil engineer.

The fourth, and final, reason for using task data pertains to operational processes. There will always be more respondent task data within an occupational sample than there will be domain data. This is due to the pre-existing survey design that requires all respondents to rate an occupation's task list, but divides the domain survey sampling into quarters (with the exception of a minority of occupations rated by occupational experts). Thus, for an occupation that has 100 respondents, all 100 will have provided task ratings, while only 25 will have ratings for each domain. Of note is that the primary information used in deviance identification is derived from the statistical results of empirical deviance detection. Sample size clearly impacts the power, generalizability, and overall confidence in results of any statistical test. Moreover, because deviance analysis identifies potential deviants by comparing respondents relative to each other, performing empirical deviance detection on small samples is generally undesirable. Task data tends to be free of these operational limitations.

## References

Cunningham, J. W., Drewes, D. W., & Powell, T. E. (1995). Framework for a revised Standard Occupational Classification (SOC). In *Standard Classification Revision Policy Committee (Eds.), Seminar on research findings* (p. 57-165). Washington, DC: U.S. Department of Labor. (U.S. Government Printing Office No. 1995-398-319/40067).

Dierdorff, E. C., & Wilson, M. A. (2003). Meta-analysis of job analysis reliability. *Journal of Applied Psychology, 88*, 635-646.

McCormick, E. J. (1979). *Job analysis: Methods and applications*. New York: AMACOM.

Mitchell, J. L., & Wilson, M. A. (1999). Using O\*NET to develop organization-specific tasks. In D. W. Drewes, M. A. Wilson, J. W. Cunningham (Eds.), *O\*NET Work Analysis Fieldbook: A guide for defining the world of work*. Raleigh, NC: National Center for O\*NET Development. Unpublished manuscript.

Shartle, C.A. (1952). *Occupational information: Its development and application* (2<sup>nd</sup> Ed.). New York: Prentice Hall.

U.S. Department of Labor, Employment and Training Administration. (1991). *The Revised handbook for analyzing jobs*. Washington, DC: U.S. Government Printing Office.



## Appendix C

**A Regression Model of Occupation Membership.** For respondent  $j$  of a surveyed occupation, the task means  $T$  can be used to predict respondent  $j$ 's task ratings. According to the linear regression model

$$\hat{T}_j = \beta_{Tj \cdot T} T$$

If respondent  $j$  is a non-member of the surveyed occupation but is inadvertently included, the regression model should predict 0-ratings for all tasks, assuming a valid rater and no task overlap across surveyed occupations and the respondent's actual occupation. This prediction is because 0-ratings imply that tasks are not relevant to the work performed in the respondent's occupation. This logical requirement necessitates that  $\beta_{Tj \cdot T} = 0$  if and only if respondent  $j$  is *not* a member of the surveyed occupation (i.e., is a deviant). Thus, an observed beta of 0 for a respondent would imply occupation non-membership.

Unfortunately, the 'real world' of occupation membership determination is seldom quite so straightforward. All respondents whose entirety of survey responses comprise 0-ratings are already excluded by O\*NET data screening procedures prior to empirical deviance detection, thereby precluding the event  $\beta_{Tj \cdot T} = 0$  for any given respondent  $j$ . Furthermore, there is likely to be varying degrees of task content overlap across O\*NET occupations. The net result is that  $\beta_{Tj \cdot T}$  for a respondent  $j$  is a *relative index* of occupation membership varying from a low of near 0 to a high of 5

$$5 \frac{\sum_{i=1}^p \bar{t}_i}{\sum_{i=1}^p \bar{t}_i^2}$$

where  $\bar{t}_i$  is the mean of the  $i^{\text{th}}$  task for the inlier group.

Fortunately,  $\beta_{Tj \cdot T}$  bears a direct monotonic relation to occupation membership. That is to say, the higher the value of  $\beta_{Tj \cdot T}$ , the greater the likelihood that respondent  $j$  belongs to the surveyed occupation. The reason for this designation is that the regression line is hinged at the 0 intercept. As the vector of respondent ratings comes to contain higher importance ratings across the occupation's tasks, the slope increases to accommodate to the upward pull of these ratings. For the situation in which a respondent's rating profile is congruent with the profile of mean ratings, then  $\beta_{Tj \cdot T} = 1$ . As the respondent's profile rises above the benchmark profile of means,  $\beta_{Tj \cdot T}$  moves from 1 toward its maximum in a steady upward progression.

The relative nature of  $\beta_{Tj \cdot T}$  as an index of occupation membership versus the binary nature of the membership decision (i.e., yes or no) necessitates that a cut-point (CP) be imposed on the range of  $\beta_{Tj \cdot T}$ ,  $j = 1, 2, \dots, N$ , where  $N$  is the number of incumbents in the occupation sample. Imposition of a cut-off allows the decision of occupation membership = "No" when  $\beta_{Tj \cdot T} < \text{CP}$  and occupation membership = "Yes" when  $\beta_{Tj \cdot T} \geq \text{CP}$ , where the cut-point value is set at the discretion of the decision maker. The empirical deviance detection procedure standardizes corr\_plus values (equivalent to beta-weights) and uses the one-tailed  $t_{(.05)} = -1.65$  as the CP.

