Research Report

Outliers in Assessments

Shelby J. Haberman

July 2008

ETS RR-08-41

Outliers in Assessments

Shelby J. Haberman

ETS, Princeton, NJ

July 2008

Abstract

Outliers in assessments are often treated as a nuisance for data analysis; however, they can also assist in quality assurance. Their frequency can suggest problems with form codes, scanning accuracy, ability of examinees to enter responses as they intend, or exposure of items.

Key words: Residual, regression

Acknowledgments

This paper has benefited from a number of conversations with Tim Davey and from helpful comments from Hyeon-Joo Oh. Helpful reviews have been provided by Tim Davey, Michael Walker, and Dan Eignor. Data were supplied by the staff in the Center for Statistical Analysis.

Outliers are often encountered in educational assessments. They can be used in a program of quality assurance to detect unusual results that suggest gross errors in test administration such as mistakes in form codes or scanning problems. Outliers may also have potential to detect problems that involve item disclosure or errors of examinees in data entry. Two types of outliers are readily considered. The first type is an unusual score on an examination. The second type is an unusual deviation from the score predicted by a regression of an examination subscore on other examination scores. Analysis of outliers in assessments is typically made more complicated by the large number of examinees. Some outliers are expected with virtually any reasonable definition of outliers; however, the fraction of observations that are outliers should be small. To investigate potential for outlier analysis, data were examined from an ETS assessment. The methods of analysis are primarily designed for use with conventional tests that are not adaptive and that are scored by adding up raw item scores to provide a total score. The analysis is not concerned with subsequent procedures to equate, link, and scale scores. In section 1, the basic methodology is discussed. In section 2, an application is made to the data under study.

1 Methodology

The basic methodology involved is quite simple. One has n examinees numbered from 1 to n and test sections numbered from 1 to $q < n - 2$, where $q \geq 2$. The test sections do not overlap. These sections may be conventional sections of a test or sections based on the format of the answer sheet. For instance, in the example under study, there are four sections based on content areas covered by the test, and there are three columns on the answer sheet. The section scores may be quite relevant in a study of whether some examinees have remarkably high or low scores in some content area. The column scores may be important if scanning errors or gridding errors are of concern.

For each test section j, $1 \leq j \leq q$, and examinee i, a section score X_{ij} is available, and the total score Y_i for examinee i is the sum of the q section scores for that examinee. Let the q-dimensional vector $\mathbf{X}_i$ have coordinates X_{ij}, $1 \leq j \leq q$. It is assumed that the $\mathbf{X}_i$, $1 \leq i \leq n$, are mutually independent and identically distributed, and it is assumed that the X_{ij} have finite means and variances. The expectation of X_{ij} is μ_j, and the covariance of X_{ij} and X_{ik} is γ_{jk}. Let $\mathbf{\Gamma}$ be the q by q symmetric covariance matrix of $\mathbf{X}_i$, so that row j and column k of $\mathbf{\Gamma}$ is equal to γ_{jk}. It is assumed that section scores are not trivially related, so that $\mathbf{\Gamma}$ is positive definite and the correlation coefficient $\rho_{jk} = \gamma_{jk}/(\gamma_{jj}\gamma_{kk})^{1/2}$ of X_{ij} and X_{ik} is defined. Analysis uses the basic summary statistics that include the section average $\bar{X}_j = n^{-1}\sum_{i=1}^{n} X_{ij}$, which estimate the population mean μ_j and the sample covariance

$$C_{jk} = (n-1)^{-1}\sum_{i=1}^{n}(X_{ij} - \bar{X}_j)(X_{ik} - \bar{X}_k),$$

which estimates γ_{jk}.

1.1 Section Residuals

When the means μ_j and covariances γ_{jk} are known, then, for each section score X_{ij}, a best linear predictor is easily found for prediction of X_{ij} by the other section scores X_{ik}, $k \neq j$. One has the predictor

$$\tilde{X}_{ij} = \alpha_j + \sum_{k \neq j} \beta_{jk} X_{ik},$$

where

$$\alpha_j = \mu_j - \sum_{k \neq j} \beta_{jk}\mu_k$$

and

$$\sum_{m \neq j} \gamma_{km} b_{jm} = \gamma_{jk}.$$

The error $\epsilon_{ij} = X_{ij} - \tilde{X}_{ij}$ then has a mean 0 and a variance

$$\tau_j^2 = \gamma_{jj} - \sum_{k \neq j} \beta_{jk}\gamma_{jk} > 0.$$

In the case of $q = 2$,

$$\beta_{12} = \gamma_{12}/\gamma_{22},$$

$$\epsilon_{i1} = (X_{i1} - \mu_1) - \beta_{12}(X_{i2} - \mu_2),$$

$$\beta_{21} = \gamma_{21}/\gamma_{11} = \gamma_{12}/\gamma_{11},$$

$$\epsilon_{i1} = (X_{i2} - \mu_2) - \beta_{21}(X_{i1} - \mu_1),$$

$$\tau_1^2 = \gamma_{11} - \beta_{12}\gamma_{12} = \gamma_{11}(1 - \rho_{12}^2),$$

$$\tau_2^2 = \gamma_{22} - \beta_{21}\gamma_{21} = \gamma_{22}(1 - \rho_{12}^2),$$

and the correlation coefficient of ϵ_{i1} and ϵ_{i2} is easily seen to be

$$\frac{\gamma_{12} - \beta_{12}\gamma_{22} - \beta_{21}\gamma_{11} + \beta_{12}\beta_{21}\gamma_{12}}{(\tau_1\tau_2)^{1/2}} = -\rho_{12}.$$

In like manner, for $q > 2$, let j be a positive integer not greater than q and let M be a nonempty subset of positive integers not greater than q such that j is not a member of M. Let $\epsilon_{ij|M}$ be the error from a linear regression of the section score X_{ij} on the section scores X_{im}, m in M. Under the assumption that $\mathbf{\Gamma}$ is positive definite, $\epsilon_{ij|M}$ has a positive variance $\gamma_{jj|M}$. If k is a positive integer not greater than q, if $j \neq k$, and if k is not in M, then one may also consider the error $\epsilon_{ik|M}$ from a linear regression of the section score X_{ik} on the section scores X_{im}, m in M. The variance $\gamma_{kk|M}$ of ϵ_{ik} is positive. The partial covariance of X_{ij} and X_{ik} given X_{im}, m in M, is $\gamma_{jk|M}$, the covariance of ϵ_{ij} and ϵ_{ik}, and the partial correlation of X_{ij} and X_{ik} given X_{im}, m in M, is the correlation

$$\rho_{jk|M} = \frac{\gamma_{jk|M}}{(\gamma_{jj|M}\gamma_{kk|M})^{1/2}}$$

of $\epsilon_{ij|M}$ and $\epsilon_{ik|M}$. Of special interest is the special case of M equal to the set of positive integers $m \leq q$ such that m is neither j nor k. In this case,

$$\beta_{jk} = \gamma_{jk|M}/\gamma_{kk|M}$$

is the partial regression of X_{ij} on X_{ik} given X_{im}, m in M,

$$\epsilon_{ij} = \epsilon_{ij|M} - \beta_{jk}\epsilon_{ik|M},$$

$$\beta_{kj} = \gamma_{jk|M}/\gamma_{jj|M}$$

is the partial regression of X_{ik} on X_{ij} given X_{im}, m in M,

$$\epsilon_{ik} = \epsilon_{ik|M} - \beta_{kj}\epsilon_{ij|M},$$

$$\tau_j^2 = \gamma_{jj|M} - \beta_{jk}\gamma_{jk|M} = \gamma_{jj|M}(1 - \rho_{jk|M}^2),$$

$$\tau_k^2 = \gamma_{kk|M} - \beta_{kj}\gamma_{kj|M} = \gamma_{kk|M}(1 - \rho_{jk|M}^2),$$

and the correlation coefficient of ϵ_{ij} and ϵ_{ik} is $-\rho_{jk|M}$ (Lord & Novick, 1968, pp. 264–269). For instance, if $q = 4$, $j = 1$, and $k = 3$, then $M = \{2, 4\}$ and $\rho_{jk|M}$ is the partial correlation of X_{i1} and X_{i3} given X_{i2} and X_{i4}.

The standardized error $d_{ij} = \epsilon_{ij}/\tau_j$ has a mean of 0 and a variance of 1. If the vectors $\mathbf{X}_i$ have multivariate normal distributions, then the errors ϵ_{ij}, $1 \le j \le q$, have a joint multivariate normal distribution. For each j, the mean of ϵ_{ij} is 0 and the variance is τ_j^2, and ϵ_{ij} is independent of X_{ik}, $k \ne j$. It follows that the d_{ij}, $1 \le j \le q$, have a joint multivariate normal distribution with zero mean. For each j, d_{ij} has variance 1, so that d_{ij} has a standard normal distribution. The covariance and correlation ρ_{jkd} of d_{ij} and d_{ik} are $-\rho_{jk}$ for $q = 2$ and are $-\rho_{jk|M}$ for $q > 2$ and M the set of positive integers not greater than q and not equal to j or k. Values of d_{ij} that are unusually large for a standard normal random variable thus suggest some deviation from the assumption of multivariate normality of the vector $\mathbf{X}_i$ of section scores. The source of the deviation is not generally evident, but the unusually large standardized error suggests that investigation is in order.

In practice, the μ_j and γ_{jk} are unknown, and they must be estimated by use of the sample statistics $\bar{X}_j$ and C_{jk}. In a standard linear regression analysis, for each section j, the score X_{ij} of examinee i, $1 \le i \le n$, is predicted by the remaining section scores X_{ik}, $k \ne j$, by use of the least-squares prediction

$$\hat{X}_{ij} = a_j + \sum_{k \ne j} b_{jk} X_{ik}.$$

Here a_j and b_{jk} are selected so that the sum of squares $S_j = \sum_{i=1}^{n} (X_{ij} - \hat{X}_{ij})^2$ is minimized. Thus

$$a_j = \bar{X}_j - \sum_{k \ne j} b_{jk} \bar{X}_k$$

and

$$\sum_{m \ne j} C_{km} b_{jm} = C_{jk}$$

for $k \ne j$. In addition,

$$S_j/(n-1) = C_{jj} - \sum_{k \ne j} b_{jk} C_{jk}.$$

The raw residuals $e_{ij} = X_{ij} - \hat{X}_{ij}$ can be used to find unusual differences between observed section scores X_{ij} and predicted section scores $\hat{X}_{ij}$. As the sample size n becomes large, analysis can exploit the well-known limit results that $\bar{X}_j$ converges with probability 1 to μ_j and C_{jk} converges with probability 1 to γ_{jk}, so that a_j converges with probability 1 to α_j and b_{jk} converges with probability 1 to β_{jk}. The mean squared error $s_j^2 = S_j/(n-q-1)$ then converges with probability 1 to τ_j^2, so that, for the standardized residual $u_{ij} = e_{ij}/s_j$, the difference $u_{ij} - d_{ij}$

converges with probability 1 to 0 for each fixed examinee i. It follows that u_{ij} then has an approximate standard normal distribution, so that large values of u_{ij} can suggest deviations from multivariate normality.

With a bit more effort, a variation on the standardized residual u_{ij}, the externally studentized residual r_{ij} (Draper & Smith, 1998, p. 208), can be used, which has the property that r_{ij} has a Student t distribution on $n-q-1$ degrees of freedom if the multivariate normality assumption holds. It remains true that $r_{ij}-d_{ij}$ converges to 0 with probability 1 as the sample size n becomes large; however, the exact distributional result may be helpful in samples of modest size, and calculation of r_{ij} is quite straightforward with standard software.

The definition of the externally studentized residuals involves estimation of the error in prediction of X_{ij} that results from use of regression coefficients computed by use of data from all observed examinees except for examinee i. Let δ_{if} be 1 for $i=f$ and 0 otherwise. Consider minimization of the sum of squares

$$S_{j(i)} = \sum_{f=1}^{n}(X_{fj}-\hat{X}_{fj(i)})^2,$$

where

$$\hat{X}_{fj(i)} = a_{j(i)} + \sum_{k\neq j} b_{jk(i)}X_{fk} + v_{ij}\delta_{fi}.$$

Because δ_{fi} is only nonzero for $f=i$ and the equation $X_{fj}=\hat{X}_{fj(i)}$ is achieved for the deleted residual

$$v_{ij} = X_{ij} - a_{j(i)} - \sum_{k\neq j} b_{jk(i)}X_{ik},$$

it follows that $S_{j(i)}$ is the residual sum of squares from a regression of the score total X_{fj} on the score totals X_{fk}, $k\neq j$, for examinees $f\neq i$. In addition, v_{ij} is the error in prediction of X_{ij} by X_{ik}, $k\neq j$, that results from use of the regression based on all examinees $f\neq i$. If h_{ijk}, $1\leq k\leq q$, $1\leq j\leq q$, $k\neq j$, satisfies

$$\sum_{m\neq j} C_{km}h_{ijk} = X_{ik}-\bar{X}_k$$

for $k\neq j$ and

$$g_{ij} = 1-n^{-1}-\sum_{k\neq j} h_{ijk}(X_{ik}-\bar{X}_k) > 0,$$

then $v_{ij}=e_{ij}/g_{ij}$ and

$$S_{j(i)} = S_j - v_{ij}^2 g_{ij} = S_j - e_{ij}^2/g_{ij}$$

(Draper & Smith, 1998, p. 207). Under the multivariate normality assumption, $g_{ij} > 0$ with probability 1, v_{ij} has mean 0 and variance $\tau^2 g_{ij}$, the estimate $s^2_{j(i)} = S^2_{j(i)}/(n - q - 2)$ of τ^2_j is independent of v_{ij}, and the externally studentized residual $r_{ij} = v_{ij}/[s^2_{j(i)} g_{ij}]^{1/2}$ has a Student t distribution with $n - q - 2$ degrees of freedom. In typical applications, n is so much larger than q that the t distribution is very close to a standard normal distribution.

In quality assurance, externally studentized residuals provide a guide to examinees who merit investigation and a guide to assessment results that warrant investigation. For example, at an individual level, consider an inspection scheme in which an examinee's responses are examined for possible processing errors if $|r_{ij}| > 4$. The examination might involve a hand examination of the original answer sheet, an image of the answer sheet, or a full list of responses stored in a database. The examiner would seek to detect possible scanning errors, accidental omission of all or part of a section, or errors in gridding. Especially in cases in which the number of examinees inspected is close to the number expected under multivariate normality, it is quite likely that most or all inspections will not indicate anything noteworthy. Some examinees will necessarily score much higher or lower on a section than suggested by performance on other sections. Not much can really be done about an examinee who omits an entire section by accident, although a notification that this situation was observed might be more useful to the examinee than a diagnostic score report.

Under the multivariate normal model, for a particular examinee i and section j, $|r_{ij}| > 4$ with probability equal to the probability $P(|T_{n-q-2}| > 4)$ that $|T_{n-q-2}| > 4$, where T_{n-q-2} has a t distribution on $n - q - 2$ degrees of freedom. As $n - q$ approaches ∞, $P(|T_{n-q-2}| > 4)$ decreases to the limit $P(|Z| > 4)$, where Z is a standard normal random variable. If Φ is the standard normal distribution function,

$$P(|Z| > 4) = 2[1 - \Phi(4)] = 0.0000633.$$

The approach to the limit is not unusually rapid. For instance, for $n - q - 2 = 1,000$, the probability is 0.0000680. For $q = 4$ and $n - q - 2 = 1,000$, the Bonferroni inequality (Feller, 1968, p. 110) bounds the probability of inspection by $0.0000680q = 0.000272$. Thus inspection may be relatively infrequent in favorable cases.

A summary of residual results for the complete cohort of examinees may also be of interest. For example, one might examine the number F of examinees i, $1 \leq i \leq n$, for whom $|r_{ij}| > 4$ for some j from 1 to q. The fraction $p = F/n$ of examinees with some externally studentized residual of magnitude at least 4 might be studied. If the administration size n is large relative to q, then p may be regarded as an estimate of the probability π that $m_i = \max_{1 \le j \le q} |d_{ij}| > 4$. In reality, exact multivariate normality is not expected, so that π is likely to differ from the value based on a multivariate normality assumption, but examination of p may provide a reasonable method to screen the examination data for unusual behavior associated with groups of examinees. For example, a problem with keys, scanners, or form codes is likely to affect a large number of examinees. Thus a large p can suggest a more thorough study of examination results.

1.2 Control Limits

Examination of fluctuations of p from administration to administration may be employed to indicate that further investigation of the results of a particular test administration is warranted. In principle, such examination is a standard problem in statistical process control (Burr, 1979; Montgomery, 2004); however, the problem in practice is often complicated by the limited number of administrations for which data are available and by the fact that unusually large or small residuals normally appear with low probability. Three basic options can be considered. The options depend on the assumptions made concerning what is regarded as the normal situation. To discuss the available options, consider a sequence of administrations $k \ge 1$. For administration k, let π_k be the probability that, for an examinee i from the population from which the administration is drawn, some standardized section error exceeds 4 in magnitude, let n_k be the number of examinees observed in administration k, and let p_k be the fraction of examinees in administration k with some externally studentized residual, which exceeds 4. Here the externally studentized residuals in the definition of p_k are computed based only on data from administration k. If the sample size n_k is sufficiently large, then the distribution of $n_k p_k$ is very close to a binomial distribution with sample size n_k and probability π_k. Under the multivariate normality assumption, the Bonferroni inequality implies that π_k cannot exceed $\pi^* = 2q[1 - \Phi(4)] = 0.0000633q$. One may then obtain the simple p-chart upper control limit on p_k that

$$p_k \le \min\{1, \pi^* + 3[\pi^*(1 - \pi^*)/n_k]^{1/2}\}.$$

Provided that $n_k \pi^*$ is large enough for a normal approximation to be reasonable, the probability that p_k is not within its control limit is about $1 - \Phi(3) = 0.00135$, the probability that a standard normal random variable does not exceed 3. If one is concerned about the normal approximation, then the binomial approximation may be used. The upper control limit is then L_k/n_k, where L_k is the smallest integer such that either $L_k = n_k$ or $1 - \Phi(3)$ is no greater than the probability that L_k is less than a binomial random variable with sample size n_k and probability π^*. The probability is then no greater than $1 - \Phi(3)$ that p_k is not within its control limit. Whether or not a normal or binomial approximation is used, a lower control limit is not likely to be of much interest given that π^* is only an upper bound. In addition, attempts to find a lower control limit based on π^* are likely to lead to a trivial limit of 0 in many cases.

If one relaxes the multivariate normality assumption but assumes that the π_k are constant for different administrations k, then an alternative approach may be used based on estimation of a common π_k. Let data from administrations 1 to $K \geq 1$ be used for examination of administration $K + 2$. Suppose that both unusual individual values of p_k and large changes in $p_k - p_{k-1}$ are to be explored. Let

$$N_K = \sum_{k=1}^{K} n_k,$$

and let

$$\hat{p}_K = N_K^{-1} \sum_{k=1}^{K} n_k p_k$$

be the estimate of the common π_k based on the first K administrations. Then the estimated standard deviation of $p_{K+2} - \hat{p}_K$ based on the first K administrations and the sample size n_{K+2} is

$$s_{pK} = [\hat{p}_K(1 - \hat{p}_K)(N_K^{-1} + n_{K+2}^{-1})]^{1/2},$$

so that p_{K+2} has control limits $\hat{p}_K - 3s_{pK}$ and $\hat{p}_K + 3s_{pK}$. Similarly, the estimated standard deviation of $p_{K+2} - p_{K+1}$ based on the first K administrations and the sample sizes n_{K+1} and n_{K+2} is

$$s_{pKd} = [\hat{p}_K(1 - \hat{p}_K)(n_{K+1}^{-1} + n_{K+2}^{-1})]^{1/2},$$

so that $p_{K+2} - p_{K+1}$ has control limits $-3s_{pKd}$ and $3s_{pKd}$. If n_{K+1}, n_{K+2}, and N_K are all sufficiently large for normal approximations to apply, then p_{K+2} is within its control limits with approximate probability $2[1 - \Phi(3)] = 0.00270$, and $p_{K+2} - p_{K+1}$ is within its control limits with the same approximate probability.

Because typical variations in the distribution of $\mathbf{X}_i$ for different administrations may result in appreciable variation in π_k for different administrations, consider the following variation on the traditional XmR chart that uses plots over time of both individual measurements and changes between successive individual measurements. Assume data have been gathered to date from administrations 1 to $K > 2$ and control limits are to be placed on administration $K + 2$. The π_k are regarded as random variables, and the p_k for $k \geq 1$ are assumed to behave as a white-noise time series, so that the p_k are independent normal random variables with positive common mean ν and positive common variance υ. Let

$$\bar{p}_K = K^{-1} \sum_{k=1}^{K} p_k$$

be the sample mean of the proportions p_k for k from 1 to K, let

$$s_K^2 = (K-1)^{-1} \sum_{k=1}^{K} (p_k - \bar{p}_K)^2$$

be the sample variance of the p_k, $1 \leq k \leq K$, and let s_K, the square root of s_K^2 be the sample standard deviation of the p_k, $1 \leq k \leq K$. Consider construction of control limits for p_{K+2}. Let t_{K-1} be selected so that the probability that a random variable with a t distribution on $K-1$ degrees of freedom is less than t_{K-1} is 0.00135, the probability that a standard normal distribution is less than 3. Because $p_{K+2} - \bar{p}_K$ has variance $(K+1)\upsilon/K$ and $p_{K+2} - p_{K+1}$ has variance 2υ, the control limits for p_{K+2} are between $\bar{p}_K - [(K+1)/K]^{1/2} t_{K-1} s_K$ and $\bar{p}_K + [(K+1)/K]^{1/2} t_{K-1} s_K$. In like fashion, control limits for $p_{K+2} - p_{K+1}$ are between $-2^{1/2} t_{K-1} s_K$ and $2^{1/2} t_{K-1} s_K$. Under the white-noise model, the probability is $2[1 - \Phi(3)] = 0.00270$ that p_{K+2} is within its control limit, and the probability is also $2[1 - \Phi(3)] = 0.00270$ that $p_{K+2} - p_{K+1}$ is within its control limit. For K large, the outlined procedure approaches the customary results from an XmR chart.

1.3 Residuals for the Total Score

The approach using externally studentized residuals can also be applied to the total score Y_i by itself; however, the analysis is much simpler. One considers prediction of Y_i by a constant μ. The optimal choice of μ for best linear prediction has μ equal to the expectation of Y_i. The prediction error $\epsilon = Y_i - \mu$, and the standardized error is $d_i = \epsilon_i/\sigma$, where σ is the standard deviation of Y_i. The mean-squared error of prediction of Y_i is then the variance σ^2 of Y_i. Let $\bar{Y} = n^{-1} \sum_{i=1}^{n} Y_i$ be the sample mean of the Y_i. The estimated prediction $\hat{Y}_i$ of Y_i is $\bar{Y}$ for each examinee i. The residual mean-squared error is the sample variance

$$s^2 = (n-1)^{-1} \sum_{i=1}^{n} (Y_i - \bar{Y})^2.$$

For examinee i, the residual $e_i = Y_i - \bar{Y}$, the standardized residual is $u_i = e_i/s$, the deleted residual is $v_i = ne_i/(n-1)$, and the externally studentized residual is then

$$r_i = \frac{[n/(n-1)]^{1/2}e_i}{\{(n-2)^{-1}[(n-1)s^2 - e_i^2 n/(n-1)]\}^{1/2}}.$$

The multivariate normality assumption for the section scores implies that r_i has a t distribution with $n-2$ degrees of freedom. For individuals, unusually large or small values of r_i may suggest problems with form codes, gridding, or scanning, so that detailed examination of the examinee record can be warranted. An unusual fraction p' of examinees with $|r_i| > 4$ may suggest more general problems with the administration. The procedures for control limits for p are readily changed for control limits for p'. The main change is that the multivariate normality case implies that p' has expectation approximately equal to $2[1 - \Phi(4)]$.

1.4 Low Scores

A second screening method may also be helpful, especially in the case of errors in recording form codes or using answer keys. One can determine the expected total score μ_0 obtained by an examinee who answers all items randomly. For example, in an examination with 50 items that is right-scored and has five alternatives for each item, the expected total score for a random responder is 10. One might consider examination of all answer sheets for examinees with an observed score less than G. In cases in which this number is large, a random sample of such answer sheets might be considered. In addition to verification of form codes, one might again check by hand for scanning errors, A simple screen is to record the number G of examinees i, $1 \le i \le n$, with Y_i no greater than μ_0. Again examination of fluctuations of the fraction $g = G/n$ from administration to administration can be used to indicate that further investigation is needed. The control limits previously described can be applied to this case as well, although the multivariate normal case would not normally be considered.

2 An Application

A multiple-choice right-scored examination produced by ETS was examined. In this assessment, 120 items are divided into $q = 4$ sections, sections 1 to 4. For examinee i, X_{ij} is the number of correct responses in section j, and $Y_i = X_{i1} + X_{i2} + X_{i3} + X_{i4}$ is the total raw score. Each item is multiple-choice, and four choices are used in each case. The test is right-scored,

Table 1
Relative Frequency of Examinees With Large Externally Studentized Subscore Residuals in Administrations of Examination

Administration	No. examinees	Frequency	Relative frequency
1	6,432	30	0.0047
2	9,087	33	0.0036
3	6,409	25	0.0039
4	9,073	31	0.0034

Table 2
Relative Frequency of Examinees With Very Low Total Scores in Administrations of Examination

Administration	No. examinees	Frequency	Relative frequency
1	6,432	2	0.0003
2	9,087	7	0.0008
3	6,409	3	0.0005
4	9,073	2	0.0002

so that random generation of responses leads to an expected total score of 30. To obtain some insight into customary variability, four administrations of the same form were analyzed separately. No access to the original answer sheets was available, so that some limitations on explanation of results necessarily exist. Note that the order of the examinations in the example is not necessarily the actual temporal order, so that the control limits here sometimes may differ from those that would be constructed in practice.

Table 1 summarizes results for the regression analyses, while Table 2 summarizes results for low total scores. Administrations are numbered rather than listed by date to prevent disclosure of dates on which the same form was used. For these administrations, for the observed sample means and sample variances of the total scores, no externally studentized residual for a total score can exceed 4; however, it is possible for such a residual to be less than -4. Nonetheless, no instance of an externally studentized residual less than -4 was observed.

The administrations do not differ markedly in terms of examinees with externally studentized residuals for section scores that are exceptionally small or large or in terms of examinees with very low scores. In the case of section scores, a Pearson chi-square test that the π_k are constant for all administrations yields a chi-square of 1.68 on three degrees of freedom. Given the large sample sizes, the problem of dependence of residuals within an administration can be assumed to be negligible, so that this Pearson chi-square indicates no obvious variability. The case of low scores involves rather low expected values under the null hypothesis of a constant rate, but the chi-square value of 3.40 on three degrees of freedom does not indicate any obvious variability by administration.

For each administration, the rate of externally studentized subscore residuals with magnitude greater than 4 is much higher than expected under normality, for π^* is $0.000253 = 4(0.0000633)$, so that the largest control limit under the multivariate normality assumption is

$$0.000253 + 3[0.000253(1 - 0.000253)/6409]^{1/2} = 0.000850.$$

Under the binomial approximation, the control limits range from 0.000440 to 0.000468, so the choice of approximations does affect the control limits. Nonetheless, the persistent failure of the multivariate normal model remains evident.

The control limits for constant π_k are much less readily violated by the externally studentized subscore residuals. For a simple example, consider the case of $K = 2$. With this case, the bounds for p_4 are 0.00154 and 0.00867. The bounds for $p_4 - p_3$ are -0.00311 and 0.00311. These bounds are readily satisfied by the observed data, so that no suggestion of a fundamental change exists. For this example, use of the bounds that do no assume constant π_k is impractical due to the very large value of t_{K-1}. For $K = 2$, $t_{K-1} = 235.801$ and $[(K+1)/K]^{1/2}t_{K-1} = 288.797$. The approach is much more reasonable for somewhat larger K. For example, for $K = 11$, $t_{K-1} = 3.957$ and $[(K+1)/K]^{1/2}t_{K-1} = 4.150$. For $K = 21$, $t_{K-1} = 3.422$ and $[(K+1)/K]^{1/2}t_{K-1} = 3.507$. Note that for quite large K, $[(K+1)/K]^{1/2}t_{K-1}$ is close to 3.

In the case of externally studentized residuals based on total score, control limits based on π^* are obviously not violated, for no externally studentized residual is unusually large in magnitude. In the case of the fraction of total scores that are not greater than the expected total score with random response, consideration of a constant probability of a very low score for each administration is complicated by the very small but positive frequency counts observed, so that more data would be desirable before a normal approximation was used. Again the data do not suffice for the case of variable probabilities for different administrations.

Both very low scores and very large externally studentized section residuals are very strongly associated with omitted responses. Generally, examinees on this right-scored test answer all items.

Table 3
Relative Frequency of Examinees With Large Externally Studentized Column Residuals in Administrations of Examination

Administration	No. examinees	Frequency	Relative frequency
1	6,432	25	0.0039
2	9,087	29	0.0032
3	6,409	25	0.0039
4	9,073	25	0.0028

Among all 31,001 examinees, 27,876 examinees, 90.0% of the total, answered all items. Among these examinees, 31, or 0.111 per cent, had patterns of responses that resulted in unusually large externally studentized residuals for section scores, and 4, or 0.014%, had no more correct answers than would be expected with random response. Of the 3,125 examinees with some missing response, 88, or 2.816%, had unusual section residuals, and 10, or 0.320%, had very low total scores. Thus 73.9% of all examinees with large section residuals and 71.4% of all examinees with very low scores had missing responses. Results are even more striking when 30 or more responses were omitted. For the four administrations, 44 examinees omitted 30 or more responses. Among this group, 25 examinees, or 56.8%, had unusually large externally studentized residuals and 7, or 15.9%, had very low total scores. Thus half of examinees with very low scores and 26.1% of examinees with unusually large section residuals had at least 30 omissions. Among the 44 examinees with 30 or more omitted responses were 27 who omitted the entire final section.

A parallel analysis divided the 120 responses in three parts of 40 items each to reflect the use of an answer sheet in which each column contained 40 items. Results are shown in Table 3. This analysis does not appear to differ much from the analysis in Table 1. The reader should remember that some reduction in examinees with unusually large residuals is associated with a reduction of q from 4 to 3. For example, π^* is now 0.000190 rather than 0.000253.

3 Conclusions

Because retrieval of original papers was not feasible, it was not possible to determine if unusual scores represented any problem in data processing. As described in section 1, effective use of outlier checks will require data collection for a significant number of administrations of a test title. Given these data, a quality monitoring procedure can be established to check new administrations to ascertain if the rate of major outliers or very low scores is unusually high relative to historical records. If a rate is sufficiently notable, then an examination of the administration is needed to determine if a significant problem related to data collection, data processing, or test administration has been found.

References

Burr, I. W. (1979). *Elementary statistical quality control.* Milwaukee: ASQC Quality Press.

Draper, N. R., & Smith, H. (1998). *Applied regression analysis* (3rd ed.). New York: John Wiley.

Feller, W. (1968). *An introduction to probability theory and its applications* (3rd ed., Vol. 1). New York: Wiley.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Montgomery, D. C. (2004). *Introduction to statistical quality control* (5th ed.). New York: John Wiley.