STATISTICS IN WORK QUALITY MEASUREMENT: LICENSING ADULT AND CHILD CARE FACILITIES

Richard Bolstein and John D. R. Cole, George Mason University
Richard Bolstein, 6118 Mountain Springs Lane, Clifton, Va. 22024

INTRODUCTION

We study the problem of measuring the intrinsic quality of work in an office environment when performance cannot simply be judged by output or productivity. Typically, one would devise a set of criteria to be used to evaluate performance by peer review. In some cases the criteria are such that a rating of excellent to poor (say, 1 to 5) would be assigned to each item (as in 'teaching evaluations', for example). In other cases, the criteria are in the form of questions which call for a 'yes', 'no', or 'not-applicable' response. This latter situation arose in the authors' work on designing a quality review system for licensing professionals in the Virginia Department of Social Services. These professionals are responsible for overseeing private child and adult residential and day-care centers in the State, including the issuance, supervision and revocation of licenses. Since there are many legal and policy requirements that should be followed, the 'yes-no' checklist lends itself very well to the problem of measuring work-quality in this situation.

After construction of the final 'yes-no-NA' checklist, the next question is how to use it to devise a numerical measure of work quality. In this paper the authors consider several measures, including the ratio of 'yes' responses to 'yes + no' responses, the ratio of 'no' responses to the total number of checklist items, and weighted versions of these. The first method is a ratio estimate and hence is technically more complicated than the latter. However, the latter has the possibly undesireable effect of not distinguishing between 'yes' and 'NA'. Another question that arises is how to combine quality measures of different operations (with different checklists) into a single measure of quality for that office. In addition, if it is desireable to compare quality among different offices both for individual operations and overall, a stratified sample design would be appropriate.

This is the situation in the application to licensing professionals presented in this paper. The problem is complicated by the fact that the number of strata is large and the sample size is constrained so only two cases are reviewed in many strata (the maximum is 15). Thus, either jackknife or bootstrap methods are needed for a good estimate of standard error. It turns out that the jackknife is of no help for within-stratum estimates and is only marginally useful for combined ratio estimates.

BACKGROUND

Organizational Setting

The Virginia Department of Social Services is responsible for assuring that social service program policies are carried out efficiently and effectively, in accordance with federal and state legal and regulatory requirements. One such program concerns the licensing and oversight of various privately owned and managed residential and day care facilities. The department regulates the operations of such facilities through its Division of Licensing Programs, whose activities entail issuing licenses to and exercising continuing oversight of providers to assure that their operations satisfy legal and regulatory requirements. The division's activities are carried out through a central office staff in the State capitol and a staff of licensing specialists in seven regional offices geographically dispersed throughout the Commonwealth.

In 1983, as part of a departmentwide project to develop and implement a system for measuring and monitoring the performance of its regional offices, the Division of Licensing Programs undertook to develop a means for assessing the efficiency and effectiveness of its regional licensing units. Efficiency measures focused on staff productivity and cost efficiency. Effectiveness measurement concerned itself with the quality of regional activities, from two perspectives: (1) an assessment of the intrinsic quality of regional staff efforts in terms of compliance with regulatory and administrative policies, by means of an internal "Quality Review" of actions taken; and (2) an evaluation of the perceived quality of regional activities by licensed providers, using survey research methods. This paper is concerned with the intrinsic Quality Review Process.

Management's interest in undertaking the "Quality Review" initiative was to establish an ongoing process for evaluating and comparing systematically the quality of regional office operations. The principal objective was to develop an asessment tool that would aid program managers both in evaluating various aggregate aspects of program quality and, when used in conjunction with other licensing performance measurement data, would better inform management decisions concerning future plans and program improvements.

The qualitative feedback mananagement hoped to gain from this process was intended to answer such overall questions as:

-- What is the overall quality of licensing operations statewide?

-- Does the quality of regional licensing operations differ among the major types of licensing activities?

-- To what extent do the various regions' overall licensing operations differ qualitatively over time and among regions?

It was management's desire to: (1) minimize the investment of staff time needed to conduct an annual review of regional case actions; and (2) optimize the reliability and validity of the review process as a basis for providing worthwhile overall assessments of program quality for management purposes. ([5, Part II].)

Focus on Outputs

A major focus of the department's performance measurement system is on eight identifiable program outputs that are the result of regional staff efforts. For the Division of Licensing Programs these include the end-product outputs, or completed case actions, resulting from the principal regional activities which are susceptible to post-hoc management review. These included: issuances (the issuance of a license to an approved facility); closures (the closing of a licensed facility due to revocation of its license); complaints

(the informal investigation of a client or citizens complaint concerning a licensed facility); allowable variances (the granting to a facility of a justifiable variance from a licensing requirement); early compliance (the conversion of a conditional license to regular license ahead of schedule, due to expedited compliance action by the licensee); supervisory visits (onsite visits by regional licensing staff for oversight and/or technical assistance purposes); allegations (the formal investigation of an alleged violation of legal or regulatory requirements by a licensed or unlicensed provider); and modifications (a change in the terms of a license due to changes in licensee operations). The regional licensing output levels for fiscal year 1986 are shown in Tables 1-3 by output type and region.

Each of the foregoing outputs entails certain work activities on the part of regional licensing staffs that are governed by State laws, regulations and/or departmental policies and procedures. Evaluating the quality of these eight end-products is best judged in terms of the extent to which the completed output and the activities which produced it satisfies applicable legal, regulatory, policy and procedural requirements. Thus, the design and development of the licensing Quality Review process focused on how best to assess these aspects of regional program performance.

QUALITY REVIEW PROCESS

An internal Monitoring and Evaluation Committee comprised of central office and regional staff specialists developed a series of eight Quality Review Checklists corresponding with the regional outputs enumerated above. Each checklist included items reflecting important legal, regulatory, policy and procedural requirements concerning which regional compliance could be reviewed on a post-hoc basis. Each checklist contained relevant items that addressed four specific aspects of compliance:

- o Factfinding
- o Procedural Compliance
- o Appropriateness of Disposition
- o Timeliness

The checklists were pilot tested by a group of licensing supervisors and specialists. Revised check-lists were prepared based on the pilot test results, and potential reviewers were trained on their use prior to the actual Quality Review process being instituted. An example of one checklist is given in the Appendix.

SAMPLING PLAN

The aggregate of 5,395 completed cases actions for 1986 (Table 1) served as the sample frame and was partitioned by region, output type, and facility type into 112 strata. Simple random samples, roughly proportional to size, were taken in each stratum. Since each case required an average 30 to 40 minutes to review and staff time was limited, total sample size was restricted to 310 cases. Many strata had a sample size of two (and some only one), making variance estimation difficult. Once the sample was selected, the case outputs were reviewed by ten central office and regional licensing specialists. To minimize rater bias, no team member reviewed cases from his/her own region. To check the extent of measurement error by reviewers, a subsample of 50 cases was selected for quality control purposes. These were re-reviewed by a team member other than the original reviewer, and the two results were compared as a means of checking the objectivity of the checklists. Sequential sampling was used in this process. Although inconsistencies in the reviews suggest the presence of measurement error, we do not take this into account in the rest of this paper.

QUALITY MEASURES

Focus now on a single stratum. The checklist is the same for all cases in the stratum but the number of applicable items varies by case. For each (population) case j, let y_j denote the number of checklist items that rate a 'yes' response (i.e., if rated it would earn a 'yes') and x_j the number that rate 'no'. The quality of case j could be measured as $y_j/(y_j+x_j)$. However, the problem is to measure the quality of the stratum as a whole. This can be done This can be done either by the average of ratios

$$(1/N)\sum_{j=1}^{N} y_j/(y_j + x_j)$$

or by the combined ratio

$$R = \sum_{j=1}^{N} y_j /(y_j + x_j) \qquad (1)$$

which represents the proportion of applicable items that would rate 'yes'. The problem with the average of ratios method is that it gives equal weight to all cases. For example, a case with 10 items rated 'yes' and zero items rated 'no' would count the same as a case with one item rated 'no' and zero 'yes'. The quality measures for these cases are 1 and 0 with an average of 0.5, even though ten of eleven applicable items (those answered yes or no) were rated 'yes'. For this reason we adopt the ratio (1) as the overall measure for the stratum. The sample estimate of R, denoted by $\hat{R}$, is of the same functional form except that the sums are over the n cases in the sample. Thus, $\hat{R}$ is a ratio estimate, and the elementary units are the checklist items. The sampling process is actually a stratified cluster sample with the cases serving as clusters of elementary units.

The ratio estimator is biased, and since within-stratum sample sizes are small, the bias is possibly not negligible and the usual approximate formula for the variance of R is probably poor (Cochran [1, p.155]). We employ the jackknife (Wolter [4,Ch.4]) method in an attempt to improve the variance estimates and reduce bias. Alternative unbiased measures of quality can be obtained by keeping the denominator constant. If M denotes the number of checklist items for the output type under consider ation, then (x_j/M) measures the proportion of items that rate 'no' and the stratum measure would be the sum of the x_j divided by NM, which is the same as the mean of the individual ratios. The drawback with this measure is that it does not distinguish between 'yes' and 'non-applicable' items. If the emphasis is on the number of things done incorrectly, this may be an acceptable measure, but for most management purposes (1) is preferred.

To distinguish between 'yes', 'no', and 'NA' and still maintain a linear measure that permits an unbiased estimator, one could assign values +1, -1, and

0 respectively (as in a true-false exam where a student is penalized more heavily for a wrong answer than an omission). Let u_{ij} denote the value +1, -1 , or 0 assigned to item i of case j, and let $u_{\cdot j}$ denote the sum of these over all items i. The 'score' for this case is $u_{\cdot j}/M$ and the stratum measure is

$$N^{-1}\sum_{j=1}^{N} u_{\cdot j}/M$$

The measure ranges from -1 to +1. The primary drawback here is that there is a penalty for an 'NA' item. For example, if M = 10 a case with 5 'yes' and 5 'NA' items would receive a score of only .5, whereas a case with 10 'yes' items would receive a score of 1. The combined stratum score (if N=2) would be .75 even though there were no mistakes.

WEIGHTED MEASURES

In summary, the ratio estimator (1) appears to be the most useful. However, a glance at the checklist in the Appendix will convince the reader that not all items should be equally weighted. For example, the disposition of the case (item 2.6) is clearly the most important item. To take this into account, we generalize (1) by assigning positive values to each item on the checklist. As there are eight different checklists, one for each output type, let M_t denote the number of items on checklist t,

v_{it} = value of item i on checklist t, i=1,2,...,M_t.

and set

$$v_t = \sum_i v_{it} , \quad \text{for } t=1,...,8.$$

Each stratum is a triple h = (t,f,r) determined by a specific output type t, facility type f (child or adult), and regional office r. Let $N_h = N_{tfr}$ denote the number of cases in stratum h. We use a dot in place of the subscript to indicate the sum over that subscript. Lower case n plays the same role for the sample as N does for the population. For a given case j in stratum h let $y_{jh} = y_{jtfr}$ be the sum of the values corresponding to checklist items that rate 'yes', let $x_{jh} = x_{jtfr}$ be the sum corresponding to items that rate 'no', and let $R_{jh} = y_{jh}/(y_{jh} + x_{jh})$ be the score for case j. The quality measure for stratum h is the ratio

$$(2)\quad R_h = \Big(\sum_{j=1}^{N_h} y_{jh}\Big)/\Big(\sum_{j}^{N_h}(y_{jh} + x_{jh}) = y_{\cdot h}/(y_{\cdot h} + x_{\cdot h})$$

This represents the ratio of the total value of items that rate 'yes' to the total value of applicable items for the region and output-facility type. The sample estimator of (2) is the ratio of sample means

$$(E2)\quad \hat{R} = \bar{y}_h/(\bar{y}_h + \bar{x}_h) = \Big(\sum_{j=1}^{n_h} y_{jh}\Big)/\Big(\sum_{j=1}^{n_h}(y_{jh} + x_{jh})\Big)$$

The next task is to combine strata to obtain certain overall measures. There are four of primary interest (see the section on BACKGROUND).:

(i) A statewide measure of quality for each output type & facility type.

(ii) A regional measure of quality for each output type.

(iii) A regional measure of quality over all output-types.

(iv) An overall statewide measure.

As regards (i), for a given output type t and facility type f, a natural statewide measure is

$$(3)\quad R_{tf\cdot} = \Big(\sum_r y_{\cdot tfr}\Big)/\Big(\sum_r (y_{\cdot tfr} + x_{\cdot tfr})\Big)$$

This is the ratio of the total value of 'yes' items to that of applicable items for all cases in the state of output type t and facility type f. (We caution the reader that $R_{tf\cdot} \neq \sum_r R_{tfr\cdot}$.) To estimate (3) from the sample we use the combined ratio estimator

$$(E3)\quad \hat{R}_{tf\cdot} = \Big(\sum_r N_{tfr}\bar{y}_{tfr}\Big)/\Big(\sum_r N_{tfr}(\bar{y}_{tfr} + \bar{x}_{tfr})\Big)$$

where $\bar{y}_{tfr} = \bar{y}_h$ denotes the sample mean for stratum h. (Note that a separate ratio estimate is not possible since the total value of applicable items in a stratum is unknown.)

For a fixed region r and output type t, a ratio measure for (ii) is constructed analogously by combining the two facility types f = 1,2. The situation for (iii) is not as straightforward. Here, for a fixed region r and facility f, we wish to combine all types of output for a composite score. But the different output types are not equally important in the performance of an office (supervisory visits are not as important as issuances, for example). The combined ratio estimate gives the most weight to those operations with a large number of cases and a large total value v_t, i.e. output type t has a contribution to the composite roughly proportional to the product $(N_{tfr})v_t$. If w_t, t = 1,...,8, is the relative importance of output type t in the statewide overall measure, then one could choose v_t so that the system of equations below are satisfied:

$$(v_t N_{t\cdot\cdot})/(v_k N_{k\cdot\cdot}) = w_t/w_k \ , \quad \text{all } t,k=1,...8.$$

This system has rank 7 and can be solved in terms of one v_t which can then be arbitrarily set. For example, if output types are listed in the order of Table 1 and if we take w = (5,10,3,2,1,5,2,1), then, from Table 1, setting $v_8 = 100$, we obtain the solution

$$v = (v_1,...,v_8) = (46,870,60,150,8,67,155,100)$$

Thus, after initially choosing the values v_{it}, they should be normalized so that $v_t = \sum_i v_{it}$ has the prescribed value (while maintaining relative values of the v_{it} within type). Observe that the number of cases of a certain type varies considerably among regions. For example, although the statewide ratio of issuances to supervisory visits is 1427/1709 = .83 (Table 1), for region #4, child facility, the ratio is .19 (Table 2), and for region #3, child facility, it is 1.89. In the combined ratio estimate for region #3, child facility, the relative weights of issuances to supervisory visits would approximately equal

$$v_1 N_{113}/(v_5 N_{513}) = (.19)(46/8) = 1.1 \neq w_1/w_5 = 5,$$

reflecting the fact that the bulk of the work in region #5 is in supervisory visits so this should contribute more (almost as much as issuances) to the total measure of quality for the region. In summary, with a proper choice of the v_t, the composite ratio can be used to provide a meaningful overall regional quality measure:

$$(4)\quad R_{\cdot\cdot fr} = y_{\cdot\cdot fr}/(y_{\cdot\cdot fr} + x_{\cdot\cdot fr})$$

This is estimated from the sample by

(E4) $\hat{R}_{\cdot fr} = \sum_t (N_{tfr}\bar{y}_{tfr})/(\sum_t N_{tfr}(\bar{y}_{tfr} + \bar{x}_{tfr}))$

For the overall statewide measure we have

(5) $R... = y..../(y.... + x....)$

The reader may wonder why we do not simply use $\tilde{R}_{\cdot fr} = \sum_t w_t R_{tfr}$ as the overall measure for region r, facility f, instead of messing with the value totals. There are two reasons: First, to estimate this quantity requires a separate ratio estimate of poor stability since the within stratum sample sizes are so small. Secondly, there is not a unique way to define an overall statewide measure since

$$\sum_r (N_{\cdot fr}/N_{\cdot f\cdot})\tilde{R}_{\cdot fr} \neq \sum_t w_t R_{tf\cdot}.$$

VARIANCE ESTIMATION

The standard sample estimate of the variance of the simple ratio estimator (E3) is given by

(V2) $v(\hat{R}_h) = (1-n_h/N_h)[\{s_{yh}(1-\hat{R}_h)\}^2+(s_{xh}\hat{R}_h)^2 -2\mathrm{cov}(y_h,x_h)\hat{R}_h(1-\hat{R}_h)]/[n_h(\bar{y}_h+\bar{x}_h)^2]$,

where s^2 and cov denote the sample variance and covariance respectively of the subscripted variables. This follows from Cochran [1, p.155]. The approximation is generally poor for the small samples in this study and jackknife (Efron and Stein [3] or Wolter [4]) or bootstrap (Efron [2]) methods are needed. We give some results in the next section. The approximate variance of the combined ratio estimator (E3) of the statewide measure for a given output and facility type follows from Cochran [1, p.166]. The result, which is usually a good estimate of mean square error for large total sample size, is

(V3) $v(\hat{R}_{tf\cdot}) = \sum_r [N_h(N_h-n_h)/n_h][\{s_{yh}(1-\hat{R}_{tf\cdot})\}^2 +(s_{xh}\hat{R}_{tf\cdot})^2-2\mathrm{cov}(y_h,x_h)\hat{R}_{tf\cdot}(1-\hat{R}_{tf\cdot})]/[\sum_r N_h(\bar{y}_h+\bar{x}_h)]^2$

The jackknife estimators for the within stratum and combined ratios (2) and (3) are, from Wolter [4, pp. 173 and 181],

(J2) $\tilde{R}_h = n_h\hat{R}_h-(n_h-1)\hat{R}_h(\cdot)$

(JV2) $v(\tilde{R}_h) = [(1-n_h/N_h)(n_h-1)/n_h]\sum_k^{n_h}[\hat{R}_h(k)-\hat{R}_h(\cdot)]^2$

where $\hat{R}_h(\cdot)$ is the average of the n_h ratio estimates $\hat{R}_h(k)$ obtained by deleting the kth case from the sample, $k = 1, \ldots, n_h$.

(J3) $\tilde{R}_{tf\cdot} = [1 + \sum_r u_h]R_{tf\cdot} - \sum_r u_h R_{tf\cdot}(h.)$

(JV3) $v(\tilde{R}_{tf\cdot}) = \sum_r [(u_h/n_h)\sum_k^{n_h}\{\hat{R}_{tf\cdot}(hk)-\hat{R}_{tf\cdot}(h.)\}^2]$

where $h = (t,f,r)$, $u_h = (n_h-1)(1-n_h/N_h)$, and $\hat{R}_{tf\cdot}(h.)$ is the mean of the n_h combined ratio estimates $R_{tf\cdot}(hk)$, $k = 1,\ldots,n_h$, obtained by deleting case k in stratum h from the sample. Variance formulas for the usual and jackknife estimators of (4) and (5) have the same form as those for (3).

SOME NUMERICAL RESULTS

To illustrate the quality measures and variance computations previously discussed, we reproduce in Table 4 below the numerical results for the group of strata t=1, f=2, r=1,...,7 (issuances for adult facilities in each of seven regions). These results are for unweighted checklist items, i.e., $v_{it} = 1$ for all i and t. (It is not easy to get a dozen people to agree on weights.) For simplicity, we abreviate h=(1,2,r) by r, and write $N_{12r} = N_r$, $R_{12\cdot} = R$, etc.. $\hat{R}$ and $\tilde{R}$ denote the usual and jackknife estimates of R respectively. Notice that the sample coefficient of variation of y_r+x_r is at most .1 in all but one stratum. Thus, the bias may be negligible (Cochran [1, p. 178]). The jackknife estimates $\tilde{R}_r$ are virtually identical with the usual estimates, and the jackknife variance estimates offer no improvement over the usual variance estimates. Since the jackknife is known to reduce bias in most samples, this supports our contention that the bias in $\hat{R}_r$ is probably negligible.

The combined ratio estimate $\hat{R}$ of the statewide quality measure for the operation "issuance of licenses for adult facilities" computed by (E3) and Table 4 turns out to be $\hat{R} = .8700$. The standard error calculated from (V3) and Table 2 is .0185. The jackknife estimate of standard error is .0176, a small improvement.

SUMMARY AND CONCLUSIONS

It appears from Table 4 that there is a difference in quality among regions for the issuance of adult facilities. From the data, regions 5 and 7 are significantly better than regions 1 and 4 provided the variance estimates are good and bias is negligible. The jackknife estimator of variance provides only a small improvement over the usual combined ratio estimator. Current work includes the implementation of Efron's bootstrap method of variance estimation (see [2]) for stratified samples and the application of it to this data. Incidentally, the authors found no differences in quality for child versus adult facilities, so future samples need not stratify by facility. This will permit larger within-stratum sample sizes.

Measurements, i.e. case reviews, were assumed perfect in this study. An open problem is how to model measurement error and make variance estimates in this situation. Analysis of the re-review sample mentioned in the section on the Sampling Plan is a first step in this direction.

REFERENCES

[1] Cochran, W.G., **Sampling Techniques**, Wiley, 1977

[2] Efron, B., "Nonparametric estimates of standard error: the jackknife, the bootstrap, and other resampling methods", **Biometrika**, 68, 589-599, (1981).

[3] Efron, B. and Stein, C., "The jackknife estimate of variance", **Annals of Statistics**, 9, 586-596, (1981).

[4] Wolter K.M., **Introduction to Variance Estimation**, Springer-Verlag, (1985).

[5] Virginia Department of Social Services, 1986 Report on Regional Performance, Richmond, Va. 22388, (1986).

TABLE 1: Licensing Program Outputs - 1986

OUTPUTS/REGION	#1	#2	#3	#4	#5	#6	#7	TOTAL
Issuances	094	366	028	093	119	332	142	1,427
Closures	008	040	037	012	007	035	011	0150
Complaints	022	135	196	074	043	099	084	0653
Variances	021	032	035	020	023	020	024	0175
Supervisory Visits	108	300	206	324	285	333	153	1,709
Allegations	011	159	462	052	023	183	091	0981
Modifications	017	030	031	030	015	029	017	0169
Early Compliance	007	009	029	014	010	050	012	0131
TOTAL OUTPUTS	0288	1,071	1,277	0619	0525	1,081	0534	5,395

TABLE 2: Licensing Program Outputs - Child Facility

OUTPUTS/REGION	#1	#2	#3	#4	#5	#6	#7	TOTAL
Issuances	047	326	187	042	075	274	081	1,032
Closures	007	037	028	009	004	033	008	0126
Complaints	012	120	107	019	007	072	007	0344
Variances	013	030	030	013	017	016	012	0131
Supervisory Visits	027	209	099	222	127	242	094	1,020
Allegations	007	150	436	027	020	165	064	0869
Modifications	006	024	019	013	004	025	011	0102
Early Compliance	007	008	020	010	007	044	009	0105
TOTAL OUTPUTS	126	904	926	355	261	871	286	3,729

TABLE 3: Licensing Program Outputs - Adult Facility

OUTPUTS/REGION	#1	#2	#3	#4	#5	#6	#7	TOTAL
Issuances	047	040	094	051	044	058	061	395
Closures	001	003	009	003	003	002	003	024
Complaints	010	015	089	055	036	027	077	309
Variances	008	002	005	007	006	004	012	044
Supervisory Visits	081	091	107	102	158	091	059	689
Allegations	004	009	026	025	003	018	027	112
Modifications	011	006	012	017	011	004	006	067
Early Compliance	000	001	009	004	003	006	003	026
TOTAL OUTPUTS	162	167	351	264	264	210	248	1,666

TABLE 4: Analysis of Data for t=1, f=2, r=1,...,7.

r	1	2	3	4	5	6	7	total
N_r	47	40	94	51	44	58	61	395
n_r	5	6	8	2	3	4	4	32
$\sum_j y_{jr}$	127	180	244	53	102	120	128	
$\sum_j x_{jr}$	46	34	26	18	5	18	9	
$\hat{R}_r$	.734	.841	.904	.746	.953	.870	.934	
s^2_{yr}	28.30	5.20	21.14	0.50	1.00	4.67	8.67	
s^2_{xr}	36.70	23.47	16.50	18.00	2.33	11.67	4.92	
$cov(y_r,x_r)$	−24.10	−10.00	−15.14	3.00	−1.00	−6.67	−6.00	
$\sqrt{v(\hat{R}_r)}$	.068	.046	.040	.058	.021	.045	.032	
$cv(y_r+x_r)$	.118	.083	.081	.100	.032.	050	.037	
$\tilde{R}_r$	.732	.839	.904	.740	.952	.870	.934	
$\sqrt{v(\tilde{R}_r)}$	.067	.047	.041	.059	.023	.045	.032	