STUDIES TO ASSESS AND IMPROVE VOCATIONAL PERFORMANCE INDICATORS

Nayak L. Polissar, Ph.D.
Moni B. Neradilek, M.S.

The Mountain-Whisper-Light Statistics
Seattle, WA 98112

14 July 2012

EXECUTIVE SUMMARY

Five studies are proposed to assess the validity of the L&I quality indicators for VRC and VRC firm work. The studies are:

Reproducibility of department decisions by vocational service specialists (VSS).
We propose a study of consistency and accuracy of determinations by the VSS. The consistency study involves comparison of VSS on their average rating of quality indicators. A method of detecting VSS with unusually high or low average ratings (outliers) is presented as well as a method for measuring variability among the VSS. An example is presented which suggests that either the VSS do vary non-trivially in their rating practices or else there are other factors that cause the noted variation. The accuracy study will involve comparison of VSS determinations with determinations made by one or more experts. The experts will be taken as a gold standard and VSS coding will be compared to the experts' coding to determine accuracy of VSS coding. Of note, the VSS consistency study will be inconclusive without the completion of the companion case-mix study.

Impact of case mix on VRC performance measures.
The goal is to determine if characteristics of closures and characteristics of claimants and the claimants' environment affect the quality indicators. The study is epidemiologic in nature, as variables (risk factors) will be analyzed to see if they are related to the quality indicators. Results will be presented as the mean of quality indicator values for different categories of closures. The case-mix study is essential to document either bias or lack of bias in the claim review process. If bias is detected, some mainstream solutions are available to reduce or eliminate bias in the review process.

VRC recommendation code compared with CM/VSS closing code.
Two studies are proposed. The focus is on closures whose closing code has been changed from that recommended by the VRC. Experts will determine the correct closing codes and the VRC and the CM/VSS closing codes will be compared to them to determine who is "right", closure by closure. In the second study, similar to the case-mix study, variables will be analyzed in relation to closures to determine the profile of closures that are likely to have a problem with closing code disagreements between VRC and CM/VSS.

Firm quality of work as a proxy for individual vocational provider (VRC) quality of work.
The statistical method of variance components will be used to determine if there are firms whose internal consistency of VRC quality of work is strong enough (consistent enough) to use the firm's quality indicator results as a proxy for the individual VRC's quality of work. This proxy use of indicators would be helpful for VRCs with very low caseloads for whom their own quality indicator scores would be very imprecise.

Current format of the indicators.

The current display format is adequate, but the statisticians recommend supplementing the current textual display with more graphics.

Some initial discussions with L&I will be needed to prioritize the studies prior to execution.

ASSUMPTIONS

The only assumption is that L&I has the data necessary to carry out the studies proposed here.

We presume that readers are familiar with our first report on the quality indicators: *Nayak Polissar and Alisa Burpee, A Statistical Evaluation of Six Indicators Used to Asses VRC Performance, The Mountain-Whisper-Light Statistics, Seattle, 30 April 2010.*

The six quality indicators that are considered here (and which were the topic of the Polissar and Burpee 2010 report) are as follows. (Table 1 from our 2010 report is also included here to give a sense of volume for each indicator.)

1) Useful Outcome. Outcomes are rated as useful, not useful or neutral. The "useful outcome" measure is based on the proportion of useful outcomes among those classified as either useful or not (excluding neutral outcomes.) L&I has adopted a simplified display of this indicator derived from the percentage of useful outcomes for each VRC (or firm). The displayed value of "useful outcome" may range by integers from 0 up to 9. L&I staff have informed us that they are using these integers to represent percentage ranges, as follows.

9: 90% to 100%
8: 80% to < 90%
7: 70% to <80%
…
…
1: 10% to <20%
0: 0% to <10%%.

The use of this simplified integer designation solves a specific problem. A large fraction of the VRCs will have a small number of cases entering into the calculation of the "useful outcome" indicator. For small sample sizes a single digit (or "10") more faithfully represents the precision of the indicator than a 2-digit percentage or a percentage with decimals. For example, A VRC who had 8 out of 11 outcomes classified as useful would have 73% useful outcomes (or 72.727…%) and an indicator value of 7. Given that the 95% confidence interval for the useful outcome percentage for this VRC spans a very wide range, 43% to 91% (adjusted Wald method), the 2-digit 73% value represents more precision than there really is. The single digit of 7 makes less claim to precision than two digits.

The other five indicators are:

2) Training plans approved after first submission to L&I.
3) AWA Eligible recommendations approved after first submission to L&I.
4) Timeliness of monthly progress reports (PR).
5) Dispute results.
6) Complaint findings.

Table 1. Approximate annual volume* (and other notes) for each indicator.

Indicator	**Approximate volume per year***	**Approximate annual volume* per VRC****
Useful outcome	10,000 closures that may be classified as either useful, not useful or neutral. Excluding "neutral", 9,000 would be classified as useful or not useful.	22 useful or not useful closures
Training plans approved first time	1,900 training plans either approved or disapproved. Approximately 60-80% of plans (~1100-1500 plans) are approved upon first submission.	5 training plans approved or disapproved
AWA Eligible recommendations approved first time	2,400 recommendations either approved or disapproved. Approximately 60-80% of recommendations (~1500-1900) are approved upon first submission.	6 recs. approved or disapproved
Timeliness of monthly progress reports (PR)	~10,000 reports per month; over 100,000 per year.	> 250 reports
Dispute results	1,000 disputes filed out of approximately 7,000 disputable closures. Approximately 69% of filed disputes (~690) are judged to have no applicable findings against the VRC.	18 disputable closures, 2-3 filed disputes and 1 dispute with findings
Complaints with findings against VRC	70 per year	0.2 complaints. Most VRCs would have 0 complaints.

**Order of magnitude only. These figures are not intended to indicate volume for any particular recent year. The figures are heavily rounded for simplicity.*
***The average count per year per VRC assumes 400 VRCs. The actual number per VRC is expected to vary widely among the 400 VRCs—from well below the average to well above the average.*

OVERVIEW OF THE METHODS USED

The statistical methods for the proposed studies are included with the description of each study.

RECOMMENDED STUDIES: DESCRIPTION AND DISCUSSION

(Note: this heading replaces the heading "Summary of findings and recommendations" in the template that was supplied to us for reports.)

The following studies are proposed as a means of improving understanding and proper use of the quality indicators. The studies may also motivate change in the definition and usage of the quality indicators.

Consistency of decisions by vocational service specialists (VSS)[1].
Due to the need to monitor VSS performance on an ongoing basis and to have a method that is not too resource-intensive, we suggest having a study of consistency—a concept also voiced by L&I staff. This study would use the VRC evaluations that are currently part of the VSS workload. There would be a need for some additional time from the Research and Data Services (RDS) staff at L&I to produce statistical reports for review. There would also need to be additional time from supervisors or experienced staff to conduct follow-up reviews and have contact with individual staff to identify and implement any modifications in practice that are needed.

As background, each VSS handles a large volume of cases, and the percentage of their decisions falling in certain categories (such as percent of training plans approved on the first submission) will reflect the decision process of each VSS. If ratings are similar among the individual VSS, then it is evidence (but not proof) that they are using the same decision process and that the VSS are "consistent" and are interchangeable in their handling of VRC submissions; equivalently, they are doing the same thing. The purpose of this consistency study is to determine if the VSS are, indeed, consistent.

If one or more VSS differ substantially from others in the average or percentage of a particular performance measure (e.g., plans approved), it is evidence (but not proof) that at least some of the VSS are handling the cases in a unique way and that the VSS are not interchangeable. Looming behind this process is the issue of "case-mix" which is covered in another section. Briefly on case-mix: if there is substantial variation among VSS in their average ratings, the variation may not be due to the VSS at all, but may be, rather, due to some variation in the composition of claims that come to the different VSS for processing. Suppose a VSS works on claims from a particular sector with "difficult" claims or with claims that, for some reason, result in a poor performance rating, on the average. Then that particular VSS may appear different from other VSS due to the material received and not due to a variant processing style of the VSS.

At this point it will be helpful to introduce a hypothetical example that shows that the VSS *may* differ in how they handle the plans submitted. (Please note the emphasis on "*may*".) This is an example of case-mix influence on outcomes, where the case-mix factor being considered is the role of the individual VSS. In this hypothetical dataset Table 2 shows the decisions made by the VSS for 1104 plans submitted. The caseload per VSS varies from 124 to 179. The percentages in Table 2 are plotted in Figures 1A-C.

[1] We will use the acronym VSS in both a singular and plural sense. E.g., the following three phrases use "VSS" in, respectively, a singular, a singular and a plural sense: "…the VSS works…", "…a VSS works…" and "…the VSS work…."

Table 2. VSS decisions for 12 months of activity. N = 1104 plans reviewed.

VSS Work Position #	% approved the 1st time	% revised	% denied
VR-A	64%	21%	15%
VR-B	79%	15%	6%
VR-C	76%	15%	9%
VR-D	77%	16%	7%
VR-E	82%	12%	6%
VR-F	59%	34%	7%
VR-G	72%	20%	8%

A perusal of Table 2 and Figures 1A-C shows some variation among the six VSS, such as a range of 59% to 79% in plans approved the first time (a span of 20 percentage points), and ranges of 12-34% for plans returned for revision (P5 letter) and 7-15% for plans denied. These appear to be wide ranges.

One might argue that there could be the same true rates of approval, revision and denial among the six VSS and it is just by chance that they differ in these data—like the random variation that occurs in flipping a coin. One might easily have seven heads out of ten flips on one series of flips of a fair coin (70% heads) and then get three out of ten on another series of flips (30%.) However, in Table 2 we are seeing a large number of "flips"—between 100 and 200—so that percentages should be more stable than for a small sample size. A formal statistical test rejects the hypothesis that all the VSS have equal rates In statistical parlance there is less than one chance in a thousand that random variation alone would yield the diverse values of *percent approved* shown in Table 2, if, indeed, all of the VSS have the same underlying approval percentage. The same can be said of *percent revised* and *percent denied*—less than one chance in a thousand that the diversity of the percentages in each column of Table 2 are due to random variation[2].

[2] The formal definition of statistical significance is slightly different than that offered here. This heuristic definition is offered for simplicity.

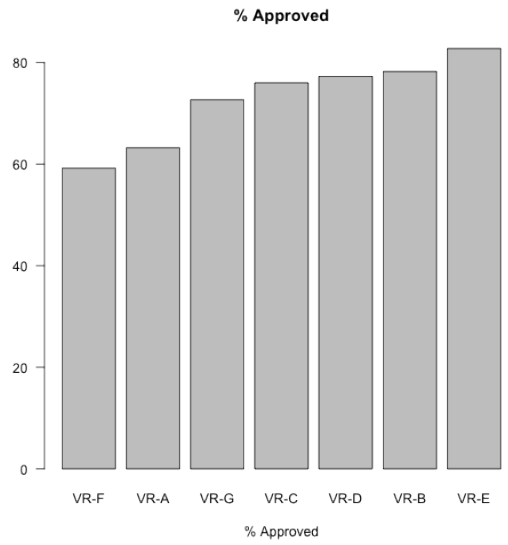

Figure 1A. % approved by VSS position #.

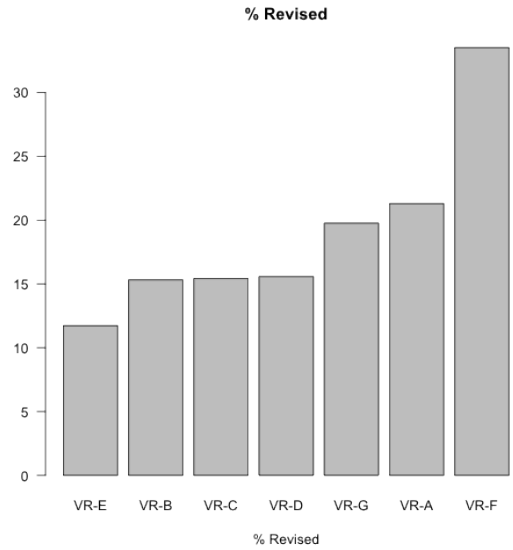

Figure 1B. % revised by VSS position #.

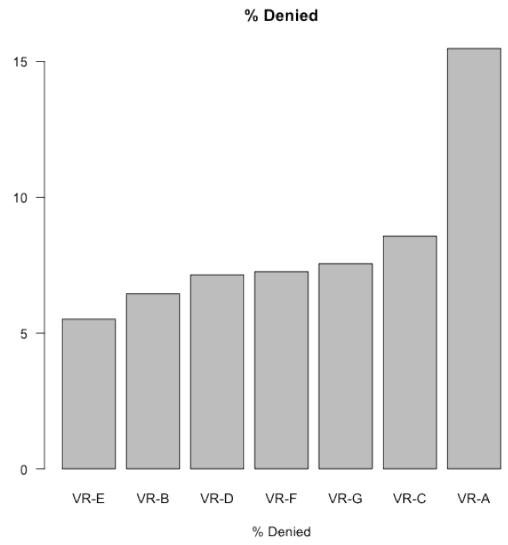

Figure 1C. % denied by VSS position #.

Note that the high revision rate for VR-F stands out in Figure 1B and the high denial rate for VR-A stands out in Figure 1C. These two work positions also show slightly lower rates of approval than the five other VSS positions in Figure 1A, but the contrast among the VSS is much less in Figure 1A than in Figures 1B and 1C.

Given the apparently non-random departure of decisions by VR-A and VR-F from the other VSS in this hypothetical example, it would be worth (with real data) L&I attempting to discover why their rates are different from rates of the other VSS.

We will suggest a more formal method for identifying which VSS may need review (see below), but there is a second form of variation of the group of VSS that should be taken into account. In another dataset there might be no “outliers” among the VSS, but the group of VSS may show unacceptable variation. For example, consider five hypothetical VSS with percent approval of 69%, 70%, 73%, 75% 75%. There is no outlier and the five VSS have quite similar rates—a span of six percentage points. However, if five other VSS have percent-approved rates of 50%, 62%, 69%, 79% and 86%, there are no outliers, but the spread is very large—perhaps unacceptably so—a span of 36 percentage points.

We propose tests for a) detecting outlying individual VSS (with unusually high or low percentage rates) and b) determining whether the VSS as a pool have an unacceptably wide spread of rates. We offer an outline of the tests, which would need to be implemented in conjunction with L&I staff. As a side note, we have suggested simple and approximate tests which can be handled by the L&I RDS personal[3].

[3] It is possible that more sophisticated and more powerful tests can be implemented, depending on the statistical training (in methods and in statistical packages) that is available within the L&I team handling this activity.

Detecting Outliers. We propose, initially, a simple, approximate test for detecting individual VSS whose performance may be outliers. We also recommend that this test be implemented in an exploratory fashion until we gain some experience with detecting outliers[4]. The test works reasonably well when there is one outlier rate among a group of VSS rates, but may not work well when there are two or more outliers. The test should be used when there are a sufficient number of claims behind each rate[5].

The test consists of comparing a rate or a mean for a particular VSS to the mean rate for the balance of the VSS group. This test is repeated for each of the VSS. Statistically, we can begin by using a two-sample, two-sided t-test. The two samples consist of a) one designated VSS, and b) the balance of the VSS group. The difference between the rate for the one VSS and the mean of the other VSS rates is divided by an appropriate measure of variability. The statistical significance of the difference will tell us how likely it is that one VSS rate could differ from the mean of the other rates by the observed amount. This test is repeated for each VSS. The statistical significance is adjusted (using the Bonferroni method) for the fact that there are quite a few tests carried out—one per VSS[6]. For the seven VSS considered earlier, the high "revised" rate for VR-F in Figure 1B is statistically significant ($p = 0.04$) in comparison to the balance of the VSS; the high denial rate for VR-A in Figure 1C is also statistically significant ($p = 0.005$.) None of the other VSS had a statistically significant difference from the balance of the VSS for any of the three categories: approvals, revisions or denials.

Measuring variation among VSS. As noted previously, there may be no outliers, but we wish to assess the variation among the VSS in their average rating of submitted claim work from the VRCs. The ideal would be that all of the VSS would give very similar quality ratings when presented with the same claim material.

The need and potential method for measuring variation among the VSS will be best decided after the case-mix study is carried out, since the model for assessing VSS variation depends on what other factors (if any) are found that affect ratings. The factors

---

[4] The applicability of the test also depends on the results of the case-mix study described in this document. The test may need to be modified if there are non-trivial case-mix effects on the quality performance measures.

[5] The number of cases needed will depend on a) the nature of the particular performance measure: binary (such as useful vs. not useful outcome) or continuous, and b) certain parameters of within-VSS and between-VSS variability of the performance measures. For the binary measures it is likely that at least 100 claims per VSS would be needed, and at least 50 per VSS for continuous data. An assessment can be carried out in advance of the analysis to ensure that there is a large enough sample size make it very likely that an unacceptably large variation among the VSS would be detected.

[6] Some references for this method of outlier detection, which uses the Bonferroni adjustment, include: Cook, R. D. and Weisberg, S. (1982) Residuals and Influence in Regression. Chapman and Hall. Fox, J. (2008) Applied Regression Analysis and Generalized Linear Models, Second Edition. Sage. Fox, J. and Weisberg, S. (2011) An R Companion to Applied Regression, Second Edition, Sage. Weisberg, S. (2005) Applied Linear Regression, Third Edition, Wiley. Williams, D. A. (1987) Generalized linear model diagnostics using the deviance and single case deletions. Applied Statistics 36, 181–191. The Hadi method of outlier detection can accommodate multiple outliers: Hadi, Ali S. (1994), "A Modification of a Method for the Detection of Outliers in Multivariate Samples," Journal of the Royal Statistical Society (B), 2, 393-396. Hadi, A. S. and Simonoff , J. S. (1997), "A More Robust Outlier Identifier for Regression Data," Bulletin of the International Statistical Institute, 281-282.

to be considered would include characteristics of the claimant, the injury, the environment (industry, occupation, etc.), geographic location and the characteristics of the VRC handling the claim. If there is no effect of any measured factor on the quality performance ratings, then quite a simple model can be used for estimating variation among the VSS in their average ratings. That variation would then be due to the VSS and to any unmeasured factors that influence ratings. If we are able to assess the effect of a very thorough set of factors and eliminate the case-mix effect for all of them, then it is quite likely that any remaining variation is due to different rating practices of the individual VSS. Related material and concepts are discussed in the case-mix section.

We discuss now a simple and approximate method of measuring variation among the VSS and use it with the VSS dataset from this section to show that either a) the VSS do vary in their quality rating of claims or b) other factors affect the rating of claims.

The standard deviation (SD) is a common way to measure variation. The standard deviation becomes larger as variation increases. One approximate way to measure variability is to calculate the VSS-to-VSS standard deviation of their average performance measures. Under one common assumption[7] a span of plus or minus two standard deviations (±2 SD) around the mean would include about 95% of the VSS in the population of VSS from which the Table 2 VSS were drawn[8].

The means and standard deviations (with and without outliers) for percentage approval of plans, percentage revision and percentage denial are shown in Table 3 for the seven VSS initially introduced in Table 2. The Table also shows the standard deviations calculated more accurately using other methodology ("random effects models".) Considering the SD with outliers removed, the variation does seem quite large in relation to the mean values. For example, the mean of the seven VSS rates of approval (the seven percentages in the first numeric column of Table 2) is 72.7%, and the SD is 8.4% so that one would expect that approximately 95% of VSS would have rates falling between 56% and 90%--two standard deviations. This "95%" range is wider than the observed range of the seven specific VSS considered, but the interpretation is that in drawing from the population from which these VSS came, it would not be uncommon to find VSS at or near the 95% limits. That is, if practices continue stably in the form that they are now, then it would not be surprising to find future VSS employees whose rates are as diverse as the width of this 95% range, ranging from 56% approval to 90% approval.

There is no statistical standard on how large the SD values can be and still be considered acceptable. SDs need to be interpreted in the context in which they are calculated. L&I will need to weigh in on an acceptable magnitude of SDs. Nevertheless, we can establish the principle of measuring diversity of VSS rating practices through

[7] the assumption of a normal distribution.

[8] The proper interval to use here is a statistical tolerance interval. The interval noted in this section will be a 95% tolerance interval for some level of confidence (and an underlying normal distribution), but the example is simplified. The specific interval to be used will depend on the statistical computing capacity of the L&I staff.

use of a variability measure such as a standard deviation estimated from an appropriate modeling process.

Table 3. Means and standard deviations among 7 VSS on their percentage approval, revision and denial of plans.

	approved*	**revision**	**denied**
Means			
with outliers	72.7%	19.0%	8.3%
without outliers	72.7%	16.5%	7.2%
Standard deviation			
with outliers	8.4%	7.3%	3.1%
without outliers	8.4%	3.4%	1.2%
from random effects models, with outliers	*7.8%*	*6.5%*	*2.4%*

*There are no outliers in the "approved" category, so means and SDs are the same with/without outliers.

The consistency study can be carried out on all of the quality indicators except the complaint findings. In addition to the challenge of a low incidence of complaints, a given complaint is not tied to a specific VSS (or claim manager) so there is no meaning to a measure of "consistency" among VSS (or CM) for complaints. In the case of disputes, if the dispute is electronically linked to either a claim manger or VSS decision, then the consistency study can be carried out for such disputes. It may take a longer calendar period of experience to provide a sufficient volume of cases for a consistency analysis of disputes than for the other quality measures.

Handling continuous quality measures. The quality indicators are currently all binary (e.g., useful/not useful), but L&I may wish at some time to create more continuous measures to evaluate for consistency. Ratings underlying two of the quality indicators, *"training plans approved/not approved after first submission to L&I"* and *"AWA eligible recommendations approved/not approved after first submission to L&I"*, can be converted to a fairly continuous value as the sum of the number of boxes checked and/or as the sum of the "yes" notations on the worksheet. For the plan development worksheet, there are 16 major items, which would lead to a score of zero to 16 items checked as "yes". Under some of the major items there are other boxes to check (e.g., 7 check boxes under item #1 of the plan development worksheet.) There are a total of 36 check boxes for the plan development worksheet, and the fairly continuous variable, number of boxes checked, would potentially vary from 0 to 36. If L&I does not currently digitize all of the box-checking activity of the VSS, they may wish to do so in the future.

In general, continuous measures provide a more powerful basis for analysis than binary measures.

For each continuous score we can calculate the mean value (and standard deviation) for a particular VSS (or claim manger) and examine these means for consistency across the VSS. We can carry out a formal test for equality of means using analysis of variance or regression and using a method to accommodate the statistical dependence of multiple claims per VRC or VRC firm.

The steps needed to carry out the consistency study are as follows. The statisticians should be consulted about each of these steps. The outcome of this consistency study will not be clear unless the case-mix analysis has been carried out.

1. Identify a calendar period from which to draw claims submitted to L&I by the VRCs. One year or more would be desirable. Choose a period to generate at least 100 cases[9] per VSS (or claim manager.)
2. Create a file of all closures during the specified period. The file will include the VSS (and claim manager) identification. The i.d. information may be changed so that individuals are not identified; the key should be retained separately (old i.d. matched with new, anonymous i.d.).
3. Calculate the percentage positive (or negative) outcome for each VSS or each claim manager, as appropriate (depending on the quality measure considered.) If the case-mix study has been carried out, this percentage may need to be adjusted for the composition of cases that were reviewed by each VSS or CM.
4. Tabulate the percentages as in Table 2 of this report and create bar charts or other graphs as in Figures 1A-C of this report. Only include VSS or claim managers who have at least 100 claims (so that percentages will be precise within ±10% or better[10].)
5. Identify outliers using the Bonferroni/t-test or other method.
6. For the balance of the VSS (or claim managers) who are not outliers, calculate the standard deviation of outcome percentages using the simple method described earlier—or a more sophisticated method.
7. Consider reviewing the work of the outlier VSS and CM.
8. Determine if the SD of outcome rates across the VSS of CM is greater than is acceptable. If so, investigate possible causes of the unacceptable diversity of VSS or CM ratings.
9. Further analysis, if needed.

[9] Ideally, a simple exercise would be carried out in advance to determine the minimum desirable sample size per VSS.

[10] The maximum width of a 95% confidence interval for an observed 50% based on 100 cases is ±10%, i.e., 40%-60%, using the normal approximation. The 95% confidence interval for any other observed percentage based on 100 cases would be narrower than ±10%.

It is important to repeat that if there are outliers or excessive variability among VSS or CM, this consistency study will not be very conclusive unless the case-mix study has been completed first. For example, if a VSS is an outlier on their rate of denying plans, is it because the VSS is doing something wrong or because the VSS is receiving the types of claims which are prone to problems that lead to a denial?

Accuracy of department decisions by VSS and claim managers (CM).

The determination of accuracy of VSS[11] decisions is a more labor-intensive project than determination of consistency. The reason that extra effort is needed for measuring accuracy of the work of VSS or of claim managers is that some kind of "gold standard"—that does not now exist—will be needed. Claims will need to be reviewed separately from (and in addition to) their review as part of the production work of L&I.

"Accuracy" refers to the closeness of ratings to the truth, and, therefore, the truth has to be known or approximated. This will be difficult for some of the quality measures. While the timeliness of progress reports is quite objective, the other five quality indicators involve more judgment. The "truth" would have to be defined by an expert panel or by an individual or individuals who are recognized as experts on a given indicator, such as return to work.

Based on discussions with L&I staff and review of an earlier draft of this report, we understand that a formal one-time accuracy study is not likely to be carried out, due to the amount of labor and resources that would be needed during a relatively short period of time. In a previous draft of this document we presented a plan for determining accuracy of VSS or claim manager work, but that section is omitted here given the substantial additional labor needed to carry out an accuracy study.

However, we do suggest that expert review of VSS and claim manager work be carried out on an ongoing basis—a practice that L&I has reportedly already begun. We recommend that the claims selected for review be those prone to having problems—a selection process that can be better specified after a case-mix study is carried out. The two classes of claims that will be of most interest are a) claims at high risk of receiving a poor quality rating; b) types of claims which show a large spread in VSS average ratings—e.g., some VSS might have a much higher rate of approval than other VSS for referrals involving mild back pain. The case mix study would result in a model which can help identify claims that have a high likelihood of receiving a low quality rating or claims for which the VSS ratings are apt to be diverse; a sample of these "high risk" claims can be reviewed.

[11] Any of the procedures that we suggest for measuring accuracy of VSS work can also be applied to the work of claim managers, and we sometimes refer only to the VSS but it should be understood that the claim managers can be reviewed in the same way, though the work of the VSS and of the claim managers differ on their impact any given quality measure. The claim mangers, for example, have relatively more impact on "useful outcome", and the VSS have relatively more impact on plan review decisions.

L&I will get the most out of these reviews if the composition of the claims reviewed is similar from year to year. It would be best to develop a summary sheet noting the results of each review and compiling the reviews into an error rate (e.g., percent of the "high risk" claims with an erroneous quality rating) which can be compared from year to year. These reviews will also be helpful if they are discussed openly at meetings with L&I staff and VRC groups in the manner of "grand rounds" where physicians gather to review interesting cases. If the reviews are carried out on an ongoing basis with continuous feedback to the staff and the VRC community, the effort should result in a decreasing error rate over time and in an increase in the quality of work of the VRCs.

Among the six indicators, we recommend that only the following three indicators be included in the accuracy study:

1) Useful Outcome (outcomes are rated as useful, not useful or neutral.)
2) Training plans approved after first submission to L&I
3) AWA Eligible recommendations approved after first submission to L&I

(We may refer to these later as, briefly and respectively, useful/not useful, training, and AWA recommendations.)

The omitted indicators would be:

4) Timeliness of monthly progress reports (PR)
5) Dispute results
6) Complaint findings

We discuss the omitted indicators later in this sub-section.

Accuracy measures something different than consistency. If the VSS are very consistent, that is good, but they may be consistently providing either the correct or the incorrect determinations. If in some setting (not necessarily L&I) the true rating for a module of work is "4" on some scale, and six raters provide scores of 2, 2, 2, 2, 2 and 2, there is consistency, but the ratings are not accurate.

The steps needed to carry out the **accuracy** study on an ongoing basis are as follows. The statisticians should be consulted about each of these steps.

1. Select an expert, experts or expert panel to determine the true ratings for cases. Ideally the same expert or experts would review cases over time.
2. Select a sample of "high risk" claims for review. The sample can be built up over time, but the ultimate composition of the to-be-completed review sample should be decided in advance. If there is more than one "high-risk" claim in one of the categories from which sampling is to be conducted, then the cases for review from that category should be selected randomly using a random number table or computer-generated random numbers.

3. Develop a data collection form to capture the results of the review—including pre-coded items and a section for open-ended comments.
4. Carry out the expert review without consulting VSS, CM or VRC involved.
5. Key enter review results and periodically compile. Error rates may be calculated at any time, but they may not be very precise until approximately 50-100 cases have been reviewed. The precision will depend on the magnitude of the error rate.
6. Provide feedback to VSS, CM and VRC on findings.
7. Track error rates over time to determine if there is a trend..
8. Additional analysis if needed.

Indicators omitted from the accuracy and consistency analysis. We earlier mentioned omitting the following indicators from this analysis.

4) Timeliness of monthly progress reports (PR)
5) Dispute results
6) Complaint findings

The reasons for this omission are as follows. The *timeliness* indicator can be omitted from our consistency and accuracy analyses because it is quite a mechanical determination. We assume that little judgment is involved in determining whether or not a progress report is late. Thus, the impact of the VSS or CM is either minimal or non-existent. It would be easy to carry out a quick analysis to verify that the percentage of on-time progress reports does not vary across VSS or CMs.

The *dispute* results are an indicator for which a consistency study is feasible (see the consistency section.) However, an assessment of accuracy of the dispute decision process is probably not feasible. The dispute resolution process can be quite complex, involving multiple staff members and reviews, and the L&I experts are already involved with resolution of the dispute. The dispute resolution process is already a study of accuracy: was the right decision made? It would be labor-intensive and perhaps not even possible to find a qualified external or internal expert who a) knows the L&I system well and b) was not already involved with L&I dispute resolution.

The incidence of complaints is so low that neither an accuracy nor a consistency study is feasible. Additionally the complaints are usually tied to the VRC and not to a specific VSS or CM. Further complicating the issue of complaints and measuring the accuracy and consistency of L&I action on complaints, we understand that a VRC can request a reconsideration of a complaint finding, and, if the VRC is not satisfied with the results of the reconsideration, they can appeal to a higher Board.

Impact of case mix on VRC performance measures.

As background, the current quality performance measures do not incorporate any adjustment for case-mix. We understand that the L&I and VRC communities feel that the current definitions of the quality measures have eliminated case-mix issues. This is an excellent ideal, and we hope that it is true. However, hope and belief do not replace evidence, and it is important to base practice upon evidence. Earlier in this document we showed that a hypothetical sample of work of seven VSS did not allow us to rule out the VSS as an influential factor in determining the value of a quality indicator. The quality indicators are supposed to measure the quality of VRC work, but if a VRC happened to have more work assigned to the VSS labeled as VR-F or VR-A in Table 2, they might have had higher rates of plans returned for revision or plans denied and lower rates of plans approved the first time than if they had had their work reviewed by other VSS. The VSS is only one of a number of factors that may affect the rating of a VRC, aside from the quality of work of the VRC. The VSS example studied here simply shows that case mix can be important.

The case-mix study is essential for establishing the validity of the performance measures and verifying that the system is giving unbiased estimates of VRC performance.

It is worth presenting two examples where a hidden case-mix issue yielded biased results when case-mix was ignored.

Kidney stone treatment[12]. Table 4 shows the success rates (and numbers of patients) for treatments involving both small and large kidney stones, where Treatment A includes all open procedures and Treatment B is percutaneous nephrolithotomy. Note that treatment B has an overall higher success rate than treatment A (see the first numeric row of the table.) Nevertheless, treatment A is actually better than B, because it has the higher success rate for both small and large kidney stones (the last two numeric rows of the table.) What went wrong? The problem is that treatment B, which, without a case-mix analysis, initially appears better than A, was carried out more frequently on patients with small stones; treatment A was carried out more frequently on patients with large stones. The two stone-size groups have quite different success rates and their combination actually reversed the apparent effectiveness of the treatments.

[12] Steven A Julious, Mark A Mullee. Confounding and Simpson's paradox. BMJ 1994;309:1480-1

Table 4. Success rate (%) by treatment and size of kidney stone.

	Treatment A	Treatment B
All patients	78% (273/350)	**83% (289/350)**
Small Stones	*Group 1* **93% (81/87)**	*Group 2* 87% (234/270)
Large Stones	*Group 3* **73% (192/263)**	*Group 4* 69% (55/80)

In this example the "lurking" variable (or confounding variable) of the stone size was not previously known to be important until its effects were included in the analysis.

A second examples (Tables 5 and 6) concludes our presentation on the importance of considering case-mix.

Berkeley sex bias case. The admission figures for the fall of 1973 showed that men applying to UC Berkeley were more likely than women to be admitted, and the difference was so large that it was unlikely to be due to chance (Table 5).

Table 5. Admission rate by gender.

	Applicants	% admitted
Men	8442	**44%**
Women	4321	35%

However when examining the individual departments, it was found that no department was significantly biased against women—examples in Table 6.)

Table 6. Admission rate by department and gender. Illustrative departments.

Department	Men		Women	
	Applicants	% admitted	Applicants	% admitted
A	825	62%	108	**82%**
B	560	63%	25	**68%**
C	325	**37%**	593	34%
D	417	33%	375	**35%**
E	191	**28%**	393	24%
F	272	6%	341	**7%**

The research paper by Bickel, et al[13] concluded that women tended to apply to competitive departments with low rates of admission even for qualified applicants (such

[13] Author(s): P. J. Bickel, E. A. Hammel, J. W. O'Connell. Sex Bias in Graduate Admissions: Data from

as in the English Department), whereas men tended to apply to less-competitive departments with high rates of admission among the qualified applicants (such as in engineering and chemistry).

In our context, there are a number of potential factors that might influence the quality performance measures, and these factors should be screened to investigate the possibility of bias through confounding.

A proper case-mix study is not difficult and will either support or refute the assumption that the performance measures are measuring VRC performance and are not unduly influenced by other factors. Other factors may include VSS practices and the personal, medical and environmental characteristics of the claimant. The "consistency" section above provided a numeric example of the potential role of the VSS in generating varied ratings of the VRC-submitted plans—over and above whatever variation there is in quality of work by the VRC. It is possible that the VSS factor and other factors result in biased performance statistics being credited to a VRC.

The VSS is only one factor that may influence the performance measures. Again, "may" has to be emphasized, since we do not yet know whether <u>any</u> factors influence the ratings. The purpose of this case-mix study is to determine if there are other factors, aside from the quality of VRC work, which influence the quality performance measures. It is likely that some case-mix factors are influential, but if the influence is small, these factors could be ignored when using the measures. A factor that raises or lowers the value of an indicator by just a small amount is unlikely to have practical importance.

This first step of the analysis (the univariate analysis) is fairly straightforward and can be carried out by L&I staff with input and assistance in interpretation from the statisticians.

The goal of the univariate analysis is to determine if each factor or variable, considered singly, may have an effect on a specific performance measure, such as useful outcome.

Table 7 is a template table for the univariate analysis for case-mix factors which are categorical (such as gender, diagnostic group, geographic regions, etc.) This is our non-expert rendering of examples of case-mix factors to be investigated. The examples are supplied merely to give an idea of how the table works. Subject-matter experts from L&I and the VRC community will need to specify the actual list of potential case-mix factors to be investigated. The list should be very inclusive. The experts should not exclude a potential case-mix factor unless they know (based on quantitative evidence) that it is not important. The list should be exhaustive, because the initial screening of case-mix factors is not difficult.

Table 7 is made up for the useful/not useful outcome, but similar tables can be made up for the other quality indictors. The last column of the table is the "bottom line."

Berkeley. Science, New Series, Vol. 187, No. 4175 (Feb. 7, 1975), pp. 398-404.

Table 7. Number of closures, exclusions and outcomes for the "useful outcome" quality indicator.

Factor and levels	N closures	N excluded from analysis*	N included in analysis**	N useful	N not useful	Percent useful	P-value+
Gender							
male							
female							
Phase							
Phase A							
Phase B							
Phase C							
etc.							
Etc. (other variables)							
Etc. (other variables)							
Etc. (other variables)							

* neutral outcome.
** N included in analysis = (N useful) + (N not useful)
+ The p-value is the statistical significance of the difference in % useful among categories of a factor. It is helpful in determining if the difference is likely to be random. There is just one p-value per factor.

After the univariate analysis is completed there will be a multivariate analysis to determine if a combination of variables has an influence on the quality measures. The multivariate analysis would also deal with a) the potential relationship among case-mix variables in affecting the quality indicators, and b) potential confounding, whereby the effect of a case-mix variable on the quality indicators may be either masked or exaggerated or be an artifact due to the relation or correlation of the specific case-mix variable with other variables.

Five of the six quality indicators can be included in the case-mix analysis. However, the analysis for each indicator will be carried out separately. The complaint indicator will be omitted at this time because a complaint is not tied to a specific closure or other module of VRC work. It is tied to the VRC or to the VRC's firm.

While the closure or other module of work will be the unit of analysis, the modules are clustered within the VRC or the VRC's firm, and the analysis will need to take that into account. Technically, we will use generalized estimating equations (GEE) to take account of the correlation among the multiple cases handled by an individual VRC.

The results of the univariate analysis can be displayed in a brief table such as the following, which includes the odds ratio as a way of gauging the impact of a variable on the outcome.

Table 8. Case mix study: results of the univariate analysis for a specified quality indicator*.

Variable and categories	N per category	% approved**	odds ratio	p-value
Gender				
male	(n)	83%	reference	0.17
female	(n)	85%	1.16	
Phase				
Phase A		83%	reference	0.002
Phase B	(n)	86%	1.26	
Phase C	(n)	90%	1.84	
etc.	etc.	etc.	etc.	
Etc. (other variables)	etc.	etc.	etc.	etc.
Etc. (other variables)	etc.	etc.	etc.	etc.
Etc. (other variables)	etc.	etc.	etc.	etc.

** Values of n and phase and other items to be filled in when calculations are actually carried out.*
***The specific heading depends on the quality indicator.*

Table 8 (hypothetical data) shows that male and female claimants have a similar rate of approval of plans and the difference is not statistically significant. $P < 0.05$ is a common threshold for declaring a comparison inconsistent with chance. We shall use this threshold. The odds ratio (OR) is a common measure of outcomes. OR = 1.0 means there is no difference in percent approval between the two categories compared, and here the value OR = 1.16 for females vs. males is very close to 1 and not of particular concern as having an impact on approval. The (hypothetical) analysis in Table 8 of phase of closure, however, does show a statistically significant effect of phase on percent approval ($p = 0.002$), and the OR = 1.84 for phase C vs. phase A may be of some concern, reflecting the 90% vs. 83% difference in approval rate.

The results of multivariate analysis would be presented in a similar table, but the effect of each variable would be adjusted for other variables included in the analysis.

Tables 7 and 8 present a paradigm for presenting case mix results for categorical factors, such as gender. Corresponding tables can be prepared for continuous factors, such as age, years with employer, time since accident, etc. These tables for continuous variables would show the sample size, mean and standard deviation for each category of an outcome variable. For example, consider the variable “useful outcome” and an analysis that investigates the time in days since the accident as potentially affecting the

outcome. The table would show—for the category of closures with a useful outcome and, separately, for the category of closures without a useful outcome—the sample size, median, mean and standard deviation of the number of days since the accident. The table would also show the p-value (statistical significance) of the difference in means between the two categories, useful and not useful outcomes. The p-value helps the user to decide whether the difference between the useful and not useful outcome categories is consistent with chance or, on the other hand, is likely to be a real difference.

A critical step in the case-mix study is the specification of a "laundry list" of variables to be investigated for their potential influence on each given quality indicator. We suggest that the L&I staff and VRCs collaborate in coming up with an exhaustive list of factors to be investigated for a case mix effect. Hopefully, all these factors will be "cleared" of any case-mix guilt and the quality indicators can be used as is.

In our original report we listed some areas which should be included in the case-mix analysis. Part of the list is repeated in Table 9, and it will need to be considerably extended and fully specified prior to the analysis being carried out. The items in Table 9 are quite general and will need to be expanded into a detailed list.

Table 9. Examples of topics and items to be included in a case-mix analysis.

Factors related to the injured worker, such as age, gender and other demographics, prior work history, nature of the injury, motivation, etc.
Factors related to the employer or the work environment
Geographic location
Variation over season and time
Information supplied by medical practitioners, occupational assessors, and other assessment personnel
Some measures of the complexity of the case, such as number of pages or electronic size (megabytes.)
Etc.

The steps to carry out the case-mix study are as follows.

1. VRC and L&I community collaborate in listing items (variables, factors) which should be considered as potentially having (or not having) an influence on any of the quality indicators. The variables listed need not all be available in L&I electronic files, but the analysis will be carried out only on variables that already exist or can be computed from data that now exist in the L&I electronic files.

2. L&I will need to create one file per quality indicator (except complaints). The rows (observations) of each file will have one closure (or one progress report) or one other module appropriate to the quality indicator being considered for analysis based on the specific file. The row will include columns (fields) indicating the outcome (e.g., useful or not useful for the file corresponding to the useful outcomes quality indicator) and including columns (fields, variables) for all of the case-mix variables which are specified in the preceding step and which are electronically available. The file will also need to include the VRC i.d. and the firm i.d. for the closure represented by each row.
3. The statisticians or L&I staff working with us will analyze the files corresponding to each quality indicator and prepare tables similar to Tables 7 and 8 of this report—both for univariate and for multivariate analysis. Case-mix variables which appear to have an important influence on a given quality indicator will be noted. These tables can form the basis for a discussion on whether or not case-mix adjustments should be considered in use of and computation of the quality indicators.

If it is apparent that there are important case-mix factors that influence quality indicators in addition to the actual quality of the VRC work, then a specific quality indicator can be adjusted for such factors. Rather than a percentage or a single-digit representation of a percentage (the current format for quality indicators, presented earlier) the quality indicator can be presented as an odds ratio where a value of 1.0 indicates the VRC is performing at the average level, and values above or below 1.0 indicate superior or inferior performance, respectively[14].

VRC recommendation code compared with CM/VSS closing code.

In some instances the VRC will recommend a closing code for a module of work and the claim manager (CM) or VSS will use a different closing code. The CM/VSS code is currently considered by L&I as the definitive ("correct") closing code in the L&I database and is used to calculate the useful outcome score for each VRC. Some fraction of these code changes will result in a case being changed from "useful" (based on the VRC code) to "not useful" (based on the CM/VSS code.) Alternatively, there may be some changes from "not useful" to "useful". Clearly the changes between the "useful" and "not useful" categories will affect the related quality indicator of useful outcomes and will affect the reported performance of any VRC involved. Further, changes from "neutral" to either "useful" or to "not useful" (or vice versa) will affect the quality indicator. Because there is a clear difference of opinion between the VRC and the CM/VSS when there is a

[14] The odds ratio for a VRC will be computed from a) the observed proportion of positive outcome closures for a given VRC (used to calculate the numerator odds of the odds ratio) and b) the expected proportion of cases with a positive outcome (used to calculate the denominator odds of the odds ratio.) The expected proportion will be the mean of individual estimated probabilities per each VRC closure from a logistic regression model.

code difference, and because both persons are looking at the same case material, something can be learned by examining these cases of code change. It seems unlikely that the CM/VSS is right 100% of the time on these code changes, and it is also seems unlikely that the VRC is right 100% of the time on their original, pre-change codes. Not all closures with code changes will have to be reviewed. The biggest payoff will be to review code changes that cross the boundaries between "useful", "not useful" and "neutral" categories, as shown in Table 10.

Table 10. Comparison of closing codes by the VRC and the CM/VSS. Code difference noted by "X" can be included in the code change study.

CM/VSS ↓ VRC→	Useful	Not useful	Neutral
Useful	---	X	X
Not useful	X	---	X
Neutral	X	X	---

A study reported to us by L&I found that only 3% of 16,087 referrals had code changes that fell in the "X" boxes in Table 10. The 3% can be broken down as follows: 2.8% of the referrals had code changes that would cause a poorer quality score for the VRC after the change, and 0.2% had code changes that would improve the VRC's quality score.

We understand that a non-trivial proportion of VRC submissions do not have a recommended closing code. It would be helpful to know if the breakdown between useful and not useful outcomes (assigned by the CM) differs between a) the group of referrals with no code change recommended by the VRC vs. b) referrals where the VRC did supply a recommended closing code.

The 3% of code changes found in the L&I study is not a large proportion of cases and this community will need to weigh the value of reviewing these cases against putting that time into the other studies. Possibly such a study could be of great value to some VRCs. It is possible that these code changes, particularly those that have a negative impact on the VRC quality rating, may be occurring among a limited group of VRCs, in which case there might be a much larger percentage of that VRC group's closures involved in code changes. If the cases with negative code changes are occurring among just half of the VRCs, then by finding the cause of and eliminating the problem, these VRCs might raise their "useful outcome" quality rating by several percentage points, and L&I would have fewer troublesome closures to deal with.

A sample of closures with a changed code could be reviewed by experts convened by L&I. The experts should be blinded to the recommended closing code of the VRC and the final closing code of the CM/VSS. The experts would designate a closing code in this blind fashion, and the new code would be compared to the closing codes of the CM/VSS and VRC, respectively. The results can be reported in a simple table, such as Table 11. For the expected small fraction of cases where both the VRC and the CM/VSS are incorrect (last column of Table 11), the correct code may or may not change the L&I classification within the three categories, useful/not useful/neutral. For

those where the category does change, the percentage of changes that increase vs. decrease the VRC's useful outcome value can be calculated, respectively.

In order to get the most benefit out of this exercise for future training and potential procedural changes, we recommend that the expert(s) (or an expanded panel including the experts) confer on the closure material after all codes by the experts have been determined and the initial closure codes by the VRC, VSS and CM have been unblinded. In this re-review, the expert or committee would come up with a list of potential causes for the coding disagreements and recommendations for changes in forms or procedures.

The number of code-change cases to be reviewed to provide meaningful percentages in Table 11 is, again, at least 50-100, and the various percentages in the table will probably not become very stable until over 100 closures have been reviewed. Nevertheless, the expert who reviews these cases may see some patterns that can not be anticipated now.

Table 11. Results of expert review of closing code differences between The VRC and the CM/VSS.

Type of code difference: VRC code/L&I code	**N of closures**	**% VRC correct**	**%CM/VSS correct**	**%both incorrect**
Changes that decrease VRC quality score				
useful/not useful				
useful/neutral				
neutral/not useful				
Changes that increase VRC quality score				
not useful/useful				
not useful/neutral				
neutral/useful				
All closures				

The steps for carrying out the study of code changes are as follows.

1. Identify all closures in the L&I database which have closing code differences between the VRC and the CM/VSS. Closures where the VRC did not supply a recommended closing code should be omitted. Closures where the code difference does not cross the boundary of the categories useful/not useful/neutral should be omitted. The study can be limited to one or more years of closures.
2. Identify an expert or experts to review and offer new closing codes, blinded to the original codes.
3. Develop simple forms to capture the results of the expert review.
4. Select a random sample of 50 or more closures.

5. The experts carry out their review and supply their determination of closing codes.
6. All codes are unblinded and the experts re-review cases and determine potential causes for code differences. Share findings within L&I and the VRC community.
7. Statisticians or L&I analyze the coding results and complete Table 11 of this report.
8. The expert panel makes recommendations on changes in forms or procedures.

The second, and easier question that can be addressed from the code change study is this: what kinds of closures are more likely to result in closing code differences between the VRC and the CM/VSS? This is a valuable analysis, as it can lead to a profile of case more likely to lead to coding differences. This profile of difficult cases can alert both the VRC and the CM/VSS to be extra-cautious in coding. It is also possible that there are no notable risk factors for coding disagreements and that they are simply a random occurrence among closures. The analysis will resolve these issues. We recommend that this sub-study be carried out before the expert study of “who is correct” noted just above.

The statistical analysis to address the issue of risk of coding disagreements is similar to that for the case-mix study, and the analysis can be carried out in the same way as for that study. The analysis would be based on data from the L&I electronic files, and no new coding would be needed. In the case-mix study we had binary outcomes such as approved vs. disapproved or yes vs. no. In this code change study the binary outcome would be VRC and CM/VSS agree vs. disagree on closing code. The variables considered as potentially leading to (or causing) code differences can be the same as those considered for the case-mix study, plus there may be a few other variables unique to this study. The results of the univariate and multivariate analysis can be presented in tables similar to Tables 7 and 8 of this report. In Table 7 the column heading “% approved” would be changed to “% with closing code difference.” The statistical analysis methodology is exactly that as presented for the case-mix study.

The steps to carry out this portion of the code change study are as follows.

1. The “laundry list” of variables used in the case-mix study can be used here, and the list can be augmented for this specific study.
2. L&I will need to create a file of all closures during a defined period, such as one year. Closures without a VRC recommended closing code will be excluded. The rows (observations) of the file will represent closures. The columns will include columns (fields) indicating the outcome (codes agree or disagree) for each closure, the numeric closing code, the category of the closing code as coded by the VRC (categorized as useful, not useful, neutral) and two additional columns indicating the numeric closing code and its category as designated by the

CM/VSS. The file will also need to include the VRC i.d. and the firm i.d. for the closure represented by each row.

3. The statisticians or L&I will analyze the data and prepare tables similar to Tables 7 and 8 of this report—both for univariate and for multivariate analysis. Variables which appear to have an important influence on the incidence of coding disagreements will be noted.

Firm quality of work as a proxy for individual vocational provider (VRC) quality of work.

Many VRCs do not have a sufficient number of closures to provide a basis for a precise estimate of the quality indicators. Can the quality of work for the firm be used as a proxy indicator for the VRC? We can carry out an analysis of L&I files to address this question. Briefly, if the VRCs within a firm are all working at a common average quality level (which might be consistently excellent, consistently fair or consistently poor), then the firm could be a good proxy for the VRC's quality of work. If the VRCs within a firm have diverse levels of quality of work, the firm would be a poor proxy. Thus we need to look at the variation in quality of work among the VRCs within a firm. The statistical method of variance components is appropriate for this kind of analysis. As the name "variance components" suggests, it is a method to study variation. We can calculate the VRC-to-VRC variation in quality of work within each firm. If that variation is small, it is good news and the firm can be a good proxy for the individual VRC in the firm.

It is likely that the firm will be a good proxy for the VRC for some firms and a poor proxy for the VRC at other firms. It seems unlikely that all firms will have consistent quality of work among their VRCs or that all firms will have inconsistent quality of work among their VRCs. It is likely to be a mix of firm types. Thus, the outcome of this study might be that for firms X, Y, Z, Q, P, etc., the firm is a good proxy for the VRC, whereas for all other firms it is not a good proxy or else there is not enough cases from the firm to tell whether or not the firm is a good proxy.

The determination of proxy status may vary among the quality indicators. For example, calculation of the percentage of progress reports delivered on time involves a large number of progress report (one per month per worker), and it is likely that almost all firms can be designated as a good or bad proxy for the VRC, due to the large volume of progress reports. On the other hand, for the quality indicator dealing with percent of training plans approved, the caseload is much smaller, and we will be able to rate many fewer firms on their proxy adequacy for this particular quality indicator. Also, we propose dropping the complaint quality indicator, as there will surely be too few complaints per VRC to rate internal consistency of complaint incidence within a firm. Currently, L&I does not use or propose to use proxy scores for the complaint and dispute quality measures.

Using the variance components method we can calculate a numeric measure of VRC-to-VRC variation in quality of work. We can express that measure as a quantity similar to a standard deviation (SD), so that a SD of 3%, for example, for the useful outcomes

indicator, would tell us that the firm has quite consistent quality of work among its VRCs on this quality measure. Most of the VRCs in the firm would have a "true" useful outcomes rate within two SD units of the average for the firm, i.e., within ±6% of the firm average. An SD of 10% or 20% for a firm would tell us that the VRCs have very inconsistent quality of work. Further, we can place confidence bounds ("margin of error") on the SD, so we will be able to distinguish between firms of which we are confident of their internal consistency and designate other firms for which we are confident of their internal inconsistency. Yet other firms may have caseloads that are too small to designate either consistency or inconsistency with confidence.

Our results may take different forms, but Table 12 is a likely way to express some of the results. Firm A in the table has fairly consistent quality from VRC to VRC as indicated by the modest SD of 4% (all hypothetical data). There would be a version of Table 12 for each quality indicator. The variance components methodology is applicable to firms with varying numbers of closures per VRC, so, for example, the method can be used for a firm where one VRC has 200 closures, another has 20 and another has 4, and so on. It would deal equally well with a firm where there are 15 VRCs, each with only 5 closures. The latter firm does not have any single VRC with enough closures to enable calculation of a reliable performance percentage for each specific VRC, yet this methodology can still be used to estimate the degree of consistency of performance among the VRCs within the firm.

Table 12. Results of firm proxy study for useful outcomes (hypothetical values)

Firm Name	Total caseload	No. VRCs	Cases per VRC: mean±SD	Min-max cases per VRC	Firm mean useful outcomes	SD of % useful outcomes across VRCs
Firm A	132	10	13.2±6.1	3 - 25	78%	4%
Firm B	etc.	etc.	etc.		etc.	etc.
Firm C	etc.	etc.	etc.		etc.	etc.
etc.	etc.	etc.	etc.		etc.	etc.
etc.	etc.	etc.	etc.		etc.	etc.
etc.	etc.	etc.	etc.		etc.	etc.
etc.	etc.	etc.	etc.		etc.	etc.

The steps to carry out this firm proxy study are as follows.

1. L&I will need to create a file of closures (or of disputable closures) for each quality indicator except complaints. Each row will have one closure and will include the unique VRC i.d. and the unique firm i.d. along with the value of the quality indicator for the closure (e.g., useful/not useful/neutral) and the date of the closure. Closures for a period of one year may be adequate for some quality indicators, while closures from two or three years may help the analysis for other closures. L&I and the statisticians will work together on the specifications of these files.

2. The statisticians or L&I will carry out the analysis and prepare tables similar to Table 12 of this report.

Combining similar indicators to increase sample size (e.g. Plans and Eligible).

The consensus from a committee of LNI staff and VRCs is that combining quality indicators does not make sense vocationally. This study will not be pursued.

Current format of the indicators.
Considerable work has gone into developing, computing and displaying the quality indicators. The displays (which one of us, NLP, has reviewed) are meaningful and clear. The presentation is primarily textual, which means that the viewer has to read numbers and other information in order to process it. We recommend some graphical aids as a means of quick communication of quality levels.

We recommend that L&I consider color and size as additional ways to represent the quality level. For example, the useful outcome indicator could have a stack of blocks that are up to 10 blocks tall. If there is one block (a score of 1), it could be red, two blocks (score = 2) could have a color (for both blocks) that is moving toward blue, nine blocks (score = 9) could be almost pure blue (all blocks) and ten blocks would be pure blue. The confidence level could be shown by the boldness of the color or by speckled vs. non-speckled color. The advantage of the color and boldness is that there is no thinking or information processing to receive the information. It is visual and instantaneous. The other indicators can also follow this color scheme.

SHORT AND LONG TERM PLANS

Our short-term plan is to discuss the various proposed studies with L&I staff and determine the sequence of studies to be carried out.

Our long-term plan is to carry out the designated studies in collaboration with L&I staff. Our longer-term plan is to design studies to assess the value of the quality indicators by finding if they are predictive of external, measurable outcomes

A SEPARATE LETTER ATTESTING TO THE CONTRACTOR'S INDEPENDENCE IN RELATION TO THE REPORT

To be attached.

UNRESOLVED ISSUES

A number of issues yet to be settled were raised in our description of the various studies. Please see each study.

LESSONS LEARNED.

These studies are a major effort and an investment by L&I. Completion of these studies will help to put the quality indicators on firmer ground. The results of the studies may lead to some modification of the indicators, to modification of procedures within L&I and to modification of procedures within the VRC community. On the other hand, the studies may show that some or all the activities covered by each study do not need further attention or fixing.