Final Report

Local Public Health Department, Objectives Evaluation

Contract 155090

May, 2003

To:
State of Wisconsin,
Department of Health & Family Services,
Division of Public Health

By:
Wisconsin Public Health / Health Policy Institute,
University of Wisconsin-Madison

Robert Stone Newsom, PhD
Principal Investigator

Table of Contents

Measure Twice, Cut Once:
Measuring and Evaluating Objectives in Performance-Based Contracting.

Executive Summary:

The State of Wisconsin Division of Public Health (DPH) in conjunction with the Wisconsin Public Health / Health Policy Institute at the University of Wisconsin - Madison has completed efforts to develop and test a measurement methodology based on collective expert knowledge that allows "non-experts" to reliably evaluate Local Public Health Departments (LPHD) and contracted provider's performance objectives. The performance objectives form the core of an ongoing state initiative to move public health financing from a traditional cost-based model of reimbursement to a negotiated, cooperative, performance-based contracting process. The LPHD providers use performance objectives to succinctly describe the proposed service for which they are requesting funding and DPH must measure and evaluate each performance objective as a part of determining which services to fund. A reliable and valid measure of each performance objective is the key to reliable evaluation of LPHD proposals and the cornerstone to DPH efforts to become an efficient and effective buyer of public health services. In 2002 the DPH leadership published this statement, "*Developing good objectives for all programs and LPHDs is still in the future and is, unexpectedly, the most difficult implementation issue."*

To address this issue, a team of LPHD representatives, DPH staff and leadership and university experts leveraged a modest CDC Region V grant ($18,236), over a 15 month period, to meet the following goals: *Provide scoring of three years of LPHD performance objectives using the SMART criteria and the CDC Logic models and analyze the scores for differences / similarities both between identified groups of submitters and for changes in scores across the three year time interval.*

During Phase I quantifiable advances and results were made in:

- Documenting the art and science of conceptualizing, defining and describing individual components of the SMART and Logic Models so a scorer can reliably determine whether a component is present in any given performance objective.
- Developing, through an iterative process of meeting, defining, testing and re-defining, a computerized application that allows non-expert scorers to reliably judge the "quality" of the performance objectives currently in the database.
- Documenting that experienced scorers test-retest reliability was significantly higher than for inexperienced scorers, particularly for the Logic Model components of Process and Output - practice counts!
- Contributing new and significantly redefined improvements to the training and contracting processes ongoing at DPH - exact definitions are critical.
- Finalizing a measurement instrument that provided quantitative analysis and evaluation of progress in performance objective writing across time, geographic locations and types of public health providers.

During Phase II the evaluation instrument developed in Phase I of the project was used to score 4874 performance objectives. Analysis of the scoring data revealed:

- As a group, Wisconsin's Local Public Health (LPHD) providers significantly improved their ability to write performance objectives that met the SMART criteria between 2000 and 2001. This improvement was maintained at the same high level (99%) for the Measurable and Timeframe components, but fell back for the Specific component from a high of 95% in 2001 to 70% in 2002.
- As a group, Wisconsin's LPHD providers were relatively unsuccessful at improving their ability to write performance objectives that met the Logic Model criteria between 2000 and 2001. Almost none (8 total) of the performance objectives met the scoring criteria for Outcomes; whereas the ratio of Output to Process objectives fell from an unimpressive 1:2.3 ratio in 2000, to a 1:3.7 ratio in 2001 and then slightly improved again to a 1:2.7 ratio in 2002.
- In total scoring (range 0-6) 70% of the objectives received a total score greater than 4; however 50% of the scores reflected objectives meeting all the SMART criteria but only the minimal "Process" Logic Model criterion.
- The relative performance rankings of individual LPHD groupings were quite mixed from year to year. Strong performance on one of the two models was only weakly associated with performance on the other. Performance inconsistency was also apparent when combining the two models into a total combined score. The total scores did continue to illustrate the drop in performance between years two and three. With current information, further analysis was unable to disentangle whether the fall in performance was a result of actual differences in learning / retention levels among the groups, or differences in their training, or how recently their training occurred, or all of the above.

Based on the results of Phase I and II the POEM research team recommends:

1. If DPH plans to continue to use the Logic Model scoring procedure internally to validate progress or the quality of the performance objectives, then further definitional work and testing should be done on the components. In addition, the state appointed "scorer" should be trained in the concepts behind both the two models and allowed 2-3 testing trials where an experience staff member reviews their scores and provides feedback on their understanding / recognition of the components.

2. Both the Phase I and II processes and outcomes reinforced the concern that there remains continuing confusion over Outcome, Output and Process criteria. If this is true, then any actual differences or similarities between LPHD groupings using the Logic Model are likely to be obscured by the measurement error that was in turn driven by the definitional ambiguity reflected in the expert committee. Since almost no "Outcome" objectives were identified during Phase II, at minimum, the expert committee should review the definitions of "Outcome" criteria and re-score the objectives on this criterion.

3. An internal study should review the time lag between training, or re-training and when the objectives were written. This review will help resolve the issue of whether differential "time since training" was the factor behind the observed performance drop in year three.

4. Given the performance differences between groupings, administrative effort should be directed towards ensuring the similarity of the content and methods utilized by the trainers over time.

5. DPH should develop and include in the database persistent LPHD provider identifiers so future research can utilize a matched-pair analytical design.

6. If the scoring methodology is put into practice without further refinement, all performance objectives with a total score of 4 or higher should be expertly evaluated.

Measure Twice, Cut Once:
Measuring and Evaluating Objectives in Performance-Based Contracting.

Robert Stone Newsom, John Chapin, Sherry Gehl
Wisconsin Division of Public Health: University of Wisconsin-Madison

Prologue:
In 1999, following years of prior administrative and political background effort, the State of Wisconsin Division of Public Health (DPH) began implementing a performance-based contracting process with the Local Public Health Departments (LPHDs). This process[1] (See Chapin & Fetter, 2002) replaces a traditional cost-based model of reimbursement with a negotiated, cooperative, performance-based model. The new model has at its core the surprisingly complex task of writing, measuring and then evaluating a "good" performance objective. The LPHDs must use the performance objective to succinctly describe the proposed service for which they are requesting funding (selling). The Division of Public Health, on the other hand, must measure and evaluate each performance objective to determine which services to fund (purchasing). This paper summarizes DPH continuing efforts to develop a measurement methodology based on collective expert knowledge that allows "non-experts" to reliably and fairly evaluate LPHD performance objectives.

In addition to a myriad of administrative challenges, the new contracting system required training state and local public health staff to be buyers and sellers, to negotiate, and to write performance objectives that describe outputs or outcomes rather than processes. In the first year (CY2000) the training and writing of performance objectives resulted in over 1200 objectives. The second year contracts (CY2001) were greatly delayed because the DPH realized that the LPHDs and DPH staff still were not able to write good objectives. For the LPHDs, regional offices and central office DPH staff to agree on what a good objective was, took months. Once the contracts were settled, a informal review of the objective in CY2001 showed that many were still weak and poorly written. The DPH attempted to edit them for style and language but this led to hard feelings from some LPHDs who felt the editing had changed the contract's intent and workload. Following ongoing training another 1200 objectives (for a total of 4874) were written and delivered to DPH for CY2002.

Chapin and Fetter published a review of Wisconsin's performance-based contracting experience in 2002. The review, among other things, concluded that one of the effects of the new system was as follows: "*Forcing the LPHDs and state agencies to produce output and outcome objectives, deliverables, and measures of impact is training experience that should help them apply this objective-based approach to public planning generally. Three year of objectives have now been drawn up for many programs and LPHDs. Developing good objectives for all programs and LPHDS is still in the future and is, unexpectedly, the most difficult implementation issue*." The authors go on to describe an evaluation completed in the summer of 2000 by all the LPHD directors participating in the program. They state "*70 percent agreed that this contracting process*

[1] Chapin, J., Fetter, "Title", 2002

was an improvement over the old method. But hundreds of details, frustrations, and bugs, still to be worked out through trial and error, have alienated many LPHD directors". They go on to say *"the opponents of this process are still numerous, as no one really likes to be held to performance standards, especially, when the implementation processes are still experimental."*

In response to this "*most difficult implementation issue"* of writing, measuring and evaluating performance objectives the Wisconsin DPH acquired a small grant in late 2001 from the Region V Office of the CDC. In turn, the DPH contracted (from January, 2002 through August, 2002 and extended through May, 2003) with the Wisconsin Public Health Health Policy Institute at the University of Wisconsin-Madison Medical School to:

GOAL "*Provide scoring of a sample of three years of LPHD performance objectives using the SMART criteria and the CDC Logic models and analyze the scores for differences / similarities both between identified groups of submitters and for changes in scores across the three year time interval."*

Funds were sufficient to contract a (5% FTE) Senior Scientist (an Educational Psychology PhD specializing in measurement and evaluation) to manage the project and direct the evaluation and measurement efforts, and a (50% FTE) graduate project assistant to do the actual scoring. The DPH and the Institute financially absorbed the associated costs of additional graduate student scoring assistance and "expert committee" assistance. The effort acquired the name "Performance Objective Evaluation and Measurement", or (POEM) project, and functionally consisted of two phases: With the assistance of an expert committee Phase I developed and tested definitions and a computer interface of the SMART and CDC Logic Model criteria that the scorer would use to evaluate the objectives. Phase II scored and analyzed three years of performance objectives.

Phase I

A POEM team was assembled consisting of a public health regional director, several LPHD local directors, four members of the DPH staff most involved in the performance contracting process, the DPH Administrator and the university contractors. The team's charge was to agree, as they applied to performance objectives, on the definitions of the different components of the *Specific, Measurable, Achievable, Realistic, Time* (SMART) and CDC Logic Model's - *Inputs, Process, Output, Outcomes*. They reasoned that the component definitions would then serve as a means of scoring each performance objective - e.g. the scorer decides whether or not each performance objective contains wording meeting the component definition (see Appendix I). In essence a "good" performance objective would contain, or meet the definition of, all the components while a really bad one would contain no components. Change in the ratio of good to bad performance objectives both within and across different groups of performance objective authors would thus provide the DPH team with measurements from which they could consequently evaluate the:

- Continued validity of the expert consensus of component definitions
- Success of past performance objective training programs over time,
- Improvements in performance objective writing by various types of agencies,
- Geographic regions and / or organization groupings where more training was required,
- Use of the refined component definitions to provide a consistent and tested base for further training at the LPHD level.

It was assumed that the refined component definitions would provide a consistent and tested base for further training at the LPHD level. Creating a reliable and valid scoring criterion allows the DPH to train, so to speak, for the test rather than current practice where the trainers and evaluators may be working from slightly different definitions.

During development a major insight was that the set of component definitions, and thus their recognition during scoring, are conceptual rather than factual. The utility to the team was in understanding why any lack of consensus regarding a component definition would, during scoring, necessarily be magnified into unreliable application / recognition of that definition. Inexact definitions will be applied inexactly. Teaching and learning to recognize a fact is much less complex than teaching a concept and then asking people to recognize that concept in action. Comparing intra-scorer trials of the same 100 performance objectives neatly illustrated how those components that gave the team the most trouble during the definition stage were identical to those components that were the most unreliably scored. The Specific component from SMART model during early testing and the Output & Process components from the CDC model throughout the testing phase were the strongest examples of this phenomenon.

The component definitions were built into the (Microsoft Access 1997) "scoring interface" developed by DPH computer personnel to allow rapid Yes-NO scoring of the performance objectives already contained in an Access database. The plan, which proved quite workable, was to have the scorer read the performance objective on the screen and then click their choice from an option set (YES, NO, UNSURE) as to whether the performance objective met each SMART or CDC Logic Model criteria. The "definitions" remain on the screen beside each scoring option set as the performance objectives change. In this fashion, the scorer can refresh their memory as to how each component (Specific, for example) was defined and consequently refresh their memory regarding what phrases, key wording, or concepts for which they are searching. In an educational psychological sense the scorer is presented with a series of concept recognition tasks imbedded within a descriptive prose sentence. As mentioned earlier, the reality that the recognition target is an example of a concept in action rather than a simple fact lies at the heart of why the training, development and evaluation of performance objectives had become the "most difficult implementation issue".

The POEM team met five times. Following three of the meeting, two different graduate project assistants scored a sample of 100 performance objectives using the interface and the results where compared for inter- and intra-reliability. The first meeting resulted in a set of goals for the scoring instrument and a set of scoring requirements (see below) plus

a first definition of the SMART components. Meeting two evaluated the results of the SMT testing and defined the components of the CDC logic model. Meeting three evaluated the testing of the POO model, reviewed the repeat testing of the SMART components and evaluated current component definitions in an effort to determine why test - retest reliability continued to be both low and inconsistent. Meeting four redefined and rationalized the combined model components in preparation for further reliability testing. Meeting five focused on the Output and Process components since their inter-scorer reliability remained elusive. Testing following the fifth meeting produced sufficient reliability (see Hypothesis page 11), with the experienced scorer, to proceed to final scoring of all 4874 performance objectives.

Much of the effort in all of the expert meetings revolved around the complicated task of agreeing on when a component definition as conceived was actually contained within a real performance objective example. Although the committee always produced a set of component definitions, which the university contractors would then test, the committee did not reach 100% consensus on every performance objective example they "scored". This lack of consensus in application of the component definition was vividly illustrated during early testing in the low degree of, and particularly inconsistent, inter- and intra-scorer reliability. The final objective scoring results also further reflected this with the "Outcome" Logic Model component.

The goals for the measuring instrument (tool) were:

- Provide an overall determination of whether the quality (defined by SMART & CDC Logic model) of written performance objective's is annually improving statewide.
- Determine which "group" (defined by a year, region, organization, location) of writers /developers may need additional performance objective instruction / assistance.
- Determine which of the individual SMART or Logic Model components could actually be applied and scored.
- Determine which component might need more work among the groups.
- Provide, in addition to these measurement and evaluation outcomes, a method for ongoing and initial *training* of performance objective authors.

The scoring requirements included:

- The scoring could be accomplished by a "organizationally naïve" scorer - that is, one who knew nothing about the organization needs and functioning of the Division, LPHDs or the performance objective process.
- The performance objective must stand alone, - that is, the scorer must be able to make their determinations without referring to any further information outside of the performance objective / testing interface itself.
- Scoring will amount to a decision of "Yes" the performance objective meets the criteria (for example, Is the performance objective "**S**pecific?") or "No" it does not. A third responses "Undecided" was added to alert DHFS staff that the performance objective in question was outside the training and experience of the

scorer; however, in practice it seemed to produce more scorer confusion than less and was later removed.
- Scoring will take place using a PC based Access application built by DPH

Pruning the criteria:
During the first meeting the team discussed at length the feasibility of building a tool that would allow a naive scorer to make decisions about whether a performance objective was "**A**chievable" and / or "**R**ealistic". After much discussion it was acknowledged that these two criteria (A&R) were prospectively a "best guess" or probabilistic statement made by experts with considerable prior knowledge. Retrospectively, A&R are fairly easily determined by reviewing whether the performance objective reached its goals, or was fulfilled, within the resource limitations (time, money, people) proposed. By definition (see scoring requirements) a naïve scorer would not have expert prior knowledge and would not have available the "result / outcomes" of the performance objective as they were reading it. Consequently the committee decided that of the five (SMART) criteria, the present POEM tool would only include **S,M & T**. A second meeting was convened to review the SMT testing and to define the logic model components of Process, Output and Outcome (POO).

Redefining criteria:
Following testing of the combined SMT + POO components observations made by the lead scoring individual regarding possible reasons for the disappointing reliability comparisons led the team to take a fresh look at the components. Indeed, it was obvious on inspection that the S component of SMT and the Output component of the POO were essentially looking for the same concepts. This observation led to a fourth meeting devoted entirely to a detailed word for word analysis and redefinition until full census was reached on all six of the SMT + POO components. These new definitions were then reprogrammed into the scoring interface and were tested on a new set of 100 performance objectives by both the seasoned scorer and by a new scorer selected from DPH staff. The fifth meeting focused only on the POO components and was again followed by another round of testing using seasoned and naïve scorers.

Reliability testing trials:
Seven different scoring trials were held between April and October. Trials one and two (T1& T2) were a test-retest situation using two scorers. Between T1 & T3 a short training session occurred. Trial 3 (T3), again using both scorers, was held two weeks later to review whether the weak correspondence across scorers might have been introduced by the scoring session. Following more expert committee development on the Specific component and new work on the POO components trials four & five (T4 & T5)) was held in early September, 2002. T4 & T5 were identical test-retest situations a week apart and included two scorers. Finally after further work on the Output and Outcome components of POO trials six and seven (T6 & T7) also consisted of identical test-retest situations a week apart with two scorers. One scorer took part in all seven trials, the second scorer was the same for trials T1-T3, and a different scorer was include in trials T4-T7. Using this mix of same and different scorers allowed testing of assumptions / hypothesis

regarding "training effects" and whether the goal of using "naïve scoring personnel" was realistic.

Trial comparison methodology:
Following each trial, intra-scorer pattern comparisons were examined for both *total* component patterns (SMT or POO) and *individua*l patterns (S or M or T, P or O or O). Inter-scorer total and individual patterns were also compared in trials T1-T3. In keeping with common test-retest reliability statistics, correlations were originally considered to be the comparison measure of choice and hypothesis were developed using them. Our observations, in concert with the published literature, however soon indicated that the 0,1 nature of the data strongly distorted such correlations. Instead we fell back to a descriptive statistic referred to as the "percentage of similarity". This statistic describes the percent of identical responses found in two lists of scores - for example with 5 sets of scores (00, 11, 01, 00, 11) only 1 in 5 or 20% are different, consequently the reliability percentage would in this case be 80%. Therefore, when using the percentage of similarity statistic, the higher the percentage the higher the similarity or "reliability" of the scoring. As anything above a correlation of .75 is considered very good and / or a high association, we arbitrarily set a percentage of similarity as >=.85 as the gold standard we would strive to meet in our hypothesis testing.

Hypothesis:
It was hypothesized that the final outcome of the testing and training would provide numerical data indicating that in a test-retest trial using the same 100 POs:

1. In comparison of <u>total component</u> SMT patterns a single scorer (intra - reliability) would show a percentage of similarity of >= 85%.
2. In comparison of <u>individual component</u> SMT patterns a single scorer (intra - reliability) would show a percentage of similarity of >= 85% for each component.
3. In comparison of <u>total component</u> POO patterns a single scorer (intra - reliability) would show a percentage of similarity of >= 85%.
4. In comparison of <u>individual component</u> POO patterns a single scorer (intra - reliability) would show a percentage of similarity of >= 85% for each component.
5. In comparison of <u>total component</u> SMT patterns, the differences *between* scorers (inter-reliability) would show a percentage of similarity of >= 85%.
6. In comparison of <u>total component</u> POO patterns, the differences *between* scorers (inter-reliability) would show a percentage of similarity of >= 85%.

Preliminary results and Support of Hypothesis:

Hyp. Number	**Accept / Reject**	**Comments**
1	Accept both	*Scorer1 = 92% **Scorer2 = 83%
2	Accept both	Scorer1: S=96%, M=99%, T=97% Scorer2: S=89%, M=95%, T=96%
3	Accept scorer1 Reject scorer 2	Scorer1 = 85% Scorer2 = 81%

4	Accept scorer1 Reject scorer 2	Scorer1: O=98%, O=88%, P=87% Scorer2: O=97%, O=80%, P=78%
5	Reject	Scorer1 vs. Scorer2 percentage of similarity = 76%.
6	Reject	Scorer1 vs. Scorer2 percentage of similarity = 62%.

*Seasoned scorer, **Naïve scorer

Reliability Conclusions:
The results offer strong evidence that both trained and naïve scorers can reliably identify the presence / absence of the SMART components in performance objectives using definitions developed by expert committees and displayed during the computerized scoring. Trained scorers, but not the naïve scorers, are also able to reliably identify the presence / absence of the CDC logic components developed and presented in the same manner.

The results also offer strong evidence that high *intra-scorer* reliability, does not also guarantee high *inter-scorer* reliability. Neither available resources nor the study design are sufficient to provide a causal explanation for this phenomenon. However, after observing the difficulties the expert committee often had in reaching consensus as to when their own definitions were present / absent in performance objectives two interrelated possibilities present themselves. One, that each scorer reliably recognizes what he or she understand to be true, but their understanding is different - same instructions, same ability but different interpretation. Two, even though they do what they do reliably, one of the scorers is simply more able (cognitively capable) to do the task than the other; this implies that one scorer is reliably accurate (right) while the other is reliably inaccurate (wrong). Were, as mentioned earlier, the task factual rather than conceptual in nature possibility number two would be easy to check. One could compare the two scorer's results with the "right" answer. While further research could clarify this issue, the right answer remains a matter of expert opinion and the experts occasionally disagree!

Inter- vs. Intra-scorer reliability
For the purposes of this evaluation effort (looking for change in the quality of performance objectives among different groups over time) the intra-scorer reliability is most crucial. In statistical terms it is a bias vs. variance tradeoff. As long as a single individual scores all of the performance objectives during a relatively short time period and that person reliably reacts the same way to the same performance objective wording, then all the scores will have the same "slant". That is, the scorer may be incorrectly applying a component description, but they will be doing so reliably. Therefore, the source of differences between groups or across time should be a function of factors residing in the groups and time not a function of scoring error / unpredictability. With high intra-scorer reliability, DPH staff can determine what, if any, concept recognition scoring errors are being committed and account for that in their evaluations.

Resource crises and solution:

Prior to the third meeting it had become clear that the uneven reliability testing results in conjunction with the necessary time delays incurred with convening and re-convening the expert team had together nearly exhausted limited project resources. DPH leadership were given the option of whether the project should proceed within the contract time frame, recognizing that would surely provide questionable / unreliable results, or allowing a contract time overrun with the more likely results of obtaining a reliable / accurate evaluation. The importance to this multi-year project of having reliable measures that would then allow consistent and fair evaluations of the performance objectives was critical. The decision was made to make the measurement tool as reliable as possible and to further reduce sources of error in the final scoring by eliminating the plan to score a sample of performance objectives in favor of scoring them all. These decisions implicitly acknowledge that insufficient funds exist for finishing the project. Out of common purpose and good will the Institute offered to absorb, within limits, the Senior Scientist overrun, and a plan was devised to use internal DPH staff to do the scoring. The final overrun was calculated at $11,400. These efforts culminated in trials T4-T7, the successful results of which were reported in the Result & Hypothesis acceptance section.

Phase I Conclusions:
SMT component scoring reliability was high and very steady with the experienced tester; the inexperienced tester reliability was not as high or steady. This finding reinforces "common sense" thinking that says, "Practice and familiarity with the scoring tool and the concepts behind it will result in more reliable scoring."

The POEM project was successful in developing a user interface that provided sufficient definitions of the SMT components to allow a naïve scorer to reliably score performance objective.

POO component scoring reliability improved for every component for both scorers indicating that changes made to the definitions by the committee produced the desired effect. The Outcome component was acceptably high for both scorers (although almost no such components were identified in actual practice).

The significantly higher Process & Output scorers for the experienced scorer over the inexperienced scorer combined with the significantly lower scores for these components compared to the Output component indicates that the definitions are still not working as well as they might be. In other words in spite of the identical definitions provided to both, the experienced scorer was bringing additional information (training / knowledge) to the task that was unavailable to the naive scorer.

The project was unsuccessful, although clear improvement was evident after each scoring- team committee meeting, in developing Output & Process definitions that allow a naïve scorer to reliably score the performance objectives. The experienced scorer's reliability with these components was significantly higher but they are not comparably as strong as the SMT component reliability.

The practical outcome of the continuing confusion over Output & Process will be that any comparisons (time, groupings or components) of overall scores in writing performance objectives will need to include a relatively high error correction. This means that any actual differences or similarities may be obscured by the need to account for measurement error - the higher the error the greater the chance that no difference will be observed, or those that are observed are inaccurate.

Recommendations:
Given infinite resources, the recommendation would have been to convene the scoring committee and test-retest "one more time" to see if improvements might be made to the Output & Process definitions. Given that the budget was exhausted, the recommendations were:

- Proceed with final scoring (all POs to date) using the experienced scorer

- If DPH plans to continue to use the POO scoring procedure internally to validate progress or the quality of the performance objectives, then further definitional work and testing should be done on the Process / Output components. In addition, the state appointed "scorer" should be trained in the concepts behind the SMT / POO models and allowed 2-3 testing trials where an experience staff member reviews their scores and provides feedback on their understanding / recognition of the components.

Executive Summary - Phase II

This paper reports on Phase Two of the Performance Objective Evaluation & Measurement (POEM) project. Methods, results, discussion and recommendations are presented based on the application and evaluation of the performance objective scoring methodology, developed during Phase One, to three years of objectives.

Broadly, the research question was "Are the different categories of local public health providers able to write performance objectives that meet the defined SMART and Logic Model criteria, and has their ability to do so improved over the years of the program?" Necessarily, the results also address the DPHs trainers and training procedures as they prepare local public health providers in the writing of objectives.

The results of the analysis and evidence suggest the followings:

- As a group, Wisconsin's Local Public Health (LPH) providers significantly improved their ability to write performance objectives that met the SMART

criteria (SMT) between 2000 and 2001. This improvement was maintained at the same high level (99%) for the Measurable and Timeframe components, but fell back for the Specific component from a high of 95% in 2001 to 70% in 2002.

- As a group, Wisconsin's LPH providers were relatively unsuccessful at improving their ability to write performance objectives that met the Logic Model criteria (OOP) between 2000 and 2001. Almost none (8 total) of the performance objectives met the scoring criteria for Outcomes; whereas the ratio of Output to Process objectives fell from an unimpressive 1:2.3 ratio in 2000, to a 1:3.7 ratio in 2001 and then slightly improved again to a 1:2.7 ratio in 2002.
- In total scoring (range 0-6) 70% of the objectives received a total score greater than 4; however 50% of the scores reflected objectives meeting all the SMART criteria but only the minimal "Process" Logic Model criteria.
- The relative performance rankings of individual LPH groupings were quite mixed from year to year. Strong performance on one of the two models was only weakly associated with performance on the other. Performance inconsistency was also apparent when combining the two models into a total combined score. The total scores did continue to illustrate the drop in performance between years two and three. With current information, further analysis was unable to disentangle whether the fall in performance was a result of actual differences in learning / retention levels among the groups, or differences in their training, or how recently their training occurred, or all of the above.

The researchers recommend that:

1. The expert committee should review the definitions of "Outcome" criteria and re-score the objectives on just this single criterion.
2. An internal study should review the time lag between training, or re-training and when the objectives were written. This review will help resolve the issue of whether differential "time since training" was the factor behind the observed performance drop in year three.
3. Given the performance differences between groupings, administrative effort should be directed towards ensuring the similarity of the content and methods utilized by the trainers over time.
4. Develop and include in the database persistent LPH provider identifiers so future research can utilize a matched-pair analytical design.
5. If the scoring methodology is put into practice without further refinement, all performance objectives with a total score of 4 or higher should be expertly evaluated.

Introduction:

As part of their ongoing effort to improve the a performance-based contracting process with the Local Public Health Departments (LPH) providers the State of Wisconsin Division of Public Health (the DPH) contracted with the Wisconsin Public Health / Health Policy Institute to assist in measuring and evaluating the objectives that form the core of the performance contracting process. This effort, know as the Performance Objective Evaluation & Measurement (POEM) project consisted of two phases: Phase I, the development of the measurement methodology and Phase II, the application and evaluation of this methodology to three years of objectives.

An earlier paper[2] summarized the phase one DPH efforts to develop a measurement methodology based on collective expert knowledge that allows "non-experts" to reliably evaluate LPHD performance objectives. Those efforts resulted in a set of definitions, agreed upon by the expert committee, and tested against actual POs until the test-retest reliability was greater than 85% for each of the three SMART (**S**pecific, **M**easurable, **T**ime) and three CDC Logic Model (**O**utcome, **O**utput, **P**rocess) criteria. The scoring itself then consisted of a computer presentation of each PO along with descriptions of each SMART and Logic Model criteria. The scorer would 'click' on a Yes answer if the PO met that criteria or a NO answer if it did not. At approximately 1.5 minutes per PO the scoring of 4,874 objectives consumed about 125 hours over a 45-day period.

This paper reviews the POEM project phase two. It presents and analyzes the results of employing a trained, but non-expert, individual to score all of the performance objectives (POs) submitted by the LPH providers from 2000 through 2002. A brief review of the task facing the scorer may be a useful precursor to reviewing the scoring results.

The scoring of POs is basically a conceptual recognition task. With exception of the "Time" criteria, the scorer is not looking for facts, figures or specific words as one would in an object recognition task. Instead the scorer is looking for concepts (meanings) that may be stated or phrased in a myriad of ways. The scorer must read the objective and then decided whether the objective contains, however worded, the concept that lies behind (was defined by the expert committee) each SMART or Logic Model criteria. The process is effectively the reverse of the process the authors of the POs go through in generating an objective. They have learned (been taught by expert teams) the concept behind each criteria and their task is to produce an objective that meets the requirements of those concepts. Clearly there are opportunities within these cycles of training and learning for differences in understanding, retention and thus performance. The expert committee, the individuals that do the training, the LPH providers and the scorer may all generally agree on what, for example an "Outcome" objective should look like. However, when it comes time to write and then score any given objective, the possibility remains that these same individuals may perceive the form or existence of that Outcome objective differently. In many ways it is the reliability of this "perception of the existence of a concept" that phase two results address.

[2] Newsom, Robert Stone, Chapin, John "Measure Twice, Cut Once: Measuring and Evaluating Objectives in Performance-Based Contracting. Final contractor Report to the Wisconsin Division of Public Health, Department of Health and Family Services, Madison, WI , 5/16/2003.

Methods:
The analysis was designed to produce two sets of statistics. The first is a descriptive analysis of the distribution of the POs as reflected in the many permutations of SMART and Logic Model, local public health (LPH) provider descriptive categories and years. The second is a comparative analysis of the SMART and CDC Logic Model scoring results across years and LPH provider criteria. Generally the question of interest was "Are the different categories of local public health providers able to write PO's that meet the defined SMART and logic model criteria, and has their ability to do so improved over the years of the program?" Necessarily, the corollary of this question is whether the DPH has improved in its ability to train the local public health providers in the writing of POs that meet criteria? Finally, the research gathers evidence as to whether this method might increase transparency and objectivity in the performance objective evaluation process. These general questions were translated into three Null Hypothesis.

Null HYP 1: There are no differences between the percentage of total POs produced across groups* for the different SMART (SMT) criteria or logical model (POO) criteria.
Null HYP 2: There are no differences between mean scores among the different groups* within a single year or across years.
Null HYP 3: There are no differences between mean scores of a single group* across years.

**A group consists of those PO that are defined as having in common one selection from each of the following characteristics*:

- YEAR of objective: (Y1, Y2, Y3)
- LEVEL of organization: (L1, L2, L3)
- Geographical setting (Rural, Urban)
- Population Density (Metro, Non-Metro)

For example, a single "group" definition might then look like (Y1,L1,R,NM) which would translate into POs from "year 2000, written by a Level 1 local public heath provider that represented or served a Rural, Non-Metro area. The observations to be made may thus be logically illustrated (Table I) where each cell shows the four independent group observations made on the intersection of Year and Level. Note however that the group observation of Rural+Metro is a contradiction of terms and no such characterized POs were found in the database. The analysis consequently focused on the 27 permutations that were available.

Table I:

	Year 2000	Year 2001	Year 2002
Level 1	Rural, Metro Urban, Metro Rural, Non-Metro Urban, Non-Metro	Rural, Metro Urban, Metro Rural, Non-Metro Urban, Non-Metro	Rural, Metro Urban, Metro Rural, Non-Metro Urban, Non-Metro
Level 2	Rural, Metro	Rural, Metro	Rural, Metro

	Urban, Metro Rural, Non-Metro Urban, Non-Metro	Urban, Metro Rural, Non-Metro Urban, Non-Metro	Urban, Metro Rural, Non-Metro Urban, Non-Metro
Level 3	Rural, Metro Urban, Metro Rural, Non-Metro Urban, Non-Metro	Rural, Metro Urban, Metro Rural, Non-Metro Urban, Non-Metro	Rural, Metro Urban, Metro Rural, Non-Metro Urban, Non-Metro

The first descriptive analysis consisted of counting and finding the percent of total for each individual criteria (e.g. all Level 1's), as well as finding the count and percent of total for PO's matching each of the 27 combinations of criteria.

The second analysis consisted of calculating the mean and variance of the various permutations of LPH characteristics for each scoring criteria. These permutations (scores X years X LPH provider characteristics) could then be arranged to illustrate:

1. Variation within a LPH provider characteristic by year and score level. For example, "Within Level 3 providers, What was the average SMT+OOP score across years"?
2. Variation within a year by score level and LPH provider characteristic. For example, "For 2002, what was the variation in average scores across provider varieties"?
3. Variation within a score level by year and LPH provider characteristic. For example, "What was the variation in average OOP scores across years for each LPH provider variety"?

The second analysis also included transforming the raw average scores into ranks and presenting these ranks in both table and graph format. By definition, ranking eliminates variability in scaling between data points in favor of equidistant scaling and clarifying relative position. For example scores of 1.6, 1.7 and 1.8 would be ranked 1,2,3 just as would scores of 4, 6 and 9. Nevertheless it was decided that since one of the research goals was to tease out differences in the relative performance of LPH provider grouping, and the overall variation in scores was limited, the ranking presentation might contribute to better understanding.

A computer program (Visual FoxPro 6.5) was written to apply the numerical values of the scoring system to the Yes-No answers recorded in the databases, and also to calculate and print the counts, percents, means and variances that make up the results tables. The raw output, prior to table development, is available in Appendix II.

A Note on Scoring:
The initial experimental design called for sampling from the population of POs for each cell. However it became clear that waiting on the DPH for access to the database to determine cell sizes and score variations within the cells would exhaust already limited time and resources that needed to be directed to the scoring itself. Consequently, the research team agreed to score the entire universe of POs. Although this decision extended

scoring time, a major advantage was gained by eliminating both the sampling error, and consequent doubt, the initial design would have included

The scoring produced a similar matrix to Table I of (Level X Year) independent group observations with the variable being either SMART (SMT) or Logic Model (OOP) scores. The SMART model components S,M & T are additive (achieving 3 components is better than 2, 2 are better than 1, 1 is better than 0) but not hierarchical. That is, it is not "better", as in the Logic Model, to achieve one component rather than another - e.g. S is not better than T. The Logic Model components, on the other hand, are hierarchical in the sense that an Outcome PO is "better" than an Output, and an Output is better than a Process PO. Therefore for scoring purposes Outcomes=3, Output=2 and Process=1. Note, as discussed later, the scoring definition process surprisingly excluded nearly all POs from reaching the Outcome level.

The scoring theoretically allows for total scores ranging from 0 -no criteria were detected, to a maximum of 6 - the PO reached all three SMT criteria, plus it was an Outcome. The SMT criteria could generate at 1 where either S or M or T was detected; a 2 where SM, ST or MT was detected or a 3 where S&M&T were detected. The OOP scores generated a 1 for Process, a 2 for Output, and a 3 for Outcome.

The SMART model criteria may also be combined with the logic model to produce SMT+OOP scores. Table II illustrates that such combined scores range from a possible high of 6, to a low of 0 where a six would indicate that the POO had yes answers for all SMT criteria and was determined to fulfill the requirement of an "Outcome" objective.

Table II: Scoring Results and Meaning

SMT	OOP	Combined	Meaning / Value
3	3	6	Best possible, All SMT + "Outcome"
2	3	5	2nd Best, 2 SMT +"Outcome"
3	2	5	2nd Best, 2 SMT +"Output"
1	3	4	3rdd Best, 1 SMT +"Outcome"
2	2	4	3rd Best, 2 SMT +"Output"
3	1	4	3rd Best, 3 SMT +"Process"
3	0	3	4th Best, 3 SMT + undecided
2	1	3	4th Best, 2 SMT +"Process"
1	2	3	4th Best, 2 SMT +"Output"
1	1	2	5th Best, 2 SMT +"Process"
0	2	2	5th Best, 0 SMT +"Output"
2	0	2	5th Best, 2 SMT + undecided
1	0	1	6th Best, 1 SMT + undecided
0	1	1	6th Best, 0 SMT +"Process"

Results:

Whether comparing LPH provider groupings, combined or individual component scores the research protocol called for observations of whether the measure improved, remained steady or worsened from 2000 to 2001 (Y1 to Y2) and from 2001 to 2002 (Y2 to Y3).

The following tables display by year the distribution of PO's and the average and variance of the SMART or Logic Model scoring generated by each permutations of LPH provider groupings.

The descriptive sections of the analysis show that over the three years 4,874 PO's had been produced for 661 unique contract ID numbers for an average of 7.10 PO's for year 2000; 8.81 PO's year 2001 and 6.13 PO's by year 2001 Table III.

Table III: Total POs, Unique and Average Number of contracts Per Year:

	Y 2000	**Y 2001**	**Y 2002**
Total POs	1263	2141	1470
Unique Contract IDs	178	243	240
Avg. # of Contracts	7.10	8.81	6.13

Further, as can be seen in Table IV, the percentage of PO's produced by each Level 1-3, Rural-Urban or Metro-NonMetro LPH providers varied little from year to year. This finding is important because the DPH-provided database contained no individual identifiers that carried over from year to year among the different public health providers. Consequently comparing the percentage of total of the same characteristic (e.g. - Level 1, Urban, Metro) over time was the only obvious way to observe whether specific groups performance, or the training for that performance, improved over time. These groupings of LPH provider characteristic permutations were the "next best thing" to having a matched pairs analytical design. Since (See Table IV) the percent of total of any characteristic grouping from year to year was very stable, it was assumed that indeed, the identity of the individual providers within the groupings were also stable over time. This assumption was further supported (Table V) when the numbers of unique ID numbers were calculated within each LPH grouping for each year. Although there was some variation from Y1 to Y2, the number of unique ID numbers was nearly identical from Y2 to Y3.

Table IV: Distribution of POs by Characteristic and Year

Characteristic	Year	Count of POs	Per Cent of total
Level I	2000	135	10.689%
Level I	2001	234	10.929
Level I	2002	141	9.592
Level II	2000	710	56.215
Level II	2001	1260	58.851
Level II	2002	861	58.571
Level III	2000	403	31.908
Level III	2001	563	26.296
Level III	2002	408	27.755
Rural	2000	524	41.489
Rural	2001	920	42.971
Rural	2002	615	41.837
Urban	2000	739	58.511

Urban	2001	1221	57.029
Urban	2002	855	58.163
Non-metro	2000	729	57.720
Non-metro	2001	1262	58.944
Non-metro	2002	869	59.116
Metro	2000	534	42.280
Metro	2001	879	41.056
Metro	2002	601	40.884

Table V : Unique ID Numbers within LPH Groupings by Year

LPH Grouping	Y 2000	Y 2002	Y 2002
level 1 + Urban + Non-metro =	2	3	3
level 1 + Rural + Non-metro =	12	16	16
level 1 + Urban + metro =	10	10	8
level 2 + Urban + Non-metro =	18	25	25
level 2 + Urban + metro =	29	34	34
level 2 + Rural + Non-metro =	59	77	77
level 3 + Rural + Non-metro =	4	5	5
level 3 + Urban + metro =	32	39	38
level 3 + Urban + Non-metro =	8	12	12
Totals	194	221	218

Table VI combines the provider groupings into nine permutations and shows the number of POs produced by each grouping across years. It is interesting to note that in all cases, more POs were generated from 2000 to 2001 and then fewer were generated from 2001 to 2002.

Table VI : PO Distribution by Year by LPH Grouping Combinations

LPH Grouping	Y 2000 N	Y 2001 N	Y 2002 N	Total N
level 1 + Urban + Non-metro	16	23	17	56
level 1 + Rural + Non-metro	89	157	86	332
level 1 + Urban + metro	30	54	38	122
level 2 + Urban + Non-metro	109	218	161	488
level 2 + Urban + metro	197	345	220	762
level 2 + Rural + Non-metro	404	697	480	1581
level 3 + Rural + Non-metro	31	49	35	115
level 3 + Urban + metro	292	413	297	1002
level 3 + Urban + Non-metro	80	101	76	257

It is noteworthy that, rightfully, the contradictory description permutation of Rural+Metro was not found at any Level among the LPH groupings.

1. Statewide, all SMT individual components improved from 2000 to 2001. From 2001 to 2002, *T*-timeframe improved, *M*-measurable remained at its high (99%) level of attainment whereas the *S*-specific component fell precipitously (from 95% to 70%) although not as low as it was in 2000 (See Table VII).

Table VII: Distribution of SMART components by years

SMART component	**Y 2000 N/% of total**	**Y 2001**	**Y 2002**	**Total PO / % of total**
S specific	508/40.2%	2039/95.2%	1027/69.9%	3574/73%
M measurable	1249/98.9%	2135/99.7%	1468/99.9%	4852/99%
T timeframe	504/40%	2038/95.2%	1452/98.8%	3994/82%

2. The percentage (Table VIII) of POs meeting all three SMT component criteria increased significantly from Y1 to Y2 (23% to 90%), but fell during Y3 to 69%.
   a. The percentage of statewide POs that reached two of the SMT component criteria fell (a sign of improvement) from Y1 to Y2 (33% to 10%) but then increased (not an improvement) to 30% for Y3.
   b. Finally the percentage of POs that reached one of the SMT component criteria fell (a sign of improvement) from Y1 to Y2 (43% to .09%) but then increased very slightly to .8% for Y3 (Table VIII).

Table VIII: Distribution of combinations of SMART components by years

SMART combinations	**Y 2000 N/% of Total**	**Y 2001**	**Y 2002**	**Total POs/ % of total**
3 components YYY	295/23.4%	1934/90.3%	1018/69.3%	3247/67%
2 components YY	418/33.1%	204/9.5%	441/30%	1063/22%
1 component Y	540/42.6	2/.09%	11/.75%	553/11%
No components	10/.79	1/.05	0/0	11/.002

3. All combinations of LPH provider descriptions (Table IX) improved their average SMT score from Y1 to Y2, and all combinations also fell back to a lower average score during Y3; although again not as low as their average during Y1.

Table IX: Average SMART score and variance by year and LPH provider groupings.

PO provider description	**Y 2000 Avg/var.**	**Y 2001**	**Y 2002**
Level 1 + Urban + Non-metro	1.63/.23	2.87/.11	2.53/.25
Level 1 + Rural + Non-metro	2.43/.40	2.92/.07	2.73/.20
level 1 + Urban + metro	1.60/.57	2.89/.10	2.47/.41
level 2 + Urban + Non-metro	1.60/.52	2.83/.15	2.66/.27
level 2 + Urban + metro	1.64/.58	2.90/.09	2.68/.22
level 2 + Rural + Non-metro	1.99/.64	2.92/.08	2.68/.23
level 3 + Rural + Non-metro	2.32/.41	2.96/.04	2.57/.24
level 3 + Urban + metro	1.71/.65	2.90/.09	2.70/.21
level 3 + Urban + Non-metro	1.08/.12	2.90/.09	2.78/.17
Any level + Rural + Metro	0	0	0

4. OOP individual components results were much more mixed. Although there were almost no Outcome POs recorded in any year, from Y1 to Y2 Outcomes improved

slightly (from 0 to 6) but fell off again from T2 to T3 (from 6 to 2). As percentage of total there were fewer Output scores recorded T1 to T2 (an improvement if the percentage of Outcomes had grown rather than the percentage of Process components) but their percentages did increase from T2 to T3 while the Process component dropped. (See Table X)

Table X: Distribution of Logic Model components by years

Component	Y 2000 N/% of total	Y 2001	Y 2002	Total POs
O Outcome	0/0	6/.28	2/.14	8
O Output	352/27.9	432/20.9	366/24.9	1150
P Process	826/65.4	1633/76.3	990/67.4	3449

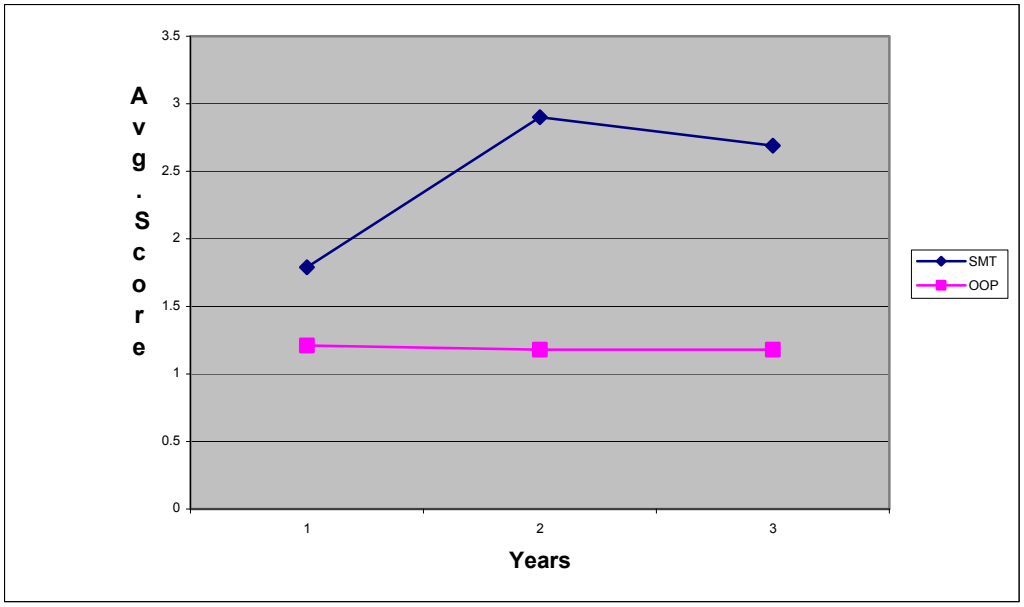
a. The average OOP score for all LPH provider groupings (Table XI) except "Level 1, Urban, Metro and Level 2, Urban, NonMetro fell (non-imporvement) from Y1 to Y2, whereas the average score improved Y2 to Y3 for "Level 1, Rural, NonMetro, Level 2, Rural NonMetro and Level 3, Urban, NonMetro groupings.

Table XI: Average OOP score and variance by year and LPH provider groupings.

PO provider description	Y 2000 Avg/var.	Y 2001	Y 2002
level 1 + Urban + Non-metro	1.31/.34	1.26/.37	1.59/.24
level 1 + Rural + Non-metro	1.24/.23	1.10/.15	1.16/.32
level 1 + Urban + metro	1.13/.32	1.15/.13	.92/.34
level 2 + Urban + Non-metro	1.07/.49	1.24/.28	1.16/.33
level 2 + Urban + metro	1.21/.30	1.13/.17	1.11/.22
level 2 + Rural + Non-metro	1.23/.28	1.17/.21	1.18/.29
level 3 + Rural + Non-metro	1.26/.26	1.22/.17	1.20/.33
level 3 + Urban + metro	1.21/.27	1.20/.22	1.18/.31
level 3 + Urban + Non-metro	1.21/.29	1.12/.24	1.20/.29
Any level + Rural + Metro	0	0	0

5. Graph 1 charts each years average SMT score and average OOP score. The graph illustrates that given a maximum possible score for either criteria of 3.0, overall performance judged by the SMART criteria was much higher than that of the Logic Model. The graph additionally illustrates the slump in performance that occurred between 2001 and 2002.

Graph 1: Average SMT and OOP Scores by Year



6. The distribution of Combined SMT+OOP scores (Table XII), illustrates the preponderance of scores at the "4" level and lower levels.

Table XII: Distribution of Combined SMT+OOP scores

Combined Score	N	% of Total	Value
6	6	0.123	Best
5	696	14.28	2nd
4	2767	56.77	3rd
3	868	17.81	
2	478	9.807	
1	49	1.005	

7. As might be expected, the combined scores reflected the individual strengths and weaknesses of the individual components making up the total. Consequently, largely on the strength of the SMT scoring patterns, average combined scores improved from Y1 to Y2 but fell off between Y2 and Y3 while still remaining "better" than Y1 (Table XIII).

Table XIII: Average Combined Scoring by Year

	Y 2000	Y 2001	Y 2002	Overall
Average Score	3.00	4.07	3.86	3.73
Variance	1.05	.31	.55	.77

a. Overall, outside of the insignificant number of perfect "6" scores, the second best scores "5s" consisted of POs meeting all the SMART criteria

and that of Output for the logic model. The majority (50% overall) of third best scores "4s" consisted of all the SMART criteria and that of Process for the logic model. The other "4" scores (6.5% overall) reflected two of the SMART criteria and that of Output for the logic model. The next highest percentage of scores (13% overall) consisted of "3s" which represented two of the SMART criteria and that of Process for the logic model (Table XIV).

Table XIV: Distribution of combined and component SMT+OOP scores

Combined Score	Component breakout	N	% of Total	Value / explanation
6	smt=3 + oop=3	6	0.123	Best (all SMT + Outcome)
5	smt=3 + oop=2	694	14.239	2nd (all smt + output)
5	smt=2 + oop=3	2	0.041	2nd (2 of smt + Outcome)
4	smt=3 + oop=1	2450	50.267	3rd (all smt + process)
4	smt=2 + oop=2	317	6.504	3rd (all smt + output)
3	smt=3 + oop=0	97	1.990	(all smt no OOP)
3	smt=2 + oop=1	633	12.987	(2 of smt + process)
2	smt=2 + oop=0	111	2.277	(2 of smt + no oop)

b. On a year-by-year basis (Table XV): "5" scores followed the similar improvement Y1 to Y2 and fall between Y2 to Y3 pattern as did the two SMART plus Output "4s".

c. Further, (Table XV) the majority "4" pattern, consisting of all SMART criteria and a Process logic model however showed that 70%, up from 17%, of the LPH providers where only able to generate Process measures in Y2 - a negative result; whereas that percentage dropped to 50% in Y3 - a definite improvement, and one that was reflected in the much increased numbers of Output scores recorded between Y2 and Y3.

d. The remaining score patterns (Table XV) also generally showed an improvement from Y1 to Y2 evidenced by a decrease in the percentage of low (3 and 2 scores) SMART and logic model POs generated in favor of higher percentage of SMT=3 and Output rather than Process scores.

Table XV: Count / percent of total component Breakout by Years

Combined Score	Component breakout	Y 2000 N / %	Y 2001	Y 2002
6	smt=3 & oop=3	0/0.00	4/0.19	2/ 0.14
5	smt=3 & oop=2	82/6.49	383/17.89	229/15.58
5	smt=2 & oop=3	0/0.00	2/0.09	0/ 0.00
4	smt=2 & oop=2	133/10.53	48/2.24	136/ 9.25
4	smt=3 & oop=1	209/16.55	1497/69.92	744/50.61
4	smt=1 & oop=3	0/0.00	0/0.00	0/ 0.00
3	smt=3 & oop=0	4/0.32	50/2.34	43/ 2.93
3	smt=0 & oop=3	0/0.00	0/0.00	0/0.00
3	smt=1 & oop=2	136/10.77	1/0.05	1/0.07
3	smt=2 & oop=1	259/20.51	136/6.35	238/6.19
2	smt=1 & oop=1	358/28.35	0/0.00	8/0.54
2	smt=2 & oop=0	26/2.06	18/0.84	67/4.56
2	smt=0 & oop=2	1/0.08	0/0.00	0/0.00

8. The next set of comparisons (Table XVI) reviews the average combined scores across years for the LPH provider groupings. All groupings improved from Y1 to Y2, but only the Level 1, Urban, NonMetro group (which also produced the fewest POs!) managed to generate an average combined score that did not fall back to a lower, but still improved over Y1, level at Y3.

Table XVI: Average combined SMT+OOP score and variance by year and LPH provider groupings.

PO provider description	Y 2000 Avg/var.	Y 2001	Y 2002
level 1 + Urban + Non-metro	2.94/.81	4.13/.55	4.12/.34
level 1 + Rural + Non-metro	3.66/.79	4.02/.24	3.90/.56
level 1 + Urban + metro	2.73/1.13	4.04/.22	3.39/.77
level 2 + Urban + Non-metro	2.67/1.05	4.07/.41	3.81/.60
level 2 + Urban + metro	2.85/.98	4.04/.27	3.79/.52
level 2 + Rural + Non-metro	3.21/1.06	4.09/.30	3.86/.51
level 3 + Rural + Non-metro	3.58/.70	4.18/.23	3.77/.63
level 3 + Urban + metro	2.92/.95	4.10/.32	3.88/.60
level 3 + Urban + Non-metro	2.29/.45	4.02/.32	3.97/.42
Any level + Rural + Metro			

a. Table XVII provides a different view of the same data by converting the LPH group's average raw scores for a year into relative rankings. The final AVG. column then provides a reasonable indication of which LPH groupings received the highest relative scores across years - that is for which grouping has the training or learning been most successful overall.

Table XVII: Ranking* of LPH providers by component scoring and years

	2000 SMT	2000 OOP	2000 Comb.	2001 SMT	2001 OOP	2001 Comb.	2002 SMT	2002 OOP	2002 Comb.	Total	AVG
l3+R+Nm	2	2	2	1	3	1	6	2	8	27	3.00
l1+U+Nm	5	1	5	8	1	2	7	1	1	31	3.44
l2+R+Nm	3	4	3	2	5	4	4	4	5	34	3.78
l1+R+Nm	1	3	1	2	9	8	2	6	3	35	3.88
l3+U+m	4	5	5	4	4	3	3	4	4	36	4.00
l3+U+Nm	9	5	9	4	8	8	1	2	2	48	5.33
l2+U+m	6	5	6	4	7	6	4	8	7	53	5.89
l2+U+Nm	7	9	8	9	2	5	5	6	6	57	6.33
l1+U+m	7	8	7	7	6	6	8	9	9	67	7.44

*Note: Rank of 1 is highest or best relative score, 9 is lowest or worst score

b. Table XVIII removes the combined component scores to show that even when considering the component schemes individually, the relative order of the LPH groupings only changes among the 2nd and 3rd ranks.

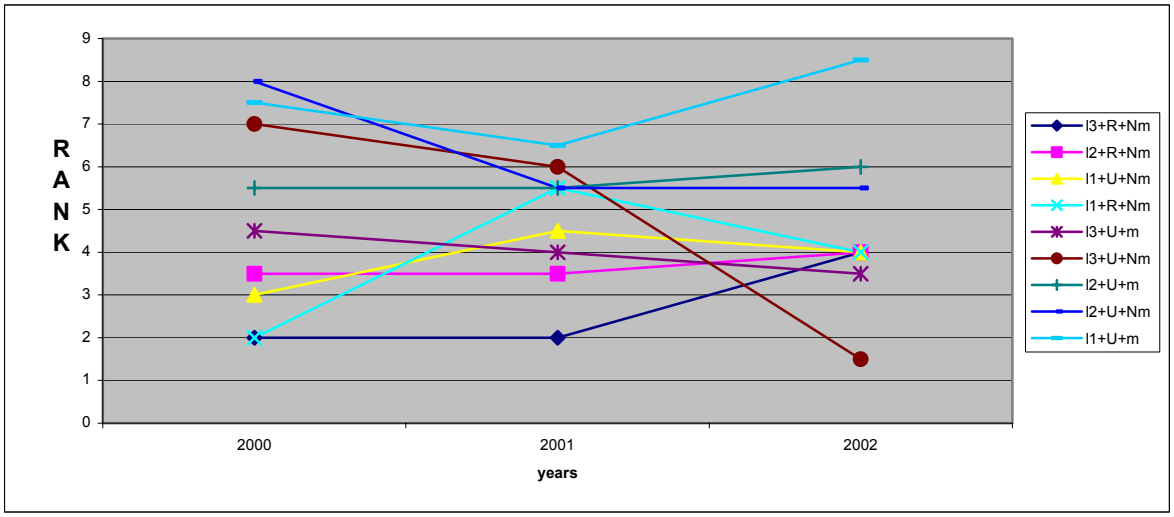
Consequently the top one-third and bottom one-third LPH provider rankings remain the same.

Table XVIII: Component Scheme Rankings without the Combined Scores

	2000	2000	2000	2001	2001	2001	2002	2002	2002	Total	AVG
	SMT	OOP	Avg	SMT	OOP	Avg	SMT	OOP	Avg		
l3+R+Nm	2	2	2	1	3	2	6	2	4	16	2.67
l2+R+Nm	3	4	3.5	2	5	3.5	4	4	4	22	3.67
l1+U+Nm	5	1	3	8	1	4.5	7	1	4	23	3.83
l1+R+Nm	1	3	2	2	9	5.5	2	6	4	23	3.83
l3+U+m	4	5	4.5	4	4	4	3	4	3.5	24	4.00
l3+U+Nm	9	5	7	4	8	6	1	2	1.5	29	4.83
l2+U+m	6	5	5.5	4	7	5.5	4	8	6	34	5.67
l2+U+Nm	7	9	8	9	2	5.5	5	6	5.5	38	6.33
l1+U+m	7	8	7.5	7	6	6.5	8	9	8.5	45	7.50

c. Graph 2 more clearly illustrates the relative performance differences over time by each grouping. For example L3+U+NM got much better over time; L1+U+M improved slightly then fell badly in year three; L1+R+NM fell over year two then improved year three; and L3+R+NM did very well for two years, and then fell off in year three.

Graph 2: Changing Rank of the LPH groupings over years



9. As mentioned previously, it was assumed that the LPH provider groupings represented a fairly stable set of individual organizations from year to year. Given that, it might also be hypothesized that their individual performance, while different from year to year, would not change their relative performance within the group. That is, a group who does relatively well in 2000 might be expected to also do relatively well in 2001. To test this, correlations were run between the average SMT and OOP scores for each set of years. The results showed an SMT correlation between Y1 and Y1 of r = .58 and between Y2 and Y3 of r=.07. The OOP averages reversed this pattern, with nearly identical numbers of Y1 to Y2 of r.042 and Y2 to Y3 of r = .57. Such results indicate a much more complicated picture than the simple "maintenance of relative position" hypothesis offered.

Conclusions:
These research findings support the following conclusions:

- As a group, Wisconsin's LPH providers significantly improved their ability to write performance objectives that met the SMART criteria (SMT) between 2000 and 2001. This improvement was maintained at the same high level (99%) for the Measurable and Timeframe components, but fell back for the Specific component from a high of 95% in 2001 to 70% in 2002.
- As a group, Wisconsin's LPH providers were less successful at improving their ability to write performance objectives that met the Logic Model criteria (OOP) between 2000 and 2001. Almost none (8 total) of the performance objectives met the scoring criteria for Outcomes; whereas the ratio of Output to Process objectives fell from a not very impressive 1:2.3 ratio in 2000, to a 1:3.7 ratio in 2001 and then slightly improved again to a 1:2.7 ratio in 2002.
- In total scoring (range 0-6) 70% of the objectives received a total score greater than 4; however 50% of the scores reflected objectives meeting all the SMART criteria but only the minimal "Process" Logic Model criteria.
- The relative performance rankings of individual LPH groupings were quite mixed from year to year. Strong performance on one of the two models was only weakly associated with performance on the other. Performance inconsistency was also apparent when combining the two models into a total combined score. The total scores did continue to illustrate the drop in performance between years two and three. With current information, further analysis was unable to disentangle whether the fall in performance was a result of actual differences in learning / retention levels among the groups, or differences in their training, or how recently their training occurred, or all of the above.

Discussion:
The evidence would lead us to reject Null Hypothesis I in favor of a finding that the percentage of total POs produced, as well as the average SMT, OOP or SMT+OOP scores, meeting different criteria were different from year to year.

The noted differences followed a strikingly similar pattern of large improvements between 2000 and 2001 followed by a decrease in performance between 2001 and 2002. The decrease was characterized by 2002 values (% of total or average) that were still an

improvement over 2000, but not equal to, or an improvement over 2001. Among the SMT scores, the drop in performance was exclusively a function of a drop in the S-Specific component that fell in year 2002 and only came back up 4 percentage points in 2003. Among the OOP components Outcome (of which there were only eight total) and the Output component improved somewhat from 2000 to 2001 but fell off in 2003; whereas 2001 was the worst year for Process components (76%) with 2002 falling back to the relatively high level (67%) found in 2001. Review of the average and percent of total combined (SMT+OOP) scores and their permutations revealed this same pattern - big improvements from 2000 to 2001 and then a falling back during 2002.

Among the LPH groupings (Null Hypothesis 2 and 3) there were actual as well as relative rank differences both across groups within years and within a group across years. Further, the now familiar pattern of improvement and then fall back was also strongly evident for two-thirds of the groups. However the Level 3+Urban+Metro and Level 3+Urban+Non-Metro groups improved their relative positions both years.

The data does not illustrate "why" the ability of LPH providers to write Specific SMART objectives, or any of the Logic Model objectives, performance generally improved between years one and two and then fell off in year three. At least two scenarios are possible. One, the intensity or content of the training may have been different between years one and two versus two and three. Secondly, there may have been no training between years two and three for some, or perhaps all, of the groups. The observed pattern of scores definitely support a classic learning model where performance improves following training, but as time goes on "forgetting" sets in and old patterns began to reestablish themselves.

Recommendations:

Since the DPH likely knows the training dates relative to the submission of the PO's for each year, further exploration of these "degradation of performance as a function of elapsed time from training" issues would either eliminate or solidify this potential explanation

If "time since training" proves to be an issue, the DPH should consider instituting re-training sessions for the LPH groups whose performance has slipped; or, if such sessions are ongoing then the materials and goals of the trainers themselves should be reviewed in an attempt to see if the concepts behind the SMART or Logic Model criteria are consistent.

The DPH should Develop and include in the database persistent LPH provider identifiers so future research can utilize a matched-pair analytical design.

At minimum the DPH should review the training materials, methods and particularly the working definitions of "Specific" from the SMART criteria and "Outcome and Output" from the Logic Model. Agreeing on working definitions of these concepts was a major

effort for the expert committee during development of the scoring procedures[3]. The results of this scoring effort clearly indicates either that a well trained scorer with high test-retest reliability ratings was unable to find POs that matched the definitions; or the procedures used to train the LPH PO writers generated conceptually different POs then the descriptions / definitions provided by the expert committee.

Finally, if the scoring methodology is put into practice without further refinement, all performance objectives with a total score of 4 or higher should be expertly evaluated. Although this recommendation means that only 30% of current objectives would have been eliminated without human scrutiny, continued issues with the Logic Model criteria preclude moving the "automatically eliminate" bar higher.

Appendix I:

Component Definitions:

Definitions POO:
Process: Does it describe actions or activities?

Output: Does it identify a product, or the number (or %) of people who will receive a service?

Outcome: Does it measure: knowledge, skills or behaviors; community practices or policies; or health status, AND does it answer "so what"?

Program logic (representing a quality hierarchy with outcome as most desirable):

Is Outcome? - IF Yes, got to Next performance objective
IF NO,
Is Output? - IF Yes, go to Next performance objective
IF NO,

[3] Newsom RS, Chapin, J, Gehl S. "Measure Twice, Cut Once: Measuring and Evaluating Objectives in Performance -Based Contracting - Phase I". Final report to Wisconsin Division of Public Health, May, 2003.

Is Process? - IF Yes, go to Next performance objective
IF NO,
Next performance objective

Definitions SMT

Specific: Should include wording that identifies: "Who, what, where". A "no" answer to any of these three questions by the scorer should result in a "NO" score for the specific criteria.

Measurable: The measurable criteria requires a "yes" answer to such questions as: 1) Is this objective quantifiable? or 2) Has the evidence of change, or improvement, or benefit been described? A positive "yes" answer to either of these questions should result in a "YES" score for the measurable criteria.

Time Frame: The time frame criterion requires only that the objective include the month, day and year by which the proposed performance objective will be completed. An objective that contains the wording (date representation) "By month/day/year" as a precursor to the proposed objective should result in a "YES" score for the time frame criteria.

Appendix II

POEM Statistics, run date: 06/03/03

```
total records scored = 4874
total year 2000=  1263
total year 2001=  2141
total year 2002=  1470

total type LPHDPBC = 3120
total type LPHDCPS =  196
total type LPHDPBS = 1399
total type PPPCR   =   76
total type PPP     =  157
total type PPCR =    2

total unique contract ID numbers =  661
total unique contract year 2000 =  178
total unique contract year 2001 =  243
total unique contract year 2002 =  240

Average # POs per unique ID
Avg per for year 2000 7.10
Avg per for year 2001 8.81
Avg per for year 2002 6.13
```

```
total level I LPHD   =  510
total level II LPHD  = 2831
total level III LPHD = 1374
Count / % total of POs by individual provider groupings

Total Level I and year 2000    135 percent of total  10.689
Total Level II and year 2000   710 percent of total  56.215
Total Level III and year 2000  403 percent of total  31.908
Total Level I and year 2001    234 percent of total  10.929
Total Level II and year 2001  1260 percent of total  58.851
Total Level III and year 2001  563 percent of total  26.296
Total Level I and year 2002    141 percent of total   9.592
Total Level II and year 2002   861 percent of total  58.571
Total Level III and year 2002  408 percent of total  27.755

total Rural, year 2000=   524 percent of total =   41.489
total Urban, year 2000=   739 percent of total =   58.511
total Rural, year 2001=   920 percent of total =   42.971
total Urban, year 2001=  1221 percent of total =   57.029
total Rural, year 2002=   615 percent of total =   41.837
total Urban, year 2002=   855 percent of total =   58.163

total non-metro year 2000=   729 percent of total =   57.720
total metro year 2000=       534 percent of total =   42.280
total non-metro year 2001=  1262 percent of total =   58.944
total metro year 2001=       879 percent of total =   41.056
total non-metro year 2002=   869 percent of total =   59.116
total metro year 2002=       601 percent of total =   40.884

Count of POs by LPH provider groupings combinations
count of level 1 + Urban + Non-metro =   56
count of level 1 + Rural + Non-metro =  332
count of level 1 + Urban +     metro =  122
count of level 2 + Urban + Non-metro =  488
count of level 2 + Urban +     metro =  762
count of level 2 + Rural + Non-metro = 1581
count of level 3 + Rural + Non-metro =  115
count of level 3 + Urban +     metro = 1002
count of level 3 + Urban + Non-metro =  257
count of rural+Metro  =     0 Contradiction in labels - should be 0


Distribution of LPH groupings by Year
                                2000    2001   2002
 level 1 + Urban + Non-metro =,   16,   23,   17
 level 1 + Rural + Non-metro =,   89,  157,   86
 level 1 + Urban +     metro =,   30,   54,   38
 level 2 + Urban + Non-metro =,   109,  218,  161
 level 2 + Urban +     metro =,  197,  345,  220
 level 2 + Rural + Non-metro =,  404,  697,  480
 level 3 + Rural + Non-metro =,   31,   49,   35
 level 3 + Urban +     metro =,  292,  413,  297
 level 3 + Urban + Non-metro =,   80,  101,   76


 SCORING results: (scoring by poem-scoring.prg
```

```
  SMT counts
3 YYYs best=  3247 percent of total =   66.62
2 YYs  2nd=   1063 percent of total =   21.81
1 Y    3rd=    553 percent of total =   11.35

 counts by S specific and year
SPECIFIC records for year 2000=   508 percent    40.22
SPECIFIC records for year 2001=  2039 percent    95.24
SPECIFIC records for year 2002=  1027 percent    69.86

 count by M measurable and year
MEASURABLE records for year 2000=  1249 percent    98.89
MEASURABLE records for year 2001=  2135 percent    99.72
MEASURABLE records for year 2002=  1468 percent    99.86

 counts by T timeframe and year
TIMEFRAME records for year 2000=   504 percent    39.90
TIMEFRAME records for year 2001=  2038 percent    95.19
TIMEFRAME records for year 2002=  1452 percent    98.78

Distributions of OOP by Year, RURAL-URBAN, METRO-NON, & LEVEL I-III

  OOP counts
outcome, best=     8 percent of total =    0.16
output,  2nd =  1150 percent of total =   23.59
process, 3rd =  3449 percent of total =   70.76



 OOP counts by year
outcome
3 best year 2000 =     0 percent of total =    0.00
3 best year 2001 =     6 percent of total =    0.28
3 best year 2002 =     2 percent of total =    0.14
output
2 2nd, year 2000 =   352 percent of total =   27.87
2 2nd, year 2001 =   432 percent of total =   20.18
2 2nd, year 2002 =   366 percent of total =   24.90
process
1 3rd, year 2000 =   826 percent of total =   65.40
1 3rd, year 2001 =  1633 percent of total =   76.27
1 3rd, year 2002 =   990 percent of total =   67.35


 OOP counts by year and RURAL-URBAN
outcome
3 best year RURAL 2000 =     0 percent of total =    0.00
3 best year RURAL 2001 =     2 percent of total =    0.22
3 best year RURAL 2002 =     1 percent of total =    0.16
3 best year URBAN 2000 =     0 percent of total =    0.00
3 best year URBAN 2001 =     4 percent of total =    0.33
3 best year URBAN 2002 =     1 percent of total =    0.12
output
2 2nd, RURAL year 2000 =   145 percent of total =   27.67
2 2nd, RURAL year 2001 =   170 percent of total =   18.48
2 2nd, RURAL year 2002 =   157 percent of total =   25.53
2 2nd, URBAN year 2001 =   207 percent of total =   28.01
```

```
2 2nd, URBAN year 2002 =   262 percent of total =   21.46
2 2nd, URBAN year 2002 =   209 percent of total =   24.44
process
1 3rd, RURAL year 2000 =   354 percent of total =   67.56
1 3rd, RURAL year 2001 =   720 percent of total =   78.26
1 3rd, RURAL year 2002 =   413 percent of total =   67.15
1 3rd, URBAN year 2000 =   472 percent of total =   63.87
1 3rd, URBAN year 2001 =   913 percent of total =   74.77
1 3rd, URBAN year 2002 =   577 percent of total =   67.49

 OOP counts by year and METRO-NON
outcome
3 best year METRO 2000 =     0 percent of total =    0.00
3 best year METRO 2001 =     4 percent of total =    0.46
3 best year METRO 2002 =     0 percent of total =    0.00
3 best year NONMETRO 2000 =     0 percent of total =    0.00
3 best year NONMETRO 2001 =     2 percent of total =    0.16
3 best year NONMETRO 2002 =     2 percent of total =    0.23
output
2 2nd, METRO year 2000 =   148 percent of total =   27.72
2 2nd, METRO year 2001 =   172 percent of total =   19.57
2 2nd, METRO year 2002 =   141 percent of total =   23.46
2 2nd, NONMETRO year 2000 =   204 percent of total =   27.98
2 2nd, NONMETRO year 2002 =   260 percent of total =   20.60
2 2nd, NONMETRO year 2002 =   225 percent of total =   25.89
process
1 3rd, METRO year 2000 =   355 percent of total =   66.48
1 3rd, METRO year 2001 =   680 percent of total =   77.36
1 3rd, METRO year 2002 =   412 percent of total =   68.55
1 3rd, NONMETRO year 2000 =   471 percent of total =   64.61
1 3rd, NONMETRO year 2001 =   953 percent of total =   75.52
1 3rd, NONMETRO year 2002 =   578 percent of total =   66.51

 OOP counts by year and LEVEL
outcome
3 best year LEVEL I 2000 =     0 percent of total =    0.00
3 best year LEVEL I 2001 =     0 percent of total =    0.00
3 best year LEVEL I 2002 =     1 percent of total =    0.71
3 best year LEVEL II 2000 =     0 percent of total =    0.00
3 best year LEVEL II 2001 =     2 percent of total =    0.16
3 best year LEVEL II 2002 =     1 percent of total =    0.12
3 best year LEVEL III 2000 =     0 percent of total =    0.00
3 best year LEVEL III 2001 =     2 percent of total =    0.36
3 best year LEVEL III 2002 =     0 percent of total =    0.00
output
2 2nd, year LEVEL I 2000 =    36 percent of total =   26.67
2 2nd, year LEVEL I 2001 =    36 percent of total =   15.38
2 2nd, year LEVEL I 2002 =    34 percent of total =   24.11
2 2nd, year LEVEL II 2000 =   198 percent of total =   27.89
2 2nd, year LEVEL II 2001 =   253 percent of total =   20.08
2 2nd, year LEVEL II 2002 =   198 percent of total =   23.00
2 2nd, year LEVEL III 2000 =   107 percent of total =   26.55
2 2nd, year LEVEL III 2001 =   120 percent of total =   21.31
2 2nd, year LEVEL III 2002 =   108 percent of total =   26.47
process
1 3rd, year LEVEL I 2000 =    93 percent of total =   68.89
1 3rd, year LEVEL I 2001 =   191 percent of total =   81.62
```

```
1 3rd, year LEVEL I 2002 =    91 percent of total =   64.54
1 3rd, year LEVEL II 2000 =   454 percent of total =   63.94
1 3rd, year LEVEL II 2001 =   964 percent of total =   76.51
1 3rd, year LEVEL II 2002 =   599 percent of total =   69.57
1 3rd, year LEVEL III 2000 =   275 percent of total =   68.24
1 3rd, year LEVEL III 2001 =   424 percent of total =   75.31
1 3rd, year LEVEL III 2002 =   267 percent of total =   65.44


Distributions of SMT by Year, RURAL-URBAN, METRO-NON, & LEVEL I-III

best (YYY) SMT outcome by year
count Best year 2000=   295 percent of total =   23.357
count Best year 2001=  1934 percent of total =   90.332
count Best year 2002=  1018 percent of total =   69.252

2nd best (=2Yes) SMT outcome by year

count 2nd Best year 2000=   418 percent of total =   33.096
count 2nd Best year 2001=   204 percent of total =    9.528
count 2nd Best year 2002=   441 percent of total =   30.000

3rd best (=1Yes) SMT outcome by year

count 3rd Best year 2000=   540 percent of total =   42.755
count 3rd Best year 2001=     2 percent of total =    0.093
count 3rd Best year 2002=    11 percent of total =    0.748


 Worst perf (NNN) by Year

count worst year 2000=    10 percent of total =    0.792
count worst year 2001=     1 percent of total =    0.047
count worst year 2002=     0 percent of total =    0.000


best (YYY) SMT outcome by year and rural / urban

count Best year 2000 rural=   182 percent of total =   34.733
count Best year 2000 urban=   113 percent of total =   15.291
count Best year 2001 rural=   848 percent of total =   92.174
count Best year 2001 urban=  1086 percent of total =   88.943
count Best year 2002 rural=   425 percent of total =   69.106
count Best year 2002 urban=   593 percent of total =   69.357


2nd best (=2Yes) SMT outcome by year and rural / urban

count 2nd year 2000 rural=   201 percent of total =   38.359
count 2nd year 2000 urban=   217 percent of total =   29.364
count 2nd year 2001 rural=    72 percent of total =    7.826
count 2nd year 2001 urban=   132 percent of total =   10.811
count 2nd year 2002 rural=   186 percent of total =   30.244
count 2nd year 2002 urban=   255 percent of total =   29.825

3rd best (=1Yes) SMT outcome by year and rural / urban
```

```
count 3rd year 2000 rural=   140 percent of total =   26.718
count 3rd year 2000 urban=   400 percent of total =   54.127
count 3rd year 2001 rural=     0 percent of total =    0.000
count 3rd year 2001 urban=     2 percent of total =    0.164
count 3rd year 2002 rural=     4 percent of total =    0.650
count 3rd year 2002 urban=     7 percent of total =    0.819

SMT scoring by metro non-metro and year
best (YYY) SMT outcome by year and metro vs non

count Best year 2000 metro=        99 percent of total =   18.539
count Best year 2000 Non-Metro=   196 percent of total =   26.886
count Best year 2001 metro=       793 percent of total =   90.216
count Best year 2001 Non-Metro=  1141 percent of total =   90.412
count Best year 2002 metro=       415 percent of total =   69.052
count Best year 2002 Non-Metro=   603 percent of total =   69.390


2nd best (=2Yes) SMT outcome by year and metro vs non

count 2nd year 2000 metro=       161 percent of total =   30.150
count 2nd year 2000 Non-Metro=   257 percent of total =   35.254
count 2nd year 2001 metro=        84 percent of total =    9.556
count 2nd year 2001 Non-Metro=   120 percent of total =    9.509
count 2nd year 2002 metro=       183 percent of total =   30.449
count 2nd year 2002 Non-Metro=   258 percent of total =   29.689

3rd best (=1Yes) SMT Results by year and metro vs non

count 3rd year 2000 metro=       268 percent of total =   50.187
count 3rd year 2000 Non-Metro=   272 percent of total =   37.311
count 3rd year 2001 metro=         1 percent of total =    0.114
count 3rd year 2001 Non-Metro=     1 percent of total =    0.079
count 3rd year 2002 metro=         3 percent of total =    0.499
count 3rd year 2002 Non-Metro=     8 percent of total =    0.921

SMT scoring by LEVEL and year
best (YYY) SMT Results by year and LEVEL

count Best year 2000 Level I =      49 percent of total =   36.296
count Best year 2000 Level II =    170 percent of total =   23.944
count Best year 2000 Level III =    76 percent of total =   18.859
count Best year 2001 Level I =     213 percent of total =   91.026
count Best year 2001 Level II =   1135 percent of total =   90.079
count Best year 2001 Level III =   508 percent of total =   90.231
count Best year 2002 Level I =      93 percent of total =   65.957
count Best year 2002 Level II =    590 percent of total =   68.525
count Best year 2002 Level III =   287 percent of total =   70.343

2nd best (=2Yes) SMT Results by year and LEVEL

count 2nd Best year 2000 Level I =      58 percent of total =   42.963
count 2nd Best year 2000 Level II =    253 percent of total =   35.634
count 2nd Best year 2000 Level III =   106 percent of total =   26.303
count 2nd Best year 2001 Level I =      21 percent of total =    8.974
count 2nd Best year 2001 Level II =    123 percent of total =    9.762
count 2nd Best year 2001 Level III =    55 percent of total =    9.769
```

```
count 2nd Best year 2002 Level I =      45 percent of total =   31.915
count 2nd Best year 2002 Level II =    263 percent of total =   30.546
count 2nd Best year 2002 Level III =   121 percent of total =   29.657

3rd best (=1Yes) SMT Results by year and LEVEL

count 3rd Best year 2000 Level I =      27 percent of total =   20.000
count 3rd Best year 2000 Level II =    282 percent of total =   39.718
count 3rd Best year 2000 Level III =   217 percent of total =   53.846
count 3rd Best year 2001 Level I =       0 percent of total =    0.000
count 3rd Best year 2001 Level II =      2 percent of total =    0.159
count 3rd Best year 2001 Level III =     0 percent of total =    0.000
count 3rd Best year 2002 Level I =       3 percent of total =    2.128
count 3rd Best year 2002 Level II =      8 percent of total =    0.929
count 3rd Best year 2002 Level III =     0 percent of total =    0.000

COMBINED SCORING Results - (oop+smt)= total

 Avg. combined score 3.733 Var = 0.770
 Avg. combined score 2000 3.002 Var = 1.045
 Avg. combined score 2001 4.076 Var = 0.314
 Avg. combined score 2002 3.861 Var = 0.550

count of combined score=6 =    6 percent of ttl =0.123
count of combined score=5 =  696 percent of ttl =14.28
count of combined score=4 = 2767 percent of ttl =56.77
count of combined score=3 =  868 percent of ttl =17.81
count of combined score=2 =  478 percent of ttl =9.807
count of combined score=1 =   49 percent of ttl =1.005

Combined Scoring variations & explanations
Rem: oop=3 is outcome, oop=2 is output, oop=1 is process

count of smt=3 & oop=3 is    6 percent= 0.123  best possible
count of smt=3 & oop=2 is  694 percent=14.239  best of smt + output
count of smt=3 & oop=1 is 2450 percent=50.267  best of smt + process
count of smt=3 & oop=0 is   97 percent= 1.990  best of smt no OOP
count of smt=2 & oop=3 is    2 percent= 0.041  2 of smt + outcome
count of smt=2 & oop=2 is  317 percent= 6.504  2 of smt + output
count of smt=2 & oop=1 is  633 percent=12.987  2 of smt + process
count of smt=2 & oop=0 is  111 percent= 2.277  2 of smt + no oop
count of smt=1 & oop=3 is    0 percent= 0.000  1 of smt + outcome
count of smt=1 & oop=2 is  138 percent= 2.831  1 of smt + output
count of smt=1 & oop=1 is  366 percent= 7.509  1 of smt + process
count of smt=1 & oop=0 is   49 percent= 1.005  1 of smt + no oop
count of smt=0 & oop=3 is    0 percent= 0.000  no smt + outcome
count of smt=0 & oop=2 is    1 percent= 0.021  no smt + output
count of smt=0 & oop=1 is    0 percent= 0.000  no smt + process
count of smt>0 & oop=0 is  257 percent= 5.273  smt, but no oop score
count of smt=0 & oop>0 is    1 percent= 0.021  no smt, but oop score

Best result (SMT=3 + output) analysis
Best result year 2000, count=   82 per cent of total  6.492
Best result year 2001, count=  383 per cent of total 17.889
Best result year 2002, count=  229 per cent of total 15.578

2nd best result (SMT=2 + output) analysis
```

```
2nd best result year 2000, count=  133 per cent of total 10.530
2nd best result year 2001, count=   48 per cent of total  2.242
2nd best result year 2002, count=  136 per cent of total  9.252

Most prevalent result (SMT=3 + process)
Most prevalent result year 2000, count=  209 per cent of total 16.548
Most prevalent result year 2001, count= 1497 per cent of total 69.921
Most prevalent result year 2002, count=  744 per cent of total 50.612

Average OOP scores by Year
avg. OOP score all years = 1.184 var=  0.263
avg. OOP score year 2000 = 1.211 var=  0.301
avg. OOP score year 2001 = 1.175 var=  0.215
avg. OOP score year 2002 = 1.176 var=  0.300

Avg. OOP by Year & LEVEL
avg. OOP score year 2000, level I = 1.222 var=  0.262
avg. OOP score year 2001, level I = 1.124 var=  0.168
avg. OOP score year 2002, level I = 1.149 var=  0.354
avg. OOP score year 2000, level II = 1.197 var=  0.322
avg. OOP score year 2001, level II = 1.171 var=  0.210
avg. OOP score year 2002, level II = 1.159 var=  0.282
avg. OOP score year 2000, level III = 1.213 var=  0.272
avg. OOP score year 2001, level III = 1.190 var=  0.221
avg. OOP score year 2002, level III = 1.184 var=  0.312

Avg OOP by year & RURAL-URBAN
avg. OOP score year 2000, Rural = 1.229 var=  0.272
avg. OOP score year 2001, Rural = 1.159 var=  0.199
avg. OOP score year 2002, Rural = 1.187 var=  0.298
avg. OOP score year 2000, Urban = 1.199 var=  0.322
avg. OOP score year 2001, Urban = 1.187 var=  0.227
avg. OOP score year 2002, Urban = 1.167 var=  0.301

AVG. OOP by Year & METRO-NONMETRO
avg. OOP score year 2000, Metro = 1.219 var=  0.287
avg. OOP score year 2001, Metro = 1.179 var=  0.208
avg. OOP score year 2002, Metro = 1.155 var=  0.291
avg. OOP score year 2000, NonMet = 1.206 var=  0.312
avg. OOP score year 2001, NonMet = 1.172 var=  0.220
avg. OOP score year 2002, NonMet = 1.190 var=  0.306

Average SMT scores by Year
avg. SMT score all years = 2.548 var=  0.488
avg. smt score year 2000 = 1.790 var=  0.649
avg. smt score year 2001 = 2.901 var=  0.094
avg. smt score year 2002 = 2.685 var=  0.231

Avg. SMT by Year & LEVEL
avg. smt score year 2000, level I = 2.148 var=  0.571
avg. smt score year 2001, level I = 2.910 var=  0.082
avg. smt score year 2002, level I = 2.638 var=  0.273
avg. smt score year 2000, level II = 1.828 var=  0.635
avg. smt score year 2001, level II = 2.899 var=  0.094
avg. smt score year 2002, level II = 2.676 var=  0.238
avg. smt score year 2000, level III = 1.630 var=  0.630
avg. smt score year 2001, level III = 2.902 var=  0.088
```

```
avg. smt score year 2002, level III = 2.703 var=  0.209

Avg SMT by year & RURAL-URBAN
avg. smt score year 2000, Rural = 2.076 var=  0.616
avg. smt score year 2001, Rural = 2.922 var=  0.072
avg. smt score year 2002, Rural = 2.685 var=  0.229
avg. smt score year 2000, Urban = 1.587 var=  0.573
avg. smt score year 2001, Urban = 2.886 var=  0.109
avg. smt score year 2002, Urban = 2.685 var=  0.232

AVG. SMT by Year & METRO-NONMETRO
avg. smt score year 2000, Metro = 1.661 var=  0.617
avg. smt score year 2001, Metro = 2.899 var=  0.100
avg. smt score year 2002, Metro = 2.686 var=  0.226
avg. smt score year 2000, NonMet = 1.885 var=  0.651
avg. smt score year 2001, NonMet = 2.903 var=  0.089
avg. smt score year 2002, NonMet = 2.685 var=  0.234

SMT Scoring by LPH grouping groupings combinations and Years
Year 2000 avg SMT of level 1 + Urban + Non-metro = 1.63 var =  0.23
Year 2001 avg SMT of level 1 + Urban + Non-metro = 2.87 var =  0.11
Year 2002 avg SMT of level 1 + Urban + Non-metro = 2.53 var =  0.25
Year 2000 avg SMT of level 1 + Rural + Non-metro = 2.43 var =  0.40
Year 2001 avg SMT of level 1 + Rural + Non-metro = 2.92 var =  0.07
Year 2002 avg SMT of level 1 + Rural + Non-metro = 2.73 var =  0.20
Year 2000 avg SMT of level 1 + Urban +     metro = 1.60 var =  0.57
Year 2001 avg SMT of level 1 + Urban +     metro = 2.89 var =  0.10
Year 2002 avg SMT of level 1 + Urban +     metro = 2.47 var =  0.41
Year 2000 avg SMT of level 2 + Urban + Non-metro = 1.60 var =  0.52
Year 2001 avg SMT of level 2 + Urban + Non-metro = 2.83 var =  0.15
Year 2002 avg SMT of level 2 + Urban + Non-metro = 2.66 var =  0.27
Year 2000 avg SMT of level 2 + Urban +     metro = 1.64 var =  0.58
Year 2001 avg SMT of level 2 + Urban +     metro = 2.90 var =  0.09
Year 2002 avg SMT of level 2 + Urban +     metro = 2.68 var =  0.22
Year 2000 avg SMT of level 2 + Rural + Non-metro = 1.98 var =  0.64
Year 2001 avg SMT of level 2 + Rural + Non-metro = 2.92 var =  0.08
Year 2002 avg SMT of level 2 + Rural + Non-metro = 2.68 var =  0.23
Year 2000 avg SMT of level 3 + Rural + Non-metro = 2.32 var =  0.41
Year 2001 avg SMT of level 3 + Rural + Non-metro = 2.96 var =  0.04
Year 2002 avg SMT of level 3 + Rural + Non-metro = 2.57 var =  0.24
Year 2000 avg SMT of level 3 + Urban +     metro = 1.71 var =  0.65
Year 2001 avg SMT of level 3 + Urban +     metro = 2.90 var =  0.09
Year 2002 avg SMT of level 3 + Urban +     metro = 2.70 var =  0.21
Year 2000 avg SMT of level 3 + Urban + Non-metro = 1.08 var =  0.12
Year 2001 avg SMT of level 3 + Urban + Non-metro = 2.90 var =  0.09
Year 2002 avg SMT of level 3 + Urban + Non-metro = 2.78 var =  0.17

SMT+OOP Scoring by LPH grouping groupings combinations and Years
Year 2000 avg SMT+OOP of level 1 + Urban + Non-metro = 2.94 var =  0.81
Year 2001 avg SMT+OOP of level 1 + Urban + Non-metro = 4.13 var =  0.55
Year 2002 avg SMT+OOP of level 1 + Urban + Non-metro = 4.12 var =  0.34
Year 2000 avg SMT+OOP of level 1 + Rural + Non-metro = 3.66 var =  0.79
Year 2001 avg SMT+OOP of level 1 + Rural + Non-metro = 4.02 var =  0.24
Year 2002 avg SMT+OOP of level 1 + Rural + Non-metro = 3.90 var =  0.56
Year 2000 avg SMT+OOP of level 1 + Urban +     metro = 2.73 var =  1.13
Year 2001 avg SMT+OOP of level 1 + Urban +     metro = 4.04 var =  0.22
Year 2002 avg SMT+OOP of level 1 + Urban +     metro = 3.39 var =  0.77
```

```
Year 2000 avg SMT+OOP of level 2 + Urban + Non-metro = 2.67 var =  1.05
Year 2001 avg SMT+OOP of level 2 + Urban + Non-metro = 4.07 var =  0.41
Year 2002 avg SMT+OOP of level 2 + Urban + Non-metro = 3.81 var =  0.60
Year 2000 avg SMT+OOP of level 2 + Urban +     metro = 2.85 var =  0.98
Year 2001 avg SMT+OOP of level 2 + Urban +     metro = 4.04 var =  0.27
Year 2002 avg SMT+OOP of level 2 + Urban +     metro = 3.79 var =  0.52
Year 2000 avg SMT+OOP of level 2 + Rural + Non-metro = 3.21 var =  1.06
Year 2001 avg SMT+OOP of level 2 + Rural + Non-metro = 4.09 var =  0.30
Year 2002 avg SMT+OOP of level 2 + Rural + Non-metro = 3.86 var =  0.51
Year 2000 avg SMT+OOP of level 3 + Rural + Non-metro = 3.58 var =  0.70
Year 2001 avg SMT+OOP of level 3 + Rural + Non-metro = 4.18 var =  0.23
Year 2002 avg SMT+OOP of level 3 + Rural + Non-metro = 3.77 var =  0.63
Year 2000 avg SMT+OOP of level 3 + Urban +     metro = 2.92 var =  0.95
Year 2001 avg SMT+OOP of level 3 + Urban +     metro = 4.10 var =  0.32
Year 2002 avg SMT+OOP of level 3 + Urban +     metro = 3.88 var =  0.60
Year 2000 avg SMT+OOP of level 3 + Urban + Non-metro = 2.29 var =  0.45
Year 2001 avg SMT+OOP of level 3 + Urban + Non-metro = 4.02 var =  0.32
Year 2002 avg SMT+OOP of level 3 + Urban + Non-metro = 3.97 var =  0.42
```

```
OOP Scoring by LPH grouping groupings combinations and Years
Year 2000 avg OOP of level 1 + Urban + Non-metro = 1.31 var =  0.34
Year 2001 avg OOP of level 1 + Urban + Non-metro = 1.26 var =  0.37
Year 2002 avg OOP of level 1 + Urban + Non-metro = 1.59 var =  0.24
Year 2000 avg OOP of level 1 + Rural + Non-metro = 1.24 var =  0.23
Year 2001 avg OOP of level 1 + Rural + Non-metro = 1.10 var =  0.15
Year 2002 avg OOP of level 1 + Rural + Non-metro = 1.16 var =  0.32
Year 2000 avg OOP of level 1 + Urban +     metro = 1.13 var =  0.32
Year 2001 avg OOP of level 1 + Urban +     metro = 1.15 var =  0.13
Year 2002 avg OOP of level 1 + Urban +     metro = 0.92 var =  0.34
Year 2000 avg OOP of level 2 + Urban + Non-metro = 1.07 var =  0.49
Year 2001 avg OOP of level 2 + Urban + Non-metro = 1.24 var =  0.28
Year 2002 avg OOP of level 2 + Urban + Non-metro = 1.16 var =  0.33
Year 2000 avg OOP of level 2 + Urban +     metro = 1.21 var =  0.30
Year 2001 avg OOP of level 2 + Urban +     metro = 1.13 var =  0.17
Year 2002 avg OOP of level 2 + Urban +     metro = 1.11 var =  0.22
Year 2000 avg OOP of level 2 + Rural + Non-metro = 1.23 var =  0.28
Year 2001 avg OOP of level 2 + Rural + Non-metro = 1.17 var =  0.21
Year 2002 avg OOP of level 2 + Rural + Non-metro = 1.18 var =  0.29
Year 2000 avg OOP of level 3 + Rural + Non-metro = 1.26 var =  0.26
Year 2001 avg OOP of level 3 + Rural + Non-metro = 1.22 var =  0.17
Year 2002 avg OOP of level 3 + Rural + Non-metro = 1.20 var =  0.33
Year 2000 avg OOP of level 3 + Urban +     metro = 1.21 var =  0.27
Year 2001 avg OOP of level 3 + Urban +     metro = 1.20 var =  0.22
Year 2002 avg OOP of level 3 + Urban +     metro = 1.18 var =  0.31
Year 2000 avg OOP of level 3 + Urban + Non-metro = 1.21 var =  0.29
Year 2001 avg OOP of level 3 + Urban + Non-metro = 1.12 var =  0.24
Year 2002 avg OOP of level 3 + Urban + Non-metro = 1.20 var =  0.29
```

```
Count / percent of total of combined Scoring variations by years
                     2000         2001       2002
smt=3 & oop=3,      0/  0.00,    4/  0.19,    2/ 0.14
smt=3 & oop=2,     82/  6.49,  383/ 17.89,  229/15.58
smt=2 & oop=3,      0/  0.00,    2/  0.09,    0/ 0.00
smt=2 & oop=2,    133/ 10.53,   48/  2.24,  136/ 9.25
smt=3 & oop=1,    209/ 16.55, 1497/ 69.92,  744/50.61
smt=1 & oop=3,      0/  0.00,    0/  0.00,    0/ 0.00
```

```
smt=3 & oop=0,      4/  0.32,   50/  2.34,   43/ 2.93
smt=0 & oop=3,      0/  0.00,    0/  0.00,    0/ 0.00
smt=1 & oop=2,    136/ 10.77,    1/  0.05,    1/ 0.07
smt=2 & oop=1,    259/ 20.51,  136/  6.35,  238/16.19
smt=1 & oop=1,    358/ 28.35,    0/  0.00,    8/ 0.54
smt=2 & oop=0,     26/  2.06,   18/  0.84,   67/ 4.56
smt=0 & oop=2,      1/  0.08,    0/  0.00,    0/ 0.00

Distribution of UNIQUE LPH groupings by Year
                                    2000    2001   2002
 level 1 + Urban + Non-metro =,      2,      3,     3
 level 1 + Rural + Non-metro =,     12,     16,    16
 level 1 + Urban +     metro =,     10,     10,     8
 level 2 + Urban + Non-metro =,      18,    25,    25
 level 2 + Urban +     metro =,     29,     34,    34
 level 2 + Rural + Non-metro =,     59,     77,    77
 level 3 + Rural + Non-metro =,      4,      5,     5
 level 3 + Urban +     metro =,     32,     39,    38
 level 3 + Urban + Non-metro =,      8,     12,    12
```