The Validity of Proficiency Testing: A Review of Data from the College of American Pathologists Laboratory Improvement Programs

Noel S. Lawson, M.D.*
Department of Pathology
St. John Hospital and Medical Center and
Wayne State University School of Medicine
Detroit, Michigan

Dan Tholen, M.S.
Statistical Consultant
Traverse City, Michigan

John W. Ross, M.D.
Department of Pathology
Kennestone Hospital
Marietta, Georgia

George G. Klee, M.D.
Department of Pathology and Laboratory Medicine
Mayo Clinic and Mayo Foundation
Rochester, Minnesota

*** Presenting Author**

Abstract: The validity of proficiency testing (PT) can be studied through comparing and correlating participant performance with that measured in other laboratory evaluation programs, and by examining consistency of grading. The College of American Pathologists (CAP) has undertaken various studies assessing the CAP Surveys PT Program. Performance correlates significantly with participation in the CAP Laboratory Accreditation Program (LAP), with LAP participants achieving overall lower rates of unacceptable results. For a set of analytes studied there is moderate positive correlation between bias and precision performance as measured by PT compared to that determined through CAP Quality Assurance Service (QAS) affiliated regional internal quality control. Furthermore, QAS participants achieve better survey performance than nonparticipants. Significant correlation has been demonstrated between performance measured in the CAP Linearity Survey and by concurrent PT. Relatively high consistency of participant survey performance ranks over time has been documented for two three year testing cycles. Finally, continuous improvement over several years has been documented for laboratories participating in the EXCEL Survey, with more experienced laboratories achieving significantly lower rates of unacceptable results. The findings, in aggregate, support the validity of PT, in the context of multiprogram characterization of laboratory performance.

Introduction - Validity of PT

Through the Surveys Program, the College of American Pathologists (CAP) is a dominant provider of professionally directed clinical laboratory Proficiency Testing (PT). This presentation focuses on contributions by CAP assessing the value of PT, emphasizing significant interprogram qualitative and quantitative relationships between PT and alternate measures of laboratory quality, consistency of performance, and effects of experience and time on results. Both previously published and new data support the validity of PT.

Material and Methods

In published studies, Proficiency Testing (PT) performance results have been compared with data from regional internal quality control, laboratory accreditation,[1] and linearity/calibration studies.[2] Tholen et al have reported the effect of experience and length of participation on PT performance.[3] We now provide additional data on the relationship between performance in PT and participation in the CAP laboratory accreditation program (LAP) and on consistency of performance in PT over time. To compare PT performance with LAP participation, survey performance from 1988-1990 was quantitated using an index which reflected aggregate score for 10 chemistry analytes over 3 years. Two sets of five common analytes each were used. (Set 1 = Calcium, Cholesterol, Creatine Kinase, Glucose, Potassium; Set 2 = Aspartate aminotransferase [AST], Bilirubin, Creatinine, Sodium, Triglycerides). Index scores P11, P12, P21, and P22 were derived from a nonparametric algorithm designed to penalize large bias and poor precision, respectively, and ignore small deviations from the target. All were derived from the same formula, with P12 and P22 giving slightly higher penalty for poor precision. P11 and P12 were for the 1st set of analytes and P21 and P22 were for the second. Index scores were prepared for five categories of laboratories, i.e., hospitals <100, 100-500, and >500 beds, independents, and others. For a laboratory to be included as an LAP participant, it must have been enrolled in the program for at least 1 year during the 1988-1990 period. In addition, a study of performance covering 1991-1994 was performed. All laboratories that were inspected (on-site) and active before 1989-1994 were included. A laboratory was considered to be accredited, i.e., "LAP" for the year of inspection and the following 2 years. Two quantitative survey performance measures were used - i.e., the Rate of Unacceptable Results and the Average Percent of Allowable Deviation (PAD), both using limits established by the Health Care Financing Administration (HCFA).[4] Unacceptable rates were obtained for chemistry, bacteriology, and immunology (qualitative and quantitative) challenges, which included all commonly performed HCFA regulated analytes.[4] The rate was obtained for all graded specimens in the specialty for each of the four years. Average PAD was determined for quantitative analytes in chemistry and immunology. Differences in means were analyzed by the Wilcoxon two sample test.

To study consistency of performance, Hematology (H1,H2) and Immunology (S,SM) Surveys data for 16 analytes from 1988-1990 were analyzed for laboratory performance. The study included 2736 participants with at least 20 challenges per survey per year, and at least 100 challenges overall. Performance cutpoints were set at the 25th and 75th percentiles of unacceptable

LAP LABORATORIES vs SURVEY LABORATORIES WITH AT LEAST ONE UNACCEPTABLE RESULT*

	No. of Laboratories				
ANALYTE	TOTAL	NOT IN LAP	WITH UNACCEPTABLE RESULTS,(%)	IN LAP	WITH UNACCEPTABLE RESULTS,(%)
AST	5071	2838	850 (30.0)	2233	606 (27.1)†
GLUCOSE	7718	4675	502 (10.7)	3043	177 (5.8)‡
PHOSPHORUS	4555	2009	305 (15.2)	2546	215 (8.4)‡
POTASSIUM	7459	4443	249 (5.6)	3016	92 (3.1)‡

*LAP indicates Laboratory Accreditation Program; AST, aspartate aminotransferase
†P<.05 by x^2 ‡P<.0005 by x^2
From Lawson et al, *Arch Pathol Lab Med* 1988; 112:454-461

Table 1

rate. The 3 years' consecutive performance was ranked 1 = top quartile, 2 = middle two quartiles, and 3 = lowest quartile. The 27 possible combinations were analyzed for actual vs. predicted performance. The study compared the observed rates of participants who were consistently in the same quartile group, with rates that would be expected if performance class were random.

The 1988-90 study was recently updated using data from Chemistry, Hematology, Immunology, and Bacteriology Surveys. These were analyzed to evaluate consistency of performance from 1992 to 1994. Participants were divided into three groups according to performance, i.e., relatively high = zero unacceptable results, relatively low = approximately 0-20 percentile, and intermediate = all others. The expected performance over 3 years was likewise the product of the percentages of participants in that group in each of the 3 years.

Results and Discussion

Using data from 1986 CAP programs, Lawson et al compared performance in PT by laboratories in the CAP LAP vs. that of nonparticipants.[1] They examined the analytes aspartate aminotransferase (AST), glucose, phosphorus, and potassium. The point of separation was one or more unacceptable results for an analyte during the year. In all cases, significantly more laboratories with unacceptable results joined the group of LAP nonparticipants (Table 1). The 1988-1990 data, using aforementioned indices, confirm that LAP participation is associated with improved PT performance. This is manifest as lower scores on multiple indices, across all categories of laboratories. Table 2 summarizes data on the relative performance of LAP and non LAP Survey participants. For all categories of laboratories as well in aggregate, LAP participants have better survey performance. The differences in performance are most

1990 SURVEYS PERFORMANCE INDICES VS. LAP STATUS
FOR EACH INSTITUTION TYPE AND ALL LABORATORIES†
(Lower index indicates better performance)

PRIVATE, COMMUNITY & FEDERAL HOSPITALS

PERFORMANCE MEASURE	1-99 BEDS NOT IN LAP	1-99 BEDS IN LAP	100-500 BEDS NOT IN LAP		100-500 BEDS IN LAP	500+ BEDS NOT IN LAP		500+ BEDS IN LAP
	n=932	n=288	n=787		n=1486	n=83		n=300
Index P11	73.1	70.6	61.3	*	55.9	61.2	*	56.5
Index P12	79.3	76.8	66.1	**	60.1	65.2		60.9
Index P21	69.0	67.9	58.3	**	54.8	70.2	**	55.3
Index P22	74.8	73.7	63.1	**	59.3	74.8	**	59.8

	INDEPENDENT NOT IN LAP	INDEPENDENT IN LAP	OTHER NOT IN LAP		OTHER IN LAP	ALL LABS NOT IN LAP		ALL LABS IN LAP
	n=243	n=146	n=251		n=52	n=2552		n=2402
Index P11	69.6	** 55.3	68.0	*	61.1	68.2	**	58.1
Index P12	75.3	** 59.5	73.6		67.1	73.7	**	62.7
Index P21	71.0	** 62.7	68.4		64.2	66.1	**	57.2
Index P22	76.9	** 67.5	73.9		69.5	71.5	**	61.9

STATISTICAL SIGNIFICANCE NOTES: * = .01 < P < .10
(Wilcoxon 2-sample test) ** = P < .01

†LAP = Laboratory Accreditation Program

Table 2

significant in medium-sized and large hospitals, and independents, and in the all laboratories grouping. Of the various groups studied, the lowest indices were noted among the medium-sized and large hospital cohorts.

The third surveys vs. LAP study, using the 1991-1994 data, has reaffirmed the better PT performance of LAP participants. In all specialties and years, LAP accredited laboratories have significantly better survey performance than non-LAP laboratories (Table 3a-d). This difference is seen both in the data reflecting rates of unacceptable results as well as in the PAD.

Thus, three different studies, each using different survey performance endpoints, yield the same conclusions. Laboratories in the CAP Surveys who are also in the CAP LAP program obtain better performance than non-participants. These studies have not been designed to evaluate the sources of improved performance. Possible contributing factors include directorship, overall attention to quality and documentation, adherence to the specific quality-related LAP questionnaire

SURVEY PERFORMANCE vs. LAP PARTICIPATION
1991-1994 CHEMISTRY

YEAR	NON LAP			LAP		
	NO.	UNACCEPTABLE (%)	ERROR	NO.	UNACCEPTABLE (%)	ERROR
1991	3358	2.62	40.0	2390	1.36	34.3
1992	3417	2.19	38.1	2597	1.08	32.2
1993	3453	1.89	36.1	2843	0.80	30.0
1994	3139	1.71	34.8	2906	0.74	29.2

All mean differences significant $p < .01$

Table 3a

SURVEY PERFORMANCE vs. LAP PARTICIPATION
1991-1994 BACTERIOLOGY

YEAR	NON LAP		LAP	
	NO.	UNACCEPTABLE (%)	NO.	UNACCEPTABLE (%)
1991	3012	7.60	2562	4.37
1992	2645	6.88	2571	3.74
1993	2482	6.52	2644	3.90
1994	2106	4.52	2679	2.94

All mean differences significant $p < .01$

Table 3b

items, and the integrated requirement within LAP that PT deficiencies be appropriately addressed.

Lawson et al[1] reported correlations between bias, precision, and total error as measured for laboratories participating in CAP Surveys and in the Great Lakes - Southeast Regional Quality Control Program, using the Quality Assurance Service (QAS) data program of CAP. Significant and moderate positive correlation of performance ranks was found for precision, bias, and total error for AST, glucose, and potassium, and of bias for phosphorus (Table 4). In addition, significant correlation was confirmed between quantitative survey and QAS bias for the four analytes, when analyzed by linear regression (Table 5). For AST, glucose, and potassium, QAS participants performed significantly better in surveys, with significantly lower bias, precision, and total error (Table 6).

Lum et al[2] in reporting on the relationship

SURVEY PERFORMANCE vs. LAP PARTICIPATION
1991-1994 QUANTITATIVE IMMUNOLOGY

YEAR	NON-LAP			LAP		
	NO.	UNACCEPTABLE(%)	ERROR	NO.	UNACCEPTABLE (%)	ERROR
1991	389	3.10	39.1	794	2.01	35.6**
1992	387	2.33	36.8	860	1.63	34.8*
1993	391	2.64	36.4	935	1.60	32.9**
1994	335	2.83	37.2	972	1.50	32.2**

Means significantly different by $p < .01$ (**) or $.01 < p < .10$ (*)

Table 3c

SURVEY PERFORMANCE vs. LAP PARTICIPATION
1991-1994 CATEGORICAL IMMUNOLOGY

YEAR	NON LAP		LAP	
	NO.	UNACCEPTABLE (%)	NO.	UNACCEPTABLE (%)
1991	2040	1.32	2008	0.96
1992	2057	1.59	2147	1.24
1993	2122	1.33	2337	0.96
1994	1786	1.01	2340	0.74

All mean differences significant $p < .01$

Table 3d

between laboratory performance in the Linearity Survey and that seen with concurrent PT, have documented a consistent and strong relationship between unacceptable survey results and calibration verification problems.[2] In addition, participants with performance-rated linear and verified calibration have lower rates of unacceptable results. Their study included some 33 analytes from the Chemistry, Ligand Assay, and Therapeutic Drug Surveys.

In studying consistency of PT performance during 1988-1990 and 1992-1994, two data sets lead to the same conclusion. The proportion of laboratories with consistent performance, i.e., 111,222,333 patterns, is greater than predicted with the blended hematology and immunology data from the earlier comparison study (Table 7) as well as within the latter specialty-specific study for chemistry, hematology, immunology, and

CORRELATION OF QAS & SURVEY DATA*					
ANALYTE	NO.	BIAS	ABSOLUTE BIAS	PRECISION	TOTAL ERROR
AST	88	.4893	.3256	.3573**	.4310
GLUCOSE	156	.7854	.4768	.3297	.4242
PHOSPHORUS	77	.5548	NS	NS	NS
POTASSIUM	64	.4206	.2570	.4758	.5505

*Spearman Correlation Coefficient
**N = 113
From Lawson et al, *Arch Pathol Lab Med* 1988; 112:454-461

Table 4

SUMMARY OF LINEAR REGRESSION RESULTS: CHEMISTRY SURVEY DATA-BIAS (y) vs GL/SE/NE QAS DATA (x)*				
ANALYTE	SLOPE	INTERCEPT,%	R	NO.
AST	0.70	0.14	.49	80
GLUCOSE	0.75	-0.83	.79	151
PHOSPHORUS	0.67	0.23	.60	72
POTASSIUM	0.64	1.20	.59	61

*From Lawson et al, *Arch Pathol Lab Med* 1988; 112:454-461

Table 5

bacteriology (Table 8). In both the earlier and latter data sets, the ratio of observed to expected performance was higher for the 333 than for the 111 pattern. This suggests that within the set of consistent performers, relatively high performance is more difficult to sustain than relatively low performance. Observed consistency of performance, from two different 3 year study cycles, lends credibility to the PT process, by implying that performance is not random, but rather related to intrinsic operational characteristics of participating laboratories.

The effect of experience and time on chemistry, hematology, and immunology PT performance in the CAP EXCEL Survey has been reported by Tholen et al.[3] The data covered the 1987-1993 period. In the group as a whole, there is a tendency for progressive decrease in the average rates of unacceptable results with increasing years of participation (Table 9). Individual participant performance was also tracked. Significant improvement over time of participants was documented for all specialties (Table 10). Findings also suggest that laboratories with more experience with PT have higher rates of acceptable results

SURVEY PERFORMANCE vs QAS PARTICIPATION*				
ANALYTE	NOT IN QAS		IN QAS	
	NO. OF LABORATORIES	VALUE,%	NO. OF LABORATORIES	VALUE,%
AST				
Survey bias	5163	6.10	919	5.49†
Survey precision	5163	8.01	919	7.70‡
Survey total error	5163	10.41	919	9.69‡
GLUCOSE				
Survey bias	6632	3.18	1104	2.58‡
Survey precision	6632	4.10	1104	3.35‡
Survey total error	6632	5.36	1104	4.36‡
PHOSPHORUS				
Survey bias	3928	3.60	767	3.83
Survey precision	3928	3.94	767	4.13
Survey total error	3928	5.55	767	5.85
POTASSIUM				
Survey bias	6468	2.07	998	1.78†
Survey precision	6468	2.79	998	2.13‡
Survey total error	6468	3.55	998	2.86‡

*QAS indicates Quality Assurance Service; AST, aspartate aminotransferase
†P< .01 by Wilcoxon's test. ‡P < .0001 by Wilcoxon's test.
From Lawson et al, *Arch Pathol Lab Med* 1988; 112:454-461

Table 6

SURVEY QUARTILE PERFORMANCE OVER TIME - (1988-1990)
HEMATOLOGY & IMMUNOLOGY GROUPS WITH CONSISTENT PERFORMANCE
OBSERVED vs. EXPECTED N=2750

3 YEAR SEQUENCE	OBSERVED %	EXPECTED %	RATIO
111	4.3	1.6*	2.7
222	15.3	12.4*	1.2
333	6.4	1.6*	4.1

* $p < .001$ by x^2

1 = Highest Quartile
2 = Middle Two Quartiles
3 = Lowest Quartile

Table 7

SURVEY PERFORMANCE OVER TIME - (1992-1994)
GROUPS WITH CONSISTENT PERFORMANCE
OBSERVED vs. EXPECTED (%)

3 YEAR SEQUENCE	CHEMISTRY, n=5017			HEMATOLOGY, n=1765		
	OBSERVED	EXPECTED	RATIO	OBSERVED	EXPECTED	RATIO
111	7.0	3.0*	2.3	16.8	10.9*	1.5
222	12.7	10.5*	1.2	2.7	2.4	1.1
333	5.1	0.8*	6.4	4.8	1.1*	4.4

* $p < .001$ by x^2
1 = No Unacceptable Results
2 = Intermediate Performance
3 = Lowest Relative Performance

Table 8

SURVEY PERFORMANCE OVER TIME - (1992-1994)
GROUPS WITH CONSISTENT PERFORMANCE
OBSERVED vs. EXPECTED (%)

3 YEAR SEQUENCE	IMMUNOLOGY, (Q) n=986			IMMUNOLOGY, © n=3480		
	OBSERVED	EXPECTED	RATIO	OBSERVED	EXPECTED	RATIO
111	23.1	19.6*	1.2	32.7	27.9*	1.2
222	1.5	0.7**	2.1	0.5	0.3***	1.7
333	2.8	1.0*	2.8	1.5	0.6*	2.5

* $p < .001$ by x^2
** $p < .01$ by x^2
*** $p < .05$ by x^2

1 = No Unacceptable Results
2 = Intermediate Performance
3 = Lowest Relative Performance

Table 8b

SURVEY PERFORMANCE OVER TIME - (1992-1994)
GROUPS WITH CONSISTENT PERFORMANCE
OBSERVED vs. EXPECTED (%)

3 YEAR SEQUENCE	BACTERIOLOGY, n=4237		
	OBSERVED	EXPECTED	RATIO
111	12.7	6.7*	1.9
222	6.9	5.2*	1.3
333	4.7	0.9*	5.2

* $p < .001$ by x^2
1 = No Unacceptable Results
2 = Intermediate Performance
3 = Lowest Relative Performance

Table 8c

RATES OF UNACCEPTABLE RESULTS AND YEARS OF PARTICIPATION IN COLLEGE OF AMERICAN PATHOLOGISTS EXCEL SURVEYS, 1987-1993

SPECIALTY	No. of CHALLENGES	No. of Years of Participation 1	2	3	4	>4
Routine chemistry	1135	7.4	6.7	6.1	5.6	5.8
Therapeutic drug-monitoring Chemistry	200	6.8	7.3	7.8	6.6	5.9
Hematology						
Categorical	292	5.6	5.5	5.1	4.9	4.6
Quantitative	371	6.0	5.6	4.9	4.9	4.4
Common immunology	188	8.4	6.6	6.1	5.2	4.9
Special immunology	49	11.2	10.8	10.3	9.5	7.5
Blood bank	52	2.1	2.0	1.8	1.1	1.5

From Tholen et al, *Arch Pathol Lab Med* 1995; 119:307-311

Table 9

PERFORMANCE IMPROVEMENT IN EXCEL SURVEYS*
TIME EFFECT p VALUES

	1987-1993		1989-1993		1991-1993	
	n	p	n	p	n	p
ROUTINE CHEMISTRY	247	<.001	632	<.001	1379	<.001
CATEGORICAL HEMATOLOGY	589	<.001	1236	<.001	2527	<.001
QUANTITATIVE HEMATOLOGY	612	<.001	1298	<.001	2668	<.001
COMMON IMMUNOLOGY	444	<.001	1009	<.001	2249	<.001

*Analysis of variance, experience vs. time effect
From Tholen et al, *Arch Pathol Lab Med* 1995; 119:307-311

Table 10

PERFORMANCE IMPROVEMENT IN EXCEL SURVEYS
EXPERIENCE EFFECT p VALUES*

	1987-1993		1989-1993		1991-1993	
	n	p	n	p	n	p
ROUTINE CHEMISTRY	247	.585	632	.487	1379	.049
CATEGORICAL HEMATOLOGY	589	.813	1236	.275	2527	<.001
QUANTITATIVE HEMATOLOGY	612	.077	1298	.883	2668	<.001
COMMON IMMUNOLOGY	444	.065	1009	.002	2249	.015

*Analysis of variance, experience vs. time effect
From Tholen et al, *Arch Pathol Lab Med* 1995; 119:307-311

Table 11

(Table 11). These results support the validity of PT by indicating that, as expected, prior experience and duration of participation are associated with performance improvements.

References

1. Lawson NS, Gilmore BF, Tholen DW. Multiprogram characterization of laboratory bias, precision, and total error. *Arch Pathol Lab Med.* 1988;112:454-461.

2. Lum G, Tholen DW, Floering DA. The usefulness of calibration verification and linearity surveys in predicting acceptable performance in graded proficiency tests. *Arch Pathol Lab Med.* 1995;119:401-408.

3. Tholen D, Lawson NS, Cohen T, Gilmore B. Proficiency test performance and experience with College of American Pathologists' Programs. *Arch Pathol Lab Med.* 1995;119:307-311.

4. Clinical Laboratory Improvement Amendments of 1988: final rule. *Federal Register.* Feb 28, 1992;55:7002-7186.