JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, May 2010, Vol. 55, No. 3
doi: 10.1111/j.1556-4029.2010.01344.x
Available online at: interscience.wiley.com

TECHNICAL NOTE

GENERAL; TOXICOLOGY

Jack Wallace,[1] *Ph.D.*

Proficiency Testing as a Basis for Estimating Uncertainty of Measurement: Application to Forensic Alcohol and Toxicology Quantitations

ABSTRACT: While forensic laboratories will soon be required to estimate uncertainties of measurement for those quantitations reported to the end users of the information, the procedures for estimating this have been little discussed in the forensic literature. This article illustrates how proficiency test results provide the basis for estimating uncertainties in three instances: (i) For breath alcohol analyzers the interlaboratory precision is taken as a direct measure of uncertainty. This approach applies when the number of proficiency tests is small. (ii) For blood alcohol, the uncertainty is calculated from the differences between the laboratory's proficiency testing results and the mean quantitations determined by the participants; this approach applies when the laboratory has participated in a large number of tests. (iii) For toxicology, either of these approaches is useful for estimating comparability between laboratories, but not for estimating absolute accuracy. It is seen that data from proficiency tests enable estimates of uncertainty that are empirical, simple, thorough, and applicable to a wide range of concentrations.

KEYWORDS: forensic science, data analysis, forensic toxicology, uncertainty of measurement, breath alcohol, blood alcohol, reproducibility, proficiency tests

Recently, there has been increased interest within the forensic community regarding uncertainty of measurement (UM), no doubt arising in part from the ISO requirement that laboratories report such uncertainties in a manner that is useful to the end users of the results. While the trend among accrediting bodies has been towards adapting ISO standards (1), aside from certain mathematically complex protocols (2,3) applicable mainly to physical measurements, there has been little discussion of this topic in the forensic literature. This article illustrates a simple and reliable approach for estimating uncertainties, which avoids such complexities. It is applicable to a wide range of chemical test results reported by forensic alcohol and toxicology laboratories.

Interlaboratory comparisons have long been recognized in the literature as an important means for estimating the range of errors (i.e., the UM) associated with a chemical analysis (4–8). Consequently, major standard setting organizations as diverse as ISO (9), Eurachem (10), the Association of Official Analytical Chemist International (11), the American Association for Laboratory Accreditation (12,13), and others (14,15) have accepted such comparison studies as a valid basis for estimating uncertainties. The forensic community, however, has been slow to embrace this approach.

For most sectors in the testing community, the application of this approach is limited by the high cost of performing interlaboratory comparisons, and the consequent scarcity of such data. However, because of its frequent participation in proficiency tests, the forensic toxicology community has an abundance of interlaboratory comparison data. This article illustrates three similar approaches for applying proficiency test data to the estimation of uncertainties for common assays: breath alcohol calibrations, blood alcohol determinations, and forensic toxicology (i.e., drug) quantitations. In each case, we will summarize the available data, explain the approach, and identify major assumptions and limitations.

Breath Alcohol

Each year since 2006, the Collaborative Testing Service (CTS; Sterling, VA) has provided a proficiency test that consists of two, 500-mL aqueous samples. Participants generate a gas-phase sample using a simulator and analyze the resulting gas sample nine times in succession. (A "simulator" is a device for generating reference gases by passing air through a temperature-controlled aqueous solution of ethanol.) Simulators can be operated in two modes: (i) In the "recirculation" mode, the gas recirculates through the analyzer and simulator, thereby minimizing evaporative losses and assuring complete saturation. In the "once-through" mode, the gas passes through the simulator and analyzer and then to exhaust. CTS requests that each participant complete this procedure using both the recirculating mode and the once-through mode, although not every participant complies with this request. To avoid effects because of incomplete saturation of the gas phase and flushing of the test chamber, the estimates in this article are based only on data obtained from the recirculation tests. At the completion of the tests, CTS provides a report giving each individual result, the average for each laboratory, and the mean and standard deviation of the entire dataset. During the period 2006–2008, between 49 and 60 laboratories participated in the calibration mode for each test. Details of

[1] Ventura County Sheriff's Department Forensic Sciences Laboratory, 800 S. Victoria Avenue, Ventura, CA 93009.

Received 10 Oct. 2008; and in revised form 20 April 2009; accepted 6 May 2009.

these surveys are available at the CTS website in the "breath alcohol" report series (16).

While laboratories are free to employ any type of analyzer, in practice nearly all participants used only infrared (IR)-based systems; and for those analyzers using both IR and fuel-cell detectors, only the results from the IR detector were included in the statistical summaries (CTS, personal communication, 2008). A few participants employed fuel-cell systems, but not in sufficient numbers to affect the summary results.

Errors associated with results from instrumental analyzers can often be considered to consist of two parts: a constant value that is independent of concentration and a variable part that is proportional to the concentration:

$$\mathrm{ERR}_{\mathrm{total}} = \mathrm{ERR}_1 + \mathrm{ERR}_2 * (\mathrm{conc}) \tag{1}$$

where ERR_1 and ERR_2 represent random sources or error (i.e., random variables, in the jargon of statisticians) and conc represents the concentration. According to this model, the variance (the standard deviation squared) of a set of results should vary as

$$\mathrm{VAR(C)} = a + b * (\mathrm{conc})^2 \tag{2}$$

providing that the ERR_1 and ERR_2 are independent. Here VAR(C) is the expected variance for measurements on samples having the specified concentration, and "a" and "b" are constants equal to the variances of ERR_1 and ERR_2, respectively.

Figure 1 shows the interlaboratory variance ("reproducibility" variance) as a function of $(\mathrm{conc})^2$ for the three CTS tests carried out during 2006–2008, and as can be seen, the relationship is as predicted. The 95% confidence ranges (CR-95, calculated as two times the standard deviation) taken from this figure can be summarized as

$$\text{CR-95} = \pm 6\%\,\text{relative},\ 85\,\text{mg/dL} < \text{concentration} \leq 230\,\text{mg/dL}$$

$$\text{CR-95} = \pm 5\,\text{mg/dL},\ 50\,\text{mg/dL} \leq \text{concentration} \leq 85\,\text{mg/dL}$$

This figure also predicts that the CR-95 will approach ±4 mg/dL as the concentration approaches zero. These values are broadly similar to the between-laboratory variations observed by Gullberg and Logan (17).

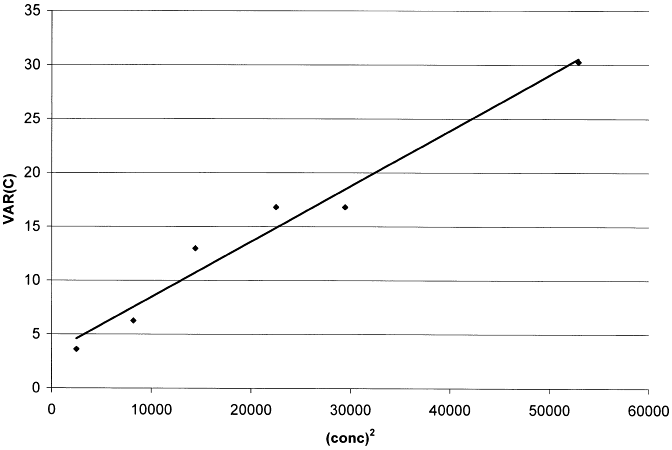

FIG. 1—*Interlaboratory variance exhibited by breath alcohol proficiency test results (Collaborative Testing Service, 2006–2008).*

These values apply to breath alcohol measurements from a particular laboratory, providing that certain broad assumptions are satisfied:

- The laboratory must be using the same type of analyzers as the participants—the CTS dataset is not applicable to fuel-cell type analyzers because of the limited number of participants using this technology. As a further refinement, individual laboratories may wish to limit their consideration to the data arising from the make of instrument used in their laboratory.
- The laboratory's internal quality control must be consistent with the CTS dataset. Laboratories typically check the calibration of their breath alcohol analyzers once a week or so, and the variation seen during such calibration checks should be less than or equal to the reproducibility standard deviation exhibited by the proficiency test results.

Given that these assumptions are satisfied and that there is no systematic difference between the consensus and true values, it is reasonable to consider the reproducibility standard deviation described as a sound measure of the accuracy of this method.

An individual laboratory may find that the standard deviation of its calibration checks is significantly smaller than the between-laboratory standard deviation, and on this basis may argue that its uncertainty is likewise smaller than that predicted by the proficiency test data. However, this is hardly a compelling argument when only a few proficiency test results are available. See, however, the following section for a more individualized approach.

As with any approach to uncertainty estimates, some limitations apply. In particular, this approach gives the range of errors one would expect when making measurements on randomly selected analyzers. This approach gives no information on a specific instrument, other than what can be expected for a set of instruments of similar type. In addition, the effect of spectral interferences resulting from concomitant compounds that might be found in breath, if any, are beyond this approach. Another limitation is that this approach may not fully account for differences that may arise from the use of bottled gas calibrators as opposed to simulators. That is, if a large number of laboratories adjust the calibration of their analyzers using a simulator (as opposed to a bottled gas) and then analyze the proficiency test sample using the same simulator, this approach would not account for variability that might arise from calibrating with an independent bottled gas. These differences should be small and should be controlled by exercising proper temperature control of the simulator.

Another consideration is that this approach applies only to gases that have been delivered into the sample chamber; it does not account for biological components or for errors associated with delivering the gas to the chamber. In other words, this approach addresses the accuracy of the analyzer *per se*. Assessing biological effects is no doubt interesting in its own right, but is well beyond the scope of the present note. In other words, the present approach assesses the errors ascribed to the measurement itself, without further implication.

Further examination of the CTS dataset reveals two additional facts. First, as described earlier, each solution was analyzed in sequence nine times in a single run by each laboratory. For the data examined, it was seen that the within-run variance ("repeatability" variance) was insignificant compared to the between-laboratory variance ("reproducibility" variance). This means that the within-run standard deviation does not reflect the analytical error associated with this method. This is consistent with the general observation that within-laboratory repeatability is consistently less than between-laboratory reproducibility (18). Second, as noted earlier, the vendor requested that participants analyze samples in both the recirculating and direct mode. Comparison of the results for each mode found that the average reading for the direct mode was *c*. 98% of the average value for the recirculating mode. This means that the gas passing through the simulator achieves at least 98% saturation in the direct mode. This is important because the direct mode is often used to calibrate other sorts of analyzers, such as fuel-cell systems, that lack the recirculation option.

Blood Alcohol

Of the many forensic specialties subject to proficiency testing, it is unlikely that any chemical test is subject to more frequent testing than is blood alcohol. Between 2004 and 2008, the College of American Pathologists (CAP; Northfield, IL) alone provided 75 blood alcohol test samples under their AL-1 series; and during the same period, CTS provided another 36 test samples under their 564 and 565 test series (19,20). In each of these tests, typically 100–200 participants contributed results from gas chromatographic analyses, the procedure employed by the vast majority of forensic laboratories.

Figure 2 shows the standard deviation as a function of concentration for those test samples in these series having a concentration greater than or equal to 10 mg/dL. Standard deviations for both the CAP and CTS datasets are seen to be similar for concentrations ≤100 mg/dL; but for higher concentrations, two distinct populations are clearly observed, with the CAP data demonstrating greater standard deviations. This phenomenon may reflect different operational pressures for participants in these surveys. Many clinical laboratories, for which speed is of the essence, participate in the CAP series, while mainly forensic laboratories, for which accuracy is paramount, participate in the CTS series. These observations suggest that the end user's requirements, as well as the method *per se*, may affect the reproducibility of results.

In view of the data illustrated in Fig. 2, a laboratory following the approach described earlier for breath alcohol might arrive at different estimates of its uncertainty depending on whether it participated in the CAP or CTS series. This situation illustrates the value of employing complementary approaches for estimating uncertainties, whenever available. Still, there would be little practical impact if forensic laboratories were to estimate their reproducibility from either dataset seen in Fig. 2, for two reasons: First, for the important range 0–100 mg/dL, there is little difference between the reproducibilities of the CTS and CAP datasets. Second, for higher concentrations, the reproducibility of either population suffices to demonstrate that the measured concentration is well above the legal threshold of 80 mg/dL, common in the United States.

The intent here, though, is to illustrate an alternative approach that avoids the ambiguities suggested by Fig. 2. Given the extent of blood alcohol proficiency testing, it is expected that many laboratories will have analyzed a sufficient number of proficiency samples to allow an individualized approach based on comparing their results to the consensus values, as follows: (i) Define the individual error (ind_err) as the difference between the laboratory and consensus value realized for each individual sample. (ii) Plot the values for ind_err against the consensus concentration to visualize any trends in performance. (iii) Summarize the average and standard deviation of ind_err using available statistical methods.

This approach is illustrated in Fig. 3*A,B* showing ind_err on an absolute and relative basis for a hypothetical dataset intended to be typical of a forensic laboratory. Here, the consensus concentrations were selected to be the same as those occurring in the CAP dataset shown earlier, but with normally distributed values of ind_err somewhat smaller than those exhibited by this CAP dataset. As can be seen, the behavior of ind_err can be roughly separated into three areas: (i) For concentrations <~100 mg/dL, the absolute error tends to be constant. The standard deviation for ind_err in this

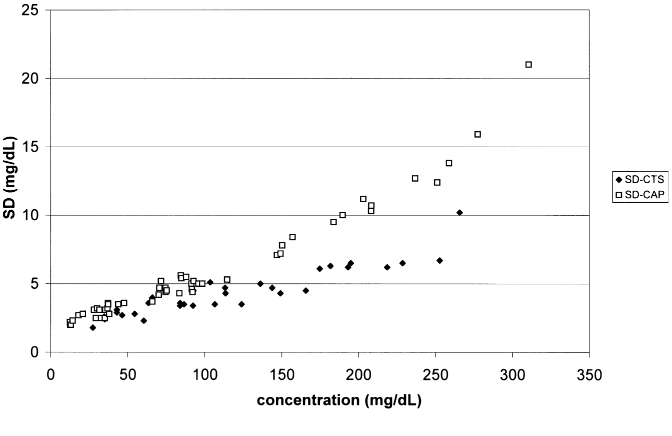

FIG. 2—*Interlaboratory standard deviations from blood alcohol proficiency tests (2004–2008). SD, standard deviation.*

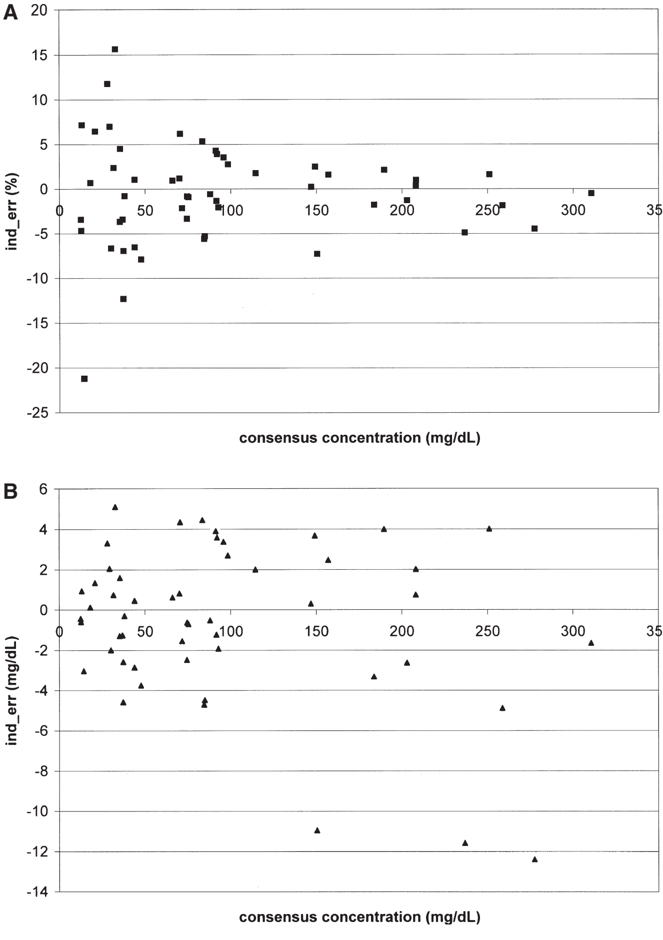

FIG. 3—*(A) Individual errors (ind_errs) for a hypothetical laboratory participating in the College of American Pathologists (CAP) AL-1 blood alcohol series, relative basis. Errors were generated to be normally distributed and typical of a forensic laboratory. (B) Ind_errs for a hypothetical laboratory participating in the CAP AL-1 blood alcohol series, absolute basis. Errors were generated to be normally distributed and typical of a forensic laboratory.*

range is 2.5 mg/dL. (ii) For concentrations >~150 mg/dL, the relative error is approximately constant. The relative standard deviation in this area is 2.4%. (iii) In the intermediate range 100–150 mg/dL, there is a scarcity of data. A conservative approach in this situation is to assign to this range the larger of the neighboring values, namely, 2.5%. Thus, the CR-95 for ind_err (after rounding) can be summarized as 5 mg/dL for the range 10–100 mg/dL and as 5% for the range 100–300 mg/dL.

Another important consideration is that the scatter of values about the consensus value appears symmetrical, and the average value for ind_err does not differ significantly from zero. This means that the method as employed by this hypothetical laboratory is unbiased with respect to the consensus value.

The principal assumptions that must be satisfied for this method to apply are (i) that the laboratory has used the same, or at least similar methods, throughout the period reflected in the estimate and (ii) that the consensus value is unbiased with respect to the true value. Unfortunately, the vendors for these proficiency tests do not provide target values with sufficient significant digits to assess the latter assumption; however, it seems unlikely all participating laboratories would somehow exhibit a bias for what is a relatively simple and well-characterized assay. Assuming that the consensus values are unbiased, then, the confidence ranges of ind_err represent a valid expression for the accuracy obtained by an individual laboratory.

One cautionary note applies here in that proficiency tests samples arrive well preserved and homogeneous, while some samples processed in a forensic laboratory, especially in support of death inquiries, may be clotted or putrefied. To address contributions because of nonideal samples, the laboratory can analyze selected samples from this population in replicate and then adjust uncertainty estimates accordingly.

Toxicology

Unlike the previous two examples, where each test contains the same compound, any given proficiency test in toxicology challenges the laboratory with only a small subset of the compounds that are actually determined by such laboratories. Another difference is that the concentrations of target compounds are much lower, typically in the range 10–100 ng/mL, meaning that issues of sample integrity and stability can be an issue.

Nevertheless, after the accumulation of sufficient data, certain patterns become apparent. This is seen, for instance, in Table 1 summarizing the last 7 years' data from the FTC surveys of the CAP (21). (Data in this table were compiled by the author from the individual reports listed in this reference.) It is seen here that the between-laboratory relative standard deviation is typically in the range 15–30%, with a root-mean-square average of 24%, seemingly regardless of the type of compound or whether the target is common or infrequent in case samples. Another observation from this series is that the consensus value (the average value reported by the participants) is on the average 20% lower than the expected (target) value; it is not clear whether this discrepancy is because of a bias on the part of the participants, or to the difficulty of preparing and preserving samples in the target concentration range. For this reason, data of this sort must be used cautiously for estimating the accuracy of toxicology methods. ("Accuracy" is used in this article in the normal sense to denote the closeness of a measured parameter to the true value of that parameter. Interlaboratory comparisons measure the comparability of a method, which may or may not be equivalent to the accuracy. In the case of this toxicology survey, comparability and accuracy differ.)

Fortunately, toxicology data are not generally interpreted in an absolute sense; rather, what matters to the data user is comparability between laboratories, or perhaps consistency within a single laboratory. Comparability to other laboratories is measured directly by comparing a laboratory's results to those of the consensus values according to either of the two approaches illustrated earlier, and the consistency within the laboratory is measured directly by its own precision control sample results. Given these considerations, the between-laboratory standard deviation of 24% seems to be one reasonable estimate of uncertainty relevant to the end user's needs. We leave it to the toxicologists among the readers to discern whether this estimate may in fact vary with compound class or whether it is improving with time. (See the penultimate column in Table 1.)

This approach assumes that the laboratory's internal quality control (QC) results are consistent with the between-laboratory standard deviations seen in the proficiency test and that the proficiency test results can be extended to compounds of similar type even when these are not included in proficiency tests. The latter assumption appears reasonable within the bounds of the data seen in Table 1; however, application to unusually difficult compounds should be approached cautiously. Another consideration is that the method in question is similar to those employed by other participants. One would not, for instance, use this approach to estimate the reproducibility of an LC/MS method for tetrahydrocannabinol (THC) when most other participants used GC/MS for this particular assay.

Another cautionary note is that the results shown in Table 1 have been purged of outliers by CAP prior to publication. As is typical of uncertainty assessments, the treatment of outliers is beyond the present scope.

It is interesting to note that the root-mean-square average of the relative standard deviation from Table 1 (24%) is significantly larger than would be predicted from internal precision control samples typical of GC/MS, the method employed for most tests represented in this table. In the author's experience, such internal samples typically exhibit a standard deviation of 10% or better, and this seems to be the general experience of the forensic toxicology community. Thus, both the Society of Forensic Toxicologists and the National Committee on Clinical Laboratory Standards (now the Clinical Laboratory Standards Institute) recommend control limits for precision control samples of ±20%, implying standard deviation for such controls of 8–10% or better (22,23). This means that the within-laboratory variance can only explain a minor part of the total between-laboratory variance. Furthermore, there are no recognizable factors that account for the difference between these values. These observations support the model that individual chemical testing laboratories develop protocols that are precise but biased and that such bias can only be recognized by the analysis of extramural samples.

Summary

To place these approaches in perspective, it is helpful to consider what is accepted in other parts of the chemical testing community, as described in the documents cited in the introduction to this article. A typical scenario described in these references is that 10 or so laboratories participate in an interlaboratory study during method validation (24). Other laboratories performing the same method at a later date may then use the interlaboratory standard deviation (the "reproducibility standard deviation") from the initial study as the uncertainty of its current results, provided that (i) the laboratory is using the same method; (ii) the laboratory's internal QC is consistent with the original performance of the method; and (iii) the original study encompassed all recognized sources of error. As is

TABLE 1—*Summary of proficiency test results for the FTC series, 2002–2008. Concentrations are in ng/mL unless otherwise indicated.*

Year	ID No.	Compound	Target	Consensus Value	Consensus/Target	Standard Deviation of Reporting Labs	Relative Standard Deviation of Reporting Labs	Number of Labs Quantitating
2008	FTC-B	Diazepam	1500	1255	0.84	164.6	0.13	92
		Nordiazepam	800	657	0.82	91.3	0.14	86
		Cyclobenzaprine	100	75.7	0.76	22.5	0.30	53
		Amphetamine	5000	4462	0.89	548	0.12	79
		MDMA	1000	791	0.79	116	0.15	76
		MDA	750	600	0.80	97.47	0.16	70
		Doxepin	8000	5898	0.74	1648	0.28	70
		Nordoxepin	1500	1037	0.69	256	0.25	49
2008	FTC-A	Hydrocodone	500	380	0.76	58	0.15	101
		Acetaminophen (μg/mL)	100	90	0.90	17	0.19	42
		Cocaine	100	61.8	0.62*	21.7	0.35*	86
		Benzoylecgonine	400	339	0.85	41.6	0.12	90
		N-desalkylflurazepam	200	158	0.79	28	0.18	55
2007	FTC-B	Hydromorphone	1000	785	0.79	106	0.14	46
		Ephedrine	1500	1177	0.78	190	0.16	31
		Ketamine	2000	1376	0.69	209	0.15	15
		Meperpidine	500	441	0.88	78.9	0.18	66
		Normeperidine	5000	3956	0.79	619.5	0.16	45
2007	FTC-A	Temazepam	500	462	0.92	150	0.32	37
		Diphenylhydantoin	5000	4100	0.82	900	0.22	49
		Oxccodone	200	152	0.76	30.7	0.20	72
2006	FTC-B	Fluoxetine	4000	2635	0.66	712	0.27	51
		Sertraline	2000	1183	0.59	510	0.43	50
		Norfluoxitine	1000	178	0.18	77	0.43	27
		Diphenhydramine	250	212	0.85	60	0.28	52
		Diazepam	1000	759	0.76	164	0.22	62
		Methamphetamine	1000	913	0.91	167	0.18	63
		Nordiazepam	100	81.8	0.82	26	0.32	52
		Amphetamine	100	96	0.96	22.7	0.24	57
2006	FTC-A	Phencyclidine	100	82.6	0.83	16	0.19	65
		EDDP	100	83.8	0.84	37	0.44	17
		MDMA	200	167	0.84	29	0.17	60
		MDA	50	41.9	0.84	7.9	0.19	36
		Methadone	1000	850	0.85	122.6	0.14	70
		Alprazolam	80	63	0.79	12.3	0.20	51
		Carisoprodol	4000	3800	0.95*	3800	1.00*	53
		Meprobamate	500	900	1.80*	600	0.67*	10
2005	FTC-B	Hydrocodone	600	474	0.79	78	0.16	64
		Tramadol	1000	901	0.90	212	0.24	49
		Phentermine	120	105	0.88	20.9	0.20	26
		Amitriptyline	2000	1598	0.80	318	0.20	59
		Nortriptyline	600	387	0.65	104	0.27	59
2005	FTC-A	THCA	100	81.2	0.81	16.9	0.21	38
		THC	25	17.1	0.68	5.1	0.30	30
		Propoxyphene	1500	1204	0.80	258	0.21	58
		Norpropoxyphene	500	295	0.59	62	0.21	33
		Cocaine	600	357	0.60	130	0.36	65
		BE	750	773	1.03	156	0.20	57
2004	FTC-B	Morphine	250	196	0.78	56.9	0.29	57
		BE	500	310	0.62	74.1	0.24	56
		Butalbital (μg/mL)	5	4.3	0.86	1.1	0.26	57
		Lorazapam	100	81.5	0.82	16.6	0.20	19
		Temazepam	900	788	0.88	153.8	0.20	45
		Cyclobenzaprine	60	53.6	0.89	17	0.32	30
2004	FTC-A	Nortriptyline	na	3078	na	1060	0.34	58
		Alprazolam	na	44.3	na	10.7	0.24	38
		Meprobamate (μg/mL)	180	153	0.85	31.2	0.20	47
		Secobarbital (μg/mL)	8	7.2	0.90	1.5	0.21	55
2003	FTC-B	Methamphetamine	300	268	0.89	43.8	0.16	70
		Amphetamine	50	45.8	0.92	10.6	0.23	41
		Methadone	200	177	0.89	32.1	0.18	63
		EDDP (methadone metabolite)	75	49.1	0.65	6.6	0.13	15
		Normeperidine	100	88.8	0.89	25	0.28	17
2003	FTC-A	Morphine	150	159	1.06	25	0.16	48
		Diphenhydramine (μg/mL)	50	42.9	0.86	9.6	0.22	54
		Nordiazepam	500	310	0.62	54.6	0.18	61
		Cyclobenzaprine	100	77.4	0.77	11.6	0.15	34

TABLE 1—*(Continued).*

Year	ID No.	Compound	Target	Consensus Value	Consensus/Target	Standard Deviation of Reporting Labs	Relative Standard Deviation of Reporting Labs	Number of Labs Quantitating
2002	FTC-B	THC	25	16.1	0.64	6.5	0.40	27
		THCA	100	70.6	0.71	28	0.40	30
		Desipramine	1000	709	0.71	200	0.28	51
		Imipramine	7000	4450	0.64	1135	0.26	54
		Diazapam	700	596	0.85	100	0.17	61
		Nordiazapam	300	206	0.69	49	0.24	55
2002	FTC-A	Codiene	3000	2468	0.82	421	0.17	59
		Butalbital (μg/mL)	30	23.8	0.79	4.1	0.17	54
		Hydrocone	150	212	1.41	72.2	0.34	48
		Alprazolam	80	71.4	0.89	19.4	0.27	27
		Sertraline	500	350	0.70	111.5	0.32	32
		Paroxetine	3000	2305	0.77	701	0.30	29

*Not used in summary statistics. THC, tetrahydrocannabinol; THCA, carboxy-tetrahydrocannabinol; BE, benzoylecgonine; MDA, methylenedioxyamphetamine; MDMA, methylenedioxymethamphetamine.

seen in the examples given above, the amount of data available to the forensic toxicology community far exceeds that which is typically available to the broader testing community.

There are, of course, other means for assessing the uncertainty of chemical measurements. For instance, to assess the uncertainty of breath alcohol systems, a laboratory that normally calibrates its breath testing instruments using a simulator might obtain and analyze a certified bottled gas standard. This would provide one estimate of uncertainty, at least at the single concentration of that reference material. In comparison, the proficiency test approaches described earlier measure uncertainties over the working range of the method. Alternately, for simple assays, a laboratory might measure its internal precision and then attempt to add any extramural sources of uncertainty. This approach might provide a credible estimate of uncertainty, providing that the method is simple and that the sources of external uncertainty can be identified. However, this approach clearly does not apply to the toxicology results discussed earlier, because in this case the extramural sources of uncertainty are large and of unknown origin. Whatever approaches are selected, the underlying assumptions must be identified and evaluated against the available data.

No discussion of uncertainty is complete without recognizing the popular "error budget" approach to estimating uncertainties (18), known traditionally as the propagation of error method. While this approach has appropriate applications, it is questionable whether routine chemical tests are one of these. In particular, the error budget approach assumes that the sources of error are known, small, and independent (24)—assumptions that are seen to be violated in a major way by the toxicology tests discussed earlier. In addition, the calculations required by the error budget approach can be daunting, and examiners will not look forward to explaining this approach to a jury. Another consideration is that the error budget approach requires the examiner to assume the shape of the distribution function for each contributing factor, when in fact the distribution function is often unknown. At best, the error budget approach for applications such as those described earlier remains controversial (5,6). Perhaps most importantly, there is no reason for engaging in questionable estimates and calculations, when direct measurements are available.

In summary, the use of proficiency test data, when available, provides an effective means for estimating the uncertainties associated with chemical measurements:

- It is complete in that it accounts for sources of error whether or not they are recognized by the local laboratory.
- The required calculations are simple and within the reach of anyone trained in the measurement sciences.
- The concept and process is easily explained to a jury.
- It is generally accepted in the chemical testing community.
- It provides a direct measure of uncertainty over a wide concentration range.

As with any approach, limitations apply, as discussed earlier.

Acknowledgment

It is a pleasure to acknowledge the advice and comments of Rod Gullberg of the Washington State Patrol on this topic.

Disclaimer

The opinions expressed in this article are those of the author alone and do not represent those of the Ventura County Sheriff's crime laboratory.

References

1. ASCLD/LAB. Available at: http://www.ascld-lab.org. Accessed April 18, 2009.
2. ISO. Guide to the expression of uncertainty in measurement. Geneva, Switzerland: International Organization for Standardization, 1995.
3. NIST. Guidelines for evaluating and expressing the uncertainty of NIST measurement results. Technical Note 1297. Gaithersburg, MD: US National Institutes for Standards and Technology, 1994.
4. Horwitz W. Evaluation of analytical methods used for regulation of foods and drugs. Anal Chem 1982;54(2):67A–76A.
5. Horwitz W. Uncertainty—a chemist's view. J AOAC Int 1998;81(4): 785–94.
6. Horwitz W. The certainty of uncertainty. J AOAC Int 2003;86(1):109–11.
7. Youden WJ, Steiner EH. Statistical manual of the Association of Official Analytical Chemists. Arlington, VA: Association of Official Analytical Chemists, 1975.
8. Boyer KW, Horwitz W, Albert R. Interlaboratory variability in trace element analysis. Anal Chem 1985;57:454–9.
9. ISO. Guidance for the use of repeatability, reproducibility and trueness estimates in measurement uncertainty estimation. Geneva, Switzerland: International Organization for Standardization, Technical Specification ISO/TS 21748:2004E, 2004.
10. Eurachem. Quantifying uncertainty in analytical measurement. Eurachem/CITAC Guide CG4, 2nd edn. 2000. Sections 7.6.3 & 7.12.2, Available at: http://www.eurachem.org/. Accessed February 8, 2009.
11. United Nations Food and Agricultural Organization (FAO). Guidance on measurement uncertainty, part J "comments." New York: FAO, 2007.

12. American Association for Laboratory Accreditation (A2LA). Guide G104. Estimation of measurement uncertainty in testing. Section 2.0. Frederick (MD): A2LA, 2002.
13. American Association for Laboratory Accreditation (A2LA). Guide G108. Guidelines for estimating uncertainty for microbiological counting methods. Maryland: A2LA, 2007; 6.
14. NordTest Report TR 537. Handbook for calculation of measurement uncertainty in environmental laboratories, 2nd edn. 2004;19. Available at: http://www.nordicinnovation.net/nordtestfiler/tec537.pdf. Accessed February 8, 2009.
15. European Directorate for the Quality of Medicines and Healthcare. Uncertainty of measurement, parts 1 and 2. 2007. Available at: http://www.edqm.eu/site/resultat_de_recherche-519.html. Accessed February 8, 2009.
16. Collaborative Testing Services. Test-568 breath alcohol simulator solution analysis. Available at: http://www.collaborativetesting.com/forensics/report_list.html. Accessed January 11, 2008.
17. Gullberg RG, Logan BK. Results of a proposed breath alcohol proficiency program. J Forensic Sci 2006;51(1):168–72.
18. Youden WJ, Steiner EH. Statistical manual of the Association of Official Analytical Chemists. Arlington, VA: AOAC, 1975;9–12.
19. College of American Pathologists (CAP). AL-1 whole blood alcohol/volatiles participant summary reports. Northfield, IL: CAP, 2004–2008.
20. Collaborative Testing Services. Reports on test series 564 and 565. Available at: http://www.collaborativetesting.com/forensics/report_list.html. Accessed January 11, 2008.
21. College of American Pathologists (CAP). FTC-A forensic toxicology participant summary reports. Northfield, IL: CAP, 2002–2008.
22. SOFT/AAFS. Forensic toxicology laboratory guidelines. Mesa, AZ: SOFT/AAFS, 2006. Available at: http://www.soft-tox.org/. Accessed February 8, 2009.
23. NCCLS. Gas chromatography/mass spectrometry (GC/MS) confirmation of drugs, approved guidelines. Wayne, PA: NCCLS document C43-A, NCCLS, 2002. Available at: http://www.clsi.org/. Accessed February 8, 2009.
24. Taylor JR. An introduction to error analysis. Sausalito, CA: University Science Books, 1997.

Additional information and reprint requests:
Jack Wallace, Ph.D.
Ventura County Sheriff's Department Forensic Sciences Laboratory
800 S. Victoria Avenue
Ventura
CA 93009
E-mail: jack.wallace@ventura.org