What Can Be Achieved in an External Quality-Assessment Scheme with a Small Number of Participants: Four Years of Experience with Thyroid-Related Tests in Israel

Luna Kahana and Haya Yechiel

The Israel External Quality-Assessment scheme (EQAS) provided service from 1985 to 1989, primarily to laboratories in Israel. Participation was voluntary and confidential, and involved 30–35 laboratories performing thyroid-related function tests. Scheme design included 36 human serum specimens distributed to the participants at the beginning of each year. For 12 months, three specimens were analyzed each month and the results were reported to EQAS on pre-agreed dates. Monthly reports sent by EQAS to participants after analysis of the results included the overall consensus mean used as target value, between-laboratory agreement, individual laboratory bias, and recovery data. The overall consensus mean calculated as all-laboratory trimmed mean was validated by reproducibility and recovery studies. Samples with undetectable concentrations of thyrotropin (TSH), obtained from patients proven thyrotoxic, were used to validate measurement of very low TSH concentrations. During the scheme, liquid serum was found superior to lyophilized specimens for distribution to the participants. The scheme helped stimulate major improvements in between-laboratory agreement, especially for low TSH concentrations, with CVs decreasing from 180% in the first year to 20% in the fourth year. Each laboratory’s performance also improved as judged by the median bias and variability of bias and by the considerable decrease in the number of laboratories with unacceptable performance for all tests. Method-related differences in performance were observed despite the small number of participants. Better methods, e.g., a radioimmunometric method for TSH, were detected and adopted by the participants, with concomitant improvement in performance. Transfer to analog procedures for free thyroxin, in addition to, but mostly instead of, thyroxin, was also documented, with no gain in analytical performance. Despite this being a small scheme, most of the goals achievable with a large scheme were realized.

Additional Keyphrases: interlaboratory performance · quality control · analog methods compared

Regional or national external quality-assessment schemes (EQAS),¹ according to the principles outlined by Whitehead (1) and Sunderland (2), have been widely used in Europe and the United Kingdom in the last few years for evaluating assays of hormones in general and thyroid-related hormones (3, 4) in particular. The main goal of an EQAS is to provide participants with information concerning their performance in comparison with that of other laboratories, so as to improve the reliability of analyses by the participating laboratories and to increase the between-laboratory agreement in reporting results. An EQAS is also a major tool for selecting better kits and improving the quality of assays (5) and can document trends in participants’ use of different assays, those affected by the market, or new commercial kits. EQASs are based on having many participants because the larger the number, the more reliable the information obtained (1). In this summary, we demonstrate that even a small scheme of not more than 35 participants can reach the majority of goals achievable by large interlaboratory surveys.

In January 1985, we organized a regional EQAS for the thyroid-related hormones thyroxin (T₄), triiodothyronine (T₃), thyrotropin (TSH), and free thyroxin (FT₄), with the participation of most of the Israeli laboratories performing thyroid tests and a few participants from Cyprus and Greece. The survey lasted four years, three of which were supported financially by the International Atomic Energy Agency. During the four years, 144 samples were sent to each participant in yearly dispatches, with three samples to be analyzed per month. From these results, we prepared monthly reports and yearly “end-of-period reports” summarizing participants’ performance and overall and method-related performance. Here, we summarize the four years of our small-scale EQAS.

Materials and Methods

Outlines of the scheme: Every year each participating laboratory received, in one dispatch, 36 identified samples of human serum nonreactive for hepatitis B (1985–89) and human immunodeficiency virus (1988–89). Instructions for storage, reconstitution of samples, and dates and units for reporting results were included. Every month, the participants analyzed three identified samples and reported the results on predefined dates to the organizer. Results were processed with a computer program we developed, based on the program of S. Sufi (6). A monthly report for individual determinations was printed, containing the overall consensus mean, median, SD, and CV for all laboratories’ results. We included both a graphic and numeric display of the par-
Denotes the results compared with the overall consensus mean and with other laboratories' results. All data accumulated during one year were used to prepare an end-of-period report containing data of cumulative mean bias, variability of bias, and recovery for each participating laboratory. Information on within-assay between-laboratory variability and method-related performance were also included, and sent to the participants.

Preparation of samples for distribution: Pooled sera from the surplus after routine analysis, covering a range of normal and pathological concentrations of thyroid hormones, were filtered through 0.22-μm pore-size filters (Whatman Ltd., Maidstone, Kent, U.K.) and supplemented with sodium azide, 2 g/L, as preservative.

Recovery samples for T₄ and T₃ were prepared by supplementing normal pooled sera thus obtained with pure T₄ or T₃ as sodium salts (Sigma Chemical Co., St. Louis, MO). The concentration of added T₄ and T₃ was determined gravimetrically and spectrophotometrically according to the method of Malan et al. (7).

Specimens with suppressed or undetectable TSH concentrations were obtained by pooling patients' sera proven clinically and biochemically to be thyrotoxic (T'Tx). Recovery samples for TSH were prepared by adding various amounts of TSH (2nd International Reference Preparation 80/558, kindly supplied by the National Institute of Biological Standard and Controls) to the above pooled sera from T'Tx patients.

For the first two years, the samples thus obtained were distributed after lyophilization (Bio-Yeda Ltd., Rehovot, Israel, and Zer Laboratories, Jerusalem, Israel), and the participants were instructed to store the samples at 4°C until reconstitution and analysis. During the last two years, the samples were not lyophilized and were distributed in liquid form with instructions to store them at −20°C until analysis.

Data analysis: The overall consensus mean is calculated as the "all-laboratory arithmetic trimmed mean" (ALTM) according to the method of Healy (8), from the participants' monthly results. The ALTM is the target value for T₄, T₃, FT₄, and TSH in our scheme. The SD and trimmed CV of all laboratory sample results are calculated for monthly reports and represent the overall between-laboratory agreement. Outliers are defined as those laboratories with a result out of the range of the ALTM ± 3 trimmed SD (8). The bias is calculated for each laboratory on each usable sample by subtracting the ALTM from the laboratory's result. Division of the bias by the ALTM and multiplying by 100 expresses the percentage of bias from the ALTM and is reported monthly for each sample. Laboratory mean bias, reported yearly for each laboratory, is the arithmetic trimmed mean of participant's biases (expressed as a ratio) in comparison with the ALTM for all of the participant's reported results. Laboratory variability of bias, reported yearly, is the CV of the participant's biases with respect to the ALTM.

The stability of the ALTM was validated by recording the reproducibility of the ALTM for specimens distributed on more than one occasion. A CV was computed from the ALTM of the same pool assayed on two or more occasions during the year. The mean CV for the ALTM was obtained by pooling the CVs of the ALTM for different pools distributed more than once, according to the method of Pilo et al. (9).

Validation of the ALTM for T₄ and T₃ by analytical recovery studies was calculated as follows: the ALTM of the pool before the addition of T₃ or T₄ was subtracted from the ALTM of the supplemented specimens, divided by the added T₃ or T₄ concentration, and multiplied by 100.

The ALTM for the recovery of TSH was calculated by simply dividing the ALTM value of the supplemented sample (TSH was added to pooled T'Tx sera with undetectable TSH concentration) by the added TSH concentration and multiplying by 100.

Results and Discussion

Stability of lyophilized vs liquid samples distributed: As stated in Materials and Methods, samples were distributed to the participants in the first two years in lyophilized form and in liquid form in the last two years. A comparison between the two forms of distribution was performed by sending the same pooled sera, both lyophilized and in liquid form, for analysis on the same dates four times during the second year of the EQAS. Table 1 shows the ALTM and its variability (CV) during the year for the two forms of sample preparation. No marked difference was observed in the mean CV of the ALTM in lyophilized vs liquid form samples for T₄, T₃, TSH, and FT₄. Thus, we decided to distribute the samples in liquid form because it (a) is cheaper and easier to prepare for distribution; (b) is more like the everyday patients' samples arriving in the laboratory; and (c) needs no reconstitution, thus avoiding errors that might be introduced by reconstituting lyophilized specimens (10).

Validation of the ALTM as a target value for T₄, T₃, FT₄, and TSH: The ALTM as a target value was validated through stability and recovery studies. The stability of the ALTM for T₄, T₃, and FT₄, recorded for specimens distributed on more than one occasion and expressed in terms of mean CV for the ALTM (see Materials and Methods), was 2.8%, 4.9%, and 3.1% in

| Table 1. ALTM and Its Variability (CV) as Obtained with Lyophilized (Ly) vs Liquid (L) Samples |
|-----------------------------------|---------|---------|---------|---------|
|        | T₄ µg/L | T₃ µg/L | TSH, | FT₄ ng/L |
|        | Ly     | L       | Ly    | L       |
| Dates  |        |         |       |         |
| 17.2.86 | 83     | 86      | 1.06  | 1.19    | 2.7   | 2.8 |
| 19.5.86 | 86     | 86      | 1.14  | 1.22    | 2.7   | 2.6 |
| 22.9.86 | 87     | 87      | 1.27  | 1.38    | 3.2   | 3.0 |
| 19.1.87 | 87     | 87      | 1.64  | 1.41    | 2.5   | 2.2 |
| ALTM CV% | 1.7    | 0.6     | 7.7   | 7.6     | 8.2   | 9.6 |
|         |         |         |       |         | 3.5   | 1.4 |
the first year of the scheme and decreased to 1.4%, 2.1%, and 2.5%, respectively, in the fourth year of the scheme. The mean CV of the ALTM for TSH was 10.4% during the first year of the scheme, and decreased to 5.0%, 2.7%, and 2.4% in the second, third, and fourth years of the scheme, respectively. Thus, the stability of the ALTM, except for TSH in the first year of the scheme, proved satisfactory.

The recovery for T₄ and T₃, as calculated for the ALTM, was rather low at 73%–77% for T₄ and 70%–82% for T₃ in the first two years of the scheme. There being no reason to believe that the methods used by the participants, mainly Amersham International plc (Bucks., U.K.) and DPC kits (Diagnostic Products Corp., Los Angeles, CA), were all negatively biased, we assumed that the problem was with the way the recovery samples were prepared. Indeed, in the next two years, when liquid instead of lyophilized samples were distributed, the analytical recovery obtained improved markedly, to 88%–101% for T₄ and 86%–91% for T₃. Perhaps through lyophilization, some of the externally added T₄ and T₃ hormone was destroyed. However, no similar phenomenon was observed for endogenous T₄ and T₃, because no difference in the ALTM for lyophilized and liquid samples was observed for unsupplemented samples (Table 1). For the same reason, we do not think that the lower recoveries obtained with lyophilized samples were due to the additional volume attributable to lyophilized powder, on subsequent reconstitution of the sample, for which in our scheme no correction was done (10).

Very good recoveries for the ALTM as a target value for TSH were obtained either with lyophilized (100%) or liquid samples (90–101%). Apparently, added TSH is less vulnerable to lyophilization than is T₄ or T₃.

Overall between-laboratory agreement was expressed as the arithmetic CV of the trimmed sample results (see Materials and Methods) for every sample. We averaged, according to Pilo et al. (9), all the CVs thus obtained during the year. Table 2 summarizes the interlaboratory mean CV for high, normal, and low ranges of concentration for each test over the four years. The most dramatic change in the interlaboratory CV was achieved for TSH, mostly in the lower concentration range, decreasing from 180% in 1986–86 to 20% in 1986–89. The main reason for the improvement was the introduction of new sensitive immunoradiometric assays (IRMA) of TSH with improved specificity. The scheme was able to demonstrate the superiority of the new methods, as shown below, and to provide this information to the participants; this, in turn, we believe stimulated the participants to shift to the IRMA (Table 3). The shift of almost 70% of the participating laboratories to the same kit could be another reason for the decrease in the interlaboratory variation for TSH in the fourth year of the scheme.

A slight improvement in the overall between-laboratory agreement of T₄ and T₃ over the years was also observed, although no changes in the methods of the participants was documented. No change in the interlaboratory CV for FT₄ was observed, which remained unreasonably high during the four years of the scheme.

Method-related performance: An EQAS with many participants is a major tool for selecting better methods or kits during updates of analytical performance of various kits. We tried to obtain that information with our more limited scheme to improve the performance of TSH analysis, which was rather poor at the beginning of our EQAS, with interlaboratory CV >180% for low TSH concentrations (Table 2). To detect biased and inappropriately standardized methods, we studied baseline validity and recovery for TSH. Baseline validity is not stressed much in general clinical chemistry but is of great importance in endocrinology, to wit: the measured TSH concentration should be undetectable in samples known to contain very low concentrations, e.g., in TTX or T₃-suppressed sera (11). Baseline for TSH was checked by distributing unmanipulated TTx samples, which should be measured as containing TSH concentrations <1 milli-int. unit/L. Analytical recovery was studied with TTx sera supplemented with known amounts of TSH, as described in Materials and Methods.

In the 1986–87 scheme, only 54% of the reported results for the TTx samples indicated TSH values <1 milli-int. unit/L, whereas the remaining 46% reported greater values. Method-related analysis of the reports
detected two methods, DPC-RIA and CIS-RIA (International CIS, Gif sur Yvette, France), dominating the positively biased reports. In addition, the recovery for TSH-supplemented samples for these two methods was rather low, 77% for DPC-RIA and 72% for CIS-RIA. Table 3 shows that, during that year (1986–87), 56% of the participants used these two kits, both of which encountered interference as expressed by positive bias in physiologically suppressed TSH samples and standardization problems as observed by the low recoveries.

This information was reported to the participants, stressing the excellent baseline validity and very good recovery obtained with other methods such as IRMA (by Serono Diagnostic SA, Coinsins, Switzerland, and DPC) and RIA (Amersham). All the reports for TTx samples, by these three methods, were <1 milli-int. unit/L for TSH, and the recoveries obtained for supplemented samples were >89%. The better methods, IRMAs by Serono and DPC, were adopted by most of the participants in the next two years, whereas the Amersham RIA, although satisfactory according to the baseline validity and recovery criteria, was not adopted, probably because of a lack of analytical sensitivity (not shown in this scheme). Consequently, only 7% of the reports for TTx-proven sera were positively biased (TSH >1 milli-int. unit/L) during the 1988–89 scheme, and all of them reported by DPC-RIA users.

**Participants' performance:** Table 4 shows the participants' performance in terms of bias and variability of bias toward the ALTm as target value. The improvement in participants' performance is shown by the much lower median bias for all the tests and considerably lower variability of bias for TSH, T₃, T₄, and FT₄ in 1988–89 as compared with 1985–86. The striking improvement in performance is also represented in Table 5, with a much lower percentage of laboratories showing unacceptable performance during 1988–89 as compared with 1985–86.

The criteria for unacceptable performance were adopted from the U.K. Thyroid-Related Hormones Scheme after approval by our local committee. The criteria used (bias and variability of bias, respectively) were T₃: >±10%, >15%; T₄: >±20%, >20%; TSH: >±25%, >25%; FT₄: >±10%, >15%. A laboratory with either or both variables outside the above-stated limits for a test was defined as having unacceptable performance for this test.

The marked decrease in the percentage of laboratories with unacceptable performance during 1985–87 occurred because, during this year, we did not include results in the laboratories' performance for TSH concentrations <5 milli-int. units/L. In the first, third, and fourth years, the results for TSH concentrations <5 milli-int. units/L were included in reports of the laboratories' performance.

**Trends in the use of different thyroid tests:** Over the four years of the scheme, about a 20% decrease in T₄ concomitant with about a 20% increase in FT₄ (analog) users was observed, although the analytical performance was considerably worse, compared with that for T₄, as shown by (a) the interlaboratory interassay CV (Table 2), (b) the participants' median variability of bias (Table 4), and (c) the percentage of laboratories with unacceptable performance (Table 5). Thus, additional factors besides analytical performance and quality-control criteria affect trends in the use of some tests.

In conclusion: During the four years of the Israel EQAS, despite its being a small scheme of 30–35 participants, the following information could be provided to the participants: (a) participants' performance in comparison with other laboratories and relative to a target value, the ALTm; (b) concentration-dependent between-laboratory agreement, and (c) method-related information identifying reliable and less reliable methods, e.g., for TSH. Under the influence of the EQAS, a marked improvement was observed in laboratory performance of all the tests, especially for TSH, and in between-laboratory agreement for T₃, T₄, and particularly TSH. The results are an example of the value of EQAS in improving the analytical performance of immunoassays.

We acknowledge the financial support of the International Atomic Energy Agency research contract no. 3695-R2/RB. We also thank Dr. S. Suft (U.K.) for his help and advice in the establishment of this scheme, Dr. Horenstein (Israel) for testing the sera for the presence of hepatitis B virus, and Mr. Askenazi (Pharmatope Pharmatrade, Israel) for distributing the quality-control samples to the participants. Special thanks to all Israeli laboratories involved for cooperation and interest.

**References**


**Table 4. Median Participants' Percentage Bias (B) and Variability of Bias (VB) as Calculated for the ALTm**

<table>
<thead>
<tr>
<th>Years</th>
<th>T₃</th>
<th>T₄</th>
<th>TSH</th>
<th>FT₄</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>B VB</td>
<td>B VB</td>
<td>B VB</td>
<td>B VB</td>
</tr>
<tr>
<td>1985–86</td>
<td>5 11</td>
<td>9 15</td>
<td>24 39</td>
<td>6 13</td>
</tr>
<tr>
<td>1986–87</td>
<td>2 8</td>
<td>1 11</td>
<td>11 24</td>
<td>0 13</td>
</tr>
<tr>
<td>1987–88</td>
<td>0 9</td>
<td>1 14</td>
<td>1 19</td>
<td>2 14</td>
</tr>
<tr>
<td>1988–89</td>
<td>1 8</td>
<td>-1 10</td>
<td>0 10</td>
<td>1 12</td>
</tr>
</tbody>
</table>

**Table 5. Percentage of Laboratories with Unacceptable Performance during 1986–89**

<table>
<thead>
<tr>
<th>Year</th>
<th>T₃</th>
<th>T₄</th>
<th>TSH</th>
<th>FT₄</th>
</tr>
</thead>
<tbody>
<tr>
<td>1985–86</td>
<td>32</td>
<td>29</td>
<td>79</td>
<td>36</td>
</tr>
<tr>
<td>1986–87</td>
<td>19</td>
<td>8</td>
<td>27*</td>
<td>18</td>
</tr>
<tr>
<td>1987–88</td>
<td>21</td>
<td>14</td>
<td>32</td>
<td>28</td>
</tr>
<tr>
<td>1988–89</td>
<td>10</td>
<td>0</td>
<td>8</td>
<td>26</td>
</tr>
</tbody>
</table>

* Only concentrations of TSH >5 milli-int. units/L were included for calculating bias and variability of bias. In other years, concentrations <5 milli-int. units/L were also included.
Effects of Storage Time and Temperature on Urinary Enzymes

Elena Matteucci, Giovanna Gregori, Luisa Pellegrini, Renzo Navalesi, and Ottavio Giampietro

We measured three renal tubular brush-border enzymes (lactate dehydrogenase, LDH, EC 1.1.1.27; gamma-glutamyltransferase, GGT, EC 2.3.2.2; and alkaline phosphatase, AP, EC 3.1.3.1) in morning urine samples from 48 healthy subjects to check whether different storage times and temperatures could modify enzyme concentrations. Short-term (24 h) storage time at room temperature or 4 °C does not affect urinary enzyme activity. A few days of freezing, at −20 or −70 °C, dramatically lowers LDH and AP values; GGT is partially preserved only at −70 °C, if the sample has been previously centrifuged. Urinary enzymes investigated in this study are extremely labile at low temperatures.

Additional Keyphrases: lactate dehydrogenase • glutamyltransferase • alkaline phosphatase • sample handling

Measurement of urinary enzymes is currently used as a noninvasive test of renal integrity in clinical nephrology (1–24). However, several methodological problems arise if an unfavorable environment for the enzymes is used, e.g., urine. Urine as a medium is extremely variable in volume, composition, pH, ionic strength, and the presence of interfering substances, blood, epithelial cells, and microorganisms. Hence, generally approved standardized procedures to measure enzymes in urine are lacking.

A decisive problem concerns the stability of enzyme activity during the collection and storage of urine samples before assay. Most published papers fail to report any details about how to treat urine samples during collection or how to store them. Some investigators (1, 3, 6, 7, 11, 14, 19, 21, 23) suggest centrifugation before the assay, with the assay to be performed promptly (22) or within 4–6 h (19, 21) after centrifugation. Others report that urine may be stored deep-frozen without enzyme loss (7).

Here we report the effects of different storage times and temperatures on some enzyme activities in urine samples, centrifuged or not, collected from healthy subjects.

Materials and Methods

Enzymes studied. We chose to study three brush-border enzymes, because we are involved in programs dealing with tubular diabetic nephropathy (7, 15, 21). These enzymes are gamma-glutamyltransferase (GGT; EC 2.3.2.2), lactate dehydrogenase (LDH; EC 1.1.1.27), and alkaline phosphatase (AP; EC 3.1.3.1).

Equipment. We used in these manual assays a DU-8 ultraviolet–visible computing spectrophotometer (Beckman Instruments, Inc., Fullerton, CA) with a temperature-controlled cuvet holder; Pipetman 1000 from Gilford Medical Electronics (France) S.A., and micropettes from Socorex (Swiss made); silica cells from Helma Italia srl; and a Model J6B centrifuge (Beckman).

Reagents. GGT activity was measured at 610 nm with gamma-glutamyl-3,5-dibromo-4-hydroxyanilide (gamma-glu-DHBA) substrate in a Sera-Pack kit (code 6679; Miles Italiana S.p.A., Ames Division, Cavenago Brianza, Italy). AP activity was measured at 405 nm with sodium p-nitrophenyl phosphate as substrate (Sera Pack kit; code 6677). LDH activity was measured at 340 nm via the conversion of pyruvate to lactate, which is proportional to the rate of NADH oxidation (Sera-Pack kit; code 6393). Control sera (Sera-Chek®; code 6656) were kindly provided by Miles Italiana. p-Nitrophenol

Cattedra di Malattie del Metabolismo, Istituto di Clinica Medica II, Università degli Studi di Pisa, Italy.

Received December 7, 1990; accepted May 10, 1991.

1 Address correspondence to this author, at: Istituto di Clinica Medica II, Via Roma, 67, 56100 Pisa, Italy.

8 Nonstandard abbreviations: LDH, lactate dehydrogenase; GGT, gamma-glutamyltransferase; AP, alkaline phosphatase; and gamma-glu-DHBA, gamma-glutamyl-3,5-dibromo-4-hydroxyanilide.