The frequency of rounds in a proficiency test: does it affect the performance of participants?

Michael Thompson* and Philip J. Lowthian

Department of Chemistry, Birkbeck College, University of London, Gordon House, 29 Gordon Square, London, UK WC1H 0PP

Received 3rd September 1998, Accepted 15th October 1998

A controlled experiment has been carried out to test whether the performance of participants in a proficiency test is improved as a result of doubling the frequency of distributing the test materials in the test. The performance of a randomly selected experimental group was compared with that of a corresponding control group after doubling the frequency of rounds in the experimental group over a period of two years. No statistically significant change in performance was detected.

The design, organisation and interpretation of proficiency tests are now regarded as well-established topics.1–3 However, certain aspects of proficiency tests are, within wide limits, left to the judgement of the organisers. The frequency with which rounds of the test are distributed is one such aspect: it varies widely among schemes in the UK, between limits as disparate as 3 and 26 per year. The frequency adopted in a scheme clearly has financial consequences for the participants, as the cost of participating will be roughly proportional to the frequency. The question naturally arises as to what is the minimum frequency that will achieve the aims of a scheme, because that would be the most cost-effective frequency. However, decisions about frequency of rounds are based only on professional judgement at present.

There was hitherto no experimental evidence to show whether higher round frequencies gave rise to better performance among the participants. Such claims, when they occur, are often anecdotal. Studies based on historical data from a number of different proficiency tests have been attempted by the authors, but the results were ambiguous because of confounding variables, lack of statistical control, or an unsatisfactory definition of performance.4 For example, differences in performance among a small number of different proficiency tests could be attributed to the numerous differences between the tests other than round frequency. Moreover, it is never clear that a measure of performance is truly comparable between proficiency testing schemes. Finally, simply observing a change in performance among laboratories in a particular scheme subsequent to an increase in round frequency is not at all conclusive: such an outcome could be the result of many factors other than round frequency.

In the present study we overcome these problems by restricting the study area to particular determinations within a single proficiency test scheme, and by using a randomly selected control group.

Experimental

Experimental data were obtained by studying the performance of laboratories in the FAPAS™ scheme (Food Analysis Performance Assessment Scheme, CSL Food Science Laboratory, Norwich Research Park, Colney, Norwich NR4 7UQ). This scheme includes a number of different types of analysis (‘series’) in the food analysis sector, and each series usually covers several different analytes. Rounds within a series are distributed to the participants mostly at a standard frequency of three per year. Data were collected in the period 1995–1997, at which time FAPAS could be regarded as a mature, stabilised scheme.

The frequency doubling was restricted to two popular series from FAPAS, Series I (proximate components of foodstuffs) and Series XIV (fatty acids). The results from four analytes from Series I (moisture, fat, nitrogen and ash) and three analytes from Series XIV (saturated fatty acids, monounsaturated fatty acids and polyunsaturated fatty acids) were considered. Popular series and uncomplicated analysis were features chosen to maximise the number of volunteers. In the event there were 56 volunteers among participants in Series I and 37 in Series XIV.

There are severe design constraints on an experiment of this type. The normal operation of the scheme cannot be disturbed because of problems for the organisers, and because the participants rely on the results of the scheme to support accreditation. This restricts possible intervention for experimental purposes to changing the round frequency to a low harmonic of the established frequency. It is also necessary to disguise the true purpose of the experiment, to avoid inducing a ‘self-fulfilling prophecy’ (that is, the participants in the experiment might try unconsciously to achieve what they think is the desired outcome).

The format selected for the present study was to double the round frequency for a subset of FAPAS participants from a particular time and, after various intervals, compare the performance of this experimental group with another subset (the control group) operating at the normal frequency of three rounds per year. The experimental group and the control group were selected at random from those volunteering to join in the experiment. The non-volunteers (abstainers) were unsuitable as a control group because of the a priori possibility that the volunteers would be more motivated in matters of data quality and would perform better even before the experiment began. In addition there might be an analogue of the ‘placebo effect’ in this experiment: merely volunteering for participation might make the control group laboratories perform better subsequently, as they would know that their FAPAS scores were being carefully scrutinised, so the control group had to be
volunteers. The experimental and control groups were told that the experiment was concerned with the effect of round frequency on internal quality control schemes.

The round frequency doubling was effected for the experimental group by the introduction of intercalated rounds at the mid-point in time between the regular rounds. (In the Series XIV experiment there was a hiatus in the distribution for reasons unconnected with the experiment: to maintain the enhanced frequency, two intercalated rounds were introduced at one point.) In order to ensure that the intercalated rounds were taken seriously, they were organised in exactly the same manner as the regular rounds, and a report was issued in identical format. However, only results from the regular rounds (undertaken by both experimental and control groups) could be used for comparison purposes. Moreover, it is valid to compare the two groups only within an individual round. Although successive rounds (both regular and intercalated) used the same types of test materials and the same analytes, it is clear from previous studies that even experienced FAPAS participants find some rounds more difficult than others, i.e., there are apparent changes in overall proficiency from round to round. This is presumably a result of variations in the matrix composition and analyte concentration in the test materials. The effects of these round-to-round variations are eliminated by considering only within-round comparisons. The regular rounds considered in the study were the three rounds preceding the introduction of the intercalated rounds and those immediately succeeding each intercalated round.

The primary measure of performance selected in this study was the robust standard deviation of the $z$-scores of the participants in any group. The rationale for using this statistic was that it represented the tendency of the results within a group to converge to the assigned value for the material. As such it gives an overall measure of the accuracy obtained by individual participants. At the same time the effect of outliers on the statistic is considerably reduced, so that the outcome represents the behaviour of the great majority of participants rather than the erratic few.

The FAPAS $z$-scores produced in each regular round were processed for each analyte by estimating the robust standard deviations separately for the experimental group ($s_{ex}$), the control group ($s_{con}$) and the abstainers ($s_{ab}$). The AMC algorithm was used to calculate the statistics. The ratios $s_{ex}/s_{con}$ and $s_{con}/s_{ab}$ were calculated along with their standard errors (under the random normal assumption). These ratios would be free from effects produced by round-to-round changes in difficulty with the test material. The same ratios were estimated from the combined $z$-scores of all of the analytes in a series, with the intention of making tests for changes more sensitive by increasing the number of observations.

These ratios were then collated as time series (Fig. 1–9) so that any changes in the ratios, subsequent on the introduction of

Fig. 1 Standard deviation ratio (experimental/control) for Series I—Moisture.

Fig. 2 Standard deviation ratio (experimental/control) for Series I—Ash.

Fig. 3 Standard deviation ratio (experimental/control) for Series I—Fat.

Fig. 4 Standard deviation ratio (experimental/control) for Series I—Nitrogen.

Fig. 5 Standard deviation ratio (experimental/control) for Series I—Combined analytes.
the intercalated rounds, would be apparent visually or in a statistical test of significance.

Results and discussion

The null hypothesis $H_0$ is that $\sigma_{\text{ex}} = \sigma_{\text{con}}$. If the frequency doubling had no consequences for the efficacy of the proficiency test (and therefore on the performance of the experimental group), we would expect to see the observations (values of the ratio $s_{\text{ex}}/s_{\text{con}}$) distributed at random (although not symmetrically) around a mean value of unity. In particular, in the regular rounds before the first intercalation (indicated by negative round numbers in the figures) there is no distinction in treatment between the experimental and control groups, because they were selected at random from the pool of volunteers after these rounds were completed, and we would expect to see this random distribution. However, if there were a detectable effect of frequency doubling it would be manifest as a trend in the ratio $s_{\text{ex}}/s_{\text{con}}$ starting after the intercalation. Prima facie we expect the trend to be negative (because the performance of the experimental group would be expected, if anything, to improve relative to that of the control group and the relative value of $s_{\text{ex}}/s_{\text{con}}$ to decrease). However, it is at least conceivable that the opposite could occur, for instance if the increased familiarity made the analysts more casual in their approach to quality, or simply less accurate because of overwork. Therefore, two-tailed tests of significance were used in the study.

Visual appraisal of the plots, taking into account the standard errors of the ratio value (shown on the figures as approximate 95% confidence intervals) and the null hypothesis ($H_0$), shows no apparent trends. Therefore, statistical tests were resorted to. Although no standard statistical test is ideal to detect such a trend, two possibilities were considered, namely (a) weighted regression of the points as a time series, and (b) a two sample test of values of each ratio before and after the onset of intercalation. In the former case, a measurable trend in the data would be manifest as a value of $b$ significantly different to zero. Over the seven separate analytes and the two sets of combined results (Fig. 1–9), only one such (fat in Series I) showed a significant regression, with a positive value of $b$ indicating a relative deterioration in performance of the experimental group after the onset of the intercalation. An examination of the plot (Fig. 3) shows the reason clearly to be the group of low values of the ratio $s_{\text{ex}}/s_{\text{con}}$ occurring before the intercalation (where the expectation of the ratio would be unity and any deviations from that purely random). After the intercalation the ratio values cluster closely around unity. This result therefore seems to be ‘pathological’. The overall conclusion here must be that there is no significant effect detectable in the experiment and, consequently, no evidence that performance in these series of FAPAS would be improved by increasing the round frequency. The two-sample tests supported these findings.

Inspection of the standard errors of the $b$ values suggests that the smallest relative change in the ratio $s_{\text{ex}}/s_{\text{con}}$, that could have been detected under the experimental conditions would have been about 20%. A greatly increased size of experiment, both in numbers of participants and numbers of rounds, would be required for a noticeably more sensitive test. An improvement in $s_{\text{ex}}/s_{\text{con}}$ of less than 20% in the ratio is unlikely to translate into a useful change for the individual participant. In the absence of a significant slope, we would expect the mean of all of the $s_{\text{ex}}/s_{\text{con}}$ results on a plot not to be significantly different to unity. That expectation was also confirmed.

As a supplementary experiment, an examination of the ratios $s_{\text{con}}/s_{\text{ab}}$ for the combined results of each series was made to compare the performance of the experimental group with that of the group not volunteering for the experiment (Fig. 10 and 11). In one instance (Series I) there was a significantly low mean...
value but no significant slope. This confirms to some extent the conjecture that the abstainer group would have a poorer performance than the volunteers and would therefore be unsuitable as a control.

In a further supplementary experiment, the performance of the experimental group in the intercalated rounds was compared with its performance in the regular rounds. This was done to test whether the experimental group took the intercalated rounds seriously, as a casual attitude to them would have mitigated any possible beneficial effect on performance in the regular rounds. The results (robust standard deviations) are shown in Fig. 12 and 13. The Series I results show no significant differences between the regular and intercalated rounds, and no trends or other patterns. The Series XIV data, apart from a striking outlier (which seems to reflect an unusually difficult test material affecting all analytes) shows a slight trend downwards but no clear differences between the regular and the intercalated rounds. The trend cannot be safely interpreted as an improvement in performance with time because the data are not controlled.

Conclusions

The overall result of the experiment was that no detectable improvement in performance was produced by frequency doubling under the given conditions. The present FAPAS frequency of three rounds per year could not be improved upon in Series I and Series XIV. Although the test was not very sensitive to changes in performance, only a much bigger (and proportionately more expensive) experiment could have provided a worthwhile improvement in sensitivity. Given the generally negative findings in the present study, there seem to be no strong grounds for conducting such an experiment.

It should be noted that the use of robust statistics ensures that the above conclusion (no significant improvement in performance) applies to the great majority of the participants. However, the small proportion of poor performers may conceivably have been otherwise affected (although there is insufficient information to test this). In any event, increasing round frequency in a scheme for the benefit of a small minority of participants is almost certainly not a cost-effective strategy.

As the experiment was restricted to Series I and Series XIV of FAPAS it is worthwhile to consider whether the above conclusions could be applied in a wider context, that is, to other series in FAPAS or, indeed, to other proficiency tests. Such extrapolations can only be tentative, but would be reasonable in FAPAS for established series where the scheme is fulfilling its aims and intentions. In contrast, one could readily envisage that a different result might be obtained in a scheme that was immature (where the majority of the participants were still improving) or in a mature scheme where a large minority of participants were consistently under-achieving. Even so, an increase in frequency would again be unlikely to be the most cost-effective way of bringing about an improvement. Such matters are, of course, best left to the advisory committees of individual schemes.

This work was supported by the Ministry of Agriculture, Fisheries and Food and conducted with the cooperation of the participants and Secretariat of FAPAS™.

References

4 M. Thompson and P. J. Lownthian, unpublished data.

Fig. 10 Standard deviation ratio (control/abstainer) for Series I—Combined analytes.

Fig. 11 Standard deviation ratio (control/abstainer) for Series XIV—Combined analytes.

Fig. 12 Standard deviations for regular rounds (open circle) and intercalated rounds (closed circle) for Series I—Combined analytes.

Fig. 13 Standard deviations for regular rounds (open circle) and intercalated rounds (closed circle) for Series XIV—Combined analytes.