**Appendix—discussion on the role of collaborative trials in determining fitness for purpose in proficiency testing**

*Referee's comments to the Author.*

The consideration of the results could also be viewed from the opposite perspective. They confirm that the proficiency testing scheme is working well using the prescribed modified Horwitz equation and this should remain. However, it may be argued that as the Horwitz equation is derived from results from collaborative trials undertaken under precisely defined conditions (single method, more experienced laboratories etc than results from a "general" proficiency testing scheme), it is to some extend surprising that the results from the scheme are so satisfactory. Thus it may now be argued that the equation is now too "lax". Recent collaborative trials often demonstrate HorRat values less than 1 and these results also suggest the primary equation should be re-examined. Perhaps that could also be commented upon?

*Authors' response.*
This comment by a referee raises some complex issues that, while marginal to the topic of the paper, are interesting in themselves.

Some recently acquired collaborative trial results display a tendency to be more precise than predicted by the Horwitz function. Even if that trend were found to be general (and it has not been rigorously tested so far), it would not in itself demand changes in food analysts' current perception of fitness for purpose. Most analytical procedures can be rendered more precise by a greater (and more costly) attention to detail, and the intense focus on data quality of the last few decades may have brought about a broad enhancement in the precision observed in collaborative trials. That does not necessarily imply that fitness-for-purpose criteria should be adjusted. Under the cost-minimisation paradigm of fitness for purpose, reducing uncertainty by the elaboration of existing procedures will not change fitness for purpose criteria (although radical improvements in analytical technology associated with lower measurement costs could make that action appropriate).

In collaborative trials a single, carefully described, procedure is used by all participant laboratories. In that respect they differ from proficiency tests where participants are nearly always free to use a method or procedure of choice. The *prima facie* expectation would be that the variation in procedure bias among the diverse methods would provide an additional source of uncertainty: that would tend to make proficiency test results more disperse than those of corresponding collaborative trials. This effect has been demonstrated in a few instances, with standard deviations inflated by up to 50%. The opposite is sometimes the case, however, because collaborative trials are carried out on newly developed procedures in which the participants are inexperienced, while most proficiency test participants would be using a routine method that had been tried, tested and refined in previous rounds of the test. It is not clear at present therefore that the average participant performance in FAPAS is unexpectedly good in relation to that in corresponding collaborative trials.

Furthermore, collaborative trials are designed to study the performance of analytical methods, so their outcomes can be reasonably regarded as converging towards fitness for purpose. Results from proficiency tests (designed to test laboratories rather than methods) cannot be seen in that light, however. Observing how well a group of laboratories *do* perform in a test is no basis for determining how well they *should* perform. Ultimately fitness for purpose depends

on the requirements of end-users in terms of the cost-effectiveness of the results, not on the current performance of the laboratories.