

Normalisation of data from allergens proficiency tests

Mark Sykes · Dominic Anderson · Bhavna Parmar

Received: 14 December 2011 / Revised: 20 January 2012 / Accepted: 22 January 2012

© Crown Copyright 2012

Abstract The problem of allergen analysis using ELISA kits from different commercial products giving significantly different results is widely acknowledged. The effect on proficiency testing results is that different assigned values have to be generated for the different kits used. Some experimental Food Analysis Performance Assessment Scheme (FAPAS) proficiency tests aimed to establish whether the use of a standardised calibrant could be used to normalise the complete data set without recourse to differentiation. Three recent FAPAS proficiency tests (2776 peanut, 2778 soya and 2781 gluten) sent out a second spiked sample, in addition to the usual spiked and unspiked samples. Further analysis of the data was undertaken after the completion of the tests. The ratio of the submitted results for the two spiked samples yielded complete data sets which could be tested for normality of the distribution. Where the raw data for each individual test sample was clearly non-normal and multi-modal, the ratio data yielded a much more normal and symmetrical distribution. The use of one of the test samples as a single-point calibrant has some limitations but the principle of applying a standardisation clearly works. The development of internationally recognised sets of certified reference calibration standards for use by allergens testing laboratories would greatly benefit the analysis.

Keywords Proficiency testing · Allergens · Calibration standard

Introduction

Our organisation previously highlighted the problem of scoring food allergens proficiency tests (PTs) [1]. In this reference, the authors demonstrated that PT results have to be separated according to ELISA kit manufacturer prior to generating the assigned values and *z* scores. The different assigned values can be significantly different, routinely up to a factor of 2. One of the major drawbacks of this approach is that sufficient numbers of results have to be submitted against each kit to be able to confidently set assigned values and *z* scores. Some *z* scores may be issued for information only (i.e., not to be taken as fully evaluative) where the assigned value has a high uncertainty. Many participants will not be scored at all due to their use of uncommon or in-house kits, where there are insufficient numbers of results to generate consensus assigned values.

The difference in results generated by different ELISA kits has been recognised for many years now. There are several reasons for this. Different antibodies are raised against different proteins, different extraction buffers or protocols may be specified, and different internal calibrants applied. The problem is not confined to food allergens testing either, but wherever ELISA kits may be used. Their speed and ease of use, particularly as a screening test, is applied in the clinical area [2, 3] and other food residues applications [4, 5]. The variability of results between laboratories applying different test kits has further implications, for example the therapeutic dose to be applied [2]. Recommendation for a specific method to be consistently followed internationally [3] would at least permit reproducible results.

Published in the special paper collection *Recent Advances in Food Analysis* with guest editors J. Hajslova, R. Krska, M. Nielsen.

M. Sykes (✉) · D. Anderson
The Food and Environment Research Agency,
Sand Hutton,
York YO41 1LZ, UK
e-mail: mark.sykes@fera.gsi.gov.uk

B. Parmar
Food Standards Agency,
Aviation House, 125 Kingsway,
London WC2B 6NH, UK

Table 1 Summary of the proficiency tests subjected to the normalisation experiment

PT 2776 peanut in chocolate	Start date 04/06/2010
2776-A	Blank (unspiked)
2776-B	Spiked test sample ~30 mg/kg
2776-C	Spiked standard sample ~10 mg/kg
PT 2778 soya in wheat flour	Start date 22/07/2010
2778-A	Spiked test sample ~60 mg/kg
2778-B	Blank (unspiked)
2778-C	Spiked standard sample ~60 mg/kg (same TM as Spiked test sample)
PT 2781 gluten in cake mix	Start date 10/02/2011
2781-A	Blank (unspiked)
2781-B	Spiked test sample ~30 mg/kg
2781-C	Spiked standard sample ~120 mg/kg

Spiking values are the approximate formulation values

The experimental design is partially dictated by the variation in test kit results [5–7]. In the case of [5], stability of test materials relied on existing knowledge, rather than a specific stability experiment. Poms et al. [6] ensured that all the test kits in their validation reported results as total peanut content and that the same production batch of each kit brand was sent to the participating laboratories. Detailed method descriptions were also provided. Fu et al. [7] were attempting to compare test kits for heat-treated foods. They discovered, however, that the three commercial kits studied reported different results even when unheated samples were used. Their discovery prompted the use of the unheated results to normalise the heat-treated results, thus allowing the comparisons to be made between kits.

Various solutions to the problem of results variation have been proposed, or at least observations made to help define the problem. The definition of the reporting determinand is a first step [6–8]. Acknowledgement of the parameters which are critical to control follows,

with carefully specified methodology [3, 6, 8]. Changes in technology are sometimes attempted to circumvent the problem, with mixed success. In the case of new ELISA methods [9], detection rates for egg increased but decreased for milk. Completely different technology is beginning to show some promise [10] with the use of mass spectrometry avoiding the root causes of immunological methods. This would be complementary to immunological methods but unlikely to supersede them. In the case of existing technology, which is in routine use for reasons of speed and cost, the desire for reference materials for direct calibration [6] or indirect concentration estimation [8] from conversion factors has been highlighted.

The normalisation of allergens results to an external standard was the subject of the work described here. The absolute results submitted would remain from the different populations but the normalised results, expressed as a ratio to the external standard, ought to be all comparable as a single dataset. Here, we describe a simple experiment in proficiency testing, how the special PTs were organised and the analysis of the results.

Organisation of the PTs

The Food Analysis Performance Assessment Scheme (FAPAS) organises some 14 PTs in allergens testing per year. In a normal allergens PT, each participant is sent two test samples—one blank and one spiked material. Participants are required to test both samples, report which one contains the allergenic material and, if possible, its concentration. A special project was set up, consisting of three PTs [11–13]. The three special PTs included a third test sample, which was also spiked and would be used as the standard. Participants were not informed of the purpose of including the third sample. Table 1 summarises the PTs and the samples. Both the spiked samples in a test were prepared from the same base

Table 2 Results of PT 2776 (peanut in chocolate), assigned values, number and percentage of z scores where $|z| < 2$ for each spiked sample and separated according to ELISA kit used

Test material	Analyte	Assigned value, mg/kg	Number of scores within $ z < 2$	Total number of scores	Satisfactory%
2776-B	Peanut 'BioKits'	23.4	17	18	94
	Peanut 'Neogen'	29.0	4	5	80
	Peanut 'R-Biopharm'	40.1	10	11	91
	Peanut 'Romer Labs'	19.0 ^a	6 ^a	7	86 ^a
2776-C	Peanut 'BioKits'	11.8	19	19	100
	Peanut 'Neogen'	15.0	4	5	80
	Peanut 'R-Biopharm'	24.2	12	12	100
	Peanut 'Romer Labs'	8.7 ^a	7 ^a	7	100 ^a
	Peanut protein 'ELISA Systems'	2.0	4	4	100

^aIndicates results issued for information only due to high uncertainty of the assigned value

Table 3 Results of PT 2778 (soya in wheat flour), assigned values, number and percentage of z scores where $|z| < 2$ for each spiked sample and separated according to ELISA kit used

Test material	Analyte	Assigned value x_a , mg/kg	Number of scores within $ z < 2$	Total number of scores	Satisfactory %
2778-A	Soya protein 'BioKits'	122 ^a	4 ^a	5	80 ^a
	Soya protein 'ELISA Systems'	17.9	12	14	86
	Soya protein 'Neogen'	122 ^a	6 ^a	7	86 ^a
2778-B	Soya protein 'BioKits'	36.0	4	5	80
	Soya protein 'ELISA Systems'	9.0	12	15	80
	Soya protein 'Neogen'	50.9	7	8	88
2778-C	Soya protein 'ELISA Systems'	18.5	13	14	93
	Soya protein 'Neogen'	154	6	7	86

^aIndicates results issued for information only due to high uncertainty of the assigned value

test materials and homogeneity testing was carried out on both test materials, according to the established procedure [14].

Other than the inclusion of the third test sample, the PTs were run as per a normal allergens PT. Samples were dispatched to registered participants on the same day and results were to be entered via the secure FAPAS website by the set deadline (approximately 5 weeks after dispatch). A mandatory question for all results entries was which ELISA kit was used for the test. Additional method questions were asked but responses were not mandatory. Following the close of the PTs, the results were analysed for all test samples in the usual way, i.e. by separation according to ELISA kit used, and participants received the PT report with their assessments in the form of z scores [15]. Further analysis of the data, by normalising to the standard sample, was carried out internally in FAPAS only.

Individual PT results

The results of the three PTs, as reported to participants, are summarised in Tables 2, 3 and 4. Results for PT 2776 (peanut in chocolate powder, Table 2) show that there is a twofold range of assigned value between kits reporting peanut concentration. A total of 75 participants returned

results for this PT, of which 40% did not receive a z score for any test material. The uncertainty of the assigned values for 'Romer Labs' kit results was higher than ideal, so z scores were issued for information only. Some participants reported results for peanut protein. An assigned value and z scores could be generated for participants using 'ELISA Systems' peanut protein kits in 2776-C but not for 2776-B, despite the higher concentration. The assigned values were all generated using the consensus medians, due to the low numbers of results for each kit.

PT 2778 (soya in wheat flour, Table 3) attracted results from 46 participants, reporting soya or soya protein. Approximately 50% of the results could be scored, for soya protein only. There were insufficient numbers of results for soya to generate assigned values. Again, the assigned values were all generated using the consensus medians and these covered an eightfold range in concentrations. Although the assigned values for 2778-A 'Biokits' and 'Neogen' were identical, these were not combined and the standard practice of generating separate assigned values was maintained. The data for 2778-A 'Biokits' and 'Neogen' kits had a high uncertainty of the assigned value, so assessments were presented as not evaluative and for information only. The uncertainty of the data for 2778-C 'Biokits' was too high to generate an assigned value, despite it actually being the same test material as 2778-A.

Table 4 Results of PT 2781 (gluten in cake mix), assigned values, number and percentage of z scores where $|z| < 2$ for each spiked sample and separated according to ELISA kit used

Test material	Analyte	Assigned value x_a , mg/kg	Number of scores within $ z < 2$	Total number of scores	Satisfactory %
2781B	Gluten 'R-Biopharm'	27.4	67	80	84
	Gluten 'Veratox (Neogen)'	42.6 ^a	5 ^a	6	83 ^a
2781C	Gluten 'Ingenasa'	124.0	4	4	100
	Gluten 'R-Biopharm'	91.6	62	75	83
	Gluten 'Veratox (Neogen)'	120.7	5	6	83

^aIndicates results issued for information only due to high uncertainty of the assigned value

Table 5 Results of PT 2776 (peanut in chocolate), combined data from all kits for individual test samples and for the ratio of spiked test sample to spiked standard sample

	2776-B peanut	2776-C peanut	Ratio B/C peanut	2776-B peanut protein	2776-C peanut protein	Ratio B/C peanut protein
Rm	29.0	14.5	2.06	15.9	8.82	1.91
$\hat{s}$	12.8	6.78	0.535	12.4	7.63	0.429
sp	7.26	3.63	0.515	3.98	2.20	0.478
Mode 1	22.3	12.1	1.84	6.11	2.92	1.91
sem 1	2.76	0.931	0.0523	0.633	0.260	0.120
Mode 2		23.6		20.7	9.37	
sem 2		6.41		1.49	1.06	
Mode 3					22.6	
sem 3					5.80	
KS	0.170	0.211	0.163	0.189	0.202	0.104
n	40	40	40	24	24	24
u of Rm	2.03	1.07	0.0846	2.54	1.56	0.0876
u/σ_p	0.279	0.296	0.164	0.637	0.706	0.183

Rm robust mean, $\hat{s}$ robust standard deviation, sp standard deviation for proficiency (using 25% RSD of robust mean), sem standard error of the mode (data for up to the first three modes are shown), KS Kolmogorov–Smirnov value (critical value=0.19 for $n>20$, <critical indicates normal distribution), n number of data points, u of Rm uncertainty of robust mean, u/σ_p uncertainty of robust mean/standard deviation for proficiency (critical value 0.3)

PT 2781 (gluten in chocolate cake mix, Table 4) attracted results from 107 participants reporting gluten. Approximately 30% of results submitted could not be scored. Most participants used ‘R-Biopharm’ kits in this test. As a consequence, the assigned value for ‘R-Biopharm’ kit results was generated from the robust mean. Assigned values for other kits were generated using the medians. The assigned value for ‘Veratox (Neogen)’ in 2781-B had a high uncertainty, so the assessments were issued for information only. Although an assigned value was generated for ‘Ingenasa’ in 2781-C, there were insufficient data points for this kit in 2781-B to set an assigned value.

Additional data analysis

These PTs were run as normal allergens PTs, albeit with an extra sample to test, from the point of view of the participants. Additional analysis was applied to the intact datasets, in the form of mode (bump-hunting [16]) and normality analysis (probability plots using Kolmogorov–Smirnov test). This analysis was applied to the complete datasets for the individual (spiked) test materials as well as the ratio data of [spiked test sample]/[spiked standard sample]. The data are shown in Tables 5, 6 and 7.

Table 6 Results of PT 2778 (soya in wheat flour), combined data from all kits for individual test samples and for the ratio of spiked test sample to spiked standard sample

	2778-A soya protein	2778-B soya protein	2778-C soya protein	Ratio A/C soya protein	Ratio B/C soya protein
Rm	58.6	22.9	64.8	1.01	0.444
$\hat{s}$	60.9	20.8	72.1	0.205	0.200
sp	14.7	5.72	16.2	0.251	0.111
Mode 1	19.9	8.82	19.4	0.999	0.378
sem 1	1.36	1.18	3.07	0.0305	0.0278
Mode 2	106	47.2	102		
sem 2	13.6	2.85	27.9		
Mode 3	137	74.0	164		
sem 3	19.3	14.2	5.15		
KS	0.272	0.231	0.3	0.213	0.252
n	30	32	30	31	31
u of Rm	11.1	3.68	13.2	0.0368	0.0358
u/σ_p	0.758	0.644	0.812	0.146	0.323

Footnotes as for Table 5

Table 7 Results of PT 2781 (gluten in cake mix), combined data from all kits for individual test samples and for the ratio of spiked test sample to spiked standard sample

	2781-B gluten	2781-C gluten	Ratio B/C gluten
Rm	28.2	92.7	0.308
$\hat{s}$	11.5	38.4	0.0910
sp	7.06	23.2	0.0771
Mode 1	25.1	89.6	0.287
sem 1	1.12	6.93	0.00771
KS	0.196	0.148	0.128
n	96	96	93
u of Rm	1.17	3.92	0.0094
u/σ_p	0.166	0.169	0.122

Footnotes as for Table 5. Only one mode was found by bump-hunting

PT 2776 (peanut) received quantitative results for peanut analysis by 40 participants, and for peanut protein by 24 participants. For peanut (Table 5), the intact datasets, i.e. not separated by ELISA kit, were largely unimodal and close to a normal distribution. 2776-C peanut data were slightly modal on the tail of the distribution, with bump-hunting identifying a distinct minor mode. Taking the ratio of 2776-B/2776-C peanut data yielded a unimodal and normal distribution.

The 2776 peanut protein data analysis, however, was not straightforward. The 2776-B protein data were bimodal, and the 2776-C data were multi-modal. The kernel density plots are shown in Fig. 1a and b. The ratio of 2776-B/2776-C, however, yielded a unimodal and normally distributed (KS=0.104, critical value 0.19) plot (Fig. 1c). A further measure of the confidence in the ratio data is the uncertainty, u , of the robust mean compared to the standard deviation for proficiency, σ_p . The value u/σ_p is 0.18, much lower than the guideline critical value of 0.3 [15]. The u/σ_p value is used by FAPAS as an indicator as to whether z scores can be issued as evaluative or not.

PT 2778 (soya) received quantitative results for soya analysis by 5 participants, and for soya protein by 32 participants. The soya results (data not shown) were not subjected to any further analysis, due to the low numbers. The intact soya protein datasets were subjected to further analysis. The intact soya protein datasets for the three individual test materials were clearly from a mixed population of results. The robust standard deviations were at about the same values as the robust means with non-normal, multi-modal distributions. The value of u/σ_p is very high (>0.60). The summary data are presented in Table 6. The distributions as kernel density plots are shown in Fig. 2a, b, and c. Using 2778-C as the standard and taking the ratios of 2778-A/2778-C and 2778-B/2778-C yields unimodal, normal distributions of much lower uncertainty. The kernel density plots are shown in Fig. 2d and e.

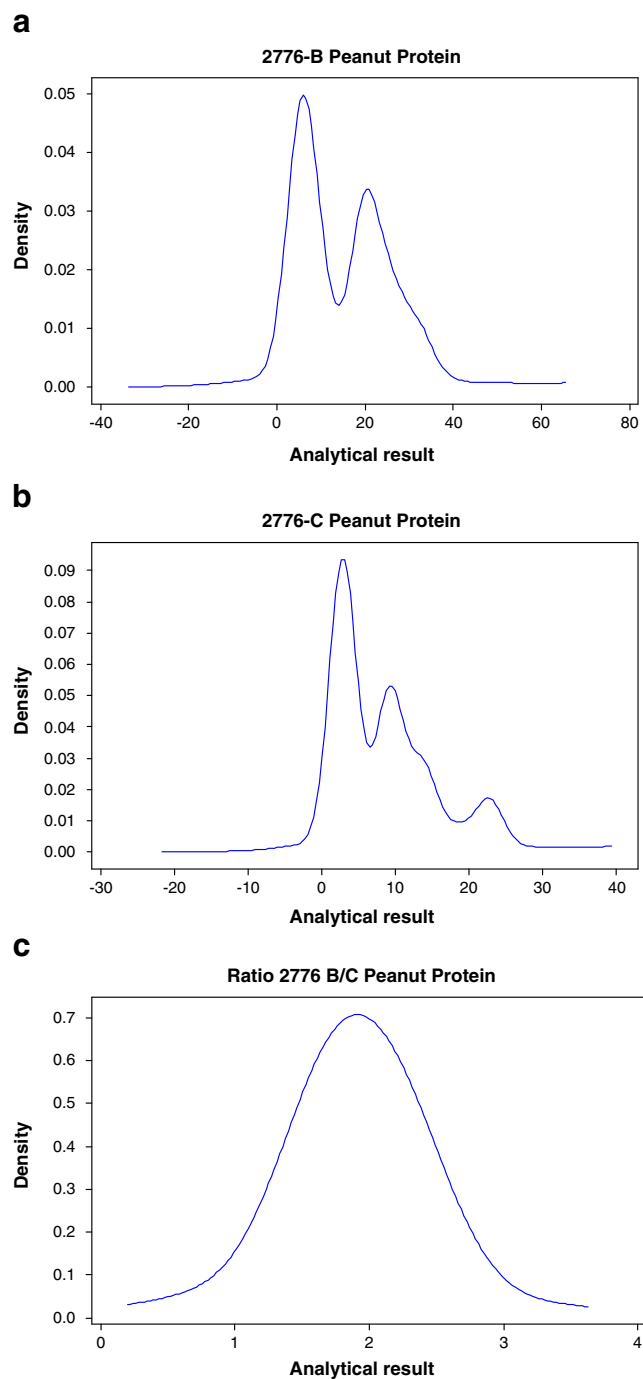


Fig. 1 a Kernel density plot of all PT 2776-B peanut protein results. b Kernel density plot of all PT 2776-C peanut protein results. c Kernel density plot of all PT 2776-B/2776-C ratio peanut protein results

PT 2781 (gluten) received a large response, with results submitted by 107 participants, most of whom used 'R-Biopharm' kits. Although the number of participants was high and results were returned for a range of ELISA kits, the proportion of 'R-Biopharm' results meant that the overall

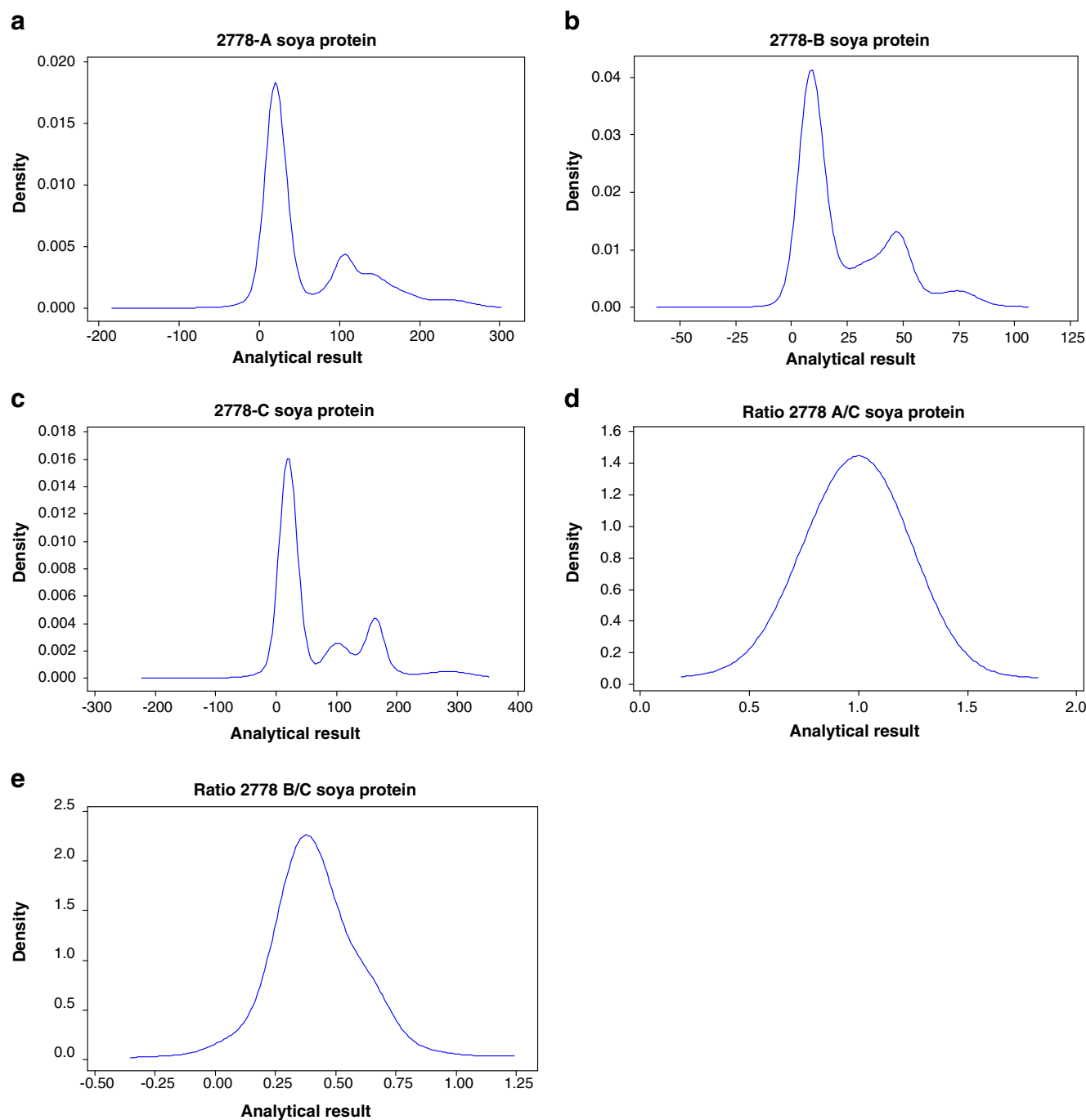


Fig. 2 **a** Kernel density plot of all PT 2778-A soya protein results. **b** Kernel density plot of all PT 2778-B soya protein results. **c** Kernel density plot of all PT 2778-C soya protein results.

d Kernel density plot of all PT 2778-A/2778-C ratio soya protein results. **e** Kernel density plot of all PT 2778-B/2778-C ratio soya protein results

distributions for the complete datasets were largely unimodal, with good agreement between the robust mean and mode. The summary data are presented in Table 7. The normal probability plots and u/σ_p values also suggest largely symmetrical distributions of low uncertainty. Taking the ratio of 2781-B/2781-C, therefore, appears to offer no obvious advantage in modelling the distribution.

Discussion

The limitations of scoring the PTs using the current practice of separation by ELISA kit is evident in the results received for these three PTs. Some participants (between one third and one half) received no z scores. A proportion of participants received a z score for one test material but not the other, and

others received *z* scores that were for information only and not to be used in an evaluative capacity. This is due to the insufficient numbers of results returned for particular kits and/or the high uncertainty of the assigned values. None of this necessarily invalidates the results which are not scored, since the true values are unknown. Even for the results which can be scored, the assigned value for one kit is not more valid or accurate than for another kit. This was particularly emphasised by the eightfold difference in assigned values (18.5 and 154 mg/kg) for soya protein in 2778-C. Occasionally, the combined raw results does produce a normal and symmetrical distribution, as in the case of 2776 peanut data and PT 2781 gluten. However, the 2776 peanut protein data were multi-modal and the 2781 gluten were artificially normally distributed due to the very high proportion of results from one kit type.

One unexpected set of results was for 2778-B (soya). This was the unspiked test sample for this PT but, nevertheless, sufficient data were returned by participants to generate consensus assigned values and *z* scores for three of the soya protein kit types used. In the same test, the ability to generate assigned values and *z* scores was not the same for 2778-A and 2778-C, even though these were actually the same test material. This highlights the high variation of results from the same ELISA analyses and the importance of testing for the presence of allergenic materials even in 'blank' samples.

The application of a standard sample against which to normalise the test sample has three clear advantages. First, multi-modal datasets become unimodal, normal and symmetrical. Second, the entire set of results can be included, regardless of the source of the ELISA kits (whether from a low volume manufacturer, high volume manufacturer or in-house). Third, the normalised distribution generally has a lower uncertainty than either the individual kit datasets or the combined but non-normalised datasets. For the purposes of scoring PTs, this is an ideal situation. In the case of PT 2781 (gluten) which, on initial inspection, appears to be already a normal distribution, this is also relevant. By including all the results regardless of ELISA kit, each result could potentially be assessed and scored. This is exemplified by the separated results for 2781-B in which 'Ingenasa' kit results could not be scored at all, and 'Veratox (Neogen)' were issued for information only.

The experiments undertaken here were necessarily simplified, with only a single level standard sample being provided. Ideally, there would be a range of standards provided; however, the cost of producing these for each PT would be high. There would also be the question of what is actually the true value. The results provided here only give a relative value (the ratio of one result to another). The best estimate of the true value could be achieved by production of the standards as certified reference materials with the reference values provided

from metrologically traceable formulation values. Confirmation reference values could also be obtained from an expert laboratory applying molecular biology techniques, such as PCR, or even from a range of techniques including mass spectrometry, once these have been fully validated. The range of matrices covered would have to be wide but initially covering the most critical allergens. The expense of their production would be partly offset by large-scale production with aliquots included in each ELISA kit or available from a central distribution point.

Conclusions

These three experimental PTs demonstrate that a matrix-matched standard can be used to successfully normalise data from different allergens ELISA kits. The normalisation also allows unusual or in-house ELISA kit results to be assessed where their uniqueness would otherwise prevent their assessment. It has also been demonstrated in these PTs that assessments issued for information only (due to limited numbers and/or high uncertainty) are more likely to be fully evaluative when analysed as a complete data set. An internationally recognised set of certified reference standard materials, globally available, would be costly to produce but would be of enormous potential benefit to the allergens testing industry.

Acknowledgement The cost associated with producing and testing the additional test materials for this project was kindly funded by the United Kingdom Food Standards Agency.

References

- Owen L, Gilbert J (2009) Proficiency testing for quality assurance of allergens methods. *Anal Bioanal Chem* 395:147–153
- Grier TJ, Hazelhurst DM, Duncan EA, West TK, Esch RE (2002) Major allergen measurements: sources of variability, validation, quality assurance, and utility for laboratories, manufacturers, and clinics. *Allergy Asthma Proc* 23:125–131
- Balloch A, Licciardi PV, Leach A, Nurkka A, Tang MLK (2010) Results from an inter-laboratory comparison of pneumococcal serotype-specific IgG measurement and critical parameters that affect assay performance. *Vaccine* 28:1333–1340
- Gaudin V, Cadieu N, Sanders P (2005) Results of a European proficiency test for the detection of streptomycin/dihydrostreptomycin, gentamicin and neomycin in milk by ELISA and biosensor methods. *Anal Chim Acta* 529:273–283
- Gaudin V, Hedou C, Rault A, Sanders P, Verdon E (2009) Comparative study of three screening tests, two microbiological tube tests, and a multi-sulphonamide ELISA kit for the detection of antimicrobial and sulphonamide residues in eggs. *Food Addit Contam Part A* 26:427–440
- Poms RE, Agazzi ME, Bau A, Brohee M, Capelletti C, Nørgaard JV, Anklam E (2005) Inter-laboratory validation study of five commercial ELISA test kits for the determination of peanut proteins in biscuits and dark chocolate. *Food Addit Contam Part A* 22:104–112

7. Fu TJ, Maks N, Banaszewski K (2010) Effect of heat treatment on the quantitative detection of egg protein residues by commercial enzyme-linked immunosorbent assay test kits. *J Agric Food Chem* 58:4831–4838
8. Dumont V, Kerbach S, Poms R, Johnson P, Mills C, Popping B, Tömösközi S, Delahaut P (2010) Development of milk and egg incurred reference materials for the validation of food allergen detection methods. *Qual Assur Saf Crops Foods* 2:208–215
9. Watanabe H, Akaboshi C, Saita K, Sekido H, Hashiguchi S, Watabe K, Tanaka K (2011) Comparison between old and new methods for detection of allergenic substances (egg and milk). *Food Hyg Saf Sci* 52:71–77
10. Heick J, Fischer M, Kerbach S, Tamm U, Popping B (2011) Application of a liquid chromatography tandem mass spectrometry method for the simultaneous detection of seven allergenic foods in flour and bread and comparison of the method with commercially available ELISA test kits. *J AOAC Int* 94:1060–1068
11. FAPAS Report 2776, The Food and Environment Research Agency, York, UK
12. FAPAS Report 2778, The Food and Environment Research Agency, York, UK
13. FAPAS Report 2781, The Food and Environment Research Agency, York, UK
14. Fearn T, Thompson M (2001) A new test for sufficient homogeneity. *Analyst* 126:1414–1417
15. Thompson M, Ellison SLR, Wood R (2006) The international harmonized protocol for the proficiency testing of analytical chemistry laboratories. *Pure Appl Chem* 78:145–196
16. Lowthian PJ, Thompson M (2002) Bump-hunting for the proficiency tester—searching for multimodality. *Analyst* 127:1359–1364