Article:

Beavis, Guy, Wilson, Julie Carol orcid.org/0000-0002-5171-8480 and Sykes, Mark (2019) Quantitative scores for binary qualitative proficiency testing. Accreditation and Quality Assurance. pp. 263-269.

https://doi.org/10.1007/s00769-019-01386-8

Quantitative scores for binary qualitative proficiency testing

Guy Beavis[1], Julie Wilson[1], Mark Sykes[2]

Keywords: Proficiency Testing, Quantitative Assessments, z-scores, a-scores

Abstract

While z-scores provide participants with easy-to-interpret scores for quantitative proficiency tests, there is no universally accepted equivalent scoring method available for qualitative testing. Under the assumption that these tests follow a binomial distribution, it is possible to calculate scores that mimic the widely used z-scores and provide participants with insight into their performance level. We show that these scores, which we term a-scores, can be combined to provide a single score for multiple tests so that participants can monitor their performance over time, and discuss the use of the exact binomial test in place of uncertainty when there is no clear consensus.

Introduction

Proficiency tests (PTs) evaluate participant performance against pre-established criteria by means of inter-laboratory comparisons [1]. Most such tests are quantitative, allowing quantitative assessment, predominantly via z-scores due to their robustness and ease of interpretation. These scores are given by

$$\mathrm{z} = \frac{x - x_{pt}}{\sigma_{pt}},$$

where x is the participants result, x_{pt} is the estimate of the true result and σ_{pt} is the standard deviation for proficiency assessment (SDPA). z-Scores can be combined from multiple tests to provide a measure of performance over time [2, 3]. In calculating the z-score, PT providers often make use of robust statistics, such as Huber's method for the robust mean [3], and fit-for-purpose criteria such as the Horwitz function for the SDPA [4, 5, 6].

Currently approximately 15% of food chemistry PTs in Fapas®, an accredited proficiency testing provider, use qualitative assessments, sometimes in combination with quantitative results and assessments. The qualitative results reported by participants are usually 'detected' or 'not detected' for the presence or absence of a target analyte in a test sample. Participating laboratories can also report a null result of 'not tested' or similar, in which case they are not assessed and do not contribute to the consensus assigned value which is assumed to agree with more than half of the qualitative results. While unlikely, it is possible that the expected outcome agrees with exactly half (or even fewer) of the valid results, in which case alternative methods can be used. Although referred to throughout as the consensus value, as described in ISO/IEC 13528:2010 [7] other methods of choosing the assigned value include using an independent reference value, formulation value, and consensus of expert laboratories. In the event that the participants' consensus does not agree with the expected outcome, the discrepancy needs to be rationalised before reaching a final decision on the assigned value. If the consensus has no expected result, such as the detection of low level naturally incurred residues (i.e. not deliberately fortified), results that agree with the consensus may be marked simply as 'agrees' (rather than 'satisfactory'). Results that do not agree with the consensus may be marked as 'disagrees' (rather than 'not satisfactory').

Unlike the scores given for quantitative PTs, qualitative results make comparison between tests difficult and offer no possibility to establish a laboratory's performance over time. Uhlig et al. [6] describe a method to provide scores for qualitative tests involving a logit model with maximum likelihood parameter estimation. They combine the probability of detection (PoD) with the level of competence of the laboratory (LCL) to derive L-scores similar to a z-score, but with a different interpretation. Although this provides a numerical measure for the competency of participating laboratories, a minimum of five different analytes are required for the statistical significance of the results to be assessed. The main difference between these L-scores and z-scores is that a positive L-score is always good, with the interpretation that the laboratory has average or higher than average competence. It is possible for a laboratory to correctly identify all pathogens in a PT, but not achieve 'higher than average' competence if all other participants also perform well, giving no real indication of how well they have done.

If results from multiple analytical tests were available for each laboratory, the trade-off between false negatives and false positives could be assessed using sensitivity and specificity [8]. However, this is not possible when assessing the competence of individual laboratories based on a single analyte. As participating laboratories will be familiar with z-scores due to their use in quantitative proficiency testing, a score that could be interpreted in a similar manner would facilitate the participants' understanding. For example, if all participants in a qualitative test give the satisfactory result, there is no deviation from the consensus and all participants should receive a score of zero. Similarly, if almost all laboratories provide

[1] Mathematics Department, University of York, Heslington, York, YO10 5DD
[2] Fera Science Ltd, Sand Hutton, York YO41 1LZ, UK
Email: gb789@york.ac.uk

the same result, the PoD is close to one and scores for these participants should be close to zero. The few that give a non-satisfactory result are either performing badly or their Limit of Quantification (LoQ) could differ from the other laboratories taking part [9]. As such an incorrect result could have serious consequences, for instance when the presence of an allergen is not detected, it should result in a negative score that can be easily interpreted as unsatisfactory. On the other hand, a laboratory that detects an analyte when most others do not, should receive a positive score, reflecting the fact the analyte should not have been detected. This paper describes a scoring system that parallels the use of z-scores for quantitative tests, with scores for unsatisfactory results increasing in magnitude as the PoD increases, being negative for non-detection of analytes and positive for erroneous detection.

Data

The data consist of 143 Fapas® Series 27 (Allergens) PTs, involving 293 individual analytes, and 47 Genetically Modified Materials Analysis (GeMMA, Fapas® GM) PTs, involving 137 analytes [10] All tests took place between 2001 and 2017. All results were qualitative, given as either 'detected' or 'not detected' and considered here as satisfactory and unsatisfactory respectively. To allow comparison with other methods, we also reproduce the data of Uhlig et al [6] (Table 1), courtesy of Springer-Verlag.

Deriving the score

The fundamental assumption for a PT is that the valid results are independent and identically distributed (i.i.d) [5] so that each qualitative result can be considered as the result of a Bernoulli trial, that is, a random variable (X) with just two possible outcomes, giving the probability mass function

$$\mathrm{P}(\mathrm{X}=1) = p = 1 - \mathrm{P}(\mathrm{X}=0).$$

Here p will be estimated by $\hat{p}$, the proportion of satisfactory results, also used by Wehling et al [11] for qualitative method validation. The sum of n such Bernoulli trials has a binomial distribution, $\mathrm{B}(n,\ p)$, which, under the central limit theorem, if the number of participants who reported valid results is large enough ($n \geq 20$, say), $np > 5$, and $n(1-p) > 5$, can be approximated by a normal distribution,

$$\sum_{i}^{n} X_i \sim \mathrm{N}(n\hat{p}, n\hat{p}\hat{q})$$

where $\hat{q} = 1 - \hat{p}$, the proportion of unsatisfactory results. Therefore, as $\mathrm{E}(nX) = n\hat{p}$ and $\mathrm{Var}(nX) = n\hat{p}\hat{q}$, we have $E(\mathrm{X}) = \frac{n\hat{p}}{n} = \hat{p}$ and $\mathrm{Var}(\mathrm{X}) = \frac{n\hat{p}}{n^2} = \frac{\hat{p}\hat{q}}{n}$ so that

$$\mathrm{X} \sim \mathrm{N}\left(\hat{p}, \frac{\hat{p}\hat{q}}{n}\right).$$

However, for a qualitative PT the conditions of $np > 5$ and $nq > 5$ do not often hold making the calculation of a z-score inappropriate. In fact, when a few laboratories disagree with the consensus, the distribution is highly skewed and, the z-scores for laboratories with unsatisfactory results would be extreme, making the usual boundaries of 2 and 3 for questionable and unsatisfactory scores meaningless.

Here we introduce a score for a participant, modelled on the z-score calculation, that we refer to as an a-score, given by

$$\mathrm{a} = I_C \frac{x - x_{pt}}{\sigma_{pt}} \qquad \mathbf{(1)}$$

where x is result for the participant and the assigned value x_{pt} is the estimated (consensus) outcome $\hat{p}$. As calculation of σ_{PT} from the approximation above would give a standard deviation dependent on n, $\sigma_{PT} = \sqrt{\frac{\hat{p}\hat{q}}{n}}$, which, due to inconsistency between tests, is not recommended for quantitative tests either [12], we use instead a fit for purpose (FFP) fixed SDPA. We use Huber's method [3] with Fapas® data from qualitative tests taking place between 2001 and 2017 to calculate a robust SDPA as σ_{pt} = 0.0524. This fixed SDPA will be a good estimator of the target standard deviation for any qualitative PT schemes.

The main difference between the z-score equation and equation (1) is the indicator, I_C, calculated as

$$I_C = \begin{cases} 1 & \text{if the concensus is 'detected'} \\ -1 & \text{if the concensus is 'not detected'} \end{cases}$$

to provide negative scores when the consensus is 'detected' and positive scores for a consensus of 'not detected'. Where the independently-derived assigned value does not agree with the consensus, the sign of the score will be changed, with 'detected' scores receiving a negative score and 'not detected' receiving a positive score.

As an example, consider a PT with 50 valid results, 40 given as 'detected' and 10 as 'not detected'. Here $n = 50$, $\hat{p} = x_{pt} = 0.8$ and $I_C = 1$. Using equation (1), for the participants reporting 'detected', $x = \hat{p} = 0.8$ and the a-score is

$$\mathrm{a} = I_C \frac{x - x_{pt}}{\sigma_{pt}} = 1 \cdot \frac{0.8 - 0.8}{0.0524} = 0$$

whereas for those reporting 'not detected', $x = \hat{q} = 0.2$ and the a-score is

$$\mathrm{a} = I_C \frac{x - x_{pt}}{\sigma_{pt}} = 1 \cdot \frac{0.2 - 0.8}{0.0524} = -11.4$$

Thus, satisfactory results receive a score of 0, and unsatisfactory results receive a large negative a-score in line with quantitative tests [5], giving a clear interpretation that investigation is required. If instead the consensus (from 40 of the 50 participants) had been 'not detected', the a-score would show the unsatisfactory results to be equally poor, but with a positive score as $I_C = -1$. As the consensus becomes less clear, the a-scores for those that do not agree reduce in magnitude. For example, in a 50-participant PT, 27 report 'detected' and 23 report 'not detected'. Then we have $\hat{p} = x_{pt} = 0.54$ and $\hat{q} = 0.46$ and the a-score for participant reporting 'not detected' is -1.5. At such a low PoD assessing the laboratories as unsatisfactory might be inappropriate and the hypothesis that there is no clear consensus could be accepted instead.

Interpretation

All results receiving non-zero scores can be considered unsatisfactory, but, as we cannot use the usual confidence intervals of ±2 and ±3 associated with z-scores, a measure of how unsatisfactory is required in order to interpret a-scores. Instead, these will be approximated from the data using the robust SDPA, $\sigma_{pt} = 0.0524$. Most qualitative PTs have a high PoD, $\hat{p}$, with the robust estimate for the mean q value, $\bar{q}$, being 0.0413 (Figure 1).

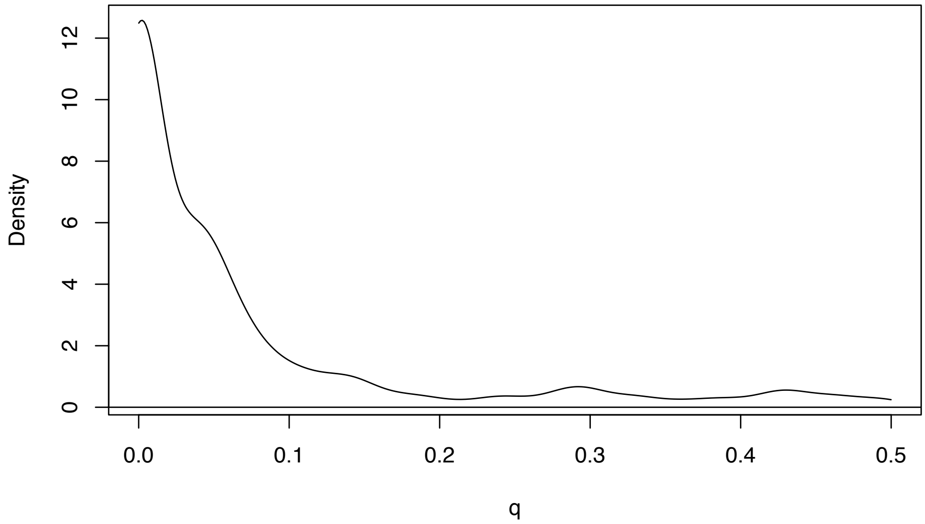

Fig. 1 Distribution of q Values. As p > 0.5, we only have values of q between 0 and 0.5. The robust estimate for the mean q value is $\bar{q} = 0.0413$

Any value of q less than 3 SDs away from the mean, i.e. $q \leq \bar{q} + 3\sigma_{PT} = 0.1986$ so that $|\mathrm{a}| \geq 11.5$ will be assessed as 'unsatisfactory', and all other non-zero results as 'questionable'. With these bounds set approximately 40 % of historical erroneous data is classed as 'unsatisfactory'.

Combining scores

It can be useful for participants to combine scores received in PTs in order to perceive their trend in performance over time. Several methods for combining z-scores have been proposed including J-scores [13], rescaled sum of z-scores [5] and others described by Medina-Pastor et al. [14] who found the SZ2 score, calculated as

$$\mathrm{SZ2} = \frac{\sum_{i=1}^{n} z_i^2}{n},$$

to be optimal, with advantages such as the normalisation producing scores on a similar scale to individual z-scores and the fact that it does not depend on external weighting values. We use the same idea here to combine our a-scores, defining SA2 as

$$\mathrm{SA2} = \frac{\sum_{i=1}^{n} a_i^2}{n}.$$

As with most combined scores, our SA2 score is always positive and so does not provide information as to whether the participant is over or under reporting. However, the combination of unsatisfactory scores with satisfactory scores will reflect the participant's overall performance and, as with the SZ2 score and z-scores, the combined SA2 score will be interpreted using the same scale as the individual a-scores to provide quick and easy comparisons between the two. Based on the distribution of the Fapas® data the average SA2 score achieved is 3.3, which is interpreted as questionable. This is in line with what is to be expected, as receiving any unsatisfactory result should trigger a level of concern about the method used. If the rate of detection for laboratories is noticeably lower than the average, the corresponding SZ2 score will increase drastically, and will be a clear sign of poor proficiency.

Species										
Lab	**HIP 1**	**HIP 2**	**HIP 3**	**HIP 4**	**HIP 5**	**HIP 6**	**HIP 7**	**HIP 8**	**HIP 9**	**ROD$_{\mathrm{lab}}$**
01	+	+	+	+	-	+	-	-	-	56 %
02	+	+	+	+	+	+	+	+	+	100 %
03	+	+	+	+	-	+	+	-	+	78 %
04	+	+	+	+	+	+	+	+	+	100 %
05	+	+	+	+	-	+	+	+	+	89 %
06	+	+	+	+	+	+	+	+	+	100 %
07	+	+	+	-	+	+	-	-	+	67 %
08	+	+	+	+	-	+	+	+	+	89 %
09	+	+	+	+	+	+	+	+	+	100 %
10	+	+	+	+	+	+	+	+	+	100 %
11	+	+	+	+	+	+	+	+	+	100 %
12	+	+	+	+	-	+	+	+	+	89 %
13	+	+	+	+	-	+	+	+	+	89 %
14	+	+	+	+	+	+	+	+	+	100 %
15	+	+	+	+	-	+	+	+	+	89 %
16	+	+	+	+	-	+	+	+	+	89 %
17	+	+	+	+	+	+	+	+	+	100 %
18	+	+	+	+	+	+	+	+	+	100 %
19	+	+	+	+	+	+	+	+	+	100 %
20	+	+	-	+	+	+	+	+	+	89 %
21	+	+	+	+	+	+	+	+	+	100 %
22	+	+	+	+	-	+	+	-	-	67 %
23	+	+	+	+	+	+	+	+	-	89 %
24	+	+	+	+	+	+	+	+	+	100 %
25	+	-	+	-	+	+	+	+	+	78 %
26	+	+	+	+	+	+	+	+	+	100 %
27	+	+	+	+	+	+	+	+	+	100 %
28	+	+	+	-	+	+	-	+	-	67 %
ROD$_{\mathrm{species}}$	100 %	96 %	96 %	89 %	68 %	100 %	89 %	86 %	86 %	

Table 1 Laboratory results. The results are shown for 28 laboratories testing for 9 highly infectious pathogens (HIP): "+" signs denote detection, while "–" signs denote unsatisfactory results (highlighted in grey and bold text). The final column and final row provide the

rate of detection (ROD) values for the laboratories and pathogen species, respectively. Previously published by Uhlig et al. [6], reproduced courtesy of Springer-Verlag.

Interpretation of scores

For comparison, we reproduce the data of Uhlig et al. [6] by kind permission of Springer-Verlag. Table 1 shows their results for 28 laboratories testing for the presence of 9 highly infections pathogens (HIPs), all of which should have been detected. Bashkansky [15] also considered this data and introduced the idea of a predefined difficulty level for each test.

Species											
Lab	**HIP 1**	**HIP 2**	**HIP 3**	**HIP 4**	**HIP 5**	**HIP 6**	**HIP 7**	**HIP 8**	**HIP 9**	**SA2(a)**	**SA2(b)**
01	0.0	0.0	0.0	0.0	**-6.9**	0.0	**-14.9**	**-13.7**	**-13.7**	**71.7**	**74.8**
02	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
03	0.0	0.0	0.0	0.0	**-6.9**	0.0	0.0	**-13.7**	0.0	**26.1**	**23.6**
04	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
05	0.0	0.0	0.0	0.0	**-6.9**	0.0	0.0	0.0	0.0	5.3	0
06	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
07	0.0	0.0	0.0	**-14.9**	0.0	0.0	**-14.9**	**-13.7**	0.0	**70.2**	**78.9**
08	0.0	0.0	0.0	0.0	**-6.9**	0.0	0.0	0.0	0.0	5.3	0
09	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
10	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
11	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
12	0.0	0.0	0.0	0.0	**-6.9**	0.0	0.0	0.0	0.0	5.3	0
13	0.0	0.0	0.0	0.0	**-6.9**	0.0	0.0	0.0	0.0	5.3	0
14	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
15	0.0	0.0	0.0	0.0	**-6.9**	0.0	0.0	0.0	0.0	5.3	0
16	0.0	0.0	0.0	0.0	**-6.9**	0.0	0.0	0.0	0.0	5.3	0
17	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
18	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
19	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
20	0.0	0.0	**-17.6**	0.0	0.0	0.0	0.0	0.0	0.0	**34.4**	**38.5**
21	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
22	0.0	0.0	0.0	0.0	**-6.9**	0.0	0.0	**-13.7**	**-13.7**	**47.0**	**47.2**
23	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**-13.7**	**20.9**	**23.6**
24	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
25	0.0	**-17.6**	0.0	**-14.9**	0.0	0.0	0.0	0.0	0.0	**59.1**	**66.2**
26	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
27	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0
28	0.0	0.0	0.0	**-14.9**	0.0	0.0	**-14.9**	0.0	**-13.7**	**70.2**	**78.9**

Table 2: Scores calculated from the data shown in Table 1. The a-scores for all unsatisfactory and questionable results are highlighted in grey with bold text, as are SA2 scores > 11.5. The penultimate column, SA2(a), shows SA2 scores calculated using a-scores for all pathogens and the final column, SA2(b) shows SA2 scores calculated without HIP 5, as suggested by the exact binomial test in the next section.

Table 2 shows the a-scores for this data set with the SA2 score corresponding to each laboratory in the penultimate column. Two of the pathogens, HIP 2 and HIP 3 have only one unsatisfactory result, giving a PoD of 96 %, which means an unsatisfactory result for these pathogens is given an a-score of -17.6. On the other hand, HIP 5 has the lowest PoD, of just 68 %, and so the 9 laboratories that did not detect this pathogen receive an a-score of -6.9.

Lab 01 has the highest SA2 score as only 56 % of the HIPs are correctly detected and 3 of the undetected pathogens are associated with an unsatisfactory result. Six laboratories receive an SA2(a) score of 5.3, interpreted as questionable, for not detecting a single pathogen (HIP 5), but Lab 20 receives an SA2(a) score of 34.4, interpreted as 'unsatisfactory', with only one pathogen, HIP 3 undetected. This reflects the fact that Lab 20 was the only laboratory to fail to detect HIP 3, whereas several laboratories did not detect HIP 5. Similarly, Lab 22 and Lab 28 both detect 67 % of HIPs, but receive different SA2 scores, as the undetected pathogens are different (HIP 5 for Lab 22, HIP 4 for Lab 28).

Participants with satisfactory results in all tests receive an SA2 score of zero, the best that can be achieved. Any deviation from zero is unsatisfactory, but as Table 2 shows, it is possible to receive a SA2 score less than 11.5, which is only questionable.

Uncertainty

In quantitative PTs a measure of uncertainty is used to determine whether the scores obtained can be used for evaluation [5]. For a normal distribution, the uncertainty is calculated by

$$\mathrm{u} = \frac{\widehat{\sigma}}{\sqrt{\mathrm{n}}}$$

where $\widehat{\sigma}$ is the internal robust standard deviation of the results. When $\mathrm{u}/\sigma_{\mathrm{pt}} \geq 0.3$ then there is concern about the uncertainty of the assigned value [7], and if $\mathrm{u}/\sigma_{\mathrm{pt}} \geq 0.4$ the option of not issuing an assessment is considered due to the high uncertainty [16]. Here, as the assumption of normality is invalid, we can use the exact binomial test to test the null hypothesis, H_0, that there is no clear consensus, i.e. $H_0: p = q = 0.5$ vs $H_A: p \neq q$. This test can be used to determine which analytes should be used for evaluation purposes, providing a pass or fail for each analyte, and by reporting the p-value (i.e. the probability of obtaining a result at least as extreme as the observed result if H_0 is true, not to be confused with p, the proportion of satisfactory results), it can be considered a rudimentary measure of the uncertainty. A p-value to determine whether the observed proportion is significantly different from 0.5 is calculated as

$$\sum_{(k,n-k)\geq k_{obs}} \binom{n}{k} p^k q^{n-k} \tag{2}$$

where k is the expected number of successes. For the data in Table 1, all pathogens other than HIP 5 result in a p-value < 0.05 so that the null hypothesis can be rejected at the 95% confidence level. For HIP 5, 9 out of 28 participants have unsatisfactory results, giving $p = 0.69$ and the associated p-value calculated as 0.08, suggesting that there is no clear consensus for this pathogen. In such cases, it is recommended that a score be issued only for information and not used for evaluation. SA2 scores calculated without HIP 5 are shown in the final column of Table 2.

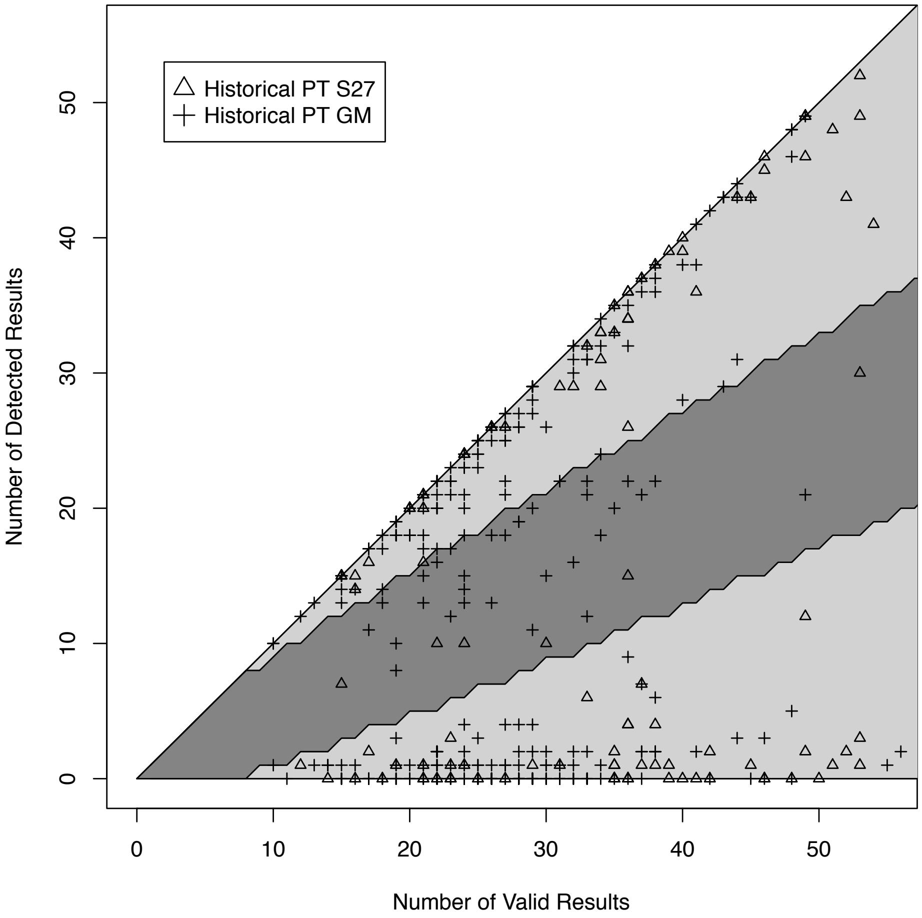

Fig. 2 Exact binomial test results. Fapas® historical data are plotted together. The light grey region represents tests where the null hypothesis should be rejected and the dark grey region represents tests where the null hypothesis should not be rejected.

Figure 2 shows the historical results from the Fapas® tests plotted together with the pass/fail regions for the exact binomial test obtained using equation (2). For each test, the number of detected results are plotted against the number of valid results, n. As $p = 1 - q$ the test is symmetric around $p = 0.5$, hence the two regions where we do not reject H_0. The majority of historical cases pass the exact binomial test with 96 % of the Fapas® Series 27 (Allergens) PTs and 90% of GeMMA PTs passing. Notice that, in the rare case that $n < 9$, all participants' results must be the same for a valid consensus to be reached.

Conclusions

By assuming a binomial distribution for qualitative PTs, scores can be calculated that mimic z-scores, widely used for quantitative PTs. These scores, that we refer to as a-scores, provide a measure of the seriousness of the disagreement with the consensus, summarised in Table 3, where any non-zero score is considered unsatisfactory. The scores can be combined to give an assessment, the SA2 score, based on multiple tests that can be interpreted using the same scale as a-scores.

a-score	Interpretation
|a| ≥ 11.5	unsatisfactory
|a| < 11.5	questionable
a = 0	satisfactory

Table 3: Interpretation of a-scores. The level of seriousness is determined from confidence intervals around the mean proportion of participants that disagree with the consensus (as determined from historical data).

As it is not possible to calculate an uncertainty in the classical sense, we use the exact binomial test to calculate the p-value associated with the hypothesis that there is no clear consensus. This p-value can then be used to decide whether the a-scores should be issued for evaluative purposes. Scores for tests without a clear consensus can then be omitted from the combined score. Our approach to qualitative PTs allows participants to assess their trend in performance over time and provides the potential to compare their overall performance with that of other participants. However, caution is still advised as the combination of scores could be misleading: the severity of an unsatisfactory result such as non-detection of an allergen, could be lessened by combining with a-scores from other tests.

References

[1] International Organization for Standardization, *ISO/IEC 17043: Conformity assessment-General requirements for proficiency testing*. ISO, 2010.

[2] M. Thompson, "Proficiency testing: assessing z-scores in the longer term," *Anal. Methods Committee. AMCTB*, no. 24, pp. 2–4, 2001.

[3] Analytical Methods Committee, "Robust statistics: a method of coping with outliers," *AMC Tech. Br.*, no. 6, p. 2, 2001.

[4] M. Thompson, "The amazing Horwitz function," *R. Soc. Chem.*, no. 17, 2004.

[5] M. Thompson, S. L. R. Ellison, and R. Wood, "The International Harmonized Protocol for the proficiency testing of analytical chemistry laboratories (IUPAC Technical Report)," *Pure Appl. Chem.*, vol. 78, no. 1, pp. 145–196, 2006.

[6] S. Uhlig, C. Bläul, K. Frost, S. Sgorzaly, B. Colson, and K. Simon, "Qualitative PT data analysis with easy-to-interpret scores," *Accredit. Qual. Assur.*, vol. 20, no. 5, pp. 347–353, 2015.

[7] International Organization for Standardization, *ISO/IEC 13528: Statistical methods for use in proficiency testing by interlaboratory comparison* ISO, 2015.

[8] F. D. McClure, "Design and analysis of qualitative collaborative studies: minimum collaborative program," *J. Assoc. Off. Anal. Chem.*, vol. 73, no. 6, p. 953—960, 1990.

[9] D. A. Armbruster and T. Pry, "Limit of blank, limit of detection and limit of quantitation.," *Clin. Biochem. Rev.*, vol. 29 Suppl 1, no. August, pp. S49-52, 2008.

[10] Fera Science Ltd (Fera), "Fapas® Proficiency Testing." [Online]. Available: www.fapas.com.

[11] P. Wehling, R. a LaBudde, S. L. Brunelle, and M. T. Nelson, "Probability of Detection (POD) as a statistical model for the validation of qualitative methods.," *J. AOAC Int.*, vol. 94, no. 1, pp. 335–347, 2012.

[12] A. M. Committee and A. No, "Fitness for purpose: the key feature in analytical proficiency testing," *Anal. Methods*, vol. 7, no. 18, pp. 7404–7405, 2015.

[13] Analytical Methods Committee, "z-Scores and other scores in chemical proficiency testing—their meanings, and some common misconceptions," *Anal. Methods*, vol. 8, no. 28, pp. 5553–5555, 2016.

[14] P. Medina-Pastor, M. Mezcua, C. Rodríguez-Torreblanca, and A. R. Fernández-Alba, "Laboratory assessment by combined z score values in proficiency tests: Experience gained through the European Union proficiency tests for pesticide residues in fruits and vegetables," *Anal. Bioanal. Chem.*, vol. 397, no. 7, pp. 3061–3070, 2010.

[15] E. Bashkansky and V. Turetsky, "Proficiency testing: binary data analysis," *Accredit. Qual. Assur.*, vol. 21, no. 4, pp. 265–270, 2016.

[16] Fera Science Ltd (Fera), "Protocol for Proficiency Testing Schemes," 2017.