INTERNATIONAL ASSOCIATION OF GEOANALYSTS

PROTOCOL

FOR THE OPERATION

OF

G-Probe

PROFICIENCY TESTING SCHEME

Revision: August 2020

International Association of Geoanalysts
Secretary: Ms J Cook, BGS, Keyworth, Nottingham NG12 5GG, UK.
(jmc@bgs.ac.uk)

CONTENTS

Foreword

Any (chemical) laboratory must implement an appropriate program of quality assurance and procedures to monitor its operations to ensure it produces consistently reliable data[1]. Proficiency testing is one of these procedures and now plays an essential role in securing the performance of analytical laboratories in many fields. One of the main purposes of proficiency testing is to enable participants to detect unsuspected errors in their analytical systems. Of course, errors will always be present — that is the nature of measurement. However, it is essential that errors are sufficiently small to make them unlikely to affect the interpretation of the data. On the other hand, we must recognise that a reduction in uncertainty may be associated with a disproportionate escalation of costs, so it is equally important to avoid the production of data with unnecessarily small uncertainties. This concept of appropriateness has long been recognised by geoanalysts, and is nowadays called 'fitness for purpose'. It is fitness for purpose that proficiency tests should strive to represent.

Proficiency tests, therefore, exist primarily to encourage laboratories to move towards fitness for purpose by instigating remedial action where error of inappropriate magnitude is detected. In addition, proficiency testing is now recognised to be an essential ingredient of accreditation. Accreditation assessors will expect to see laboratories participating in a relevant proficiency testing scheme, if one exists in the sector, and will expect to see evidence of mainly satisfactory performance and of documented remedial activity in response to occasional lapses. Moreover, participants will want to demonstrate their capabilities to potential clients by showing that their proficiency test results have been largely satisfactory. While not part of the original ethos of proficiency testing, these later uses are simply a fact of current analytical life.

After successful implementation of the Geo*PT*™ proficiency testing scheme[2] for bulk analysis of predominantly silicate rock materials that has now been running for more than twenty years, and the establishment of the G-Probe proficiency testing scheme in 2008 as a cooperative venture between the United States Geological Survey (USGS) and the International Association of Geoanalysts (IAG) that has been initiated and master-minded for many years by Stephen Wilson of the USGS it was now decided to harmonize both schemes being operated through online systems. They were designed with all of the foregoing requirements in mind, and the IAG is confident that these proficiency testing programs will continue to fulfil a need in the geochemical community. Both schemes are undertaken on a non-profit basis within the IAG, and much of the work is done on a voluntary basis. Nevertheless, there are significant costs involved in running the scheme, including the preparation, packaging and checking of the test material, posting the material around the world, and preparing and distributing reports. These costs have to be passed on as a fee for participation. The fact that so many laboratories worldwide participate in IAG’s proficiency testing programs demonstrates that the enterprise is worthwhile to them and fulfils a major objective of the IAG, to serve the global geoanalytical community.

The G-Probe microanalytical proficiency testing programme is now organised primarily by the IAG with contributions from the USGS. It is designed to evaluate the performance of those laboratories specialising in the use of microanalytical techniques such as laser ablation ICP-MS, electron probe microanalysis (EPMA) or micro-X-ray fluorescence (μ-XRF) for the analysis of minerals and other geological and environmental materials. Test materials for the programme include natural and synthetic glasses as well as pressed powder samples made from oxides, limestone, corals, bones, polymetallic sulfides, organic materials etc..

Participating laboratories are provided with two test samples a year and are asked to send the organisers their results acquired under routine measurement conditions. The data submitted are evaluated using guidelines similar to those established by the Geo*PT* ™ programme, with an assessment of accuracy based on the z-score approach. Laboratories are provided with feedback on each element for which measurement results are reported, from which the laboratory can decide whether their reported data were satisfactory or possibly affected by unsuspected bias. The goals of the G-Probe programme are to (i) evaluate the routine analytical capabilities of microanalytical laboratories on a diverse range of sample types commonly encountered in the field of geochemical analysis, and (ii) enable participants to evaluate their performance relative to the scheme's fitness-for-purpose criterion and the performance of other participating microanalytical laboratories using the same or similar techniques.

To enhance the experience of participating laboratories and improve performance feedback, the IAG has invested in replacing the original paper-based, then digital spreadsheet-based schemes for reporting measurement results with a much more efficient and effective web-based management system. Laboratories receive a full report of each round whereby they can evaluate their performance in relation to the consensus of results as expressed by z-scores.

G-Probe test materials can be re-used as reference materials. Eventually, test materials from G-Probe rounds may be further characterised as certified microanalytical reference materials, thereby enhancing the value of these materials significantly. ISO Guide 35:2017[3] provides, for the first time, an alternative way of characterizing certified reference materials using proficiency testing. It has been shown that the well-established Geo*PT*™ proficiency testing scheme is considered to be competent for the certification of bulk geological reference materials[4] and it is the IAG's goal to develop the G-Probe proficiency testing scheme to the same status of competence.

Dieter Garbe-Schönberg, August 2020

Scheme Administrator, Chair of G-Probe Proficiency Testing Steering Committee

1 Overview of G-Probe

G-Probe provides a proficiency testing service for analytical laboratories employing microbeam techniques and operating in the areas of both pure and applied geochemistry. It is concerned mainly with the analysis of geological materials, especially minerals and glasses, but also synthetic and other natural materials.

The scheme offers one or two appropriate test materials for analysis per year. The principal test material may on occasion be accompanied by a supplementary test material. Participants are required to report their measurement results by a published deadline, and the results of that proficiency test round are made available to all participants in the form of a report once the data have been processed. The report enables participants to compare their quantity values, identified by code to maintain anonymity, with the scheme provider's best estimate of the true value for each measurand and to evaluate their performance both in relation to the scheme’s fitness-for-purpose criterion, and their peers’ current performance. Participants are also encouraged to review their results in relation to their own past performance.

Operation of the scheme is managed by a Steering Committee within the IAG. Feedback from participants is encouraged and should be referred in the first instance to the Scheme Administrator.

2 Principles of proficiency testing

Proficiency testing[1,2,4–8] is a widely accepted quality assurance tool developed for analytical chemists. It provides an opportunity for analysts and quality managers to test the reliability of their analytical procedures. In its usual form, proficiency testing involves the distribution of identical samples of a test material to participant laboratories for analytical measurement, usually by a method of the participant’s own choice, employing their routine procedures. Results must be reported by a specified deadline. The scheme providers compare each participant’s measurement result with the best available estimate of the true value for each measurand, and present the outcome as a score that represents the participant's analytical performance in terms of accuracy. The score is calculated on the basis of a performance criterion specified in advance and known to the participants. The test is repeated at regular intervals.

The main objective of proficiency tests is to provide a regular, independent and external check on the accuracy of measurement results, thereby allowing participants to detect, investigate and subsequently correct any unexpected sources of error in their routine analytical procedures. To achieve this objective the measurement results submitted should reflect the performance of the laboratory operating under normal routine conditions. Participants who employ non-routine procedures, or specially chosen analysts, or unusually careful methodology for measurement of proficiency testing samples are undermining the purpose of the scheme: they will not be able to discover deficiencies in their routine practices.

Participants are encouraged to have in place a system for responding to unsatisfactory results identified in a round of a proficiency test. Where necessary, further diagnostic tests should be carried out to determine the specific source of any unexpected error. Accreditation assessors will look for evidence not just of the overall successful participation in proficiency tests, but also of an appropriate response to unsatisfactory results, including investigation of the root cause and extent, implementation of corrective action, and follow-up to ensure that the corrective action had been effective.

An important aspect of proficiency testing is that it should encourage participants to achieve a standard of performance in the quality of their routine results that is fit for purpose. 'Fitness for purpose' implies that the uncertainty in a result is of a magnitude appropriate to the use to which the data will be put.

3 Organisation of G-Probe

G-Probe is managed for the IAG by a Steering Committee appointed by Council. Normally, one member of the Steering Committee acts as Scheme Administrator. At least half of the members of the Steering Committee must be members of the IAG and at least one member should be a member of Council. Members of the Steering Committee will usually be experts in geoanalysis and microanalysis and at least one member must have expertise in statistics. Details of the current management team are provided in Appendix A.

3.1 Terminology

Each distribution of material in G-Probe is known as a ***'round'***. The material sent to participants in a particular round is called the ***'test material'***. Individual packets of test material sent to participants are called ***'distribution units'***. The quantities of the test material analysed by participants are called ***'test portions'*** and are operationally defined by the analytical technique employed (e.g., EPMA, LA-ICP-MS) and the respective beam diameters and energies applied. The principal test material is intended to satisfy the proficiency testing requirements for the majority of participating laboratories and is generally referred to as the ***'routine'*** test material. An additional test material may, on occasion, be distributed, and is designed either to test the performance of participating laboratories on a wider range of geological matrices, or to act as a ***'traceability standard'*** for certification purposes. It is normally referred to as a ***'supplementary'*** material and is treated as a separate round for purposes of reporting and data handling.

3.2 Test materials

The test materials may comprise glasses (volcanic glass or synthetic glasses obtained from vitrification of silicate rock powders,) or alternatively natural or synthetic minerals and pressed powder pellets produced from a wide variety of materials. Test materials are delivered as one or more mm-sized individual fragments or chips, or pressed powder pellets. A test for sufficient homogeneity of the test material is made before distribution, broadly conforming with recommendations described in ISO Guide 35:2017[3] and in the International Harmonised Protocol[1]. However, while between-unit homogeneity is checked to be sufficient, the participant should not assume that the distribution unit itself is sufficiently homogeneous for their particular analytical procedure. It is the responsibility of the participants to ensure that the test portion or beam diameter used for analysis is representative of the whole of the respective test material in the distribution unit[1].

3.3 Distribution of materials

Test materials are distributed by an appropriate means, normally at least 10 weeks before the reporting deadline in order to secure delivery in sufficient time to allow analytical measurements to be undertaken. Test materials are accompanied by a ***Letter of Invitation*** to participants with ***Instructions to Analysts*** that provide full instructions for handling, measurement and reporting of results. Participants are notified when test materials have been dispatched and are required to inform the administrator if they have not arrived after 2 weeks have elapsed in Europe and North America, and 4 weeks have elapsed in South America, Africa, Middle East, Asia and Australasia.

3.4 Analysis by participants and handling of PT distribution units

Analytical measurements are conducted in the participant's laboratory by any procedure or procedures selected by the participant, but the analytical protocol used should reflect the routine practice in that laboratory. Routinely two sets of results are required when two glass chips A and B of material are provided. Only one quantity value per analyte may be reported for each item of the test material even if multiple measurements have been performed on a single item. More detailed information about reporting is given in the ***Instructions to Analysts.***

The scheme organisers recommend that participants adhere to good laboratory practice when handling and measuring the test materials provided by G-Probe. Material Data Sheets applicable to the test materials supplied are available on request from the Scheme Administrator. Occasionally, when the scheme organisers are aware that materials identified as being derived from mining operations or tailings sites, or believe that they could contain noteworthy amounts of toxic elements, a specific warning will accompany the test material. However, the IAG cannot accept responsibility for any damage or misadventure occurring from handling or processing test materials.

3.5 Reporting of analytical results by participants

Participants must report their measurement results online via the G-Probe website according to the specifications stated in the ***Instructions to Analysts***. Before recording their measurement results, participants are required to provide details of their analytical procedures, including the analytical technique employed, the beam size used and other operational parameters. Different procedures can be used for different analytes. Analytical results must be reported in the quantity units specified for the species identified, e.g. Au in mg kg^{-1}, CaO in g 100g^{-1}, and be submitted by the deadline specified in the ***Letter of Invitation*** in order to take part in the proficiency test. Once submitted, results cannot be changed, either by the participant, or by the organisers, even if they are clearly in error. The reporting of measurement results is regarded as part of the measurement process and, therefore, part of the proficiency test.

3.6 Confidentiality

Participants' identity will normally remain confidential to the scheme administrator alone. In exceptional circumstances when assistance is required from other members of the G-Probe Steering Committee or IAG Officers, those individuals will also be bound by confidentiality. Participants will be identified by a numeric code in all published tables of analytical results and z-scores, also on charts in reports and in any other publicly available document. Code identifiers are changed for every round of G-Probe. The scheme organisers will not disclose the code identity of a participant to any third party without the written approval of the participant.

3.7 Assessment of laboratory performance

To assess laboratory performance, the scheme organisers usually convert each measurement result reported by participants into a z-score (see Section 4 Scoring and statistical methods), which provides a means of assessing performance for each analyte in relation to the fitness-for-purpose criterion employed by G-Probe. For some analytes no z-scores are produced. This happens when it is not possible to derive either an assigned value (i.e., the best estimate of the true value of a measurand), or a provisional value (i.e., an estimate of the true value of a measurand that carries a greater degree of uncertainty, as described in Section 4).

3.8 Reporting of laboratory performance by the scheme organisers

For each round of the G-Probe proficiency testing scheme each participant is provided with online access to a full report of results which contains:

(a) a description of the details relating to the particular round, including the type of material and its source, a summary of the data submitted and an outline of the type of statistical analysis carried out by the scheme organisers;

(b) a table of raw results as supplied by all participants – listed by laboratory code which changes from round to round to ensure anonymity;

(c) a table listing all assigned and provisional values for the test material with corresponding uncertainty estimates, target precision and statistical details;

(d) a table of z-scores corresponding to the results supplied in (b) above, for those analytes credited with assigned or provisional values;

(e) sigmoidal charts of submitted data in which results are ordered incrementally, identified according to analytical technique, and shown relative to the optimal consensus value and appropriate z-score benchmarks, and

(f) a multiple z-score chart highlighting results for participant laboratories that may not be satisfactory and so require reviewing.

There may also be comments from the scheme organisers on particular analytical issues that have arisen. Reports will be made available to participants to download from the G-Probe website normally not more than 7 weeks after the reporting deadline.

The scheme organisers may publish G-Probe results and comment on them in other media. In any such publication, participants' results will remain anonymous and where necessary, identified only by their confidential code number, unless participants grant specific approval in writing for their identity to be disclosed.

3.9 Review by the scheme provider

On behalf of the IAG the Steering Committee will periodically review the efficacy of the scheme in print, or at Geoanalysis Conferences, and, should it be necessary, take any appropriate action to revise practices.

3.10 Correction of mistakes

Mistakes by participants

Mistakes in reporting made by participants will not be corrected. The resultant z-score will stand regardless of the nature of the mistake. The z-score represents the performance of a participant's whole analytical system, which includes not only the analytical measurement result, but the whole measurement process, including maintaining the identity and integrity of the sample, and the reporting of results.

Mistakes by G-Probe

Every reasonable effort is made by the scheme organisers to avoid mistakes in conducting each round of proficiency testing, from the provision of test materials to the calculation of z-scores. Participants should communicate any perceived problems to the scheme administrator, who will deal with them immediately if at all possible. If this is not possible, they will be subject to a thorough investigation by the Chair of the Steering Committee. G-Probe will issue a correction statement to the affected participant relating to any mistake that is substantiated.

3.11 Disclaimer

Neither the IAG, as scheme provider, nor individuals involved in the management of G-Probe or in the processing of contributed data, accept liability for the outcome of any mistakes in the operation of G-Probe. Participation in G-Probe implies that the participant accepts this condition. Full terms and conditions of participation in G-Probe are provided in Appendix D and are available from the G-Probe website.

4 Scoring and statistical methods

Scoring and statistical analysis in G-Probe is compliant with the ISO 13528:2015 Standard[8] relating to statistical methods used in proficiency testing which is largely based on the earlier recommendations of the IUPAC International Harmonised Protocol[1].

4.1 The z-score

Participants' reported results (x_i) for each analyte will be converted into a '*z*-score', defined by $z = (x_i - x_{pt}) / \sigma_{pt,}$ where x_{pt} is the organisers' best estimate of the true value of a measurand, and σ_{pt} is the corresponding standard deviation for proficiency testing (SDPT), a value based on a G-Probe fitness-for-purpose criterion as detailed below in Section 4.3. The SDPT is more conveniently referred to as the ***'target precision'*** in this document. Thus, $(x_i - x_{pt})$ is the measurement error and the *z*-score provides a measure of the accuracy of the result submitted, scaled according to the SDPT in a manner similar in function to a standard deviation that describes the acceptable range of variation among the results. Accordingly, a *z*-score outside the range ±3 implies that an unacceptable source of error *may* be present in the participant's analytical system and that the need for remedial action should be considered. *z*-scores that exceed ±2 carry the same message to a lesser degree, but will occur by chance with reasonable frequency (about one in twenty results for a participant complying exactly with the scheme's fitness-for-purpose criterion), so isolated values may not be especially noteworthy.

4.2 The assigned value

For a particular analyte, the assigned value is the scheme organisers' best estimate of the true value of the measurand in the test material and is evaluated as a consensus derived from all contributed measurement results. The consensus is recognized as the location on the measurement scale at which the density of contributed results is greatest. The function of the assigned value is to enable an estimate of error in a participant's measurements to be made. Estimation of the optimal consensus value and its associated uncertainty take account of the following:

- When the dataset is approximately symmetrical apart from a small proportion of outliers, the Huber H15 robust estimates of mean $(\hat{\mu})$ and standard deviation $(\hat{\sigma})$ of the *n* data are the statistics of choice. The consensus is taken as $\hat{\mu}$ and its uncertainty is taken as the standard error of $\hat{\mu}$, namely $\hat{\sigma}/\sqrt{n}$. (*Note*: in some instances it may be preferable to replace *n* by a slightly smaller value to account for the downweighting in the robust algorithm[10].)
- When the dataset is less symmetrical, but there is nevertheless a well-defined consensus, the median may be preferable to the Huber H15 robust mean as an estimate of the consensus. The uncertainty on the median can be taken as the simple standard error of the mean multiplied by $\sqrt{\pi/2}$, i.e. 1.2533.
- When the distribution is skewed, sometimes more strongly, but there remains a clear consensus, and the asymmetry is judged to originate from recognised technical deficiencies in measurement procedures, a mode may provide a suitable location estimate. Modes may be estimated in various ways, among them several described by Thompson[11]. A procedure involving resampling techniques ('bootstrapping'), as approved by the ISO Standard[8], provides an estimate of the standard error of the mode, a value that can be taken as the uncertainty of the consensus. In such circumstances, the mode provides a better definition of the consensus location than either the median or the robust mean.

The choice of location estimator to provide the optimal consensus value is made by expert judgement of the Steering Committee.

Criteria taken into account for a consensus value to be credited with 'assigned' status normally include:

- At least 15 valid measurement results (i.e. excluding outliers) contribute to recognition of the consensus.
- These data conform closely to a random sample from a normal distribution.
- The ratio of the uncertainty in the location estimate to the standard deviation for proficiency testing, i.e. the target precision, calculated as $u(x_{pt}) / \sigma_{pt}$ is an acceptably small value (usually less than 0.5, see Section 4.3).
- An evaluation of measurement results by procedure indicates no detectable procedural bias among measurement results from which the consensus is derived. If procedural bias is detected and its origin understood, it may be eliminated by using a judicious choice of procedure for deriving the consensus.

Where these criteria are not fully met, but a well-defined consensus value can be derived from the dataset, it may be credited with 'provisional' rather than ‘assigned’ status to provide laboratories with some useful *z*-score feedback. Instances of provisional status may be recorded when:

- A relatively small number of measurement results (but at least 8) contribute to the consensus.
- The measurement results are unduly dispersed in relation to target precision.
- The distribution of values is clearly skewed, but a meaningful consensus is still obtainable.
- An evaluation of measurement results by procedure indicates that bias is present and is either of unknown origin or cannot be entirely eliminated by judicious choice of procedure for deriving the consensus.

In some instances, it is not possible to estimate either a satisfactory assigned or provisional value, and then no *z*-scores are calculated. Charts of the results can still be useful, however, and are provided for information when more than 6 results are available. Circumstances where this is likely to happen are when:

- Too few measurement results are contributed and the uncertainty on the assigned value is high enough to affect the value of the *z*-scores (see Section 4.3). This commonly occurs when the number of results is less than about 10.
- The dispersion of the measurement results is unusually wide in relation to the target value, or the distribution of results is multimodal.
- The dispersion of the measurement results is grossly skewed and no meaningful consensus can be identified.
- An evaluation of measurement results by procedure indicates that bias is present and cannot be eliminated by judicious choice of procedure for deriving the consensus.

4.3 The 'standard deviation for proficiency testing (SDPT)' or 'target precision'

The SDPT or target precision, σ_{pt}, is a scaling factor which enables the bias in individual measurement results reported by a participant to be represented as a score. In G-Probe its value is based on a fitness-for-purpose criterion; the principle on which it is based is described below. The target precision effectively describes what is judged to be the optimal acceptable level of uncertainty in measurement results taking account of fitness for purpose and factors such as cost.

It must be emphasised that σ_{pt} is not a descriptor of the results. In G-Probe this parameter is not derived directly from the participants’ results. Consequently, there is *no prior expectation* that the participants’ results will be normally distributed or that about 95% of the *z*-score results for an analyte should fall within the range of ±2.

The value of σ_{pt} used in G-Probe is derived from the Horwitz function[12,13], $\sigma_H = 0.02c^{0.8495}$, where σ_H is the reproducibility between laboratory standard deviation observed at a mass fraction c, and both are expressed as mass ratios, for example, 1 mg kg^{-1} (i.e. 1 ppm) = 10^{-6}. The Horwitz function was originally derived from empirical observations that were found to apply over a wide range of measurands, test materials, analytes and physical principles underlying the bulk analytical measurement procedure[12].

In G-Probe, only one level of uncertainty is recognised as fitness-for-purpose criterion, where $\sigma_{pt} = \sigma_H / 2$, which is judged to be appropriate for high precision measurement in 'pure' geochemical research, where care is taken to provide data of high precision and accuracy, sometimes at the expense of a reduced sample throughput rate. Some values of relative standard deviation based on σ_{pt}, over the range of mass fractions routinely encountered, are given in Table 1.

The value of σ_{pt} also acts as a benchmark for judging the uncertainty on the assigned value $u(x_{pt})$. If $u(x_{pt}) / \sigma_{pt} > 0.6$ the *z*-scores may be unduly affected by the relatively high uncertainty, so they are not usually calculated. Sometimes, however, when considered justified, *z*-scores are calculated and credited with 'provisional' status, at the discretion of the Steering Committee.

Table 1. Relative standard deviations implied by the target precision σ_{pt}.

Quantity values	**Mass fraction**	**σ_{pt} %RSD**
100 g/100g	1	1
10 g/100g	0.1	1.4
1 g/100g	0.01	2
1000 mg/kg	0.001	2.8
100 mg/kg	0.0001	4
10 mg/kg	0.00001	5.7
1 mg/kg	0.000001	8
0.1 mg/kg	0.0000001	11.3
0.01 mg/kg	0.00000001	16
0.001 mg/kg	0.000000001	22.6

5 Testing for sufficient homogeneity

Materials for microanalytical proficiency testing will typically be manufactured from raw bulk materials either by melting (vitrification) or ultra-milling of rock powders for pressed powder pellets and then split into numerous smaller distribution units. Conventional homogeneity testing is geared to ensuring that no statistically significant difference can be detected between distribution units. Such tests (experimental homogeneity studies) are described in the IUPAC Harmonised Protocol[1] or ISO Guide 35: 2017[3] and are outlined in Appendix B. They are designed to ensure that participants receive identical distribution units of essentially the same material.

In a second step, a microanalytical testing scheme is applied to a statistically defined number of individual distribution units to investigate their homogeneity in the micrometre range (Appendix B).

6 Using the information and materials supplied by G-Probe

Proficiency test results are primarily for the use of participants to:

(a) identify unexpected sources of error in their results;

(b) establish whether any remedial action previously taken to reduce errors had been successful; and

(c) check, in general, that the laboratory is working to an expected level of uncertainty.

For the increasing proportion of laboratories becoming involved in accreditation, there is an obligation to participate in a relevant proficiency test if one is available and, moreover, to demonstrate both overall appropriate performance and the effectiveness of procedures to deal with occasional inappropriate performance. It is now commonplace for a laboratory to use proficiency test results to demonstrate that a particular level of analytical performance can be achieved. All of these circumstances require the judicious use of the results.

Such activities are essentially the responsibility of the participant. G-Probe does not have the resources for activities beyond the preparation of reports and certificates as detailed above. However, some suggestions for optimal use of the data are provided here.

6.1 How to assess your results

If nearly all of the *z*-scores obtained in a round are within the range ±2 and the remaining few are outside the range by a small margin, then probably all is well with the analytical system. If only a small proportion of measurands give rise to *z*-score results outside the range ±2, it should first be considered whether they could plausibly have arisen by chance. For example, although G-Probe *z*-scores cannot be interpreted in terms of strict confidence limits, it would be reasonable to expect about one in twenty results to be outside the range ±2, and no further investigation would be necessary. A control chart approach is best practice for the long-term assessment of *z*-scores. Two successive *z*-scores outside the range ±2, or one result somewhat more extreme, would call for action.

A higher proportion of such results falling outside the ±2 range would call for further investigation and possibly the installation of a more comprehensive internal quality control system. The form that corrective action might take depends on the nature of the error. Accreditation bodies would expect to see a mechanism for responding to the outcome of each round, so participants should adopt and document a systematic way of investigating any unsatisfactory results.

G-Probe is not designed to be diagnostic: it provides no direct information for the participant to determine the sources of error within the analytical system. The participant will need to devise additional tests taken from the normal range of quality assurance practices to obtain such information. Nevertheless, because the scheme involves multiple analytes, some limited diagnostic indicators can be derived from the results themselves. Further advice on diagnostic indicators is provided in Appendix C.

6.2 Proficiency testing in the overall context of quality assurance

Proficiency testing, being an occasional check on the accuracy of measurements from an analytical system, must not be confused either with internal quality control (IQC) or with validation, both of which provide ways of monitoring routine analytical operations[14] on a routine basis.

6.3 Use of excess test material

G-Probe test material remaining after analysis can be used in a number of ways by the participant. Following G-Probe characterisation, such materials are likely to have the status of reference materials, with respect to the assigned value and its uncertainty.

G-Probe accepts no responsibility for the outcome of any use to which a test material is put. Although there are good grounds for believing assigned values to be reliable indicators of composition, it is generally considered that they do not yet, without further evaluation, have the same status as certified reference values. Future G-Probe rounds may be designed in such a way that the resulting proficiency testing data can be used for subsequent certification[4,9,14] of the test material. Some test materials will be available for sale through IAGeo Ltd at reduced rates for IAG members.

6.4 Comments on classification and ranking

There is no merit in converting *z*-scores into named classes—it destroys the information content. However, it is useful to define warning and action limits in terms of *z*-scores. G-Probe does not rank the performance of participating laboratories. A discussion of the serious shortcomings of ranking methods can be found in the IUPAC Harmonised Protocol[1].

7 Ethical considerations

G-Probe is offered on the understanding that participants are using the results to check their routine analytical activities and that the results submitted reflect that usage. Therefore, the results should incorporate errors from all of the normal sources. This means that the proficiency test material should be treated exactly like a routine sample, with no special attention paid to improving the results, no particular analyst assigned to its handling, and no more than the routine number of separate results averaged to form the submitted result.

Collusion amongst participants must be avoided. Whilst not suspected in G-Probe, it has been detected in other proficiency tests where accreditation puts pressure on participants to perform well. Measures are in place to check submissions for signs of collusion. The IAG reserves the right to exclude from the G-Probe scheme any laboratory for which a *prima facie* case of collusion can be established.

Participants must be careful to avoid giving any false impression when promoting the results of proficiency tests to advertise their services. Wider decisions based on *z*-scores should be made only with expert consideration of the analytical and statistical principles involved.

The situation sometimes arises that a member of the Steering Committee is a member of staff at a participating laboratory. G-Probe undertakes to ensure that such laboratories are treated in exactly the same way as other participants, and do not receive any privileged information regarding the test material, or other participants' results.

Appendix A

Constitution of the G-Probe management team

As of April 27, 2020, the permanent members of the Steering Committee are: Dr C-D Garbe-Schönberg - Scheme Administrator and Chair (CAU Kiel University, Germany), Dr S Wilson – Past Scheme Administrator (Mineral Resources Program; USGS Geology, Geophysics, and Geochemistry Science Center, Denver, USA), , Prof. P J Potts (The Open University, UK), Dr P C Webb – Website Administrator (formerly of The Open University, UK), Prof. M Thompson – Statistician (Birkbeck College, University of London, UK), Dr C J B Gowing (British Geological Survey, UK), Prof. L. Danyushevsky (CODES Laboratory, Hobart, Australia), Dr R. Mertz-Kraus (University of Mainz, Germany), Dr A. Kronz, Electron-Microprobe Laboratory, University of Göttingen, Germany). Temporary members involved in the provision of test materials are co-opted for specific rounds. The Scheme Administrator may be contacted by email: *gprobe.iag@gmail.com*. The G-Probe Subscriptions Manager is Mr C Jackson, who may be contacted by email: *iag-treasurer@virginmedia.com*.

Appendix B

Testing for sufficient homogeneity

Heterogeneity contributes to the uncertainty on derived values. It is tested separately and also related to the value of σ_{pt}. The term 'sufficient homogeneity' recognises that all materials including glasses will be compositionally heterogeneous at some scale. Pressed powder pellets or sintered materials are multi-phase materials that are heterogeneous in the true sense, but grain size of individual phases controls whether this heterogeneity can be detected by a given microbeam technique. Homogeneity refers to both between-unit variation and within-unit heterogeneity. Sufficient homogeneity means that the contents of the distributed units of the test material do not differ among themselves, and that the contents measured at any point on a single unit do not vary sufficiently to affect the outcome of a proficiency test for bulk analysis: that is, the z-scores will not be affected to any noticeable degree. Clearly, participants in a proficiency test must be confident that the material they are dealing with is sufficiently homogeneous. It should be noted that a material can be sufficiently homogeneous for some analytes and not for others and, hence, multi-analyte homogeneity tests are needed for G-Probe.

Concern has been expressed relating to usefulness of homogeneity tests of bulk materials after milling and re-homogenization, as the tests seldom detect significant heterogeneity because experimental designs that are economically feasible have insufficient power[16]. This applies, in principle, also to between-unit heterogeneity of test materials obtained from melting or ultra-milling of homogeneous powders. In contrast, homogeneity can be an issue for *in situ* microanalysis with high spatial resolution. While between-unit homogeneity is checked to be sufficient, a participant should not assume that the distribution unit itself is sufficiently homogeneous for their particular analytical procedure. It is the responsibility of the participant to ensure that the test portion or beam diameter used for analysis is representative of the whole of the test material in the distribution unit[1].

The procedure applied here for investigating homogeneity of test materials for *in situ* micro-analysis comprises two steps for testing individual distribution units of glass chips, pressed powder pellets and other materials. It is assumed here that further processing – e.g., embedding of glass chips into resin plugs and subsequent polishing, or pressing powders into tablets – will not significantly change the properties of the test material and, hence, will not contribute to the observed between-unit variation.

Step 1: Bulk analysis. Tests for sufficient homogeneity between single test material units are based on the bulk analysis of a number of distribution units. The Harmonised Protocol[1] and ISO Guide 35:2017[3] outline specific methods for carrying out this procedure.

- Split the material into distribution units (e.g. one or more mineral or glass fragments) or their equivalent (e.g., ultra-milled sample powder equivalent to the mass needed to manufacture a single pressed powder pellet.
- Select at random a number ($n > 10$) of the distribution units as specified in ISO Guide 35:2017 under section 7.4.1[3].
- Analyse the distribution units using a suitable experimental design that allows for separate estimation of within-run, between-run, and between-unit variances by a bulk analytical method with a sufficiently good precision (typically a solution-based method given the small sample mass available for analysis). Ideally, the analytical procedure used in this step should be traceable to the SI system. Record the result with sufficient significant figures to represent adequately the variability of the measurement. If in doubt, collect more significant figures than is normally justified. Ideally, the analytical repeatability standard deviation, σ_r, should be smaller than $0.4\sigma_{pt}$: if it is not, heterogeneity that is significant in the interpretation of G-Probe results may be undetectable.
- Inspect the results graphically, paying attention where necessary to:
  (a) outlying analyses, indicated by an exceptionally large difference between duplicated results for a distribution unit, which indicates analytical blunders. Such results, after confirmation by an outlier test, should be deleted from the data. Failure to delete could cause a heterogeneous material to pass the test.
  (b) outlying distribution units, which indicates that the material may really be heterogeneous. Such outliers must never be deleted before the statistical test.
  (c) non-random patterns among the results, which should be referred to the statistical expert, but may mean that the data should be abandoned and the whole test repeated.

- Calculate, by analysis of variance, MSW (the mean square within samples, i.e, between analyses), MSB, the mean square between samples, and calculate the estimated analytical standard deviation, $s_r = \sqrt{\text{MSW}}$, and the sampling standard deviation component, $s_s = \sqrt{(\text{MSB} - \text{MSW}) / m}$, where $m = 2$ for duplicate analysis.
- If the probability associated with the value $F = \text{MSB} / \text{MSW}$ is greater than 0.05, then no significant heterogeneity has been detected. So long as $s_r < 0.4\sigma_{pt}$, the material is taken as sufficiently homogeneous. Even if the material is significantly heterogeneous, it is taken as sufficiently homogeneous if $s_s < 0.4\sigma_{pt}$. If the analytical method has poor precision, the test may be incapable of detecting an important degree of heterogeneity. If the analytical method is very precise, even very small and unimportant heterogeneities could be statistically significant.

Step 2: Microanalysis. If the test material can be characterised as being sufficiently homogenous with respect to between-unit variation then within-unit heterogeneity is tested in a second step by appropriate *in situ* microanalytical methods (e.g. EPMA, SEM, LA-ICP-MS). It should be re-emphasised here that a material can be sufficiently homogeneous for some measurands and not for others, and that a material can be sufficiently homogeneous when analysed with a wide micro-beam diameter but not with a narrower beam. The definition of homogeneity depends strongly on the spatial resolution of the *in situ* microanalytical technique employed.

- Select at random a number ($n \geq 10$) of the already prepared distribution units (e.g., mineral or glass fragments, pressed powder pellets, etc.).

- Analyse the selected distribution units using a suitable experimental design that allows for separate estimation of uncertainty of the measurement of the within-run, between-run, and between-unit variances by an analytical method with a sufficiently good precision. If possible, analyses should be repeated with different beam sizes accounting for spatial heterogeneity that may vary with a measurand, and allowing for defining a minimum beam size (minimum test portion). Where measurements cannot be repeated at the same spot for estimating the measurement uncertainty, variances of the individual spot analyses can be compared against an independent estimate of the measurement repeatability[17].
- Further evaluation of the homogeneity study in step 2 will be done by analogy with Step 1.

Having determined the between-unit (Step 1) and within-unit (Step 2) standard deviation, it can be confirmed if the variation within and between units is sufficiently small for the use as a material for G-Probe. G-Probe recognizes materials as being “sufficiently homogeneous” if heterogeneity is not noticeable (i.e. $s_s << \sigma_{pt}$) with common and widely used microbeam techniques like e.g., EPMA, LA-ICP-MS.

Appendix C

Advice for investigation of unsatisfactory results

When there is evidence as indicated in Section 6.1 How to assess your results, that *z*-score results are unsatisfactory, the participant should attempt to discover whether the error is due to a systematic or a random effect. This can be ascertained by obtaining a few repeated measurement results for the test material in successive runs. Analysis of a matrix-matched certified reference material at the same time will help to reinforce the interpretation. Variability of measurements from such a test suggests a random effect, which could be due to a number of problems, such as determining quantity values too close to the detection limit of the method under the chosen conditions of operation, or taking insufficient care with the manipulation of the test material or the operation of the instrument. A persistent deviation from the assigned value of roughly the same magnitude over several runs suggests a systematic problem. This could be due to a number of causes and should be investigated further. One possibility is calibration of the instrument with measurement standards that are not matrix matched, i.e., having a significantly different composition or internal structure when compared to the test material. Other possibilities include improper set-up and operating conditions of the instrument, or unidentified processes during the measurement causing elemental fractionation. It should be noted that some matrix interferences can affect different elements to different degrees.

If, in a multielement analysis, a number of measurands are simultaneously suspect, the fault is probably systematic and must arise at that part of the analytical system where all of the affected analytes are involved. For example, if the errors are nearly all in the same direction, a problem with internal standardization might be the reason.

For accreditation purposes it is important to document the procedures used for investigating any problems, to keep records of actions taken and the consequential effect(s) of such actions.

In the long-term, it is beneficial for a participant to record their *z*-scores graphically so that results can be compared both by round and by element. This can be very effectively done with a chart similar in form to the 'Multiple *z*-score chart' used by G-Probe for comparing within-round results.

Appendix D

Terms and Conditions of Participation in G-Probe

The G-Probe programme is operated by the International Association of Geoanalysts (IAG) for the benefit of the geoanalytical community. By taking part in G-Probe, participants must accept the following:

General terms and conditions

A. The IAG shall not be liable for any loss, damage, personal injury or death (other than death or personal injury suffered as a result of negligence on the part of the IAG) which results from the operations of the participant whether or not in relation to G-Probe.

B. The IAG shall not be liable to the participant for loss (whether direct or indirect) of reputation, profits, business or anticipated savings or for any indirect or consequential loss or damage whatsoever even if previously advised thereof and whether arising from negligence, breach of these Terms and Conditions or howsoever.

C. In any event, and notwithstanding anything contained in these Terms and Conditions, IAG's liability in contract, tort (including negligence or breach of statutory duty) or otherwise arising by reason of or in connection with these Terms and Conditions shall be limited to the price for the proficiency test giving rise to such liability.

D. The IAG does not grant any warranties in relation to G-Probe products or the supply of analytical services or distribution of the proficiency test, and all other conditions, warranties, stipulations or other statements whatsoever, whether express or implied, by statute, at common law or otherwise howsoever, relating to the G-Probe products, analytical services or proficiency tests are hereby excluded. In particular, (but without limitation to the foregoing) no warranties are granted regarding the fitness for purpose, performance, use, quality or merchantability of the G-Probe products, whether express or implied, by statute, at common law or otherwise howsoever.

Specific terms and conditions

1. Each round of the G-Probe programme is conducted as far as possible in accordance with the published *Protocol for the Operation of* G-Probe *Proficiency Testing Scheme* (2020), available for download at http://www.geoanalyst.org/documents/G-Probe-protocol.pdf. If variations arise in a particular round, they are documented in the relevant report. Whilst every effort is made to ensure that the operation of G-Probe conforms to the published protocol and that results appearing in G-Probe reports provide an accurate account of the results submitted, neither the IAG nor any individuals undertaking activities on behalf of the IAG can be held liable for deficiencies in the operation of G-Probe nor for errors made in the reporting of results nor for the consequences of any errors that might occur. Participation in G-Probe implies acceptance of this condition.

2. Participation in G-Probe is open to any commercial enterprise, academic institution or government organisation making advance payment to the IAG at the current rate. The administrators of G-Probe reserve the right to exclude laboratories whose subscription remains unpaid by the reporting deadline. G-Probe will provide paid-up subscribers with a report that expresses their results in the form of a *z*-score for each oxide/element that has been given an assigned or provisional value.

3. G-Probe undertakes to supply subscribers twice per year with a test sample that is suitable for in situ-microanalysis - typically a glass - and has the composition of a common geological material; this may be accompanied by a supplementary sample such as carbonates and sulphides in pressed powder pellets or any other suitable material. Test samples are dispatched to addresses recorded by participants on the G-Probe website, allowing ample time for delivery and for analysis to be undertaken. Participants are notified by email at the time of dispatch. If a sample has not arrived at a destination in Europe or North America by 2 weeks following dispatch, or in the rest of the world by 5 weeks following dispatch, the organisers should be informed (gprobe.iag@gmail.com) and a replacement will be sent by courier. While G-Probe will make every effort to ensure that samples are received in good time, G-Probe cannot be held responsible for non-arrival of test samples. However, if the non-arrival of a sample prevents participation in a round of testing, G-Probe will apply a credit to the subscriber's account.

4. The identity of laboratories submitting results to the G-Probe programme and their account details are maintained as confidential by the IAG. The IAG reserves the right to publish reports or any other investigations involving G-Probe data (containing anonymised details of analytical results and/or procedures undertaken) by any appropriate means. Neither the identity of a participating laboratory nor the results submitted will be communicated to any third party without the formal approval of that laboratory. Full details of the data protection policy of the IAG are available at: http://www.geoanalyst.org/data-protection-policy/
Note: To maintain confidentiality, it is incumbent upon participating organisations to inform us when contact personnel are no longer active. Only when we have such notification can we ensure that former contact personnel cannot continue to access results.

5. Paid-up participants of the G-Probe programme have the right to use the G-Probe website to enter and/or import data, to access reports, to amend their own details for communication and for delivery of samples and to receive news from the organisers. The organisers reserve the right to deny access to any participant who abuses the system.

6. Participants are expected to ensure that their personal and institutional data as recorded on the G-Probe website are correct and up to date so that they continue to have access to the website and are contactable by the organisers. In addition, they should ensure that receipt of emails from gprobe.iag@gmail.com for direct contact with the administrator, and noreply@gprobe.info for automated notifications is permitted and they are not intercepted or trapped as junk or spam.

7. Participants are expected to supply via the G-Probe website analytical data and information about their procedures that are correct to the best of their knowledge, and to ensure that the data have been obtained and reported in the manner requested in the *Instructions to Analysts*. Participants are requested to check that the results they have recorded on the system are correct before submitting them. Should there be any apparent discrepancies in the data recorded in the report, participants should notify the G-Probe administrator immediately. If incorrect data have been submitted, *z*-scores can be provided for the revised data. However, results cannot be corrected in the report produced for any round as the reporting of data is considered to be part of the proficiency test.

8. Participants are expected to adhere to good laboratory practice when analysing the test samples provided by G-Probe. In particular, test samples should be handled with appropriate care in respect of health and safety that is compatible with geological samples. This includes taking precautions against the inhalation or ingestion of dust when handling test samples. This holds especially true for ultra-

milled pressed powder pellets containing significant amounts of nanoparticles. Samples identified as being derived from mining operations or tailings sites or certain categories of environmental sample could contain significant amounts of toxic elements that are hazardous to health. IAG cannot accept responsibility for any damage or misadventure occurring when handling or processing the test samples.

The International Association of Geoanalysts, July 2020

References

1 M Thompson, S L R Ellison and R Wood, 2006. The international harmonised protocol for the proficiency testing of analytical chemistry laboratories. (IUPAC Technical Report) *Pure Appl. Chem.*, **78**, 145-196.

2 Protocol for the operation of Geo*PT*™ proficiency testing scheme. - International Association of Geoanalysts, 2017.

3 ISO Guide 35: Reference materials — Guidance for characterization and assessment of homogeneity and stability. – Geneva, 2017-08.

4 P J Potts, P Webb, M Thompson, 2019. The GeoPT Proficiency Testing Programme as a scheme for the certification of geological reference materials. *Geostandards and Geoanalytical Research,* **43**, 409-418. doi: 10.1111/ggr.12261

5 P J Potts, M Thompson and P C Webb, 2015. The Reliability of Assigned Values from the GeoPT Proficiency Testing Programme from an Evaluation of Data for Six Test Materials that have been Characterised as Certified Reference Materials. *Geostandards and Geoanalytical Research*, **39**, 407-417.

6 R E Lawn, M Thompson and R F Walker, 1993. Proficiency testing in analytical chemistry. *The Royal Society of Chemistry*, Thomas Graham House, Science Park, Milton Road, Cambridge CB4 4WF, UK, 110pp.

7 ISO/IEC 17043, 2010. Conformity assessment - General requirements for proficiency testing, *ISO*, Geneva.

8 ISO 13528, 2015. Statistical methods for use in proficiency testing by interlaboratory comparison, *ISO*, Geneva, 2015.

9 ISO 17034, 2016. General requirements for the competence of reference material producers. International Organization for Standardization (ISO), Geneva, p 24

10 M. Thompson, 2006. The variance of a consensus. *Accred. Qual. Assur.*, **10**, 54-575.

11 M Thompson, 2017. On the role of the mode as a location parameter for the results of proficiency tests in chemical measurement. *Analytical Methods*, **9**, 5534–5540.

12 W Horwitz, L R Kamps and K W Boyer, 1980. Quality assurance in the analysis of foods and trace constituents. J. *Assoc. Off. Anal. Chem.*, **63**, 1344.

13 M Thompson, 2000. Recent trends in inter-laboratory precision at ppb and sub-ppb concentrations in relation to fitness for purpose criteria in proficiency testing. *Analyst*, **125**, 385-386.

14 M Thompson and R Wood, 1995. Harmonised Guidelines for Internal Quality Control in Analytical Chemistry Laboratories, *Pure Appl. Chem.,* **67**, 649-666.

15 P C Webb, P J Potts, M Thompson, S A Wilson, and C J B Gowing (2019). The Long - Term Robustness and Stability of Consensus Values as Composition Location Estimators for a Typical Geochemical Test Material in the Geo*PT* Proficiency Testing Programme. *Geostandards and Geoanalytical Research* 43, 397–408.

16 M Thompson, 2015. Is your ‘homogeneity test’ really useful? *Anal. Methods*, **7**, 1627.

17 M H Ramsey and M Wiedenbeck, 2017. Quantifying Isotopic Heterogeneity of Candidate Reference Materials at the Picogram Sampling Scale. *Geostandards and Geoanalytical Research* **42**, 5–24.

Additional references relevant to the G-Probe programme

Analytical Methods Committee. z-Scores and other scores in chemical proficiency testing—their meanings, and some common misconceptions. *AMC Technical Briefs* No. 74. (*Anal. Methods*, 2016, 8, 5553)

Analytical Methods Committee. Fitness for purpose: the key feature in analytical proficiency testing. *AMC Technical Briefs* No. 68. (*Anal. Methods*, 2015, 7, 7404)

Analytical Methods Committee. The amazing Horwitz function. *AMC Technical Briefs* No. 17.

Analytical Methods Committee. Proficiency testing: assessing z-scores in the longer term. *AMC Technical Briefs* No. 16.

Analytical Methods Committee. The J-chart: a simple plot that combines the capabilities of Shewhart and cusum charts, for use in analytical quality control. *AMC Technical Briefs* No. 12.

Analytical Methods Committee. Understanding and acting on scores obtained in proficiency testing schemes *AMC Technical Briefs* No. 11.

Analytical Methods Committee. Robust statistics: a method of coping with outliers. *AMC Technical Briefs* No. 6.

(*AMC Technical Briefs* are short articles covering a wide range of technical issues affecting the analytical chemist. They are produced by the Analytical Division of the Royal Society of Chemistry. They can be downloaded gratis from the website www.rsc.org/amc. From issue No. 50, they can also be downloaded gratis from the website of the RSC journal *Analytical Methods.*)