NPL Report CISE 22/99

Report to the National Measurement System Policy Unit, Department of Trade & Industry

SURVEY REPORT ON TESTING SPREADSHEETS AND OTHER PACKAGES USED IN METROLOGY

FROM THE UK SOFTWARE SUPPORT FOR METROLOGY PROGRAMME

BY

M G COX, M P DAINTON, P M HARRIS

and B A WICHMANN, NPL

March 1999

ISSN 1361-407X

Authorised by Dr Dave Rayner,
Head of the Centre for Information Systems Engineering

National Physical Laboratory, Queens Road, Teddington, Middlesex, TW11 0LW

Contents

1. Introduction

The NMS Software Support for Metrology Programme (April 1998 to March 2001) is a UK government-supported programme designed to tackle a wide range of generic issues arising from the use of software in measurement science. One of the projects within this programme is concerned with the use of spreadsheets and other software packages in metrology.

We recognise that the numerical correctness of software is of great concern: a poor software implementation can lead to inaccuracy that could have been avoided, and as a consequence the accuracy of measurement results is compromised. Spreadsheets are widely used in the day-to-day operation of many measurement and calibration laboratories, and yet the facilities they provide are usually taken on trust without any testing or validation.

The purpose of this survey report is to inform those interested in the NMS Software Support for Metrology (SSfM) programme about existing work worldwide on spreadsheet quality and reliability with recommendations on its applicability in the metrology field.

Information for the survey was gathered by:

1. canvassing spreadsheet users at NPL,
2. contacting BIPM and the following National Measurement Institutes (NMIs):
    a) PTB (Berlin and Braunschweig, Germany),
    b) IMGC (Italy),
    c) DFM (Denmark),
    d) Van Swinden Laboratory (The Netherlands),
    e) NIST (United States),
3. canvassing potential members of the SSfM Club and calibration laboratories through other NMS Clubs,
4. canvassing other contacts, such as those working in the safety critical area, known to have an interest in this area of work, and
5. literature and web searches.

The report is organised as follows. The results of the survey, including comments made by those who replied to our request for information, are summarised in Section 2. The results are classified according to the type of organisation for their source: NMIs, industrial (scientific), industrial (business), academia and others. Section 3 contains our general conclusions and recommendations. In Section 4 we acknowledge those who provided information to this survey, and Section 5 contains references.

2. Results

2.1 BIPM and the National Measurement Institutes

2.1.1 BIPM

The BIPM is at present heavily engaged in putting together an operational and administrative system to support, maintain and record a world wide mutual recognition agreement (MRA) related to national measurement standards. It has been agreed that there will be a single database to record the results of the international comparisons of national measurement standards that form the technical basis of the agreement. Within the database there will be the results of individual comparisons and in these there will often be a spreadsheet giving detailed results. One of the matters now under discussion is the method of ensuring that only authorized persons can enter data at the various levels and also making sure that the data included is correct. It is clear that databases are going to be of increasing importance in international metrology and BIPM are, therefore, most interested in the project.

2.1.2 NPL

The software used at NPL within its measurement and calibration activities is generally written using high-level languages such as those based on the Basic language (including HP-Basic, HT-Basic and Quick-Basic), and Fortran. However, because spreadsheets readily allow the manipulation and presentation of data, the use of spreadsheet packages is increasing and can be expected to continue to do so. Spreadsheet packages used at NPL include Microsoft Excel and, to a lesser extent, Lotus 1-2-3.

There are a variety of applications of spreadsheets including, in NPL's scientific activities, the processing of experimental data and the construction of uncertainty budgets, and in its management activities and application of the quality system, the processing of financial and resourcing information. Moreover, applications exist across all areas of metrology.

The complexity of these applications ranges from the numerical processing of data using in-built functions for performing, for example, mean and standard deviation calculations, evaluation of statistical distributions and regression analysis, to the use of "macros" implemented in the internal macro and programming language (for Excel, this language is Visual Basic for Applications). Furthermore, in (at least) one application automatic links between spreadsheets are used to enable a hierarchy of measurement standards to be easily defined.

Few problems are reported with the in-built functions, but generally little formal testing of these functions is undertaken by users, their correctness being taken on trust. One application has been identified in which "spot-checks" of the in-built logarithm function were undertaken as part of the testing of the spreadsheet. In other applications, there are examples of good-practice, including locking cells and using printouts of the spreadsheet formulae to document the calculations performed. Furthermore, spreadsheet users do try to follow NPL's internal quality procedures on software development with regard to specification, design, implementation and testing.

2.1.3 PTB

Dieter Richter, Head of the Software Quality Assurance and Measurement Data Processing Section at PTB (Berlin), believes that spreadsheets are not used by many people at the PTB. We believe that this is not the result of any deliberate decision to avoid the use of spreadsheets because of their possible fallibility.

2.1.4 DFM

DFM uses programs written in C to automate their measurement systems and to perform most of the routine calculations associated with their measurements. However, it is reported that their use of Excel worksheets[1] is increasing, because it provides a simple way of presenting results in graphs and tables. Since Excel includes matrix operations and a large range of mathematical and statistical functions it is very easy to test new ideas without having to write a complete C program. Excel is used to verify that C programs used in mass measurement are providing the correct results: the values of the calibrated masses and their associated uncertainties. By calculating the relevant covariance matrices, which for physical reasons should have a number of

[1] In Excel, the worksheet is the primary document used to store and manipulate data. Worksheets consist of cells organised into rows and columns that may contain data or the outputs of worksheet functions.

zero elements, they believe that the accuracy of results returned by Excel is significantly lower than what would be obtained using C because of the presence of values that differ appreciably from zero.

Recently, DFM has produced an Excel worksheet for the evaluation of measurement uncertainty in accordance with the ISO Guide [1] and EAL-R2 [2]. In this worksheet, the user types in information about the input quantities and the equation that relates the input quantities to a single output quantity. A macro then performs numerical differentiation (to estimate the sensitivity coefficients), calculates the uncertainty contribution from each input quantity, calculates the combined uncertainty and the associated effective degrees of freedom and finally calculates the expanded uncertainty at the confidence level specified by the user. Since DFM sells this program they have been concerned with the protection of the worksheet. Unfortunately, locking a cell means that the format of the number cannot be changed, which is annoying if the user wishes to adjust the number of digits in the number to give a more appropriate presentation. In general, the more a worksheet is protected the more features appreciated by a user are lost. DFM reports that a systematic way of testing that a worksheet is foolproof seems to be very difficult to find.

Except for the well known example in Excel where the standard deviation of a few very large, but equal, numbers sometimes proves to be different from zero, DFM does not report any other inaccuracies in the built-in functions. Normally, extreme parameters as input to these functions are avoided. For example, the function "TINV" for evaluating the inverse of the Student's t-distribution for a specified number of degrees of freedom fails to return a value when the number of degrees of freedom becomes too high. This situation is trapped in their worksheet for the calculation of measurement uncertainty, and the normal distribution is used instead.[2]

In general, testing of software, including worksheets, is achieved by typing in input data and comparing the output of the program with the "true" result. Sometimes the "true" result is calculated from the input by other means, e.g., manually. Other times the true result is selected, and a set of input quantities is calculated to give the true result (a general methodology for generating reference data and results in this way is described in, for example, [3, 4, 5]). Testing is generally not performed in a systematic way.

2.1.5 NIST

NIST is very interested in the issue of numerical accuracy of statistical software. In response to concerns of US industry about this very issue, the Statistical Engineering and Mathematical and Computational Sciences Divisions of NIST's Information Technology Laboratory developed the Statistical Reference Datasets (StRD) Web service [6]. This service provides reference datasets with certified computational results for a variety of statistical methods. Currently datasets and certified values are provided for assessing the accuracy of software for univariate summary statistics, linear regression, nonlinear regression, and analysis of variance. The collection includes both generated and "real-world" data of varying levels of difficulty. Since its public release in August 1997, response has been favourable, averaging 1500 distinct "users" per month between January and June of 1998.

Generally, the methodology applied in StRD is to use multiple precision calculations (accurate to 500 digits(!)) using the preprocessor and Fortran subroutine package in [7] available from NETLIB. Data is read in as multiple precision numbers and all calculations are performed to this high precision. The results are output in multiple precision, and only then rounded to 15 significant digits. The results are an idealization: they represent what would be achieved if calculations were made without roundoff or other errors. Any typical numerical algorithm (i.e., not implemented in multiple precision) will introduce errors, and will produce results which differ

[2] The normal distribution is the limiting form of the Student's t-distribution as the number of degrees of freedom tends to infinity.

from the certified values. The extent to which the results differ depends on the difficulty of the problem solved and on the quality of the solution algorithm and its implementation. Some guidance on ways of remedying the performance of poor software is provided. For example, for linear regression, centring the data is suggested as a way of generating better results.

The web-site [8] describes NIST activity in the area of testing special functions or higher transcendental functions (as they are also known) including, for example, Bessel functions, error functions and exponential integrals. In this area of numerical computation, input to its software tends to consist of the values of a small number of parameters (*eg*, the argument z of the error function $erf(z)$), and consequently high accuracy reference software that produces reference results with guaranteed error bounds is feasible. NIST is experimenting with the development of a web-based interactive software testing service for special functions. This project is part of a larger effort to develop tools for evaluating mathematical and statistical software.

2.2 Industrial (scientific)

2.2.1 Glaxo-Wellcome

Glaxo-Wellcome is undertaking a company-wide initiative with the aim of producing procedures for controlling the use of spreadsheets as part of their measurements. The company has provided NPL with draft documents describing proposals for the development, validation, control and use of spreadsheets. These documents include proposals for discussion, a validation manual, and a standard operating procedure.

Relevant points to note are:

1. Spreadsheets are included under category 3 of the GAMP categorisation [9]. This indicates the level of validation that needs to be demonstrated.

2. The first stages of the development process include “the decision to develop” and “the decision to validate”, and require review.

3. There are mandatory requirements in the design that are required to make the security and control system work effectively. These include:

    a) locking of cells,

    b) password protection for individual sheets and the workbook[3],

    c) instructions for use within the workbook,

    d) standard test data to verify the correct function of the spreadsheet at the time of use,

    e) display of standard information on the operator display (input screen) such as spreadsheet name and version, application name and version for which the spreadsheet was validated, author, etc., and

    f) display of standard information on the report produced by the spreadsheet.

4. In addition, there are some recommended practices for achieving some degree of consistency of the user interface. These include:

    a) shaded areas for operator entry cells,

    b) meaningful titles for worksheets,

[3] A workbook is a collection of worksheets, chart sheets and modules. A chart sheet is a sheet in a workbook containing a chart. A module is a sheet in a workbook containing sets of instructions (or macros) for accomplishing specific tasks.

c) hiding rows/columns/sheets of calculations and data that do not need to be visible to the operator,

d) disabling gridlines, row and column headings, status bar, formula bar,

e) disabling ‘drag and drop’ to prevent accidental data movement, and

f) reducing menu options to the minimum required to operate the spreadsheet successfully.

5. A flow chart of the proposed validation process is given. The process defines the documentation that is required, the roles and responsibilities of those involved, and how version control is managed.

NPL has arranged to hold a meeting with Glaxo-Wellcome, at which there will be broad representation from the divisions of the company. NPL will present its activity on the use of data sets for testing and Glaxo-Wellcome will present its work in the areas indicated above, There will be detailed following discussion on the way ahead and in particular on further collaboration. Glaxo-Wellcome has good relationships with its competitors and through this NPL and the project will benefit from a broad and influential user base.

2.2.2 NAG Ltd

NPL is represented on the NAG Technical Policy Committee, which is a large body (some 20 members from industry and academia in the UK, Europe and US) which steers the company’s research and commercial interests. NAG Ltd is a world leader in the provision of numerical software, a considerable proportion of which is applicable or potentially applicable to metrology. NPL made a presentation of its data-sets testing activity at the June 1998 meeting of the NAG Technical Policy Committee. It is clear that NAG Ltd and the other main providers of numerical software carry out detailed regression testing of their software. This testing involves confirming that the software performs repeatably between versions and across hardware platforms. However, there is little if any impartial explicit testing of the numerical correctness per se of its library software. Nevertheless, NAG Ltd is very interested in this area, and would wish to be involved in future collaborative activity. In a subsequent discussion, a Framework V Esprit project with appropriate European partners was seen as a viable approach. NAG Ltd stated that the testing of interactive products is considerable harder than libraries of routines.

2.2.3 Calibration Laboratories

The feedback from calibration laboratories received through the following NMS Clubs,

1. Pressure Measurement Club (NPL),
2. RF and Microwave Measurement Forum (NPL),
3. Time and Frequency Club (NPL), and
4. Flow Measurement Forum (NEL),

indicates that many laboratories use spreadsheet packages to process their data. However, little or no checking of the in-built functions contained within these packages is undertaken, and only superficial testing of the workbooks developed may be performed.

2.2.4 Others

The following organisations were contacted: Elekta, GEC-Marconi, Praxis, SEI, HSE, Ford, IBM, Cambridge Consultants, Rolls-Royce, BAe-SEMA, Nuclear Electric, AEA-Technology, TUV, and Brown & Sharpe. Positive responses were received from Elekta, GEC-Marconi, Praxis, SEI and HSE, and these described the following spreadsheet applications:

1. In reliability analysis. The application meets ISO 9000, and a spreadsheet was used to be cost effective.

2. In the calculation of safety targets. The integrity of the spreadsheet is not assumed: the spreadsheet is used to ensure the calculation is consistent, followed by hand checking.
3. In the generation of safety critical data. The spreadsheet was scrapped (before any of the data was actually used) when it was discovered that there was a total absence of high integrity processes such as design, test evidence and configuration control.
4. In the calculation of the decay chain of isotopes in processed waste. The approach taken to auditing the application was to treat it in the same manner as any other software, i.e., to require documented requirements and design, verification and validation plans, etc. The approach seemed to have a reasonable degree of success, and it is believed that from the documents produced, and assuming the plans are followed and compatibility is maintained, the application should remain usable and has sufficient documentation to be maintainable. An important issue for the application was its future-proofing: there was a requirement for approximately 30 years of data retention.
5. In the mechanisation of risk-assessment calculations. The spreadsheet offers a convenient implementation tool requiring none of the compilation stages associated with a conventional programming language. A problem is that a good deal of engineering judgement goes into a non-trivial assessment, and it is a practically difficult matter to know how much numerical precision can be left in the spreadsheet without giving a false appearance of precision.

2.3 Industrial (business)

2.3.1 Coopers and Lybrand

In an examination [10] of blue-chip companies using large spreadsheets (more than 150 rows), Coopers and Lybrand found that more than 90% of the models contained at least one calculation error. In 21 of the 23 models reviewed, results were inaccurate by more than 5%. Coopers and Lybrand have developed a robust spreadsheet modelling methodology which it employs in the development of models for clients and which is encapsulated in a one-day training course for clients and staff.

2.3.2 Institute of Accountants

The Institute of Accountants has produced an IT Briefing [11]. The briefing concentrates on design aspects of spreadsheets rather than specification, testing, documentation or implementation. The briefing was prepared in conjunction with Microsoft, manufacturers of Microsoft Excel. The design principles discussed include:

1. use a vertical, modular design,
2. use a standard layout (title/instructions/content, assumptions, data, calculations, results),
3. document assumptions,
4. write clear formulas,
5. include controls and checks, and
6. make the model easy to use.

2.4 Academia

2.4.1 University of Greenwich

The Information Integrity Research Group, within the School of Computing and Mathematics, was set up at the University of Greenwich in 1996 to research common errors in business information systems. A current project is the prevention and detection of spreadsheet errors, and some articles have appeared in New Scientist (16 August 1997), Computer Weekly (17 July

1997), and the professional journals of the British Computer Society Computer Audit Group (1997 summer edition) and the Institute of Internal Auditors (May 1997 edition of Internal Auditing). The project reached the conclusion that a strategy was needed to address the following:

1. error awareness,
2. improved spreadsheet building methodologies/techniques,
3. improved administration/control of spreadsheet developments, and
4. improved methods of costing errors and audit time.

2.4.2 University of Hawaii

The web-site [12] contains information listing

1. studies that have attempted to measure spreadsheet errors,
2. studies surveying spreadsheet developers,
3. studies surveying spreadsheet developers to learn about their organisation's control policies,
4. related web-sites (principally the Information Integrity Research Group at the University of Greenwich: see Section 2.4.1),
5. general spreadsheet studies and summary papers,
6. reported incidents of spreadsheet errors in practice,
7. audits of real-world operational and pre-operational models,
8. reported errors found in spreadsheet development experiments,
9. spreadsheet-auditing experiments,
10. questionnaires and interview research,
11. cognitive, behavioural and numerical research,
12. prescriptive writings, and
13. the Year 2000 problem for spreadsheets.

Ray Panko, the author of this web-site, makes the following comments.

Firstly, the UK seems, in general, to be far ahead of the US in appreciating problems with spreadsheet programs.

Secondly, spreadsheet error rates look very similar to error rates in other human cognitive activities, including programming. To expect lower error rates than have been seen to date actually would be unreasonable. Error rates in some other human cognitive domains are summarized in the companion web-site [13].

Thirdly, there are several types of errors: mechanical, logical, and omission [14]. Ray Panko refers to these as being like three "poisons", each of which is separately "fatal", and all of which must be removed. Yet the ways that different types of errors are eliminated may be quite different. For instance, synchronous development is good at reducing simple logic errors and omission errors but is less effective at reducing mechanical errors than is expected and is poor at reducing a category of logic errors called "Cassandra" errors. In any discussion of errors and error reduction, it is important to distinguish among types of errors.

Fourthly, surveys of developer behaviour and of corporate policy indicate that current spreadsheet development practices resemble those of programming in the 1960s, before proper disciplines were developed. In particular, testing is rarely done well and is often not done at all.

Finally, probably the biggest problem with getting corporations to take spreadsheet errors seriously is a lack of real-world field audits using a good methodology. Ray Panko knows of only a single study based on a sample larger than 20 spreadsheets, and a single study that attempted to follow rules of conventional programming code inspection. He believes that there is a clear requirement to get a sample of about 100 real-world organizational spreadsheets of decent size and inspect them in a canonical way.

2.5 Other

A literature search identified a number of relevant papers including [15–19].

[15] is concerned with recommending a structured approach to the design of spreadsheet models. Formal approaches to the "Life Cycle" of a spreadsheet model are indicated, but rather than advocating that the development follows all suggestions for every development, it is shown to be important to determine methods specific to each design problem. Identification of which specific methods are necessary can be shown to be an important part of the development procedure. Other recommendations made include the need for a block-structured format for models, and the importance of flow diagrams in the design stage.

[16] gives an introduction to the validation of spreadsheet models that have already been developed. This should be seen as advice on a structured approach to the testing of given models, although the scope of this covers only validation of logical procedures and minimisation of the potential for data entry errors.

[17] describes a framework for determining the relative importance of errors in data input to a spreadsheet model, and in particular, how potential errors can affect output results.

The three papers [15, 16, 17] provide advice on structured procedures in the design and testing of spreadsheet models. In contrast, [18] and [19] address problems with the in-built functions of spreadsheets, and indicate a need for an established testing procedure if the outputs of these functions are to be used extensively.

[18] deals with the numerical accuracy of statistical functions in Microsoft Excel (Excel version in Microsoft Office 97 for Windows). It is shown that the computation of some discrete distributions (Binomial, Poisson, Hypergeometric) fails even for probabilities in the central range between 0.01 and 0.99 and for parameter values that cannot be judged as too extreme. Also, it is demonstrated that the computation of some continuous functions and their quantiles (such as Normal, Chi-square, F, and t) give unreliable results in the tails of the distribution. A scientist or statistician using Microsoft Excel intrinsic functions in scientific work might expect results given to a specified number of significant figures to be reliable, but this work shows that this is not necessarily the case. It is demonstrated in the paper that it is possible that the user could only, in some cases, rely on one significant figure of accuracy when the result is apparently given to ten, a factor that is liable to lead to extremely misleading results if the user is not fully aware of the problem. This is clearly a severe warning to all users of Excel built-in functions and demonstrates the need for a clear and thorough testing procedure. For the purposes of this paper, testing against reference software was employed using SAS [20], the results of which were checked against ELV [21].

[19] reports on properties of the random number generator in Lotus 1-2-3, and expresses concern for the overall adequacy of the generator.

3. Conclusions

Spreadsheets are used in a wide variety of application areas covering the scientific and business sectors and including metrology. In terms of recognising that there can be problems associated with developing spreadsheet applications and doing something about it, the business sector would appear to be ahead of the scientific sector. Most anecdotes, such as those reported at [12], refer to financial data and any guidelines that have been produced, such as the Briefing Note [11] issued by the Institute of Accountants, are aimed at financial applications. The notable exception to this is the activity being undertaken by Glaxo-Wellcome to adapt procedures for developing "traditional" software to apply to spreadsheets. Very little work seems to be reported anywhere concerning the numerical accuracy of the in-built functions provided with spreadsheets. This suggests that there is a requirement for work to be undertaken that is aimed at the scientific sector and which explicitly excludes financial applications.

It is accepted that spreadsheets are a useful and popular tool for manipulating and presenting data. This is due, at least in part, to their easy interface, the graphical facilities that they provide and the fact that results (whether correct or not) may be generated quickly. Consequently, the aim of future work must be to provide an infrastructure, comprising supporting information and guidelines, to ensure that their use is made as effective as possible.

Detailed consideration needs to be given to spreadsheet-model *design* and *validation* (*cf* Section 2.2.1). If the design is made apparent, then any documentation that is required to support the design can, in principle, be produced. However, the use of spreadsheets does not encourage that the design is made visible, and this is one source of the problems encountered when using such software packages.

Account must be taken of the specially structured or extreme nature of the inputs to spreadsheet in-built functions that arise in metrology. In areas such as finance, econometrics and the social sciences, the statistics of relatively disparate data is frequently required; in other words the data may be quite scattered, reflecting the inherent high variability of quantities of concern in these areas. Further, when such data is to be represented, for subsequent purposes, by an "explanatory" model, nothing more complicated than, say, a straight line or log-logistic can usually be justified. For this type of data, calculations such as standard deviation, geometric mean and simple linear or nonlinear regression can usually be carried out to more than adequate accuracy by the in-built functions, even if they are based on potentially unstable algorithms.

In contrast, metrology computations frequently require statistics of *specially structured data*, e.g., the arithmetic mean and standard deviation of a set of nearly identical numbers (such as arise in a series of replicate weighings). Another instance is the determination of a calibration curve for a measurement process; a function with many parameters may be needed to provide a representation to the accuracy of the measurement data. An instance would be a polynomial of moderately high degree. For such circumstances, unstable algorithms are likely to deliver unreliable results [5], and thus considerations of this type need to be taken into account in developing special methods of test for spreadsheets used in metrology.

We believe that the results of this survey support requirements for work in the following two main areas:

1. Information on the numerical accuracy of the in-built functions within spreadsheets and other packages used in metrology. This includes an indication of the parameter ranges over which these functions are accurate, and guidelines on how the functions may be used in an effective manner with, for example, pre-processing of the input data and post-processing of the results.

2. Guidelines to help users design spreadsheet applications in such a way that these applications can be validated and tested, controlled and maintained. This will build on [11], the work of Glaxo-Wellcome and other material, such as [22], and take account of the initiatives described in Sections 2.3.1 and 2.3.2.

In addition, we believe there is scope for activity in the following areas:

1. To support the activity of BIPM (Section 2.1.1), particularly with regard to a flexible spreadsheet model for the treatment of key comparison data that meets a consensus

requirement. NPL has already taken an initiative in this respect by proposing in July 1998 an SMT Accompanying Measures project, “Workshop on Uncertainty Aspects of Interlaboratory Comparison”. This proposed workshop has been discussed with the EUROMET Consultative Committees, BIPM, the NMSPU and others.

2. On the mechanisms by which spreadsheet users can have flexibility in the way they employ and present the results from spreadsheet models (*cf* Section 2.1.4). This aspect needs to be researched very carefully. One approach would be to divorce the calculations from the presentation. Calculations could be performed by tested add-ins implemented as “dynamically linked libraries” (DLLs) and those presentational aspects that are not critical to the integrity of the model could be made flexible for the convenience of users. We note that NAG Ltd are now providing DLL versions of their software routines to enable, with a modest amount of VBA programming, these routines to be called from an Excel worksheet.

4. Acknowledgements

We would like to thank all those who provided information for this report, including Terry Quinn (BIPM), David Blackburn (BIPM), Eric Lagergren (NIST), Lars Nielsen (DFM), Dieter Richter (PTB, Berlin), Ted Scorer (Glaxo-Wellcome), Russell Davies (Elekta), Viv Hamilton (GEC-Marconi), Andy Ashworth (Praxis), Ed Fergus (HSE), Ray Panko (University of Hawaii), and our colleagues at NPL.

5. References

[1] Guide to the Expression of Uncertainty in Measurement. International Organization for Standardization, 1995.

[2] EAL-R2, Expression of the Uncertainty of Measurement in Calibration. European Co-operation for Accreditation of Laboratories, Rotterdam, April 1997.

[3] S.L.R. Ellison, M.G. Cox, A.B. Forbes, B.P. Butler, S.A. Hannaby, P.M. Harris, and S.M. Hodson. Development of data sets for the validation of analytical instrumentation. *Journal of AOAC International*, **77**(3), pp 1-5, 1994.

[4] Statistics Software Qualification. Edited by B.P. Butler, M.G. Cox, S.L.R. Ellison and W.A. Hardcastle. The Royal Society of Chemistry, 1996.

[5] M.G. Cox and P.M. Harris. The design and use of reference data sets for testing scientific software. Presented at the International Conference of Optical Spectrometry: Applications and Instrumentation into the 21st Century, Royal Holloway College, UK, 28 June - 2 July, 1998. To appear in Analytica Chimica Acta.

[6] NIST Statistical Reference Datasets Web Service: http://www.nist.gov/itl/div898/strd/

[7] D.H. Bailey. Algorithm 719: Multiprecision translation and execution of Fortran programs. *ACM Trans. Math. Software*, **19**, pp 288-319, 1993.

[8] Tools for Evaluating Mathematical and Statistical Software: Special Functions Focus Area: http://www.nist.gov/mcsd/Staff/DLozier/projects/stssf/

[9] Supplier Guide for Validation of Automated Systems in Pharmaceutical Manufacture. Good Automated Manufacturing Practice (GAMP) Forum, May 1996.

[10] http://www.uk.coopers.com/financialservices/junglebriefing/spmodel.html

[11] Spreadsheet Design. IT Briefing Number 6, The Institute of Chartered Accountants in England and Wales, 1994.

[12] Spreadsheet Research: http://www.cba.hawaii.edu/panko/ssr/Welcome.htm

[13] Human Error Website: http://www.cba.hawaii.edu/panko/HumanErr

[14] Panko, R.R. and Halverson, R.P., Jr. Spreadsheets on Trial: A Survey of Research on Spreadsheet Risks. *Proceedings of the 29th Hawaii International Conference on System Sciences, Vol. II*, Kihei, Hawaii, Los Alamitos, CA: IEEE Computer Society Press, pp 326-335, January 1996.

[15] Ronen, B., Palley, M.A. and Lucas, H.C., Jr. Spreadsheet Analysis and Design. *Communications of the ACM* **32**(1), pp 84-93, January 1989.

[16] Borthwick, A.F., Validating Spreadsheets: Minimizing the Potential for Errors, EDPACS, **17**(3), pp 1-8, 1989.

[17] Ballou, D.P., Pazer, H.L., Belardo, S., and Klein, B., Implication of Data Quality for Spreadsheet Analysis, DATA BASE, **18**(3), pp 13-19, 1987.

[18] Knusel, L. On the accuracy of statistical distributions in Microsoft Excel 97. *Computational Statistics and Data Analysis*, **26**(3), pp 375-7, 1998.

[19] Modianos, D.T. and Scott, R.C. Deduction and test of the parameters of the Lotus 1-2-3 random number generator. *Collegiate Microcomputer*, **10**(3), pp 155-9, 1992.

[20] SAS for Windows. Release 6.12, SAS Institute Inc., Cary, NC, USA, 1996.

[21] Knuse, L. Computergestutzte Berechnung statistischer Verteilungen. Oldenberg, Munchen-Wien, 1989.

[22] Wichmann, B.A. Measurement Good Practice Guide No 5: Software in Scientific Instruments. National Physical Laboratory, May 1997.