Mullen, Carrie (2014) *Quality Assurance of forensic investigations in toxicology and traffic safety.* PhD thesis.

http://theses.gla.ac.uk/5465/

Copyright and moral rights for this work are retained by the author

A copy can be downloaded for personal non-commercial research or study, without prior permission or charge

This work cannot be reproduced or quoted extensively from without first obtaining permission in writing from the author

The content must not be changed in any way or sold commercially in any format or medium without the formal permission of the author

When referring to this work, full bibliographic details including the author, title, awarding institution and date of the thesis must be given

Enlighten:Theses
http://theses.gla.ac.uk/
theses@gla.ac.uk

University of Glasgow

Quality Assurance of forensic investigations in toxicology and traffic safety

Carrie Mullen

BSc (Hons), MSc

Thesis submitted in part fulfilment of the requirements of the University of Glasgow for the degree of Doctor of Philosophy

College of Medical, Veterinary and Life Sciences

Department of Forensic Medicine and Science

University of Glasgow

June 2014

Summary

The work described in this thesis deals with three aspects of quality assurance in the field of forensic toxicology: proficiency testing schemes, validation of analytical methods for the piperazine group of abused drugs and validation of the police field impairment test, used at the roadside to test drivers for drug-induced impairment.

Proficiency Testing

Long term reviews were performed for two forensic external quality assurance schemes. Rounds 30 (in 2007) to 48 (in 2012) of the UKAS-accredited commercial Quartz Forensic Blood Toxicology Proficiency Testing Scheme (PTS), and a ten year period from 1999 to 2009 of the freely-available United Nations Office on Drugs and Crime (UNODC) International Collaborative Exercises (ICE). Only limited ICE data could be made available as much of the original data had been stored on a database which had become obsolete, hence the data were only available as the original results forms provided to UNODC by the ICE participants. Data was entered to Microsoft Excel® spreadsheets and Microsoft Access® databases from the original forms for the years 1999, 2001 (2 rounds), 2003 (2 rounds) and 2005 (2 rounds), and summary data was extracted from the UNODC round reports for the years 2007, 2008 and 2009.

Four methods of scoring quantitative performance were reviewed and the most suitable, a z-score using an assigned 'true' value and a percentage of the true value as acceptable deviation, was applied to reanalyse the participants' results and assess their performance. Methods of scoring proficiency which relied upon participants' data to determine acceptable variation were found merely to describe the data rather than challenge participants on whether or not they were performing fit-for-purpose analyses.

Factors such as participation, analytes tested, participants' methods of analysis and participants performance were summarised for each scheme before the performance of the two schemes, and that of their participants, were compared.

ICE tested more analytes per annum but from a smaller test menu than Quartz. This resulted in more repetitive testing and allowed for some trend analysis and performance monitoring. It was not possible to observe performance trends with Quartz due to the wide variety of analytes tested. The smaller array of potential analytes and more repetitive nature of ICE testing also meant that performance monitoring and detection of bias were easier to perform, and ICE was shown to be more effective as external quality assurance (EQA). Quartz provided a good educational resource as it incorporated the wide range of drugs which a forensic toxicology laboratory could realistically encounter. Following the review, however, it was recommended for QUARTZ that, to provide a safeguard against bias, more repetitive testing was required and this has now been adopted.

Piperazines

All piperazine analogues are now illegal in the UK, registered as Class C of the Misuse of Drugs Act (1971) and schedule 2, part III of the Misuse of Drugs Regulations (2001). Piperazines can elicit similar effects to some ATS and methods for their detection should be available in forensic toxicology laboratories.

In the present study, methods were developed for the detection of a range of piperazines in blood using LC-MS/MS (p-MeOPP, p-FPP, BZP, o-MeOPP, p-MPP and TFMPP) and GC-MS (p-FPP, BZP, TFMPP, p-MPP, o-MeOPP, m-CPP, p-MeOPP and p-CPP). Quality assurance required both methods to be validated. For all piperazine analytes accuracy was within ±15% (20% at low concentrations) and precision was within 15% (20% at low concentrations). For both methods LLOD of all analytes was 5 ng/ml of blood and upper limit of quantification was 2 µg/ml of blood. For the GC-MS method lower limits of quantification (LLOQs) were in the range 20 to 30 ng/ml of blood. For LC-MS/MS, LLOQs ranged from 50 to 60 ng/ml of blood, although quantification by the LC-MS/MS method was restricted by the lack of availability of appropriate internal standards. There were no apparent significant matrix effects and recovery by both methods was >60 % and, therefore, acceptable.

Short term stability of the piperazine analytes was investigated. Piperazines remain sufficiently stable when stored in the fridge for at least one week, and are stable through three freeze-thaw cycles. There was no detectable degradation when blood samples were left on the bench-top or when extracted 'in-process' samples were left in the autosampler for up to 72 hours.

The LC-MS/MS method could provide a readily applicable screening method. A small aliquot of a basic drug extract could be screened by LC-MS/MS for the presence of piperazines, leaving the majority of the extract for other analyses, for example, piperazines confirmation or amphetamines analysis. The GC-MS method was suitably validated to provide quantification but application to casework samples remains to be evaluated. It is recommended that piperazine testing be performed for all suspected MDMA or 'club drug' intoxication cases.

The Field Impairment Test

The detection of drugged drivers primarily depends on the current method which is the driver field impairment test (FIT). FIT comprises measurement of pupil diameter and four physical tasks (the Romberg balance test, walk and turn test, one legged stand and finger to nose test) intended to simultaneously test comprehension, short term memory, balance and motor function. Despite FIT having ISO accreditation, it has been recognised that police officers lack confidence with the protocol and do not apply the test as often as is necessary. The main difficulty arises from the requirement to make a subjective judgement of impairment and officers lack confidence in their ability to do so.

FIT has never been fully validated. The present study was designed to meet the urgent requirement to develop FIT into an objective measurement, by determining what constitutes "normal" performance in FIT by unimpaired adults of different ages. FIT performance was recorded for 79 individuals, a statistically determined cohort size, confirmed by breath and oral fluid analysis not to be under the influence of impairing substances. Each error made during FIT, as defined by the FIT standard operating procedure, was recorded and collated in a Microsoft Excel® spreadsheet for analysis. It was found that the definition of 'errors' was too stringent as many which are required to be recorded are normal physiological or behavioural characteristics, such as body

sway, and most subjects would be unable to complete the task without displaying them. A less stringent, evidence-based definition of “error” was developed which allowed statistically more significant analysis to be performed on the FIT results.

A statistically significant difference (P=0.00578) was shown to exist between the FIT performance of individuals under the age of forty years and those aged forty and over. Based on the principles of a PTS, robust mean and standard deviation were used to determine what constituted acceptable performance. Those in the younger age group could be considered impaired if the police officer witnessed more than seven errors, or, in the older age group, more than fifteen errors. Using these criteria the frequency of false positives, i.e. unimpaired drivers being assessed as impaired is estimated to be (less than 3%). Also, the ranges of errors observed in both groups was large and overlapped, such that it may be possible for an impaired person to appear unimpaired. This requires further investigation.

Table of Contents

List of Tables ... xiii
List of Figures ... xv
Acknowledgements ... xvii
Author's Declaration ... xix
List of Abbreviations ... xx

Chapter 1: Introduction ... 1
1.1 Introduction to Forensic Toxicology ... 1
1.2 Systematic Toxicological Analysis (STA) ... 2
1.2.1 Gas Chromatography – Mass Spectrometry (GC-MS) ... 3
1.2.2 Liquid Chromatography Tandem Mass Spectrometry (LC-MS/MS) ... 4
1.2.2.1 Electrospray Ionisation (ESI) ... 6
1.2.2.2 Triple Quadrupole Mass Analysers (QqQ) ... 6
1.2.3 Identification Criteria ... 7
1.2.4 Reporting Results ... 8
1.3 Quality ... 8
1.4 Fitness for Purpose ... 9
1.5 Accreditation ... 10
1.6 Quality Management System (QMS) ... 12
1.7 Validation ... 13
1.8 Internal Quality Control ... 17
1.9 External Quality Assessment (EQA) ... 18
1.10 Research Background ... 20

Chapter 2: Proficiency Testing ... 21
2.1 Introduction ... 21
2.2 Performance Scoring in a PTS ... 22
2.2.1 Determining the true value ... 23
2.2.2 Defining Acceptable Variation (σ Value) ... 24
2.2.2.1 Prescribed Value ... 24
2.2.2.2 Statistical Model ... 25
2.2.2.3 Precision Experiment ... 25
2.2.2.4 Participants' results ... 25
2.2.2.5 Perception ... 27
2.3 Performance Appraisal ... 28
2.4 Availability of Proficiency Tests ... 29
2.4.1 United States ... 30

2.4.1.1 College of American Pathologists 30
2.4.1.2 RTI International (North Carolina, US) 30
2.4.1.3 CTS – Collaborative Testing Services, Inc. US 31
2.4.2 Asia and Pacific 31
2.4.3 Europe 31
2.4.3.1 QUARTZ – LGC Standards Proficiency Testing UK 32
2.4.3.2 UKNEQAS 34
2.4.3.2.1 Drugs of abuse in urine 34
2.4.3.2.2 Quantitative blood scheme 34
2.4.3.2.3 Toxicology cases 34
2.4.4 Oral Fluid PT Schemes 35
2.4.4.1 UKNEQAS Oral Fluid Scheme 35
2.4.4.2 ORALVEQ 35
2.4.5 HAIR PT Schemes 36
2.4.5.1 Society of Hair Testing PT 36
2.4.5.2 HAIRVEQ 36
2.4.6 Specialist Forensic Toxicology Testing 37
2.4.6.1 Federal Aviation Association (FAA) PT, USA 37
2.5 Alternative EQA available to all: UNODC ICE 37
2.6 PTS Quality Assurance: Long Term Review 38

Chapter 3: United Nations Office on Drugs and Crime (UNODC) International Collaborative Exercise (ICE), within the International Quality Assurance Programme (IQAP) 40
3.1 Introduction 40
3.2 International Quality Assurance Programme (IQAP) 41
3.3 International Collaborative Exercises (ICE) 42
3.3.1 Test Materials 44
3.3.2 Import/Export Certificates 44
3.3.1 Returning Results 45
3.3.2 Reporting Performance 46
3.4 Aim 47
3.5 Long Term Review 47
3.5.1 Summary of Available Results 47
3.5.2 General protocol 48
3.5.3 Participation 49
3.5.4 Qualitative Summary 51
3.5.4.1 Biological Specimens 51
3.5.4.2 Seized Materials 54

3.5.5 Methods 57
3.5.5.1 Biological Specimens 57
3.5.5.2 Seized Materials........ 58
3.5.6 Quantitative Summary 60
3.5.6.1 Biological Specimens 66
3.5.6.2 Seized Materials........ 66
3.5.7 Variation in Results........ 67
3.6 CONCLUSIONS 74
3.6.1 Qualitative Summary........ 74
3.6.2 Methods 74
3.6.3 Quantitative Summary 74
3.6.4 ICE Review Summary 75
3.6.5 VARIATION 75
3.7 Future Work 76

Chapter 4: LGC Quartz Forensic Blood Toxicology Proficiency Testing Scheme; Long Term Review........ 78
4.1 Introduction 78
4.1.1 Scheme Details........ 78
4.1.2 Returning Results 80
4.1.3 Calculating Proficiency 81
4.1.3.1 True Value 81
4.1.3.2 Acceptable Variation 82
4.1.3.3 Scoring Samples 2 and 3 82
4.1.4 Efficacy Measures........ 83
4.1.5 Review 83
4.2 Aims of the study 84
4.3 Methods 84
4.4 Results 88
4.4.1 Participation 88
4.4.2 Analyte frequency 89
4.4.3 Qualitative Sample 1 90
4.4.4 Quantitative Sample 2 95
4.4.5 Participant Methods 98
4.4.5.1 Qualitative Sample 1 98
4.4.5.2 Quantitative Sample 2 99
4.4.6 Standard Solution Sample 3 100
4.4.7 Alcohol Sample 4 102
4.5 Summary 102

4.6 Discussion 104
4.6.1 Scheme Organisation 104
4.6.2 Proficiency with Benzylpiperazine 107
4.6.3 Performance Scoring 108
4.6.4 General Performance 110
4.7 Conclusion 112

Chapter 5: Comparison of ICE and QUARTZ EQA Schemes 114
5.1 Review 114
5.2 Conclusion 119

Chapter 6: Validation of a Method to Detect Piperazines in Blood 122
6.1 Piperazines 122
6.2 Current Piperazine Trends 125
6.2.1 Legal status 126
6.2.2 Relationship to Amphetamines 127
6.2.3 Toxicity 128
6.2.3.1 Acute 128
6.2.3.2 Risk of overdose 130
6.2.3.3 Chronic effects 130
6.2.4 Metabolism 131
6.3 Aim 132
6.3.1 Relevant piperazines 132
6.4 Experimental Section 133
6.4.1 Introduction 133
6.4.2 Reagents and Standards 134
6.4.3 Preparation of Blank Blood 134
6.4.4 Preparation of Working Standards and Quality Controls 135
6.4.5 Sample Preparation 136
6.4.6 Optimisation of GC-MS 136
6.4.6.1 Derivatisation 136
6.4.6.2 Internal Standard 137
6.4.6.3 Injection Conditions 139
6.4.6.4 Oven Temperature 139
6.4.7 GC-MS Method 140
6.4.8 LC-MS/MS Method Development 141
6.4.9 LC-MS/MS Method 142
6.4.10 Validation 143
6.5 Results and Discussion 145

6.5.1 GC-MS Validation 145
6.5.1.1 Specificity 145
6.5.1.2 Calibration Response Function....146
6.5.1.3 Accuracy 148
6.5.1.4 Limits of Detection and Quantification....148
6.5.2 LC-MS/MS Validation 149
6.5.2.1 Specificity 149
6.5.2.2 Calibration Response Function....150
6.5.2.3 Accuracy 152
6.5.2.4 Limits of Detection/Quantification 153
6.5.3 Method Performance 154
6.5.3.1 Extraction Recovery (ER) 155
6.5.3.2 Apparent Matrix Effects (ME) 155
6.5.3.3 Process Efficiency (PE)....156
6.5.4 Stability 157
6.5.5 Concentration Range 158
6.5.6 Specificity 162
6.6 Discussion of Further Work....163
6.7 Conclusion 165

Chapter 7: Further Validation of The (Preliminary) Field Impairment Test 166
7.1 Introduction 166
7.1.1 Driving Under the Influence of Drugs (DUID) 166
7.1.1.1 Drugs Detected....167
7.1.1.2 Prevalence of Drug Driving in the UK 168
7.1.2 Portable Detection Devices 172
7.1.3 Law Related to Driving Under the Influence of Drugs (DUID) 173
7.1.3.1 Impairment 174
7.1.3.2 Zero Tolerance 174
7.1.3.3 UK Legislation 175
7.1.4 The Field Impairment Test 178
7.1.4.1 History....179
7.1.4.2 Conducting the Field Impairment Test 183
7.1.4.3 Pupillary Examination....184
7.1.4.4 The Modified Romberg Balance Test (RBT) 184
7.1.4.5 Walk and Turn (WAT) 185
7.1.4.6 One Legged Stand (OLS) 186
7.1.4.7 Finger to Nose (FTN) 186
7.1.4.8 Other Observations....187

7.1.5 Interpreting FIT performance 187
7.1.5.1 Officer's Competence 189
7.1.6 FIT Accreditation 193
7.2 Aim of the Present Study 194
7.3 Experimental 194
7.3.1 Participant Recruitment 194
7.3.2 Oral Fluid Samples 196
7.3.2.1 Enzyme-Linked Immunosorbent Assay (ELISA) 196
7.3.2.2 GC-MS Basic Drug Screen 198
7.3.3 FIT Observations 198
7.3.4 Data Analysis 199
7.3.4.1 Mann-Whitney U-test 200
7.3.4.2 Kruskal-Wallis H-test 200
7.3.4.3 One Way Analysis of Variance (ANOVA) 202
7.4 Results and Discussion 202
7.4.1 Toxicology Screening 202
7.4.2 FIT Observations 206
7.4.3 Pupillary Examination 208
7.4.4 Modified Romberg Test 208
7.4.5 Walk and Turn Test 210
7.4.6 One Leg Stand 211
7.4.7 Finger to Nose Test 212
7.4.8 Complete FIT Performance 212
7.4.9 Influence of Age 213
7.4.10 Omission of Common Errors 216
7.4.11 Updated performance with conservative assessment 218
7.4.12 Scoring System 224
7.4.13 Effectiveness of FIT 229
7.5 Future Work 230
7.6 Conclusion 231

Chapter 8: Conclusion 232

Appendix 1: Quartz False Positives in Identification Samples 236
Appendix 2: Quartz Participant Identification Methods 237
Appendix 3: Quartz Participant Quantification Methods 239
Appendix 4: GC-MS Fragmentation of HFBA Derivatised Piperazines 241
Appendix 5: LC-QqQ Fragmentation of Piperazines 243
Appendix 6: Participant Information 244

Appendix 7: Participant Questionnaire 246
Appendix 8: Scottish Field Impairment Test Form 249

List of References 254

List of Tables

Table 1-1: Validation exercises 15
Table 2-1: European Proficiency Testing Schemes 33
Table 3-1: Available data by year 48
Table 3-2: Consistency of laboratories' participation 50
Table 3-3: Overview of BS results 52
Table 3-4: Summary of correct identifications for substances included in biological samples testing 53
Table 3-5: Overview of SM results 55
Table 3-6: Summary of correct identifications for substances included in seized materials testing 56
Table 3-7: Analytical methods used for biological specimens. 58
Table 3-8: Analytical methods used for seized materials. 59
Table 3-9: Laboratories performing quantification by BS substance group 61
Table 3-10: Laboratories performing quantification by SM substance group 62
Table 3-11: Performance Statistics 65
Table 3-12: BS laboratories acceptable scores 69
Table 3-13: SM laboratories acceptable scores 70
Table 3-14: Laboratory performance scores by various calculation methods for BS Morphine 2001 round 1 71
Table 3-15: Laboratory performance scores by various calculation methods for BS Delt-9-THC-COOH 2003 round 2 71
Table 3-16: Laboratory performance scores by various calculation methods for BS Oxazepam 1999 round 1 72
Table 3-17: Laboratory performance scores by various calculation methods for SM Cocaine 2001 round 2 73
Table 4-1: Drugs included in LGC QUARTZ Forensic Blood Toxicology Scheme 80
Table 4-2: Datasets not considered in the long term review 85
Table 4-3: Round 35, Zopiclone results. 86
Table 4-4: Round 40, Thioridazine. 86
Table 4-5: Result sets considered for long term review. 87
Table 4-6: Participation from rounds 30 to 48 88
Table 4-7: Participation based on sample/analysis type 89
Table 4-8: Analyte Testing Frequency: (substance groups not tested: Anticholinergic, Bronchodilator, Opioid Antagonist and Carboxyhaemoglobin) 91
Table 4-9: Analytes with poor detection: those highlighted require attention 94
Table 4-10: Participant performance with quantification samples 97
Table 4-11: Performance in standard solution samples 101
Table 4-12: Results of alcohol in blood proficiency testing 102
Table 4-13: Performance with methadone and morphine 104
Table 6-1: Structures of popular piperazines 124
Table 6-2: Relevant piperazines for toxicological detection * [149] ** [210] 133
Table 6-3: Materials and suppliers 134
Table 6-4: Preparation of calibrators 135
Table 6-5: Ions for SIM method 140
Table 6-6: Piperazine analyte identification transitions for LC-MS/MS 141
Table 6-7: Preparation of stability standards 145
Table 6-8: GC-MS Accuracy at low, medium and high concentrations 148

Table 6-9: Limits of quantification for piperazines 149
Table 6-10: LC-MS/MS Accuracy at low, medium and high concentrations 153
Table 6-11: Statistically-calculated LLOD and LLOQ for LC-MS/MS method 154
Table 6-12: Recovery, matrix effects and process efficiency by GC-MS method. 154
Table 6-13: Recovery, matrix effects and process efficiency by LC-MS/MS method. 155
Table 6-14: Percentage deviation of peak height area ratios for fridge and freezer stored samples to freshly prepared and extracted samples. 157
Table 6-15: Percentage deviation of peak height area ratios for in-process autosampler and bench-top samples compared to refrigerated samples. 158
Table 6-16: Published forensic toxicology piperazine methods 159
Table 7-1: Road traffic accidents for which drugs or alcohol were a contributing factor. 170
Table 7-2: Alcohol breath tests and FITs conducted during police summer and Christmas campaigns 171
Table 7-3: Proposed limits and recommended *per se* limits for drugs in blood. 177
Table 7-4: Drugs of abuse screened with cut-off concentrations 196
Table 7-5: Preparation of ELISA calibrators 197
Table 7-6: Preparation of ELISA controls 197
Table 7-7: Participant screening results (part 1) 204
Table 7-8: Participant Screening Results (part 2) 205
Table 7-9: Error count for each FIT task 207
Table 7-10: Total errors observed in the modified Romberg balance task 208
Table 7-11: Total errors observed in the walk and turn task 210
Table 7-12: Total errors observed in the one legged stand task 211
Table 7-13: Total errors observed in the finger to nose task 212
Table 7-14: Number of errors for complete FIT by participant age 213
Table 7-15: Median error count and error range for by age groups 215
Table 7-16: Participant error counts by task with conservative assessment ...220
Table 7-17: Participant total error count by age group with conservative scoring 221
Table 7-18: Number of significant errors by age group: medians and dispersion.. 222
Table 7-19: Kruskal Wallis post hoc Tukey analysis data 224
Table 7-20: Acceptable numbers of errors for unimpaired persons 227

List of Figures

Figure 1 Frequency of testing of specific substances in biological specimens... 48
Figure 2: Frequency of testing of specific substances in seized materials..... 49
Figure 3: Number of laboratories participating in ICE for each round.......... 49
Figure 4: Number of laboratories participating in BS testing, SM testing and total number of participants. 50
Figure 5: Correct identifications of biological specimens from 1999 to 2005 . 52
Figure 6: Correct identifications of seized materials from 1999 to 2005....... 54
Figure 7: Percentage of participants each round that correctly identified an analyte........ 92
Figure 8: Percentage of acceptable z-scores for each sample 2 analyte in rounds 31 to 48........ 96
Figure 9: Efficacy values for each dataset by round. 96
Figure 10: Methods used for the analysis of identification samples (test sample 1) 98
Figure 11: Methods used by participants in each round of quantitative testing 99
Figure 12: Placement of internal standards in GC-MS method. BZP-D8 used for mMPP, BZP, pFPP and TFMPP, mCPP-D8 used for pMPP, oMeOPP, pCPP, pMeOPP and mCPP. 138
Figure 13: Placement of internal standards in LC-MS-MS method. BZP-D8 used for pMeOPP, pFPP, mMPP, oMeOPP and BZP; mCPP-D8 used for pMPP and TFMPP. 139
Figure 14: Calibration curve for p-FPP (mean R^2 0.9994) 146
Figure 15: Calibration curve for BZP (mean R^2 0.9993)........146
Figure 16: Calibration curve for m-MPP (mean R^2 0.9958)........146
Figure 17: Calibration curve for TFMPP (mean R^2 0.9991)........147
Figure 18: Calibration curve for p-MPP (mean R^2 0.9990) 147
Figure 19: Calibration curve for o-MeOPP (mean R^2 0.9986) 147
Figure 20: Calibration curve for m-CPP (mean R^2 0.9988) 147
Figure 21: Calibration curve for p-MeOPP (mean R^2 0.9989) 147
Figure 22: Calibration curve for p-CPP (mean R^2 0.9990)........148
Figure 23: Calibration curve for p-MeOPP (mean R^2 0.9962)........150
Figure 24: Calibration curve for m-MPP mean (R^2 0.9971)........151
Figure 25: Calibration curve for p-FPP (mean R^2 0.9970) 151
Figure 26: Calibration curve for BZP (mean R^2 0.9972) 151
Figure 27: Calibration curve for o-MeOPP (mean R^2 0.9972) 151
Figure 28: Calibration curve for p-MPP (mean R^2 0.9972)........152
Figure 29: Calibration curve for TFMPP (mean R^2 0.9976) 152
Figure 30: DRE 12-step procedure........180
Figure 31: Officer performing a pupillary examination........184
Figure 32: Officer conducting Romberg test 184
Figure 33: Officer demonstrating walk and turn test........185
Figure 34: One legged stand test 186
Figure 35: Finger to nose test........186
Figure 36: Participation by age and gender.........206
Figure 37: Dispersion of errors in the age groups over and under 40 years. ...214
Figure 38: Dispersion of significant errors across groups aged over and under 40 years 218

Figure 39: Comparison of the percentage of participants making at least one error by the original scoring system and with the more conservative scoring 218
Figure 40: Relationship between participation age and FIT error count........223
Figure 41: LC-MS/MS fragmentation of BZP 243
Figure 42: LC-MS/MS fragmentation of TFMPP.....................................243
Figure 43: LC-MS/MS fragmentation of MPP..243
Figure 44: LC-MS/MS fragmentation of MeOPP 243
Figure 45: LC-MS/MC fragmentation of CPP..243
Figure 46: LC-MS/MS fragmentation of *p*-FPP.......................................243

Acknowledgements

There are many people who have contributed to this thesis in one way or another, it is only possible to give particular mention to some, although I would like to extend my heartfelt thanks to all.

Firstly I must thank my generous sponsor Professor Allan Jamieson, Director of The Forensic Institute, Glasgow, for trusting me with the opportunity of such an enriching experience. I recognise the sacrifices of The Forensic Institute to fund my research and provide me the time and emotional support required. I am forever grateful.

Special thanks are offered to my wonderful supervisor Dr Robert A Anderson for his immense knowledge, support, time, and wit. The numerous people who have contributed to this thesis are a testament to your generous spirit; I am fortunate to have had the pleasure to develop as a scientist under your guidance. Your devotion to helping me is very much appreciated!

I have had the support and encouragement of my colleges at The Forensic Institute and I thank them for their patience and assistance in helping me to make time for my studies. I also thank my fellow students and the staff at Forensic Medicine and Science for creating a stimulating and friendly work environment. I am especially thankful to Denise McKeown for her time to provide LC-MS/MS tuition.

I would like to extend my gratitude to the UNODC Laboratory and Scientific Section for allowing me to access and review the results of their ICE programme; especially Paramita Doubek and Iphigenia Naidis for their warm hospitality in Vienna and for donating time to help recover data. I am also most appreciative of LGC Standards for allowing access to the results of their Forensic Toxicology Blood Proficiency testing scheme, and Wayne Gaunt for his kind assistance.

I am very grateful for the involvement of Police Scotland, in particular Sergeant Brian Poole; his generous donation of time and experience to conduct filed impairment testing and provide guidance on these procedures was invaluable.

As was the kind contribution of all those who volunteered to perform a field impairment tests.

Finally I wish to offer my deepest gratitude to my much-loved family and beloved fiancé. They were a constant solace and their sacrifices and offerings go far beyond possible description. Thank you!

Author's Declaration

I declare that, except where explicit reference is made to the contribution of others, that this thesis is the result of my own work and has not been submitted for any other degree at the University of Glasgow or any other institution.

Signature: ________________________________

Printed name: ________________________________

List of Abbreviations

A2LA	American Association for Laboratory Accreditation
AAFS	American Academy of Forensic Sciences
ACMD	Advisory Council on the misuse of Drugs
APCI	Atmospheric pressure chemical ionisation
APLAC	Asia Pacific Laboratory Accreditation Cooperation
AQA	analytical quality assurance
ATS	Amphetamine type stimulants
BS	Biological specimens
BZP	benzylpiperazine
CI	Chemical Ionisation
CNS	Central nervous system
CPP	Chlorophenylpiperazine
CV%	coefficient of variation
DBZP	Dibenzylpiperazine
DECP	drug recognition and classification programme
DIE	Drug Influence Evaluation
DRE	Drug recognition expert
DRT	drug recognition training
DUI	Driving under the influence
DUID	Driving under the influence of drugs
EI	Electron ionisation
ELISA	Enzyme-Linked Immunosorbent Assay
EMa	First efficacy measure (correct identifications of total attempted identifications)
EMb	Second efficacy measure (closeness of participant results)
EMCDDA	European Monitoring Centre for Drugs and Drug Addiction
ENSFI	European Network of Forensic Science Institutes
EQA	External quality assessment
ERR	Standard error
ESI	Electrospray ionisation
EWS	Early warning system
FAA	Federal Aviation Association
FEWS	Forensic Early Warning System

FIT	Field Impairment Test
FME	Forensic medical examiner
FMS	Forensic Medicine and Science, University of Glasgow
FPIA	Fluorescence Polarization Immunoassay
FPP	Fluorophenylpiperazine
FSS	Forensic Science Service
FTIR	Fourier transform infrared spectroscopy
FTN	Finger to nose
GC-ECD	Gas Chromatography with Electron Capture Detector
GC-FID	Gas Chromatography with Flame Ionisation Detector
GC-MS	Gas chromatography with mass spectrometry
GC-NPD	Gas chromatography with nitrogen phosphorous detector
HFBA	Heptafluorobutyric anhydride
HGN	horizontal gaze nystagmus
HPLC	High performance liquid chromatography
HPLC UV-Vis	High performance liquid chromatography with ultra violet and visible light spectrophotometry
HPLC-DAD	High performance liquid chromatography with diode array detection
ICE	International Collaborative Exercise
IEC	International Electrotechnical Commission
ILAC	International Laboratory Accreditation Cooperation
IQAP	International Quality Assurance Programme
IQC	Internal Quality Control
I.S.	Internal Standard
ISO	International Standards Organisation
LC-MS/MS	Liquid chromatography with mass spectrometry
LLOQ	Lower limit of quantification
LOD	limit of detection
LSD	Lysergic acid diethylamide
MAD	Median of absolute deviations
MADe	Robust standard deviation
MDMA	methylenedioxymethylamphetamine commonly known as ecstasy
ME	Matrix effects
MeOPP	methoxyphenylpiperazine

MPP	methylphenylpiperazine
MRM	Multiple reaction monitoring
MS-MS	Tandem mass spectrometry
MU	Measurement uncertainty
NPS	New psychoactive substances
NSAID	Non-steroidal anti-inflammatory drugs
OF	Oral Fluid
OLS	One legged stand
PBSS	Phosphate buffer solution
PE	Process Efficiency
PFPA	Pentafluoropropionic Anhydride
PORTAL	Proficiency Online Reporting and Trend Analysis
PT	Proficiency test
PTS	Proficiency testing scheme
PTV	programmed temperature vaporisation
QC	Quality control
QMS	Quality management system
QqQ	Triple quadrupole mass analyser
RBT	Romberg balance test
RSD	Relative Standard Deviation
RSDR	Relative standard deviation for repeatability
RTA	Road Traffic Act
RTI	Research Triangle Institute International
SDI	Standard deviation index
SDPA	Standard deviation for proficiency assessment (LGC Quartz scheme)
SFST	Standardised field sobriety test
SIM	Selected ion monitoring
SM	Seized materials
SOFT	Society of Forensic Toxicologists
SOP	Standard operating procedure
STA	Systematic toxicological Analysis
TFMPP	trifluoromethylphenylpiperazine
THC	Tetrahydrocannabinol
TLC	Thin layer chromatography

UK	United Kingdom
UKAS	The United Kingdom Accreditation Service
UKIAFT	UK and Ireland Association of Forensic Toxicologists
UKNEQAS	Network of United Kingdom National External Quality Assessment Schemes
UNODC	United Nations Office on Drugs and Crime
VAM	Valid analytical measurement
WAT	Walk and turn

Chapter 1: Introduction

1.1 Introduction to Forensic Toxicology

Forensic toxicology has changed dramatically over the last 20-30 years. Previously the forensic toxicologist dealt almost exclusively with post mortem samples in which substances were present in relatively high concentrations (fatal levels) and there was a limited number of substances of interest.[1] Today forensic toxicology laboratories also receive a large number of clinical samples from which a large number of 'relevant substances' can be detected at very low concentrations. These samples can take many forms, traditionally blood or urine, but less conventional matrices such as hair and oral fluid (OF) are also routinely encountered and research is ongoing with substances such as sweat, nail clippings and exhaled breath (for drugs other than alcohol).[2-4] Newer matrices may be less useful for quantitative work due to the lack of interpretation data. Literature relating to the detection of various drugs in hair and OF is expanding.[5,6]

OF is an ultra-filtrate of plasma and substances present in blood are transferred to OF across epithelial membranes by passive diffusion,[7] and some to a lesser extent by active transport.[8] Drug compounds present in blood can be detected in oral fluid without the necessity of invasive sampling. As a clean matrix OF analysis is relatively simple, although sensitive detection is required as analytes are present in very low concentration. Quantification of analytes in OF is complicated by many factors including method of collection and rate of saliva production. Large inter- and intra-individual variations in OF-to-plasma ratios mean that a reliable estimation of blood concentration from OF is usually not possible. There are a number of detailed reviews of OF as a forensic toxicology specimen, including Drummer,[9] Aps and Martens,[10] and the DRUID evaluation,[11]and a detailed account is not provided here.

OF can, however, indicate that a psychoactive substance was present in blood at the time of sampling and is an ideal sample for point-of-care testing, especially for the detection of driving under the influence of drugs (DUID).

1.2 Systematic Toxicological Analysis (STA)

Samples are generally submitted for forensic toxicology by pathologists, forensic medical examiners, police officers or from workplace drug testing schemes. These samples have most often been taken from individuals amid suspicion of drug use e.g. driving in an unsafe manner, pre-mortem behaviour, or post mortem observations. The information supplied upon submission, together with inter- and intra-laboratory statistics, can help determine which substances may be present and which analyses should be performed. In some instances there may only be one (or limited) analyte(s) of interest and testing can be directed towards this. Most often, however, specimens are required to be analysed for a wide range of substances to obtain maximum information.

No single analysis can yet scrutinise a specimen for all possible drugs. A series of analyses, or systematic toxicological analysis (STA), is necessary. As the presence of toxic substances is uncertain and their identities unknown, STA is designed to simultaneously detect numerous intoxicants (or confirm absence).[12-17] STA should be applied even when *a priori* information is available in order to screen for other 'unknown' toxic substances, particularly as poly-drug use is regularly encountered.[1,18]

After checking paperwork and ensuring integrity of the specimen sample, pre-treatment is usually required. This could simply be dilution or mixing of the sample but depending upon the matrix, analyte properties and analytical technique to be applied other preparation such as filtration, centrifugation, homogenisation or pH adjustment may be necessary.[19,20] To direct and narrow the number of detailed analyses required a relatively quick and inexpensive screening technique (usually one or more immunological tests) will usually be applied. Caution is required with interpreting screening results. Whilst these tests can be sensitive, generally they lack specificity and some analytes, particularly emerging designer drugs, may not be detected. Positive results

must be confirmed by more detailed instrumental analysis to provide identification and concentration.

To maximise discriminating power a combination of analytical systems is required. The probability of correct identification is increased with the number and range of techniques applied. Courts in the United Kingdom (UK) recognise this and require all evidential results to have been subjected to at least two analyses. Integrated techniques are, therefore, beneficial for satisfying the court's requirement for selectivity with minimal effort and confirmation is usually by gas chromatography-mass spectrometry (GC-MS) or liquid chromatography-tandem mass spectrometry (LC-MS/MS).[17]

1.2.1 Gas Chromatography – Mass Spectrometry (GC-MS)

In order to improve chromatographic properties (peak shape, resolution, retention time, etc.) or increase the sensitivity of detection, samples may require derivatisation prior to GC-MS. This is primarily aimed at decreasing polarity (making the analyte less reactive with the stationary phase) and hence increasing volatility and/or thermal stability of the analyte molecule.[21]

The prepared sample is introduced via an injection port, maintained at high temperature to cause volatilisation. In order to prevent sample overloading the volume of sample injected is in the order of approximately 1 µl (containing up to 50 ng of individual analytes), although a larger volume can be injected if a splitter valve within the injection port is programmed to vent a defined proportion.[22] Volatilised analytes enter the chromatographic column and are swept through by an inert carrier gas. They become separated according to their boiling point and partition coefficient in the column stationary phase: volatile and inert compounds with few polar groups elute first. The column oven temperature is fundamental in ensuring that adequate separation is attained. Temperature programmes allow the oven temperature to be steadily increased at a designated rate and provide the necessary separation and resolution in the shortest possible time. The column is connected to a mass spectrometer which records presence and abundance of eluting analytes as a chromatogram. Compounds are identified by their retention time, ion fragmentation patterns (mass spectra) and molecular weights.

There is a range of mass spectrometers available, all with the same three basic requirements: an ion source to create charged ions, a mass analyser to separate the ions by mass to charge ratio (m/z) and a detector to count the ions. It is preferable with GC-MS to have hard ionisation in which fragmentation is induced by bombardment with a stream of electrons (electron ionisation, EI). EI is generally favoured as it results in greater fragmentation for more reliable identification. Alternatively, soft ionisation involving a reagent gas (chemical ionisation, CI) can be used to enhance the molecular ion region of the mass spectrum.

The most common and versatile mass spectrometer uses the quadrupole mass analyser.[23] Ion filtration is by application of an oscillating quadrupole electric field along four precisely parallel rods, equally spaced around a central axis.[23] Precisely controlled alternating voltages are applied to opposite sets of rods such that only ions with a particular *m/z* will reach the detector. As the positive ions are introduced to the analyser they are directed at the negative rod. The electric potential is then changed and ions which have not yet collided with the rod (and depleted) are repelled towards another rod. This continues along the length of the analyser until the ions reach the detector. At any given time only ions of one particular *m/z* will be in resonance the full length of the analyser and reach the detector. The radio frequency (RF) of the voltage oscillation is continuously adjusted over a set range to bring ions of different *m/z* into focus on the detector at different times and build a mass spectrum. When in full scan mode the analyser performs rapid scans through a range of m/z values to allow the full range of ions present to be detected, producing a full mass spectrum and total ion chromatogram (TIC). Alternatively, selected ion monitoring (SIM) can be performed. In this mode the analyser performs repeated cycles through selected RF values. As more time is devoted to the ions of interest, whilst the others are ignored, a more sensitive detection is achieved.[19,24]

1.2.2 Liquid Chromatography Tandem Mass Spectrometry (LC-MS/MS)

LC-MS/MS can detect lower concentrations than GC-MS,[15,25] and is applicable to a broader range of analytes.[17,19,26,27] Analytes are not required to be derivatised which reduces uncertainty of the result as well as sample preparation time.[19,28]

A limiting factor of LC-MS/MS, however, is that, as fragmentation depends upon geometric configuration of the interface and chromatographic conditions, the spectra produced can vary widely;[27] even when operating parameters appear consistent.[17] LC-MS/MS spectra are, therefore, considered laboratory- and instrument-specific and cannot be easily shared (making searching of external libraries unreliable).[16,17,27]

LC-MS/MS operates on similar principles to GC-MS. Rather than a carrier gas, analytes are swept through the system in a liquid mobile phase of highest grade solvents and buffers. The mobile phase must be degassed to ensure dissolved gases are not present which interfere with the chromatography by producing bubbles. Mobile phase is delivered at a constant rate but not necessarily constant (isocratic) composition. Similar to controlling oven temperature in GC-MS, elution can be controlled by using more than one pump, or a mixing chamber, to establish a mobile phase gradient system. Properties of the mobile phase such as the organic solvent content, pH or ionic strength are changed.

The sample is introduced at the injection valve and carried by mobile phase through a guard column to the analytical column where it interacts with stationary phase, which can also influence elution. The column used will depend upon the properties of the analytes. The mobile phase should be complementary to the stationary phase such that the chosen interactions strongly dominate in the stationary phase and are minimised in the mobile phase. LC-MS/MS columns have smaller internal diameter than HPLC-DAD columns to allow for a lower flow, and less mobile phase to be evaporated at the MS source.

Solvent evaporation (desolvation) and ionisation take place at the interface by Atmospheric Pressure Chemical Ionisation (APCI), Electrospray ionisation (ESI), Particle Beam ionisation (PB) or Thermospray ionisation (TS). The interface involves evaporation and pressure reduction as well as ionisation to transform dissolved analytes to gas phase ions.[24] ESI is most suited to polar compounds it has wide applicability and is the most frequently used system.[29,30]

1.2.2.1 Electrospray Ionisation (ESI)

Analytes in mobile phase are introduced to the interface through a heated hollow needle (nebuliser) with high electrical potential (typically from 2.4-5KV) with respect to the cone.[19] As eluent emerges from the nebuliser tip it forms a fine mist of highly charged droplets (positive or negatively charged depending upon the voltage applied). Droplet formation is aided by drying effects of a flow of sheath gas, usually nitrogen. Repelled from the charged needle tip the droplets migrate towards the ion sweep cone into the ion transfer capillary.[31] During this transition auxiliary gas can be used to encourage evaporation. The droplets become increasingly smaller and eventually the electrostatic repulsion of the ions within them exceeds their surface tension (Rayleigh limit), and a series of coulombic explosions occurs as the droplets produce smaller droplets which disintegrate further until charged ions are ejected. Ion formation can be influenced (enhanced or suppressed) by co-eluting substances. These are known as ‘matrix effects’ (although they may arise from matrix components, mobile phase additives, salts, ion-pairing agents, internal standard, metabolites, etc.).[32]

Once in the gas phase the charged analyte molecules pass through the ion transfer capillary to the mass spectrometer. ESI is a soft ionisation technique which predominantly yields only quasi-molecular ions. Fragmentation required for structural elucidation necessitates tandem mass spectrometry (MS/MS).

1.2.2.2 Triple Quadrupole Mass Analysers (QqQ)

The single quadrupole analyses externally fragmented ions but analytes are fragmented and detected within the triple quadrupole.[24] The first mass analyser, the first quadrupole, passes ions into the reaction zone, the second quadrupole. Here, precursor ions are exposed to collision gas resulting in fragmentation to form product ions. Product ions undergo second mass analysis in the third quadrupole. There are a number of operating modes depending upon whether ions are filtered or scanned at the first and third quadrupole. Non-scanning modes are most sensitive. Multiple reaction monitoring (MRM) where only selected precursor ions enter the reaction chamber, and only selected product ions from each precursor are detected provides sensitive

quantification which can be useful in forensic toxicology for the detection of low concentration analytes, particularly in alternative matrices.[33]

1.2.3 Identification Criteria

In order to have a reliable identification, analyte signals must be compared to a contemporaneously analysed reference substance.[34] In chromatographic-mass spectrometric techniques the factors which ought to be compared are relative retention time and presence and abundance of monitored ions.[16,17,35]

Acceptable relative retention is usually when the analyte value falls within ±2% of the reference value.[36-38] In order to share retention time data amongst different instruments and laboratories, however, GC retention times can be standardised by comparison to n-alkanes. With isothermal methods this is by conversion to Kováts index (KI).[35] Linear retention index (LRI or RT) in Equation 1-1 is required for programmed temperature analyses.[39] These are approximate values, as it is unlikely that they would be identically reproduced between laboratories.[39]

$$I^T = 100n(t^T_{Ri} - t^T_{Rz} / t^T_{r(z+1)} - t^T_{Rz}) + 100z$$

Equation 1-1: Linear Retention Index
n = difference in carbon atom number of the two n-alkanes that bracket an analyte's retention time, t^T_{Ri} = retention time of the analyte, t^T_{Rz} = retention time of the n-alkane eluting immediately before the analyte, and $t^T_{r(z+1)}$ = retention time of the n-alkane eluting immediately after the analyte.

Identification from full scan mass spectra is traditionally considered to be best, although SIM is acceptable. SIM should include the molecular ion where possible.[34] Other ions selected should originate from different parts of the molecule and be sufficiently characteristic of the compound structure.[34,40] Generally with GC-MS there should be three diagnostic ions giving two ratios within ±20% of the reference value.[14,16,34,36,37,41-43] The use of two MRM transitions in LC-MS/MS is considered acceptable where the abundance ratio is as expected.[16,44] LC-MS/MS is accepted to be less reproducible than GC-MS and ratios should agree to ±25-30%.[38]

1.2.4 Reporting Results

Upon identification/quantification the toxicologist needs to prepare a report which explains the analytical results within the context of the case i.e. which attempts to provide answers to the questions asked upon submission of the sample. All information gained should be reviewed and checked by the person(s) who will sign the report. Interpretation needs to be done with utmost care for a variety of reasons e.g. the data may only relate to one point in time, there may be concentration differences depending upon sampling site, results could be complicated by post-mortem redistribution or decomposition and more than one substance may be present.[1] The greatest difficulty in evaluating the results of forensic toxicological analysis arises from inter-individual variations. Adsorption, elimination and hence concentrations of drugs in various bodily specimens are affected by a range of factors, many of which are unpredictable. These include, but are not limited to, gender, race, age, illness, frequency and duration of substance use, tolerance, route of administration, presence of food, drug interactions and metabolic polymorphism (especially the polymorphic enzymes of the cytochrome P-450 system). The toxicologist can usually, therefore, only give a general indication of how a drug might affect a person, based upon the changes induced in the central nervous system (CNS) (or effects upon other body systems). Stimulants (e.g. amphetamine and cocaine) arouse the CNS to increase brain activity. Depressants (e.g. benzodiazepines and heroin) inhibit brain activity, and hallucinogens (e.g. ecstasy and cannabis) neither stimulate nor inhibit but cause alterations in perception and mood.[20]

Trends in abused drugs are witnessed to change as drugs move in and out of fashion according to changes in popular culture and social economics. New drugs are frequently being introduced. As new drugs are encountered new analytical methods are required for their detection. It is important that all new methods are rigorously tested to ensure quality requirements are upheld.

1.3 Quality

A number of recognised and respected bodies produce guidelines of best practice which laboratories can incorporate in the development of an effective quality system. Those directed at forensic toxicology laboratories include the

UNODC Recommended Guidelines for Quality Assurance and Good Laboratory Practice,[45] The Society of Forensic Toxicologists (SOFT) and American Academy of Forensic Sciences (AAFS) Forensic Toxicology Laboratory Guidelines,[38] European Laboratory Guidelines for Legally Defensible Workplace Drug Testing,[46] and the United Kingdom and Ireland Association of Forensic Toxicologists (UKIAFT) guidelines.[47] These guides are designed to help laboratories to deliver reliable results in accordance with 'best practice'. Best practice, however, will differ between laboratories as a result of available resources and the requirements of test results. UNODC recommends 'best practice' is interpreted as "*the best achievable practice within the laboratory and jurisdiction concerned*" provided these meet minimum requirements.[45] The UK Government Valid Analytical Measurement (VAM) Initiative, however, like many others, interprets best practice to be procedures sufficient to '*satisfy an agreed requirement*'; in other words 'fitness for purpose'.[48]

1.4 Fitness for Purpose

It is of utmost importance to consider why the data to be generated is required when judging whether the chosen method will be adequate. Use of the most sophisticated techniques and equipment does not guarantee quality, and may not be necessary to provide sufficient quality for the intended purpose. The laboratory must liaise with the customer to gain a thorough understanding of requirements (necessary standards, legislation, sample handling, timing, etc.), and agree how these will best be accomplished, focusing on cost-effectiveness.[49] Fitness for purpose, therefore, answers a specific aim by selecting a method which can consistently detect with acceptable measurement uncertainty, the most appropriate target analytes to meet relevant legislative and regulatory demands and satisfy the needs of the client.

Forensic toxicology investigations are conducted for a range of purposes, from monitoring (e.g. workplace, prisons or rehabilitation) to legal investigations (e.g. detection of drug-facilitated offences or cause of death). Results entering the legal system must be reliable beyond doubt and command higher quality and expense. To reassure clients and legal authorities that results can be trusted to be fit for purpose, a laboratory may wish to obtain formal accreditation.

Accreditation is considered to be the best means of demonstrating dedication to quality.

1.5 Accreditation

Accreditation is said to provide laboratories with increased confidence in their procedures by indicating that data produced is reliable for its intended use and comparable with data of other organisations.[48] Accreditation can only be awarded by an approved accrediting body. Most countries have a national body (e.g. United Kingdom Accreditation Service, UKAS), although those which do not are free to select one in any other country.[48] Accrediting bodies require to be approved by a higher body (e.g. the European Co-operation for Accreditation in Europe or the Association of Crime Lab Directors in the US which is specific to Forensic Science Laboratories).[48] This structure enables standardisation of the accreditation and assessment process (including conformance monitoring) to ensure correct application of the internationally agreed quality standards within the sector.[50] Confirmed compliance with such standards facilitates acceptance and reliable sharing of results and information between laboratories.[48] Accreditation is, therefore, often equated with confidence in laboratory output.

With such meaningful inference, achieving accreditation is a difficult process. A laboratory may, therefore, seek only to accredit a select few of its services and perhaps gradually include more.[48] This is possible as accreditation is awarded for a defined 'scope' which specifies particular combinations of analyte, matrix and test method.[49,51] The scope can be altered, activities added or removed, to reflect a laboratory's changing capabilities. There are also different standards to which a company can become accredited, each with varying levels of demand. A company can select one or more standards to adhere to depending upon the nature of its work and the level of quality they wish to demonstrate. In England and Wales it is a requirement of forensic toxicology laboratories providing services to the police authorities (and elsewhere an aim) to be accredited to ISO/IEC 17025 (general requirements for the competence of testing and calibration laboratories).[48,52] This standard is published by the International Organisation for Standardisation (ISO) and the International Electrochemical Commission (IEC). Within the UK, forensic science laboratories work towards the Forensic Science Regulator Standard.[53] Laboratories must

already be accredited to ISO17025 then adhere to some additional requirements specific to their area of analysis (including forensic toxicology). The European Network of Forensic Science Institutes (ENSFI) also requires members to be accredited to, or working towards ISO 17025.[52,54] Despite many ENSFI laboratories having quality management systems only 13 of the 53 members have any accreditation and only 9 of these had ISO 17025 in 2007.[52] This could reflect the expense and effort accreditation requires or it could be a lack of confidence in the merit of accreditation within a sector which does not have standard methods.

With many criteria to satisfy, the International Laboratory Accreditation Cooperation (ILAC) provides procedures to help laboratories embarking on accreditation.[55] In general terms the process is straightforward. The laboratory will define their accreditation scope and contact their national body to arrange a pre-assessment visit. A lead assessor will visit and make recommendations related to the overall quality management system within the laboratory, whilst one or more technical assessors will comment on the activities to be accredited.[48] The same assessors will return six to twelve months later to perform a detailed audit. Anything found lacking is termed a 'non-conformity'. Non-conformities are listed and the laboratory given a timescale in which they must be addressed. When corrected to the satisfaction of the assessors, accreditation is awarded. Regular announced assessments will continue to ensure the required standards are maintained, and re-assessment is required every four to five years. Within forensic toxicology, however, the assessor cannot monitor the application of standard methods (there are none), and is unqualified to judge the 'fitness-for-purpose' or acceptability of methods (although they may try). Conformity with the laboratory stipulated standard operating procedure is what will be confirmed.

There is increased confidence in analytical output, however, due to the requirements for Quality Assurance (QA), including internal quality control (IQC) activities, participation in external quality assessment (EQA) and the necessity to define measurement uncertainty (MU).[56] The degree of rigor applied to calculation of MU will depend upon the purpose of the testing,[56] although it is typically determined through internal quality control studies.[57-59] This is regarded as a simple indication of whether or not a protocol is fit for purpose, as

it describes the range of values within which the true value is asserted to lie.[49,59,60] It is compulsory for all QA procedures and systems to be meticulously documented within the various manuals of the laboratory Quality Management System (QMS).

It is possible for procedures to be accredited to non-technical standards, typically ISO 9001 (Quality Management Systems). Laboratories are required to demonstrate that management, operations, personnel, procedures, equipment, accommodation, security and health and safety procedures meet defined quality standards.[48,56] In support of this the laboratory must operate a stringent, documented QMS. As the ISO 9000 series standards do not address technical competence, accreditation to these standards does not provide confidence in analytic output, rather, dedication to quality.[48,52,56]

1.6 Quality Management System (QMS)

The QMS is a series of manuals which clearly stipulate criteria, methods and systems of laboratory operation. It is a means of safeguarding confidence in consistent production of fit-for-purpose results, and is therefore subject to periodic review.[56] It recognises that every action within the laboratory can impact upon quality and is therefore designed to encompass every aspect of laboratory activities where errors could be introduced.[38,51,61] Ferrara et al [62] provide a good summary of QMS, suggesting it operates on three levels: logistic (e.g. laboratory environment, reagents, equipment, personnel and chain-of custody), analytical (e.g. sampling, storage, reference materials and report format), and control (monitoring of laboratory performance and the Quality Assurance system). With technical standards the QMS must additionally ensure and document the validity, reliability, accuracy, precision, and performance characteristics of each test.[63] This requires that all analytical methods have suitably validated standard operating procedures (SOPs) [13] and mechanisms for monitoring of the overall performance of the laboratory by both internal and external proficiency testing programs.[62]

Analytical Quality Assurance (AQA) is the documented programme of activities necessary to verify that the system is operating within acceptable limits.[50,64] It

provides a configuration for correct implementation and documentation of a laboratory's Quality Control (QC) to ensure (and demonstrate) continuous production of data within defined reliability;[51] QC being the practical activities performed day-to-day to test the system and ensure the quality of individual batches/samples.[45,49-51,64] Such monitoring of the stability of the measurement process determines whether results are reliable enough to be released.[45,50,51]

It is important to recognise that although AQA ensures measurements are valid and fit-for-purpose it does not guarantee every result will be reliable. Good AQA will manage the frequency of these failures such that they are rare but they cannot be completely eliminated.[49] There are two main sources: mistakes (e.g. sample mix-ups), and random or systematic errors.

To establish AQA procedures for a method it must first be established what the acceptable performance criteria are for the intended purpose, and it must be ensured that the selected method can reproducibly attain them. Such method evaluation is achieved through validation studies. All ISO 17025 accredited procedures must be fully validated before application.[56]

1.7 Validation

Not every method based on sound theory will, in practice, produce results which satisfy acceptance criteria. In order to accurately identify an analyte, its response must be related to a single chemical species and in order to quantify an analyte its response must be consistent.[24] Method validation demonstrates suitability for an intended purpose and ability to provide the correct result; precision, limits and susceptibility to variation in parameters must be scrutinised.[50,55,65-68]

The specific end use of the protocol should be well defined in advance of validation. Experiments are designed to appraise only the aspects of analytical performance agreed to be relevant for provision of fitness for purpose. Results are documented to support attainment of necessary analytical performance for identification (or quantification) of analytes in the given matrix with suitable reliability and reproducibility, and to additionally demonstrate the exploration and characterisation of any parameters which could influence this. Validation

can, therefore, require a large series of analyses and be very time consuming and costly.

Careful planning can minimise the effort and cost required but it is imperative, particularly where high quality is demanded, that reliability of the validation is not compromised. Essential parameters for validation of quantitative analyses are largely agreed to be specificity, calibration model (e.g. linearity), lower limit of quantification (LLOQ), accuracy (bias), precision (repeatability, intermediate precision) and analyte stability, outlined in Table 1-1A and B. There are additional, optional parameters such as limit of detection (LOD), recovery and ruggedness. Validation cannot consist of spiked samples alone. [69] The analysis of real samples is necessary to demonstrate that the method works in practice, is selective, sufficiently sensitive, and has the required dynamic range to cover the concentrations encountered. Unfortunately this is not always possible and it is important to perform many additional replicates during method validation to test as many assumptions as possible.

Where the determined fitness-for-purpose requirements are not satisfied the analysis needs to be adapted until confirmation is achieved.[70] Pre-validation, or first phase validation, can help to avoid the requirement for adaptations following (or during) validation. It is focused on method optimisation, identifying and minimising sources of error.[71] Validation measures the remaining error [72] and must be performed for every method used, however, if this has previously been done but the method is adapted to be suitable for a different use (e.g. different matrix) only measurements sufficient to demonstrate that the factors changed do not bring the analytical results outside of the acceptable limits are required.[54,73]

Measurement	Description	Ref
Selectivity	Selectivity is the ability of an analytical method to differentiate the analyte in the presence of other components in the sample (including the IS). The generally accepted test for selectivity is to analyse six sources of blank matrix. As specificity is of primary importance, however, it is recommended that this be increased to 10-20 matrix sources. Dosing of subjects with the substance of interest is regarded as the best method of investigating selectivity. Often this is not possible and the favoured alternative is the analysis of real casework samples submitted for routine testing. Samples determined to contain other drugs (and their metabolites), together with negative ('blank') samples, are spiked with the analyte of interest and analysed to demonstrate there is no interfering response.	49,57,58, 68,72-76
Calibration (Response Function)	Calibration during validation is intended to verify a reproducible relationship between instrument response and analyte concentration. This enables calibration with fewer determinations (concentration and replications) in subsequent routine analyses. A calibration response function is determined for each analyte using blank matrix spiked with known concentrations. Where possible authenticated reference standards of known purity should be used to spike calibrators. Recommendations for the number of concentration levels vary, most are for a minimum of 6 to 8 concentrations (although 5 may be sufficient), a 'blank' (matrix) and 'zero' (matrix plus internal standard), with no fewer than two independent determinations of each. Duplicate analysis may be insufficient, however, statistical evaluation would be improved with more replicates (perhaps with fewer concentration levels to reduce the total number of analyses required); six replicates at six levels could be more appropriate. Outliers will be more reliably detected. No more than two outliers can be removed (and must leave at least five non-zero standards).	54,57,58, 60,67,68, 70,73-75, 78,79,84
Lower lever of Quantification (LLOQ) Level of Detection (LOD)	LOD is the lowest concentration of analyte which can be differentiated from background noise although not quantified as an exact value. LLOQ is the lowest concentration that can be quantified with acceptable precision (20% RSD) and accuracy (±20% RSD). LOD and LLOQ are often defined by signal to noise ratios of less than or equal to three and ten, respectively. Visual measurement of s/n ratio is not recommended because it is generally not possible to obtain a peak area for a blank sample (and some analytes may be present in the blank matrix). Rather, it is recommended that the equation below be applied to a calibration produced from ten independent samples with concentrations close to LOD/Q. Calculated concentrations need to be validated with 5 replicates within ±20% of nominal value and CV<20%. $LLOQ = k.SD_{low}/S$ SD_{low} = the standard deviation of the response (residual standard deviation of the regression line or the standard deviation of the y-intercepts), S = the slope of the calibration curve, and k = k factor (LOD 3.3 and LLOQ 10)	34,38,41, 50,57,58, 60,67,68, 70,80-83
Bias	Bias describes the closeness of mean test results to the true concentration. It should be determined at a low, medium and high concentration through 3 to 5 replicates at each, depending upon the variability expected. Bias is typically reported as the percent deviation from the accepted reference value which should be within ±15% (20% at LLOQ).	50,57,58, 67,68,73, 75,79,80

Table 1-1(A): Validation exercises…/Cont.

Measurement	Description	Ref
Precision	Precision is closeness (degree of scatter) of individual measures to each other, not a reference value. It relates the extent of random error components in the analytical technique; expressed as relative standard deviation (RSD), or increasingly as coefficient of variation. There are three levels of precision; repeatability, intermediate-precision and reproducibility. Repeatability is precision over a short period of time, under the same operating conditions. Between-batch variation requires measurement of at least three concentrations (low, medium and high) over a number of rounds of testing (variability should be < 15%, 20% near LLOQ). Intermediate precision describes variations under changing conditions (e.g. different operators or days). Reproducibility is precision achieved between different laboratories (only necessary for methods used across different laboratories.)	54,57,58, 68,79,80
Stability	Stability studies are required for various stages of sample analysis and storage. Investigated at low and high concentrations in *Bench-top/In-process stability:* Stability at laboratory ambient temperature over time required for sample-preparation (or longer) *Freeze/thaw stability:* Samples may be required to be removed from the freezer and re-analysed. It must be ensured that the analyte concentration is not significantly affected by freezing and thawing (over 3 freeze/thaw cycles). *Processed sample stability:* Stability in the auto-sampler is investigated for the maximum time specimens might be expected to be present (e.g 24 hours for overnight and 72 hours for a weekend). *Long-term storage:* Maximum expected storage time is the period for which the laboratory is required to retain samples.	57,73
Recovery	Recovery relates the efficiency of a method to detect all of the analyte present. It is the percentage of the detector response obtained from analytes extracted from spiked matrix compared to the response of a reference standard not prepared in matrix and, therefore, not subject to any sample pre-treatment. (Thus it cannot be determined where derivatisation is required.) It is investigated at low, medium and high concentration, but is expected to be an overrepresentation as there would be greater affinity between analyte and matrix within real samples. Where it is not possible to directly measure recovery, acceptable recovery is demonstrated in attaining good bias and precision. It is therefore often accepted that recovery is not required as part of a validation study. It is essential that the presence and magnitude of matrix effects are evaluated for LC-MS/MS methods. This can be determined during recovery studies.	50,54,57, 58,67,68, 73,74,79
Ruggedness (Robustness)	Ruggedness determinations establish susceptibility to small changes (e.g. mobile phase composition, reagent supplier or pH of a solution). The factors selected for study should be those that are most likely to be changed during routine application of the method. Changes should be made individually and the effects of each on bias and precision investigated. Any factors which are found to show a significant alteration in performance need to be investigated further. These factors will require the greatest care in performing the method routinely and it may be necessary to define a permitted operating interval. It is generally accepted that ruggedness need only be investigated if the method is intended to be transferred to another laboratory.	50,54,57, 60,68,80, 84,85

Table 1-1(B): Validation exercises continued

1.8 Internal Quality Control

Internal quality control (IQC) is an expansion of validation. It monitors method performance during routine processing, rather than the performance achieved through vigilant application at validation. Routine performance measures are compared with acceptance criteria to ensure fitness for purpose is maintained.[50,54,62] Simple IQC procedures are incorporated in every analytical batch. Minimum requirements are a negative control (to confirm signals produced are originating from the samples), and at least one positive control at LLOQ (to confirm correct working order).[38,62] It is good practice to have more quality control samples (those at high concentration should be followed by a blank to detect carry-over).[38,62] Where possible these should be prepared with certified reference standards.[62,86]

Results are monitored within batches as well as over time. Statistical criteria are applied to the IQC determinations in order to determine stable variations which can be expected within the system and acceptability is based on where a result lies relative to these.[12,50,62] For example, when the co-efficient of variation is selected this should generally remain below 15%.[38] A common method for evaluating precision and immediately identifying departures from statistical control is to plot the results of control measurements on a Shewhart control chart (variable of interest on the y-axis, run-number on the x-axis). The chart is marked with guidelines to denote the expected value for the control, the control limits within which the measured values should fall and the warning limits outside of which the analytical result must be rejected.

If results are found to consecutively plot outside of the warning limits this indicates a problem with the protocol and remedial action must be taken.[62] The QMS will stipulate corrective actions appropriate to the error identified.[56] Any problems experienced, remedial actions, servicing and maintenance are logged.[38]

IQC monitors consistency of results rather than quality of results.[48,50,57,62,68] This is particularly relevant for analyses which do not make use of certified reference standards. Satisfying IQC alone does not directly imply good quality, rather, it

demonstrates that the analytical protocol is stabilised.[48,62] A system is under control if no more than 5% of the measured values exceed the control limits.[50] Close agreement of results does not infer that this is a 'true' or accurate value, however, as biased data can be consistently produced. Systematic errors arising from permanent factors (e.g. wrong instrument set-up, source of reagents or reference material) will be consistent and will not be identified as no statistical variation will be expressed.[62,86,87] Thus, although IQC ensures day-to-day consistency of results external quality assurance (EQA) is an additional requirement as a less frequent but regular check of method accuracy.[48,49,62] Undetected errors lead to over- or under-estimations which can have serious consequences.

1.9 External Quality Assessment (EQA)

EQA encourages development of IQC. It is an independent measure of testing capability, intended to guard against bias.[62] EQA should ideally take the form of a proficiency test (PT) [88] as there are ISO guides available which aid standardisation of testing (ISO 17043:2010,[89] intended to replace ISO Guide 43: 1997 Parts one [90] and two [91]). Standardisation ensures adequate scrutiny of results and an interpretation and reporting format which allows ready comparison. The ISO guide to proficiency testing, however, offers a selection of methods and the testing body must extract the recommendations relevant to their measurement.[86] Thus the International Harmonised Protocol for The Proficiency Testing of Analytical Chemistry Laboratories, compiled in collaboration between ISO and IUPAC (revised 2006) [86] is typically the guide of choice.

A proficiency test is ideally designed and co-ordinated by an independent body.[50,62,92] In designing the PT organisers should strive to adopt the VAM principles of good practice for proficiency testing.[48] The aims of the scheme (e.g. to test that participants reach a specific required level of performance or to educate participants) should be established to ensure that the final protocol will be fit-for-purpose. This may also influence whether or not the testing is declared to the laboratory or performed 'blind'.

When samples received by the laboratory are known to them to be PTS samples, it can only be assumed that test materials are truly being processed as routine samples.[61] There is some suspicion, however, that not all laboratories are committed to the PT philosophy of an honest check of performance. Doubt arises from laboratories having good performance in declared trials yet failing blind trials indicating that declared test materials frequently receive a degree of special treatment (more determinations performed than usual or a more experienced scientist replacing the routine scientist).[62]

PTS providers UKNEQAS demonstrated disloyalty amongst participants to routine method application through monitoring of false positives (inclusion of drugs similar to the target in 'negative' samples e.g. ephedrine for an amphetamine test).[18] A high incidence of false positives was encountered. Participants noted this from feedback and the following year there was a significant reduction in false positives when the same similar analyte was included. The analyte was then omitted from subsequent test rounds. It was later re-tested and a high occurrence of false positives was reported once again. This demonstrated that rather than learning from the feedback and improving QC to prevent reporting of false positives laboratories were simply extra vigilant in the round immediately following their poor performance. Undeclared PT is, therefore, a more realistic indication of performance. Testers can be assured that the laboratory will not depart form routine protocol as the test material cannot be distinguished from routine samples.[61,63]

Where there is no suitable PTS available laboratories should organise inter-laboratory testing and submit samples to each other as a useful means of comparative performance evaluation [12,93] (or round-robin testing where the same sample is analysed by each laboratory in turn and the results compared [94,95]). The main function of EQA is error diagnosis but it additionally enables laboratories to evaluate their performance against their own targets, and the capabilities of their peers.[49,62,63,88] Where performance is unsatisfactory the laboratory should have a written procedure of action to be taken. As a requirement of accreditation ISO guide 13528 provides a checklist for error diagnosis.[96]

1.10 Research Background

The results of forensic toxicology analyses can have far-reaching consequences. They must, therefore, be legally defensible which demands analytical techniques (and reports generated) be demonstrably reliable, with verifiable sample integrity. Quality Management is, therefore, vitally important in forensic toxicology. In the absence of accreditation standards specific to forensic toxicology it is difficult to establish whether quality management is adequate within this sector. Additionally, as standardised analytical methods are not applied across laboratories there must be a means of verifying that methods are being suitably validated. This could perhaps be assessed through an evaluation of the performance of laboratories in PTSs. This would require sufficient availability of PTSs which effectively test routine forensic toxicology procedures. It is unknown whether laboratories have access to PTS which can effectively reflect their work-load to demonstrate fitness-for-purpose in an adequate range of their analyses. Are laboratories able to be discerning in their choice of PTS to usefully monitor and safeguard their results, or has participation become a function of accreditation? And does the accreditation of a PTS assist selection by endorsing its fitness-for-purpose?

Performance monitoring and quality control is more difficult for procedures which are not performed within the laboratory. Issues have been raised in relation to the fitness-for-purpose of one such test, the driver field impairment test for the detection of drugged drivers. Results from these tests can be presented as evidence in court, often as ‘scientific’ findings. Results must, therefore, be scientifically reliable. Tests based in science are standardised to ensure reproducibility; i.e. the same principles are being applied, and tested to the same extent during every application (by different operators). Interpretation of results must be part of this standard operating procedure. This is where the driver field impairment test is known to be lacking. It has not been fully validated and an objective means of interpreting results is required.

Chapter 2: Proficiency Testing

2.1 Introduction

Participants in PTS are expected to process test materials as they would an everyday sample using their laboratory routine methods. It is not expected for laboratories to be required to make special arrangements to participate, although if the laboratory does not regularly perform the required analysis, it is accepted that education rather than quality monitoring is the objective, as the laboratory can gain from staff training, maintenance of expertise, etc.[48] As the chemical properties of the target analyte are unknown the choice of detection methods to employ can be difficult. Regular PT is, therefore, recommended to ensure the credibility of the laboratory's systematic toxicological analysis (STA).[14]

Successful STA should identify, beyond a reasonable doubt, all substances of relevance regardless of their structure or chemical properties.[12] In order to uphold cost and time efficiency, sustain good sample recovery and to keep contamination potential low, the STA method should involve as few steps as possible to first establish if there is a potentially harmful substance present (screen) then confirm its presence, identity and if necessary provide quantification. It is usual for laboratory performance to be most dramatic in early rounds of a PTS and to improve with length of participation before reaching a plateau.[48]

Laboratories are supplied with individual portions of a homogeneous test material.[48,50,62] Homogeneity and stability over the test period must be ensured.[95,97] It is recommended that at least ten random samples are analysed, in duplicate and under repeatable conditions.[97] Some analytes deteriorate with time. In these instances the closing date for completion of analysis must cover only the period for which the analyte concentration is known to remain relatively stable.[38,85,86]

To monitor true performance, test materials which very closely resemble routine samples are essential,[18,48,62,98] especially where analytes are not available as certified reference standards.[48,62] IQC for these analytes is less reliable and external monitoring is of increased importance. Provision of 'routine' specimens is particularly difficult within forensic toxicology: most samples are spiked drug-free matrices. To more accurately imitate routine samples, matrices should be spiked with the drug of interest together with relative proportions of metabolites and other likely interferences such as other therapeutic or abused drugs.[62] These could still be unlike casework samples, however, as they will be of good quality compared to those which might normally be received (e.g. degradation of the analyte, condition or volume of matrix supplied, the presence of many interferences). This is particularly true of post mortem samples, the analysis and interpretation of which can be very complicated. Results of PT should, therefore, be regarded as a guideline and diminished performance may be experienced from real samples.

2.2 Performance Scoring in a PTS

A qualitative and, where possible, quantitative statistical evaluation is performed. Qualitative summary gives the proportion of correct identifications with some account of false positives. Quantitative summary gives the proportion of acceptable measurements: this requires each participant to be assigned a competency score which scales the difference between their measured value and the true value. This can be achieved by forming a z-score, q-score or standard deviation index (SDI). None of these have been found to be entirely suitable for competence scoring, however, z-score is the generally preferred method. The main advantage of z-score is that a suitably fit-for-purpose 'standard error' value can be determined in advance. By translating quantifications to a z-score the adequacy of laboratories' measurements are immediately apparent (irrespective of the nature of the test e.g. different analytes, units of measurement). This makes it much easier for participants to monitor their performance through time.[99]

Z-scoring rates performance by using the maximum variation within which data is deemed valid to define the limits of satisfactory measurements.[48,62,93] Z scores are calculated using Equation 2-1 where 'x' is the participant measured

value, 'x_a' is the true value and σ represents acceptable deviation. $|z| \leq 2$ is satisfactory; $2<|z|<3$ is questionable; and $|z| \geq 3$ is unacceptable, perhaps symptomatic of a non-random event for which corrective action is necessary.[48,87] This is based on the expectance that in a well-behaved system, in which data will be normally distributed around the mean, less than 5% of the data will be ±2 standard deviations from the mean and the chance of falling outside ±3 standard deviations is remote (0.3%).[48,100] This assumes PT data is well-behaved, with central tendency around the true value (which may not be true when different methods, with varying biases, are applied). In order that competence is not directed by the data, but fitness-for-purpose, the true value and acceptable variation are not necessarily equal to the mean and standard deviation of the data. These can be derived by a number of routes, each of which has its particular strengths and weaknesses.

$$z = (x - x_a)/\sigma$$

Equation 2-1: formula for calculating z-score

Participants should be advised of the scoring system and target of variance in advance.[48,86,95] It is essential that participants are provided details as to how these have been calculated, as the various methods of assigning them mean that the performance of a laboratory can be good in one PTS but unacceptable in another.[101]

2.2.1 Determining the true value

The use of the participants' mean as the true value is discouraged as a PTS should be strictly external and independent of the test laboratories. If large proportions of participants are operating biased methods, the 'true' value generated by their mean will be biased (unless there is an equal share of different methods and the biases are normally distributed around the true value).[48,62,94] Using the participants' mean monitors consistency across laboratories rather than accuracy, and impedes error detection.[48,62] If one or two participants operate with greater accuracy than the majority, this would not be recognised. As minorities, their measurements might be further from the mean and their performance score would not reflect their excellent performance.[48,62] If these laboratories then adapt their methods to improve

their score in the PTS, they could be introducing bias to their method.[94] Participant mean is, therefore, likely to be unsuitable for performance assessment, with a reference value being preferable.[97,102]

The reference value may be a consensus of measurements from specialist reference laboratories with a proven ability to carry out the measurements of interest to a high level of performance, using validated methods and reporting their uncertainty with their results.[48,62,94,97] Reference laboratories must be accepted as such by all PTS participants.[95] It can be difficult to gather a group of laboratories all agreed by the participants to perform measurements beyond doubt, and the expense of reference measurements may be too great for some PTSs.[94]

If there is no alternative but to use the participants' mean value, however, it must be ensured that they are sufficient in number that the uncertainty associated with their mean is acceptable (more than 20-30 participants;[92] below 30 the statistics applied become increasingly unreliable [102,103]). Also, it is recommended that results are separated into 'peer groups' according to measurement procedure (mean calculated and acceptable variation assigned for each group), due to different methods having various associated biases.[48]

2.2.2 Defining Acceptable Variation (σ Value)

There are many methods by which acceptable variation can be defined as due to the diverse nature of analytical laboratories, one universal method would not be appropriate. ISO surmises that these methods fall into five categories:[96]

1. Prescribed value
2. Statistical Model
3. Precision Experiment
4. Participants' results
5. Perception

2.2.2.1 Prescribed Value

In regulatory schemes the acceptable variation is defined by the regulating body as the level of performance required in order for a laboratory to be certified as competent, or permitted, to perform a particular analysis.

2.2.2.2 Statistical Model

Acceptable variation is derived from a general model for the reproducibility of the measurement method, usually the Horwitz Equation or Thompson's modification of the Horwitz function. This model is intended for use in inter-laboratory collaborative studies to test the performance of a particular method, although it is applicable to datasets which do not have normal distributions (which may be true of PT results). The Horwitz equation, Equation 2-2, predicts reproducibility relative standard deviation (RSD_R), which should be achieved with analyte concentration 'C' as a mass fraction. The RSD_R (as a percentage of the true value) or Thomson's σ value [104] can be used as acceptable standard deviation.

RSD_R [%] = 2(1-0.5log10 C)

Thompson's modification:[104]

$\sigma = 0.22c$ if $c<1.2 \times 10^{-7}$

$\sigma = 0.02c^{0.8495}$ if $1.2 \times 10^{-7} \leq c \geq 0.138$

$\sigma = 0.01c^{0.5}$ if $c > 0.138$

Equation 2-2: Horwitz Equation

Horwitz predictions are contradicted by everyday laboratory experiences which indicate that reproducibility is not solely dependent upon concentration but on factors such as analyte, matrix, method and time; significant deviations from the predicted values are frequently experienced.[95] Additionally, with such a model, participants would have an indication of analyte concentration from the acceptable variance provided. Horwitz values are, therefore, not ideal for a forensic toxicology PTS.

2.2.2.3 Precision Experiment

Acceptable variance is established from reproducibility and repeatability studies by a standard method. This is not appropriate for forensic toxicology where participants use a variety of methods.

2.2.2.4 Participants' results

Participants' standard deviation is not suitable for PT. Often results are non-normally distributed,[98] and when the aim is to establish variability, outlier

rejection, essential for calculation of standard deviation, should not be performed.[105] The more acceptable alternative is a robust standard deviation, which incorporates outliers but assigns them less importance, and is thereby insensitive to deviating results or non-normal distribution.[94,105,106] Removal of outlying values due to a recognisable error (e.g. transcription) is still performed.[107]

The robust measurement of variance is typically the median of absolute deviations, MAD; the median of the differences between the data and median. For robust standard deviation to be an equivalent measure to classic standard deviation MADe is calculated which equals 1.483 x MAD [99] (alternatively MAD/0.6745).[108] Robust statistics require an appropriately large dataset to minimise the influence of the spread of results and have a true picture of laboratory performance.[107] Even so, as the acceptable variation is not an independent value, robust statistics are said to describe a data set rather than assess participant performance. Regardless of the relative deviations from the true value around 95% of participants will always have acceptable scores.[86] Thus an acceptable z-score derived from a robust standard deviation is not an indication of good performance, rather it supports that the laboratory performance is in-line with the ‘norm’.

Depending upon participants’ performance, acceptability can vary substantially between rounds, for example, a higher z-score could be achieved for a determination closer to the true value than the z-score achieved in another round for a value which is further from the true value, depending upon the spread of participant results. This complicates comparison of results between rounds, making trend analysis impossible.[104] Thus, determination of acceptable variation from participant results should be avoided if possible.[99]

Despite these concerns, a robust calculation remains the ISO recommended method for use in an accredited PTS. ISO provides instruction on their iterative method for robust mean and robust standard deviation in ISO Standard 13528.[96]

2.2.2.5 Perception

Ideally acceptable variation is stipulated by the PTS steering committee, through relevant professional experience, at a level generally accepted within the sector to represent fitness-for-purpose.[85,99] In some PTSs a level which reflects the organiser's perception of participants' capabilities may be selected instead. Only a value determined by the end use, however, can truly demonstrate proficiency. Acceptable variation should challenge participants and encourage improvement.

A fixed variation is preferable to allow performance comparison between rounds. The use of a fixed value would be inappropriate in toxicology, however, where drug/metabolite concentrations vary considerably. Rather, a fixed percentage of the true value which describes the level of accuracy to be achieved, is more suitable.[86,99,109] Generally a measurement within ±10% of the true value would be desirable (higher at low concentration), therefore, setting acceptable variation at 5% of the true value (such that an acceptable result would have a z-score of less than 2).[99]

A z-score based on standard uncertainty which represents fitness-for-purpose is the only method recommended by The Harmonised Protocol For Proficiency Testing in Analytical Chemistry Laboratories 1995 revised in 2006.[86] Other models are expected to produce falsely reassuring results. This has been partially demonstrated by Rosario *et al.*[110] However, they recommend more work be performed to include non-normally distributed (heavy tailed/skewed) data with outliers.

Whilst z-scoring with perceived acceptable variation provides a good estimate of proficiency and enables round-to-round comparison for performance monitoring, it does not take account of measurement uncertainty (MU) for either the true value or the laboratory values. ISO suggests five means of performance assessment: z-score, E_n-score, zeta-score, z'-score and E_z-score.[96] Although the z-score remains the most widely used,[97,104] there are some suggestions that the others may be better measures of competence as there is some recognition of uncertainty.[104] This may be true in theory, however, in practice laboratories that have not accurately estimated their MU (expected to be quite frequent due

to the lack of standardised methods for calculation) would not receive a true reflection of performance.[104] It would also be inappropriate to apply these ISO methods where participants' MU values differ by several orders of magnitude.[111] Thus, it would be necessary for the PTS to not only consider the accuracy of results returned but also the reliability (i.e. the consistency and accuracy of the measurement). With all ISO accredited laboratories required to calculate and report MU, this would be a logical future progression.

2.3 Performance Appraisal

It is not enough to participate in a PTS: attention must be given to the feedback. Laboratory directors should review summary reports to check that the information submitted has been properly transcribed and to question performance, comparing it to previous rounds. The review should be thorough and address whether results deemed acceptable for the conditions of the PTS are truly acceptable for maintenance of fitness for purpose within the laboratory. A z-score can be re-calculated if necessary (particularly if a participant-dependent scoring method has been applied).

PTS results provide limited information as a single result simply reflects the performance of the laboratory on that particular day. Also results may be received a significant time post testing, during which there may have been changes to the SOP.[48,61] A laboratory receiving a satisfactory result in a single round cannot use this to claim reliable results have been produced on any other occasion. Where a laboratory operates a QMS and consistently achieves satisfactory PT results, however, they can use this as evidence of their competence to produce reliable data in that particular test.[96] Laboratories are therefore encouraged to frequently participate in PT and monitor their performance over time. Where the same analysis or the same analytes are being tested in each round, rolling performance indicators (RPIs) may be reported with each set of results to convey performance over a number of rounds.[94] Recording z-scores as a graph is thought to be the best monitoring method as it provides a clear overview and is less prone to misinterpretation than numerical methods.

Performance monitoring is particularly important where MU has not been considered. Performance z-score slightly >3 (or <-3) should be considered

together with MU as when the uncertainty is incorporated the range of values, and hence possible z-scores, could extend to less than 2 (or >-2), an acceptable performance. Reference to previous performance scores will help to decide whether a z-score in excess of 3 (or >-3) is truly unacceptable.[97,112] Unacceptable results should always be investigated, although immediate corrective actions may not be necessary. When the source of the error cannot be traced (clerical error, method problem, equipment, operator, the proficiency testing material, etc.) the best course of action is to repeat the test, as trying to correct an unidentified problem could introduce further errors.[85] An aliquot of retained test material can be re-analysed to determine whether a random error has occurred. If corrective actions are taken, the test material should be re-analysed to check they have been effective.[38,85] Serious errors (for example, false positives) may require re-validation.[38]

2.4 Availability of Proficiency Tests

In some countries, to practice within a particular discipline a laboratory requires a license which they are only awarded if they can demonstrate that they maintain appropriate quality standards; participation and good performance in a PTS are essential.[18,62] This is a legal requirement of toxicology laboratories in the United States.[38] Elsewhere, participation in a PTS is frequently a requirement, or at least a strong recommendation, for accreditation to technical standards.[48,62,88] SOFT/AAFS Laboratory Guidelines stipulate that toxicology laboratories should participate in quantitative proficiency testing.[38] The Asia Pacific Laboratory Accreditation Cooperation (APLAC) recommends participation in one round of testing per year. ISO accreditation does not have any particular requirement, although they suggest biannual testing is usually appropriate.

Most proficiency testing schemes (PTS) are voluntary and have an educational ethos. Their aims are usually along the lines of facilitation of learning and improvement, through sharing of participants' experiences and knowledge, and provision of technical and scientific guidance.[62,94] With the majority of laboratories free to select the PTS in which they participate, it must be noted that some are better (more appropriate to the laboratory routine, or more rigorous) than others. In selecting a proficiency testing scheme it should be ensured that a reliable test protocol is being applied and that the method of

performance scoring is clearly defined and appropriate. Accredited schemes and those which achieve good peer review are generally preferred.[49] Identifying suitable schemes in which to participate can be troublesome, however, as there are relatively few; some countries having no opportunities for participation.[113]

2.4.1 United States

The American Association for Laboratory Accreditation (A2LA) stipulates that PTSs approved and accredited by them are required where available. There do not appear to be any forensic toxicology schemes listed. As an alternative the website for ASCLAD LAB lists PTSs which they accredit.[114] There are very few schemes for toxicology. It appears that there are only three registered providers: College of American Pathologists (CAP) (which can provide ISO accreditation), Research Triangle Institute (RTI) International, and Collaborative Testing Services (CTS).

2.4.1.1 College of American Pathologists

CAP is recognised to be a worldwide leader in providing quality improvement programs to pathologists and laboratory professionals. Forensic toxicology is not their main focus, however they do offer Forensic Toxicology (Criminalistics) and Forensic Urine Drug Testing (Confirmatory) programmes.[115]

2.4.1.2 RTI International (North Carolina, US).

Since 2001 a drugs of abuse in 'neat' (synthetic) oral fluid PTS has been available in the US. Laboratories receive five frozen specimens three times per year and are given two weeks to submit their results. A European programme has been in operation since 2007 (only two rounds per annum). Samples may contain substances from a large selection of drugs of abuse.

Toxicology quality assurance/quality control samples prepared with relevant drug analytes at appropriate concentrations for the analysis of drugs in human specimens (urine, oral fluid, hair, sweat or serum) are available to purchase. These are intended to be used by laboratories as a check of quality/calibration although they are responsible for monitoring their own results and comparison with peers is not possible.[116]

2.4.1.3 CTS – Collaborative Testing Services, Inc. US

In 2014 CTS introduced two forensic toxicology proficiency testing schemes - drugs of abuse in blood and drugs of abuse in urine. These are intended to mimic casework samples with realistic drug/metabolite compositions in natural human matrices, routine sample vials, and a case scenario. These tests are directed at US laboratories but could be accessed anywhere (provided relevant import and export documentation is obtained, thus, not blind but an independent check of bias).

A breath alcohol analysis programme (with a breath alcohol quality control simulator solution) is also available.

2.4.2 Asia and Pacific

Asia Pacific Laboratory Accreditation Cooperation (APLAC), covering the entire Asia Pacific area but based in Melbourne Australia, accredits laboratories, inspection bodies and reference material producers in the Asia Pacific region. One of APLAC's primary roles is to organise overseeing a large number of proficiency tests. They list 84 different PT schemes, however forensic toxicology is not included as an approved scheme.[117] They also produce a list of available, not necessarily APLAC endorsed, PTSs.[118] There are two relevant to forensic toxicology. Austox Urine Toxicology Proficiency Program Drugs of Abuse – Qualitative/Quantitative provides monthly urine samples and is accredited to Australian Standard AS4308.[119] This is a general laboratory standard, however it is not specific to PTSs. The second PTS featured in the directory is blood alcohol and drivers toxicology from Division of Analytical Laboratories, Australia. This is an unaccredited PTS and no detailed information could be found.

2.4.3 Europe

PTSs available to European laboratories have become increasing accessible thanks to the European Proficiency Testing Information System, EPTIS,[59] funded by the Standards, Measurement and Testing (SMT) Programme and supported by the European Cooperation for Accreditation (EA), European Federation of National Associations of Measurement, Testing and Analytical Laboratories (EUROLAB) and Analytical Chemistry in Europe (EUROCHEM). This database lists

more than 1200 schemes from 300 providers in 24 countries worldwide. There are 31 different schemes listed under the product group 'drugs'. With the majority of these aimed at therapeutic drug monitoring, there are few which are of interest to laboratories providing forensic toxicology, demonstrated by the much reduced list of only fourteen on the ENFSI website (listed in Table 2-1).[59] The schemes represented on the ENSFI website can be considered to have good peer review, however few are accredited.[120] A list of accredited PT schemes is available from the UKAS website.[120] There are only two; QUARTZ (drugs in blood) and Tackler Analytical PT Scheme for Point of Collection Testing (PoCT) Devices (drugs of abuse in urine). RANDOX is listed as an accredited provider of proficiency testing, however their toxicology programme does not currently (as of May 2014) appear to be part of their accreditation scope.

2.4.3.1 QUARTZ – LGC Standards Proficiency Testing UK

The forensic blood PTS is UKAS accredited (originally based on ISO Guide 43-1) and has been in operation since 2000, providing quarterly assessments. The list of drugs included in samples is extensive, with around eighty possible analytes.

Participants' performance is rated by a z-score derived from a perceived fixed value for acceptable variation. Participants are provided an electronic report for each round which contains the composition of test materials, the assigned values, and tabular and/or graphical representations of participants' results.

Provider		Scheme	Accreditation	x/a
ENSFI approved schemes				
DGKC Reference Institute for Bioanalytics	Germany	Drugs and Alcohol Analysis	In process	
Institute of Legal Medicine and Traffic Medicine for Society of Toxicological and Forensic Chemistry (GTFCh)	Germany	Forensic Alcohol Analysis, Forensic Toxicology, Forensic Hair Testing,	In preparation	
Institut Municipal d'Investigacio Medica (IMIM)	Spain	Modality A: 3 urine samples for screening Modality B: 6 urine samples for identification and quantification analysis	In process	6 4
KKGT	The Hague	Drugs of Abuse (urine) and alcohol analysis		
Norwegian Institute of Public Health Division of Forensic Toxicology and Drug Abuse	Norway	Quantification of 15 drugs in whole blood	No	2
Universität Zürich (Scientific Forensic Service Zurich) Members only	Switzerland	Alcohol Analysis, Toxicology, Drugs (Accredited as a member of the Swiss Society of Legal Medicine but not to an international standard)		1 or 2
UKFSLG (United Kingdom Forensic Science Drugs Liaison Group), LGC Members only	UK	Alcohol Breath Analysis Toxicology Drugs	No	1 1 2
U.O.C. Tossicologia Forense Antidoping	Italy	Alcohol Analysis Drug of abuse in urine	In Progress	4 4
Equalis AB	Sweden	Drugs of abuse in urine Drugs in blood and autopsy blood	No	
Other European Schemes				
NORDQUANT: Norwegian Institute of Public Health	Norway	Drug quantitation in blood		1

Table 2-1: European Proficiency Testing Schemes

2.4.3.2 UKNEQAS

The Network of United Kingdom National External Quality Assessment Schemes (UKNEQAS) provides educational schemes, mainly for laboratories monitoring therapeutic drug levels.[18] Although they do not appear on the UKAS accreditation list, their schemes are fully approved in the UK by Clinical Pathology Accreditation (EQA) Ltd (a subsidiary of UKAS) and have over 450 participants from 36 countries. UKNEQAS offers a forensic toxicology PTS for ethanol (established 1993), paracetamol, salicylate and carboxyhaemoglobin (1991) and various drugs of abuse schemes as detailed below.

2.4.3.2.1 Drugs of abuse in urine

UKNEQAS for drugs of abuse in urine commenced in 1986 and now has over 250 participants from 26 countries (149 UK-based).[18] Participants are quarterly provided 3 specimens of lyophilised urine which reflect real specimens containing complex mixtures of interferences, and therapeutic and abused drugs/metabolites (including emerging drugs of abuse).

Abilities of participating laboratories range from offering point of care testing to more accurate legal or clinical analyses. Participants can select how their performance will be rated, qualitatively or quantitatively. There are two thresholds by which quantitative determinations can be scored: cut-off values assigned by UKNEQAS for routine clinical investigations or the more rigorous values outlined in the UK Laboratory Guidelines for Legally Defensible Workplace Drug Testing.[121] Participants should also report notes on sample integrity (qualitatively or through measurement).

2.4.3.2.2 Quantitative blood scheme

Each quarter two blood samples are provided with named analytes for quantification. Analytes are those typically associated with death by overdose (many requested by participants).

2.4.3.2.3 Toxicology cases

With the majority of PTS a results form is supplied to participants which prompts for information required. Only the level of detail routinely provided to clients

should be given, although the method used (including LLOQ or cut-off) should be completed.[62] A results proforma aids evaluation but can impede valuable comment and may not reflect everyday performance.[95] In 1995 UKNEQAS launched a special *'toxicology cases'* PTS designed to monitor interpretation of analytical results which is a key aspect of forensic toxicology.[18] Participants are supplied with a test material (serum/urine) accompanied by a case scenario. The laboratory considers time and cost requirements before following what they believe to be the best course of action. They are free to make enquiries, including requests for further samples as they would with real cases. Results are returned as a case report and interpretation is also scored out of a maximum of twenty by an independent panel of toxicologists. This type of testing more closely satisfies the requirement of a PTS, to obtain a genuine perception of routine performance. However, forensic toxicology laboratories most often work with blood and do not routinely process serum samples.[18] Thus, this scheme is more applicable to clinical toxicology but with the inclusion of interpretation it is taking PTS to the next level. Forensic toxicology schemes could benefit from the inclusion of interpretation but the most beneficial advance would be a PTS catering for the non-traditional matrices which are increasingly being encountered.

2.4.4 Oral Fluid PT Schemes

2.4.4.1 UKNEQAS Oral Fluid Scheme

In 2009 UKNEQAS commenced an accredited scheme for drugs of abuse in oral fluid. They have quarterly distributions of three oral fluid samples (collected from volunteers, patients and drug misusers) which may be spiked with additional drug compounds and metabolites. Participants are expected to complete integrity tests.

2.4.4.2 ORALVEQ

The first external Quality Assessment performed for drugs of abuse in oral fluid was ORALVEQ. Organised in cooperation between the Institut Municipal d'Investigacio Medica (Spain) and the Instituto Superiore di Sanita (Italy), the first round of testing was performed in 2007 with three samples sent to seventeen European and four US laboratories (including three reference

laboratories). The results of this round of testing have been summarised and published by Ventura *et al.*[122] Participants provided quantitative analysis and evaluation was performed by converting measurements to z-scores using robust mean and robust standard deviation of participants' results. Participants' dispersion was high (CV around 40%, reference laboratories <10%) although accuracy was quite good (the standard error, ERR, was never above 20%). High dispersion could be the result of using robust statistics for z-scoring as no outliers were rejected. High dispersion may have influenced the high percentage of satisfactory z-scores.[122]

2.4.5 HAIR PT Schemes

2.4.5.1 Society of Hair Testing PT

The Society Of Hair Testing (SOHT) provides an annual proficiency test (qualitative/quantitative), which is for their members only.[123]

2.4.5.2 HAIRVEQ

The same institutions which were involved in ORALVEQ collaborated to provide an external assessment scheme to evaluate testing of illicit drugs in hair within the Italian National Health Service and Institutes of Forensic Medicine. First testing was performed in 2002, with 23 participant and three reference laboratories, and a round of testing was conducted once in each of the three subsequent years before long term review revealed that participants' performance had deteriorated. As a definite source of error could not be uncovered, method validation and evaluation of analytical results was deemed responsible. Support and training in these areas was therefore given over three test rounds in 2006. Organisers held workshops and supplied SOPs with hair samples of known drug composition to help improve participant performance. Participation had increased in 2006 to 32 participants and four reference laboratories. Although a small group of laboratories persisted in showing unsatisfactory qualitative performance, overall there was improvement. Quantitative performance was also much improved in terms of dispersion of results (although this remained relatively high).[124]

For z-score calculation the robust mean of the participant values (not reference labs) was used in combination with two values of acceptable deviation to provide two scores. One by robust standard deviation of participants' results and another by the expected relative standard deviation given by Horwitz's equation. This was considered necessary as by the first method all participants received satisfactory z-scores (-2 to 2) which was unrealistic. By Horwitz's method only a few laboratories received satisfactory scores. This highlights the importance of having an appropriate scoring system. By using an independent scoring method it was clear that, although there had been improvement, participants were still not performing adequately. A further workshop identified two analytical method problems responsible for the laboratories struggling with qualitative analysis. The scheme continues to support participants by providing guidance in these areas.

2.4.6 Specialist Forensic Toxicology Testing

2.4.6.1 Federal Aviation Association (FAA) PT, USA

FAA provides specialist proficiency testing for forensic toxicology laboratories in the USA. The scheme was created to provide 'true' post-mortem samples as other toxicology programmes do not take account of tissues or putrid samples.[125] They supply whole blood, plasma and urine samples as well as tissues (kidney and liver), to be more reflective of the samples routinely received by participants. Test material matrix and analyte/concentration are selected on the basis of current analytical/toxicological issues. These could be informed by published research, suggestions from participants/other toxicologists, including those at the Civil Aerospace Medical Institute providing forensic toxicology, or data from aircraft accident research.[125] Testing is quarterly and free of charge although information exchange is expected in return. It is performed on a voluntary basis and participants need only do as they would routinely - qualitative or quantitative analysis.

2.5 Alternative EQA available to all: UNODC ICE

In the 1990s the United Nations Office on Drugs and Crime (UNODC, previously the United Nations Drug Control Programme, UNDCP) investigated global provision of PTSs for forensic toxicology laboratories. It was concluded that they

were significantly lacking and many member states were without access to any form of EQA, for example, in South America, the Middle East and Africa. Many of the schemes discussed above have since been created, although the laboratories that could benefit most from participation, those in developing countries, remain without.

In recognition of the power of PTS to improve quality, and the absolute necessity for all laboratories to have access to EQA, UNODC created free-of-charge, accessible to all, International Collaborative Exercises (ICE) for drugs of abuse in urine and for drugs of abuse in seized drug materials. Not a collaborative trial in the traditional sense, ICE follows the principles of proficiency testing to assess trueness (closeness of agreement between the measured and true values) achieved from various protocols operating within laboratories of differing competencies and resources. It is not yet a true PTS but biannually provides laboratories four test specimens to allow an independent check of performance.

2.6 PTS Quality Assurance: Long Term Review

It is recommended that every PTS should monitor the performance of their participants and the effectiveness of the PTS through long-term review. This is essential for accredited schemes. These are conducted after each year or twelve rounds of testing, whichever is longer.[48] The percentage of participants producing acceptable results (known as the first efficacy measure EM_a) is calculated and reported for each round, often as a chart for ease of interpretation.[48] To determine the degree to which a PTS satisfies the objective of bringing participants' performance closer to the target performance level, a second efficacy measure EM_b can be calculated. This is equal to the defined target variation divided by the observed inter-laboratory robust standard deviation.[48] The target values to suggest that the majority of participants are performing satisfactorily are EM_a >95% and $EM_b \geq 1$.

As previously mentioned, Ferrara *et al* [62] caution that whilst proficiency testing is the best means of assessing laboratory performance it should not, in itself, be considered an endorsement of quality. The limitations of the protocol previously discussed (genuine routine processing of test materials, uncertainty in statistical interpretation of data and delay in reporting results) can be too great for the

PTS to be effective.[62] Hence the encouragement for laboratories to aim to participate in a good quality PTS. With accreditation often associated with quality, many laboratories will opt for schemes which carry some form of endorsement, particularly ISO accreditation. This begs the questions of whether this approach is correct and of whether fitness-for-purpose is ensured by accreditation. Results of long term review can provide participants and laboratories considering participation with an indication of the scheme's fitness-for-purpose.

With relatively few forensic toxicology PTS available, long-term review was performed in the present study for two quite different schemes, the LGC Standards International UKAS-accredited commercial scheme, "Quartz Forensic Blood Toxicology PTS", and the freely-available International Collaborative Exercises (ICE), which does not evaluate performance against a quality standard such as ISO/IEC 17025. Detailed reviews of both schemes are given in the following chapters. The two were compared to evaluate effectiveness of the schemes in improving laboratory performance and to establish the level at which participants perform.

Chapter 3: United Nations Office on Drugs and Crime (UNODC) International Collaborative Exercise (ICE), within the International Quality Assurance Programme (IQAP)

3.1 Introduction

In 1988 a meeting was held by the United Nations Commission on Narcotic Drugs (CND) which directs the United Nations Office on Drugs and Crime, UNODC, to discuss with member states mounting concern over the expansion of drug abuse in developing countries. At this meeting the requirement for member states to be educated, trained and provided with resources for the testing of drugs (particularly in body fluids, considered the more difficult task [63]) was recognised as an essential factor in reducing drug abuse at national and international levels. In developing countries at this time there was a lack of technical knowledge and support services required for effective control.[63] UNODC aimed to ensure that every member state had adequate drug testing facilities. The necessity for proficiency testing to improve laboratory performance was quickly identified.

In the 1990s UNODC investigated the provision of EQA and PTSs. They concluded that available schemes were very much lacking with many of their member states without access. Thus, UNODC created an initiative to promote improvements of quality in routine analytical measurements, the International Quality Assurance Programme, which in turn instituted the International Collaborative Exercises (ICE). These are not collaborative trials in the traditional sense, as they do not study precision of specified methods performed by equally competent laboratories. Rather, ICE follows the principles of proficiency testing to assess trueness achieved from various protocols within laboratories with differing competencies and resources. It is intended, that in time, ICE will develop into an international PTS, available to all.

3.2 International Quality Assurance Programme (IQAP)

Rapid dissemination of information internationally between laboratories, and from laboratories to the various authorities engaged in drug control, is considered critically important for the effective control of international illegal drugs activities.[63] This information must be reliable and comparable between nations.

Laboratories with strong IQC cultures demonstrate a greater level of reliability.[79] The IQAP of UNODC, therefore, helps drug testing laboratories to raise their level of performance by teaching Analytical Quality Assurance and promoting its use. IQAP encourages laboratories to expand their capabilities and strive to perform quantitative analysis, as this has been shown to achieve consistently better laboratory performance.[126] Developed in 1995 the Standing Panel of IQAP aimed to:

> *"Develop a programme of work to harmonise the standards of quality assurance for national laboratories and to set up an international proficiency testing programme".*[63]

Development of a PT programme was an obvious aim as participation in PT is known to improve laboratory performance [50,61] and many member states of the UN were without such a resource. Creating an international PTS is a massive endeavour, however, and perhaps not the best means of IQAP achieving its target of improving quality in the poorest performing laboratories. These laboratories would have been unlikely to participate in a PTS with their performance often far below acceptable limits. Thus, UNODC offered help which included manuals, training equipment, standards, regional meetings, and the introduction of ICE, as a preliminary to a PTS. ICE was intended to help rather than regulate participants, eventually to evolve to a PTS when participant performance reached sufficiency. The overall aim of ICE is to determine an

> *"internationally accepted standard against which the performance of laboratories could eventually be monitored and assessed".*[63]

It was anticipated that ICE would aid laboratory quality improvement by enabling laboratories to demonstrate and compare the quality of their work. Participant laboratories' recognition of their improved competence, together

with their awareness of how they perform relative to their peers was expected to inspire laboratories to improve. Improvement would not be possible, however, without creation of laboratory quality management systems.

Setting up a QMS is not simple but participants would be directed in the development of effective system by using their performance results to assess the strengths of their systems and identify necessary corrective actions. IQAP was available to assist. It was hoped that ultimately participants would seek to have their efforts and commitment to quality internationally recognised by becoming accredited. The disadvantage of not being a formal PTS is that ICE may not be recognised as such for laboratories required to participate in a PTS for accreditation. Where PTSs are unavailable, however, ICE could be used as an alternative form of EQA, with verification to demonstrate commitment to quality.

The process of certification is a commercial activity which can be performed by any organisation (unlike accreditation). Certification can be accredited by an accrediting body on International Certification Standards such as ISO 17043. UNODC seeks to become an accredited Certifying Body which will enable them to provide a Certificate of Participation in ICE.

3.3 International Collaborative Exercises (ICE)

Although ICE was primarily instigated for laboratories in developing countries, it welcomed participation from all laboratories recognised (“gazetted”) for the analysis of drugs of abuse by the appropriate governing body in their country.[63] ICE was the only EQA system available to many participants and, therefore, their only means of detecting if performance was less than fit-for-purpose. Poor performance from these laboratories would have a significant negative effect on the Human Rights of those facing drugs charges and also on worldwide exchange and co-ordination of drugs information. Despite this, participation was voluntary and partial participation was possible with no penalties for opting out of certain analyses, with participation at any level appreciated as ambition to improve quality.

In order to allow maximum participation from laboratories in developing nations the introductory requirements of ICE were simplistic. Organisers were aware, however, that an over-simplified scheme would not have any developmental or educational benefits. Participants were to be challenged to demonstrate that the scheme was sufficiently discriminating. In order to achieve this, the requirements of the scheme were gradually increased to reflect the advancement of the laboratories. Initially there was no requirement to identify a specific substance; screening was the only requirement, with confirmatory testing introduced in later stages.

Analytical methods for analysis were never stipulated. Some laboratories return results obtained from simple methods such as thin layer chromatography and ultra violet spectrophotometry whilst others employ sophisticated techniques like LC-MS/MS.[127] If confirmatory analysis is not routine in a laboratory only a screening result will be returned. The provision of more reliable confirmatory results is a fundamental step in quality improvement, however, many ICE participants' resources could be so limited that they are unable to attain reference standards, rendering confirmatory testing impossible. UNODC demonstrated their commitment to these laboratories by distributing drug standards and publishing recommended analytical methods for the most prevalent illicit drugs to aid method development.[63] Initially the range of drugs included in test materials was limited to those for which manuals were available. Progressively the range has expanded and more substances have been added to the core test menu (and an optional supplementary list created) to reflect international and regional trends in illicit use of controlled substances.

The experts on the ICE Standing Panel travel to UNODC (Vienna) annually to review the programme and advise recommendations for its future development and management. Resources are required for this yet participation is free of charge, to appeal to laboratories truly in need of help. Participation is therefore restricted to 150 per round. If demand for participation were to exceed this limit, fees could be imposed on laboratories in developed countries, particularly commercial laboratories.

With such financial demands on UNODC compromises have been necessary. Biannual provision of samples was cut to annual and for a period no samples

were sent out and there were no meetings of the Standing Panel. Biannual testing returned in 2009 as the panel agreed that laboratory performance should be tested more than once a year. In order for ICE to become a PTS for the purposes of ISO 17025, a minimum of two rounds per year must be maintained.

3.3.1 Test Materials

There are two types of testing available, the analysis of seized materials (SM) or biological specimens (BS). Laboratories can participate in either or both. A round of each test comprises four samples. One bottle can contain an analyte which is not on the main menu and some bottles may be drug-free. SM are usually samples of drugs which have been recovered from illegal trade. As the precise nature of these materials is unknown they are analysed by 5-6 accredited reference laboratories prior to testing.

BS are prepared as spiked urine samples. To avoid bacterial growth on standing and transportation these are supplied freeze-dried for reconstitution (which could introduce errors not encountered in routine casework [85]). With less uncertainty related to their content, external verification of the drug/metabolite concentrations is performed in parallel with the test laboratories by three to five reference laboratories.[63]

As an international scheme, variations between laboratory routine protocols have caused some problems. In many countries testing is prioritised and is stopped at the first controlled drug detected, while other countries identify everything present above trace levels and others still report everything regardless of abundance. Additional constituents and trace substances are therefore not included in the assessment of performance but are simply reported in the laboratory feedback. There is some concern, however, that laboratories operating at this level may be disappointed that their additional work does not get formal recognition.

3.3.2 Import/Export Certificates

One of the difficulties which must be overcome with an international scheme is obtaining permission for SM and reference standards to be transported between

countries (no such requirement exists for biological specimens). Participants must obtain an import certificate to permit the entry of the substance into their country and an export certificate to allow the drugs samples to leave Austria. Import certificates must be issued by the national competent authority under International Drug Control Treaties. When problems are experienced obtaining import certificates UNODC can contact the International Narcotics Control Board (INCB) which writes to the appropriate authorities to establish what problems exist and help to rectify them.

As the composition of the test materials is unknown, permission for all possible substances must be obtained. The Austrian export certificate, however, is more complicated. In order to satisfy international drug control conventions it must indicate the full contents of each sample. Provision of truly unknown samples is therefore difficult and sometime impossible. It is recommended to participants that the ICE test materials are delivered to an intermediary. Where possible the intermediary will be an UNODC Field Office, alternatively an independent laboratory. Upon receiving the test materials they are immediately forwarded to the participant whilst the export certificate is retained. However, if only the receiving laboratory can be named as the importer, a manager should receive the export certificate and agree not to disclose the contents to the analysts.

This system was demonstrated to be effectively maintaining confidentiality when, in 2001, LSD coated on codeine and paracetamol tablets was released as an SM sample. Only five of sixty laboratories correctly identified the presence of LSD.

3.3.1 Returning Results

Laboratories are given eight weeks from notification of dispatch of samples to return their results. Results are returned on a special form provided with the test bottles. The results form has a separate area to record the results for each of the three phases of testing:

Phase 1: Screening (classification only)
Phase 2: Identification (specific compound)

Phase3: Quantification (only reported if this forms part of the laboratory routine procedure)

In order to make the task as simple as possible for participants, the results form contains a list of the substances which are potentially contained within the bottles – this is referred to as the 'menu'. It is possible that other substances may be present in one of these bottles but they do not appear on the menu. This is because the menu only contains those substances that would be expected to be detected and reported in general routine work. Participating laboratories should treat the test materials as they would routine samples, thus all three phases of testing are not compulsory. If they detect a substance not on the menu following their standard procedure, there is a comments area where this can be recorded and this is a bonus point for that laboratory. This comments area can also be used to record any other relevant details e.g. difficulties encountered during the analysis of the test materials.

The results sheet also prompts the laboratory to detail the analysis methods used. This is recorded as a code number according to the list provided in the test documentation. Laboratories can use up to two methods at each stage. It is expected that a different method will be applied to screening and confirmation, thus at least two methods are expected to be reported by each laboratory.

3.3.2 Reporting Performance

At the end of each round individual laboratories are provided with an evaluation of their own performance and are advised of the correct content of the test materials. Participants also receive an annual certificate of competence in the analysis of SM or BS, provided the following criteria have been fulfilled:[126]

- Return of all results for all samples (for both rounds where applicable)
- Correct identification of at least one controlled drug (and/or its metabolite where appropriate) in all samples that contained controlled drugs.
- At least 90% correct results

Z-scores are not calculated for participants, only for reference laboratories. These are anonymised and supplied to participants to allow them to become

familiar with the data presentation format. Participants' performances are not shared, although any pertinent information arising from the comments section is disseminated to alert laboratories for new substances likely to be encountered.

UNODC uses ICE results in determining how best to focus their support. Laboratories which fail to return results or present high instances of incorrect results are approached and offered assistance. Mentoring and advice is available to any laboratory that seeks this. It is hoped that this long term review will provide greater focus of assistance.

3.4 Aim

The University of Glasgow Department of Forensic Medicine and Science has actively supported UNODC in the global improvement of drug testing facilities since 1989. This exercise aimed to further UNODC's target to develop an accessible PTS through evaluation of ten years of data generated by ICE (1999 – 2009). It was intended to appraise whether ICE had improved participants' analytical quality, and establish the level at which they operate. In order for ICE to advance to a PTS recognised by ISO 17025, criteria for acceptable quantitative results must be stipulated. Long term review of ICE aimed to inform what constitutes an international measure of achievable, yet challenging, acceptable variation against which to measure performance.

3.5 Long Term Review

3.5.1 Summary of Available Results

The results received over the first ten years of testing were recorded to a database which used software now obsolete. Unfortunately it was impossible to reliably retrieve the required data and everything had to be manually entered from the original hard-copy participants' results forms to a specially designed Excel spreadsheet. This was extremely time-consuming and only some years could be made available for re-entry. Data for the more recent rounds were provided as summary reports prepared for coordinators' review and were, therefore, more limited. Data reviewed and summarised is listed in Table 3-1.

Year	Round	Samples	Data
1999	1	BS	Complete
2001	1	BS and SM	Complete
2001	2	BS and SM	Complete
2003	1	BS and SM	Complete
2003	2	BS and SM	Complete
2005	1	BS and SM	Complete
2005	2	BS and SM	Complete
2007	1	BS and SM	Limited
2008	1	BS and SM	Limited
2009	1	BS and SM	Limited

Table 3-1: Available data by year
(BS = Biological Specimens, SM = seized materials). No testing was performed in 2006 (2000, 2002 and 2004 data is unavailable)

3.5.2 General protocol

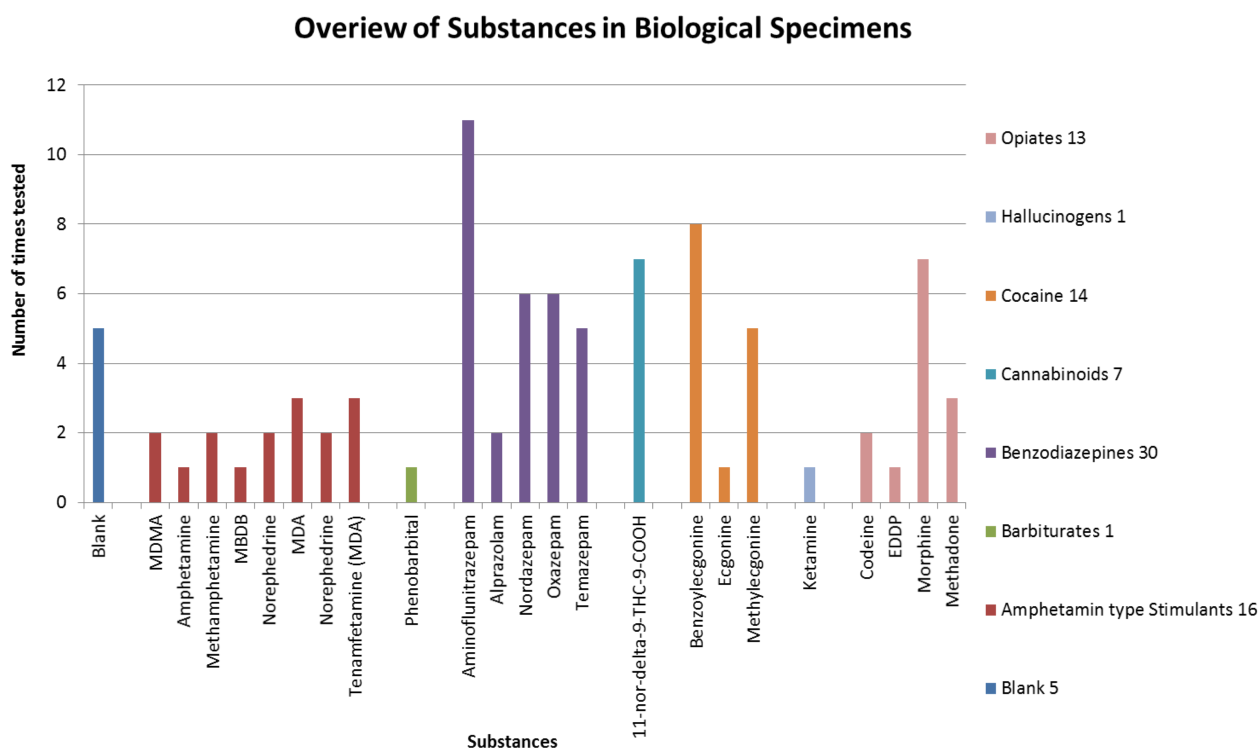


Figure 1 Frequency of testing of specific substances in biological specimens

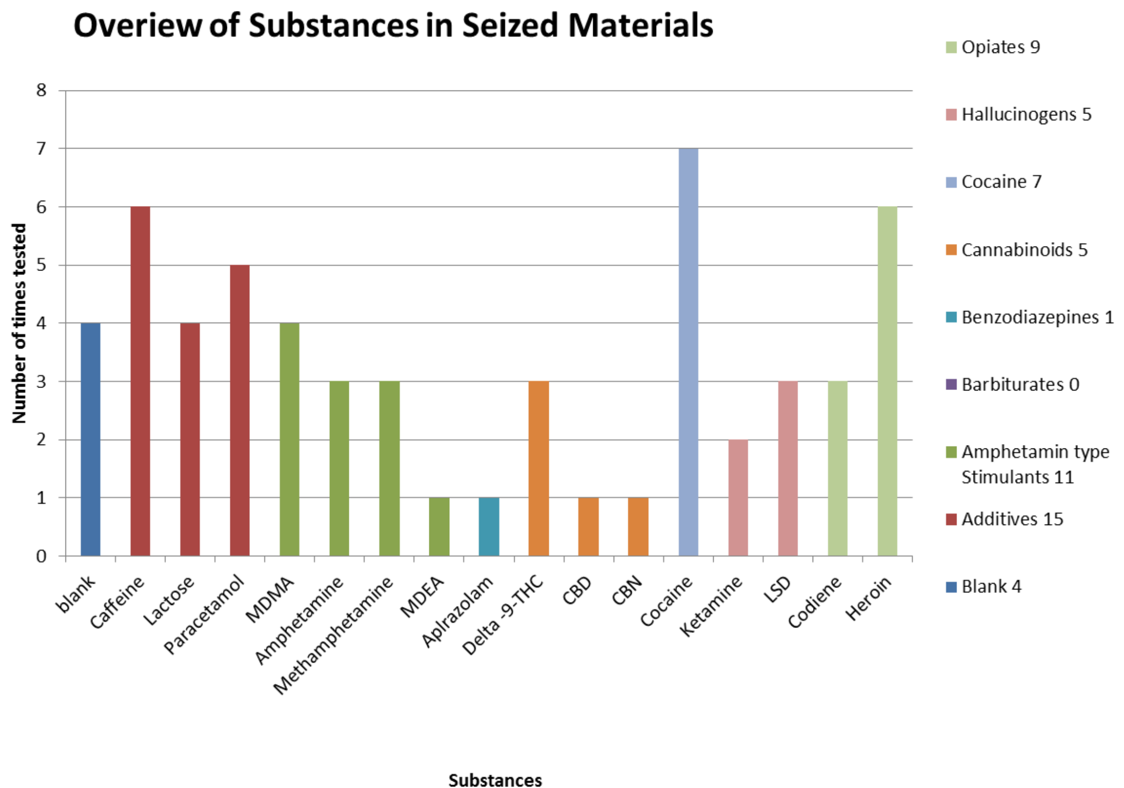


Figure 2: Frequency of testing of specific substances in seized materials

Figure 1 (BS) and Figure 2 (SM) depict the frequency with which specific drugs were included in the programme over the years profiled.

3.5.3 Participation

In general participation was unstable (Figure 3), with SM tending to be favoured over BS, Figure 4 (2009 data did not allow laboratories participating in both SM and BS to be distinguished).

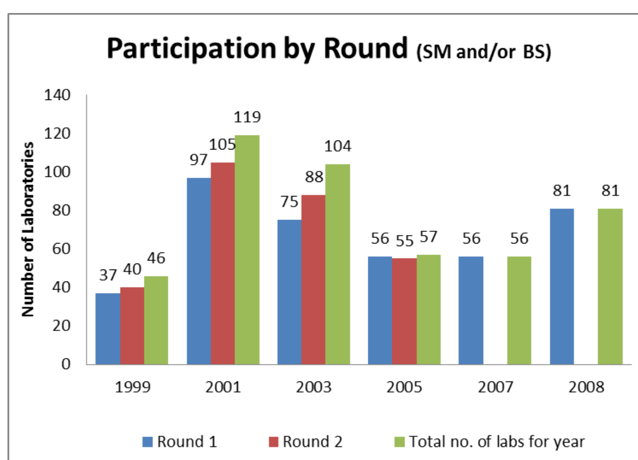


Figure 3: Number of laboratories participating in ICE for each round
There was only one round in 2007 and 2008

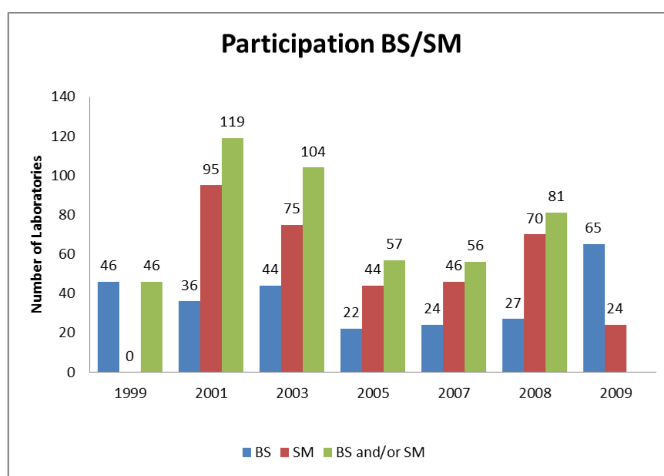


Figure 4: Number of laboratories participating in BS testing, SM testing and total number of participants.

Continuous participation in ICE could not be fully assessed with information available only for the 8 BS and 6 SM rounds in years 1999, 2001, 2003 and 2005 (Table 3-2). 65 laboratories performed at least one ICE BS round and 111 laboratories participated in at least one SM round. Only 10 laboratories participated in all 8 BS rounds and 14 in all 6 SM rounds.

BS		SM	
Number of rounds	Number of Laboratories	Number of rounds	Number of Laboratories
8	10		
7	1		
6	10	6	14
5	2	5	13
4	4	4	24
3	13	3	19
2	12	2	25
1	13	1	16

Table 3-2: Consistency of laboratories' participation

There was no data to relate the number of laboratories that were invited to participate but declined, or that enrolled in a round but did not return results. Review of such data could have diagnosed issues which prevented keen laboratories from participating (for example, difficulties with export certificates or insufficient time allotted for completion of testing) and any requisite procedural adaptations to maximise participation.

3.5.4 Qualitative Summary

Figure 5 (BS) and Figure 6 (SM) illustrate proportions of correct detections for years 1999, 2001, 2003 and 2005.

In the absence of individual laboratory data for the whole period of the review a compromise of looking at detections for each particular substance was used as a general performance measure (summarised in Table 3-4 BS and Table 3-6 SM). As a consequence of inconsistent laboratory participation, however, it is unknown whether the percentages of correct identifications given in these tables are a true reflection of participants' general performance through time. It is unknown whether any dips in performance are due to poorer performance by new or periodic participants bringing down performance as a whole, whilst improvement in regular participants is perhaps obscured.

Details for occurrences of false negatives and positives could only be deciphered from 1999, 2001, 2003 and 2005 data, with details of the analyte errors and the laboratories that reported them. All that was provided for 2007, 2008 and 2009 was the sum total of 'participants with a wrong response' per round. It was unclear whether they had failed to detect the substance of interest or if they had identified a substance which was not present. Incorrect results are presented in Table 3-3 BS in and Table 3-5 for SM.

Overall, identification was better for SM than BS. A laboratory failed to identify an analyte present in a BS sample on 137 occasions (some samples had multiple analytes) compared with 23 unidentified analytes in SM samples.

3.5.4.1 Biological Specimens

BS codeine false negatives were very frequent. As this drug is not controlled worldwide, it may not have been part of a large number of participants' routine test battery. Failure to detect codeine could not, therefore, be confidently attributed as a quality issue and was not given further consideration.

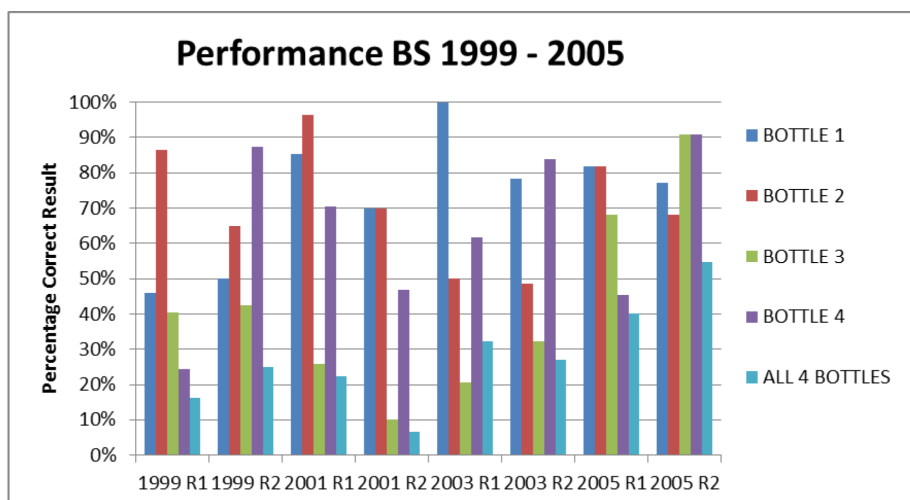


Figure 5: Correct identifications of biological specimens from 1999 to 2005

	1999_1		1999_2		2001_1		2001_2		2003_1		2003_2		2005_1		2005_2		Total tests	Total failed to ID	
Participants	37		40		27		30		34		37		22		22				
New participants	37		9		4		5		2		1		4		0				
No of analytes	296		200		135		240		136		296		110		132				
Failed to identify:																			
ATS	*	*	20	50%	9	33%	16	53%	17	50%	8	22%	5	23%	8	18%	234	83	35%
Barbiturates	*	*	*	*	*	*	*	*	9	26%	*	*	*	*	*	*	34	9	26%
Benzodiazepines	68	46%	*	*	16	59%	71	59%	*	*	48	43%	12	55%	5	23%	484	220	45%
Delta-9-THC - COOH	14	19%	*	*	9	33%	*	*	*	*	6	16%	4	18%	*	*	160	33	21%
Cocaine Group	37	50%	32	40%	*	*	8	27%	17	25%	27	36%	*	*	2	9%	348	123	35%
opiates	*	*	19	24%	10	19%	15	25%	*	*	12	32%	10	23%	24	55%	319	90	28%
Total	119	40%	71	36%	44	33%	110	46%	43	32%	101	34%	31	28%	39	30%			
False positives																			
ATS	1		2		3		3		3		2		2		1			17	
Barbiturates	2		2		*		*		*		*		*		*			4	
Benzodiazepines	4		3		6		3		*		4		4		6			30	
Delta-9-THC - COOH	2		1		*		*		*		*		*		*			3	
Cocaine Group	1		*		1		1		*		*		*		*			3	
Hallucinogens	*		*		*		1		*		*		*		1			2	
opiates	*		3		6		8		*		4		1		*			22	
Total	8		11		10		16		3		16		7		8				

Table 3-3: Overview of BS results

Benzodiazepines had the highest incidence of incorrect BS results (although this could be because it was the most frequently tested group). Investigation of the laboratories failing to detect benzodiazepines identified 3 laboratories with a particularly high share (laboratory 330 had 12, 349 had 10 and 521 had 9 errors). These were recurrent participants (330 and 521 participated in all 8 BS rounds and 349 in 7) which could explain their higher rate, although there was a common feature of their analytical method. They all applied TLC.

	Amphetamine Type Stimulants	MDMA	Amphetamine	Methamphetamine	MBDB	Norephedrine	Tenamfetamine (MDA)	Barbiturates	Phenobarbital	Benzodiazepines	Alprazolam	Nordazepam	Oxazepam	Temazepam	Cannabinoids	11-nor-delta-9-THC-9-carboxylic acid	Cocaine	Benzoylecgonine	Ecgonine	Methylecgonine	Hallucinogens	Ketamine	Opiates	Codeine	EDDP	Morphine	Methadone
1999 R1												71.8%,	54.0%			86.5%		78.4%		62.2%							
													59.5%			75.7%											
1999 R2					50.0%						33.3%					66.7%		72.5%	77.5%					100.0%		77.5%	
2001 R1			70.4%																							85.2%	74.1%
2001 R2						46.7%						23.3%	26.7%	70.0%				70.0%								83.3%	63.3%
														43.3%													
2003 R1									38.2%			50.0%						85.3%		94.1%							
2003 R2				78.4%								59.4%	54.1%	54.1%		83.4%		72.9%		54.1%						64.9%	
2005 R1							81.8%				45.5%					81.1%								77.3%		70.6%	
2005 R2		21.8%					21.8%				68.2%							91.0%									91.0%
2007				91.7%												91.7%										91.7%	
2008												96.2%	81.4%	81.4%				70.3%		85.1%		92.5%					
2009		91.3%					85.7%					91.3%	78.3%	78.3%		87.0%		82.6%		65.2%				79.2%		95.8%	

Table 3-4: Summary of correct identifications for substances included in biological samples testing

3.5.4.2 Seized Materials

Results for 2001 Round 1 SM involving LSD were disregarded as this Round probably had an error in the sample composition.

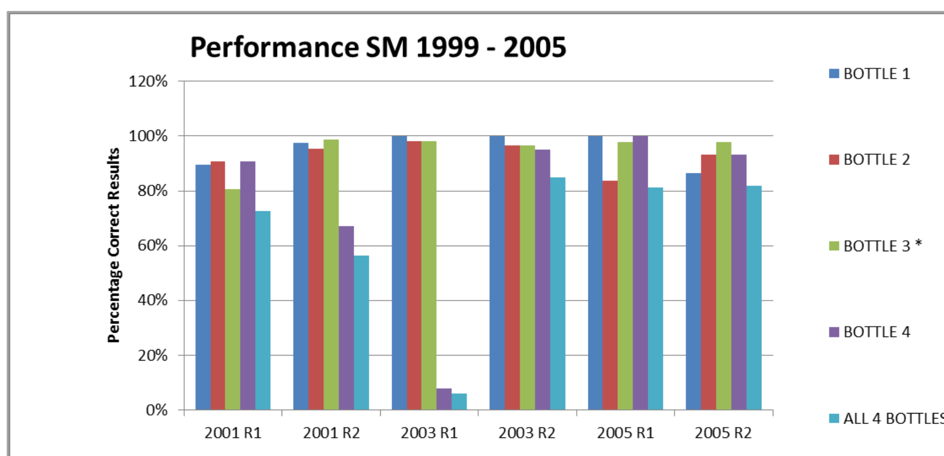


Figure 6: Correct identifications of seized materials from 1999 to 2005
*** LSD in bottle 3 2001 Round 1 has been omitted. 2003 R1 bottle 4 contained EDDP which only four laboratories correctly identified**

Amphetamine-type stimulants (ATS) accounted for 11(65%) of all relevant SM false negatives. Opiate false negatives were all for the adulterant codeine and all but 3 of the hallucinogen false negatives related to the disregarded LSD sample in Round 1. With the highest proportion of false negatives, ATS proved to be the greatest challenge to SM participants (Table 3-5 counts those laboratories who did not attempt the analysis together with false negatives as 'failed to identify', Table 3-8 differentiates between these). The false negatives were dispersed across different laboratories and therefore appeared to be random errors (although 3 laboratories had 2). In 2 instances no screening had been performed and in 6 screening was negative (5 using colorimetric reactions and 1 TLC). In the remaining 3 instances another substance was identified (2 correctly and 1 falsely) and it may have been the routine procedure of the laboratories concerned simply to stop the analysis when the first drug was identified rather than continue to test for all substances present.

All bar one of the other 8 false negatives (4 hallucinogens, 3 benzodiazepines,1 opiate) were the result of negative screening (3 by colorimetric reactions, 3 by TLC and one GCMS, the other was not screened).

	2001_1		2001_2		2003_1		2003_2		2005_1		2005_2		Total tests	Total failed to ID	
Participants	77		85		51		60		43		44				
New participants			18		5		5		0		0				
number of possible identifications	308		340		459		240		172		220				
Failed to identify:															
ATS	8	10%	5	6%	*	*	5	8%	4	9%	10	23%	508	32	6%
Benzodiazepines	7	0.09	*	*	*	*	*	*	*	*	*	*	34	7	21%
Delta-9-THC - COOH	*	*	*	*	*	*	4	7%	*	*	*	*	484	4	1%
Cocaine Group	*	*	2	2%	1	2%	0	0%	0	0%	0	0%	160	3	2%
Hallucinogens	72	47%	*	*	*	*	*	*	8	19%	3	7%	348	83	24%
opiates	*	*	32	19%	56	55%	*	*	0	0	*	*	319	88	28%
Total	87	28%	39	11%	57	12%	9	4%	12	7%	13	6%			
False positives															
ATS	1		1		*		2		3		6		13		
Barbturates	1		*		*		*		*		*		1		
Benzodiazepines	1		1		*		*		*		*		2		
Delta-9-THC - COOH	*		1		2		*		*		*		3		
Cocaine Group	*		3		*		*		*		*		3		
Hallucinogens	1		*		*		*		*		*		1		
opiates	*		1		1		*		*		2		4		
Total	4		7		3		2		3		8				

Table 3-5: Overview of SM results

	Amphetamine Type Stimulants	MDMA	Amphetamine	Methamphetamine	MDEA	Benzodiazepines	Alprazolam	Cannabinoids	Delta -9-THC	CBD	CBN	Cocaine	Cocaine	Hallucinogens	Ketamine	LSD	Opiates	Codeine	Heroin	Additives	Caffeine	Ephedrine	Lactose	Paracetamol
2001 R1			89.6%				90.9%								90.9%	15.6%		80.5%					2.3%	94.8%
2001 R2			95.3%										98.8%					67.1%	97.6%		1.2%	3.5%	4.7%	92.9%
2003 R1													98.0%					7.8%	100.0%		88.2%		47.1%	96.1%
2003 R2				96.7%	95.0%				96.7%				100.0%										3.3%	
2005 R1		90.7%											97.7%		86.0%				100.0%		46.5%		20.1%	97.6%
																					95.3%			
2005 R2		97.7%		90.9%									97.7%			93.2%							6.8%	
		93.2%																					22.7%	
2007			91.3%						73.4%		97.8%		100.0%						97.8%					
2008				97.1%												95.7%			100.0%		77.1%		71.4%	64.3%
																					90.0%			
2009			98.4%						98.4%	93.8%	95.3%		100.0%						100.0%		23			

Table 3-6: Summary of correct identifications for substances included in seized materials testing

3.5.5 Methods

3.5.5.1 Biological Specimens

The analytical methods used for BS determinations are detailed for each substance group in Table 3-7. As barbiturates were only tested once method assessment would have been unreliable. With the greatest proportion of correct identifications (87%), HPLC appeared to be the best method for benzodiazepine identification, closely followed by GC-MS (82% correct). GC/NPD was generally acceptable (greater than 80% correct identifications), however, GC-FID as well as TLC were unreliable.

At 7%, cannabis had the lowest BS frequency of false negatives (others ranged 13% to 16%); participants were therefore most competent with this identification. With a 100% identification rate, HPLC transpired to be the best identification method. There were fewer HPLC identifications than GC-MS, however, which was most popular and extremely effective (97% correct). TLC was also frequently applied. Overall TLC had unacceptable performance as just over half of analyses resulted in a correct identification. However, with cannabis it was relatively good (74%). Perhaps a consequence of the fewer cannabinoid analytes impacting on specificity. If this were true, TLC would be expected to perform well with other substance groups with few analytes, for example cocaine. However, TLC cocaine identifications were poor (55%).

All instrumental methods applied to cocaine, ATS, and opiates provided acceptable performance with the exception of GC-FID with ATS. It was clear that BS identification was most successful by LC-MS(/MS) (infrequently used methods were not considered). Only 1 of 40 analyses was a false negative. However, LC-MS/MS accounted for only a small fraction of analyses (3%), which reflects the lower availability of instruments in poorer countries. GC-MS was the most popular method for a large proportion of analyses (63%) and its proficiency was, therefore, more extensively tested. GC-MS attained 87% correct identifications, relatively good given the mixed abilities of the participants, although a higher rate would be desirable.

Substance Group	Result	Method												Total
		Unknown	EIA	FPIA	TLC	HPLC	GC/NPD	GC/FID	GC-ECD	GC-MS	FTIR	UV-VIS	LC-MS	
ATS														
	No Analysis	41	-	-	-	-	-	-	-	-	-	-	-	41
	c	2	-	-	13	3	7	6	-	113	-	-	9	153
	c +	-	-	-	1	-	1	-	-	2	-	-	-	4
	fn	-	-	-	7	1	1	5	-	10	-	-	-	24
	fn+	-	-	-	3	-	1	2	-	5	-	1	-	12
	Total	43	-	-	24	4	10	13	-	130	-	1	9	234
Barbiturates														
	No Analysis	7	-	-	-	-	-	-	-	-	-	-	-	7
	c	-	-	-	3	1	-	2	-	17	-	-	1	24
	c +	-	-	-	-	-	-	1	-	-	-	-	-	1
	fn	-	-	-	-	-	-	-	-	2	-	-	-	2
	fn+	-	-	-	-	-	-	-	-	-	-	-	-	0
	Total	7	-	-	3	1	-	3	-	19	-	-	1	34
Benzodiazepines														
	No Analysis	142	-	-	-	-	-	-	-	-	-	-	-	142
	c	6	-	-	22	20	9	3	2	153	1	1	10	227
	c +	-	-	-	3	-	-	1	-	1	1	-	-	6
	fn	2	-	-	21	3	1	4	-	28	-	-	-	59
	fn+	1	-	-	7	-	-	3	1	4	-	-	-	16
	Total	151	-	-	53	23	10	11	3	186	2	1	10	450
Cannabinoids														
	No Analysis	19	-	-	-	-	-	-	-	-	-	-	-	55
	c	2	1	-	23	11	1	7	-	80	1	-	-	91
	c +	-	-	-	-	-	-	1	-	1	-	-	-	2
	fn	-	-	-	8	-	-	2	-	1	-	-	-	11
	fn+	-	-	-	-	-	-	1	-	-	-	-	-	1
	Total	22	1	-	31	11	1	11	-	82	1	-	-	160
Cocaine														
	No Analysis	72	-	-	-	-	-	-	-	-	-	-	-	72
	c	8	1	-	20	14	9	11	-	157	-	-	7	227
	c +	-	-	-	-	-	1	-	-	-	-	-	-	1
	fn	-	-	-	16	1	2	4	-	25	-	-	-	48
	fn+	-	-	-	-	-	-	-	-	-	-	-	-	0
	Total	80	1	-	36	15	12	15	-	182	-	-	7	348
Opiates														
	No Analysis	44	-	-	-	-	-	-	-	-	-	-	-	44
	c	4	3	1	25	7	9	17	-	148	2	1	12	229
	c +	0	-	-	1	-	-	-	-	-	-	-	-	1
	fn	0	-	-	12	-	-	4	2	21	-	-	1	40
	fn+	0	-	-	2	2	1	-	-	-	-	-	-	5
	Total	48	3	1	40	9	10	21	2	169	2	1	13	319
All Groups														
	No Analysis	325	-	-	-	-	-	-	-	-	-	-	-	325
	c	22	5	1	106	56	35	46	2	668	4	2	39	986
	c +	-	-	-	5	-	2	3	-	4	1	-	-	15
	fn	2	-	-	64	5	4	19	2	87	-	-	1	184
	fn+	1	-	-	12	2	2	6	1	9	-	1	-	34
	Total	351	5	1	187	63	43	74	5	768	5	3	40	1545

Table 3-7: Analytical methods used for biological specimens.
(c = correct, c+ = correct identification with an additional false positive, fn = failed to identify, fn+ = failed to identify the correct substance but wrongly detected another)

3.5.5.2 Seized Materials

SM analytical methods are detailed in Table 3-8 (poor hallucinogen performance resulted from sample preparation error and poor opiate performance was due to codeine, as mentioned earlier). In general terms no SM methods appeared to

have unacceptable performance. GC-MS was the most frequently applied method (69% of analyses) with a relatively good rate of correct identifications (92%)

Substance Group	Result	Method									Total
		Unknown	Colour Reactions	TLC	HPLC	GC/NPD	GC/FID	GC-MS	FTIR	UV-VIS	
ATS											
	No Analysis	8	-	-	-	-	-	-	-	-	8
	c	3	-	14	10	3	60	312	19	4	425
	c +	-	-	-	-	-	1	9	-	-	10
	fn	-	-	4	-	-	5	2	-	-	11
	fn+	-	-	-	-	-	2	1	-	-	3
	Total	11	-	18	10	3	68	324	19	4	457
Benzodiazepines											
	No Analysis	4	-	-	-	-	-	-	-	-	4
	c	1	-	3	2	1	12	49	2	-	70
	c +	-	-	-	-	-	-	-	-	-	0
	fn	-	-	-	-	-	1	1	1	-	3
	fn+	-	-	-	-	-	-	-	-	-	0
	Total	5	-	3	2	1	13	50	3	-	77
Cannabinoids											
	No Analysis	3	-	-	-	-	-	-	-	-	3
	c	3	-	5	-	1	10	37	-	-	56
	c +	-	-	-	-	-	-	-	-	-	0
	fn	-	-	-	-	-	-	1	-	-	1
	fn+	-	-	-	-	-	-	-	-	-	0
	Total	6	-	5	-	1	10	38	-	-	60
Cocaine											
	No Analysis	2	-	-	-	-	-	-	-	-	2
	c	4	-	7	5	2	47	198	14	1	278
	c +	-	-	2	-	-	1	-	-	-	3
	fn	-	-	-	-	-	-	-	-	-	0
	fn+	-	-	-	-	-	-	-	-	-	0
	Total	6	-	9	5	2	48	198	14	1	283
Hallucinogens											
	No Analysis	37	-	-	-	-	-	-	-	-	37
	c	2	-	13	16	1	18	103	7	-	160
	c +	-	-	-	-	-	-	1	-	-	1
	fn	1	-	11	2	-	3	22	-	-	39
	fn+	1	-	1	-	-	-	2	-	-	4
	Total	41	-	25	18	1	21	128	7	-	241
Opiates											
	No Analysis	61	-	-	-	-	-	-	-	-	61
	c	3	1	5	2	3	25	135	10	2	186
	c +	-	-	-	-	-	2	-	-	-	2
	fn	3	-	3	4	1	18	32	4	1	66
	fn+	-	-	-	-	-	-	-	-	-	0
	Total	67	1	8	6	4	45	167	14	3	315
All Groups											
	No Analysis	115	-	-	-	-	-	-	-	-	115
	c	16	1	47	35	11	172	834	52	7	1175
	c +	-	-	2	-	-	4	10	-	-	16
	fn	4	-	18	6	1	27	58	5	1	120
	fn+	1	-	1	-	-	2	3	-	-	7
	Total	136	1	68	41	12	205	905	57	8	1433

Table 3-8: Analytical methods used for seized materials.
(c = correct, c+ = correct identification with an additional false positive, fn = failed to identify, fn+ = failed to identify the correct substance but wrongly detected another)

3.5.6 Quantitative Summary

To evaluate if ICE influenced participants to progress to quantification, a detailed study was required which identified whether the same laboratories quantified in each round or laboratories progressed from qualitative participation to returning quantifications. This was conducted by dividing the laboratories into three groups according to 1999-2005 participation in less than four rounds, participation in four rounds, or participation in more than four rounds of testing. When values unsuitable for performance evaluation were returned (e.g. zero, 'less than', 'more than', 'trace', incorrect units and errors where an explanation was apparent), laboratories were still considered to have performed quantitative analysis.

Quantification was least common amongst laboratories that did not participate often (23% of 'less than four rounds' and 56% of 'four rounds' participants), and greatest amongst those that participated in more than four rounds (79%). Regular participants, therefore, appeared most likely to quantify, although their quantification could not be linked to their regular participation. There were 106 laboratories performing quantification. Only 6 started to return quantitative results subsequent to their first round and these were equally distributed amongst the three levels of participation frequency. Regular participation did not appear to progress participants towards quantification.

Thus, it had to be addressed whether or not ICE improved quantification performance of regular participants. This required performance scoring. To avoid the possibility of sporadic participation influencing performance statistics, the appraisal only considered laboratories that regularly participated. This was defined as more than four rounds for which complete information was available. A further 5 SM laboratories were included as they had participated in four such rounds and quantified in two or more of the later rounds. This equated to 21 BS and 26 SM participants, of which 13 BS and 26 SM had returned quantifications, as detailed in Table 3-9 (BS) and Table 3-10 (SM).

Substance Group	Year/Round	No of Laboratories	Comments
Amphetamines		10	
MDMA	2005/2	3	
Amphetamine	2001/1	9	
	2003/2	8	no reference value
Methamphetamine	2003/2	10	
	2007/1	4	
MBDB	1999/2	6	
Norephedrine	2001/2	6	one wrong units
Phenylpropanolamine	2003/1	7	no reference value
Tenamphetamine	2005/1	2	
	2005/2	3	
Barbiturates		4	
Benzodiazepines		10	
Alprazolam	2001/1	4	
	2005/2	3	
Nordazepam	1999/1	9	
	1999/1	9	
	2003/2	7	Does not correspond to ref value
Oxazepam	1999/1	8	
	1999/1	8	
	2003/2	7	one “traces”
Temazepam	2001/2	8	Does not correspond to ref value
	2001/2	3	one wrong units
	2003/2	5	one “traces”
Cannabinoids		12	
	1999/1	10	Does not correspond to ref value
	1999/1	9	Does not correspond to ref value
	2001/1	8	
	2003/2	11	
	2005/1	3	
	2007/1	4	
Cocaine		13	
Benzoylecgonine	1999/1	9	
	1999/2	9	Does not correspond to ref value
	2001/2	9	
	2003/1	9	no reference value
	2003/2	11	
	2005/2	4	
Opiates		12	
Codeine	1999/2	11	Eight wrong units
Methadone	2001/1	7	
	2001/2	6	Does not correspond to ref value
	2005/2	4	
Morphine (free)	1999/2	8	no reference value
	2001/1	8	
	2001/2	6	
	2003/2	7	
	2007/1	4	
Morphine (total)	1999/2	10	no reference value
	2001/1	8	no reference value
	2001/2	9	no reference value
	2003/2	7	no reference value
	2007/1	4	no reference value

Table 3-9: Laboratories performing quantification by BS substance group

Substance Group	Year/Round	No of Laboratories	Comments
Amphetamines		22	
MDMA	2001/2	21	
	2005/1	19	one wrong units
	2005/2	22	one wrong units
	2005/2	21	one wrong units
Amphetamine (Base)	2001/1	17	
	2007/1	12	
	2009/1	15	
Methamphetamine	2003/2	20	
	2005/2	13	Does not correspond to ref value
	2008/1	11	
MDEA	2003/2	19	Does not correspond to ref value
Cocaine		22	
	2001/2	22	
	2003/1	16	
	2003/2	20	
	2005/1	21	one wrong units
	2005/2	23	one wrong units
	2007/1	13	
	2009/1	15	
Opiates		21	
Codeine	2001/1	9	no reference value
	2001/2	8	no reference value
Heroin	2001/2	20	
	2003/1	15	
	2005/1	19	one wrong units
	2007/1	11	no reference value
	2008/1	12	Does not correspond to ref value
	2009/1	15	
Hallucinogens		6	

Table 3-10: Laboratories performing quantification by SM substance group

Data sets were only further analysed when they contained at least eight estimations and there was a known reference value. They were each plotted and evaluated for central tendency of the reference value. Datasets where the given reference value did not appear to be reasonably central to the measured values were not considered further. The reference value may have been incorrect and none of the datasets were large enough to reliably apply a consensus value from the participants' results. It was also impossible to assign consensus values to the total morphine data for which no reference values were provided (many laboratories had provided this rather than free morphine which

was the given reference value but could not be used as the metabolite morphine-6-glucuronide was also present in these morphine samples).

No quantitative scoring had been performed by UNODC. For review, a proficiency score (z-score) was calculated by five different methods:

- Classical statistics using the known reference value with the standard deviation and the mean value with the standard deviation. (A 2-sided-Grubbs test was used to detect outliers. The Grubbs test considers the difference between the extreme values of the dataset and the mean together with the standard deviation to determine whether the extreme values are likely to belong to the same population. A table of critical values will identify whether the difference is significant which depends on the size of the dataset. Outliers at a 95% confidence level were not considered in the determination of the mean/standard deviation but were scored using the values calculated from the other data in the set, in accordance with ISO guidance.[90])
- Using the known value with a perceived variation of 5% (10% for very low concentrations),
- Robust statistics using the known value with robust standard deviation (Median absolution deviation estimate of standard deviation, MADe), and the median value with MADe
- ISO iterative robust method.[96]

Considerable variation was observed in the scores when calculated by each method. The International Harmonised Protocol states that the only acceptable means of PT scoring is by using a fixed percentage of the true value as the acceptable variation.[86] Due to the low participation in each data set the best method according to the ISO guide [90] was fixed acceptable deviation. In addition this was the only suitable means of performing round-to-round comparison and trend analysis and negated any requirement for data to be separated by peer groups (for example, because of the various methods applied having differing influences upon participant dependant scoring methods). Thus, scores derived by the reference value with variation up to 5% (or 10%) of this were considered further.

The z-scores for each dataset were plotted to view distribution of performance and substances for which there were multiple datasets were plotted together to try and identify whether general performance had improved with time. The data were too few to reliably draw conclusions, however there were indications of improvement with slightly reduced range of z-scores with progressive rounds

(with the exception of some ATS SM analyses, most notably MDMA). Performance parameters (coefficient of variation (CV%), accuracy as standard error (ERR%) referred to the true value and first (EM_a) and second (EM_b) efficacy measures) for each suitable round were calculated, Table 3-11. BS laboratories appeared to have struggled more with quantification than SM participants.

Grouped participant performance proved unsuitable for evaluation of improvement. Individual participant performance, however, was also insufficient. Only 11 BS laboratories had quantified 6 or more of the data sets. Of these, 5 laboratories regularly produced unacceptable z-scores, with the z-score of one being excessively high (>15). It would seem that the quality management system within these laboratories needs to be strengthened, although with the others it could be that a less stringent scoring system is necessary if so few laboratories satisfy these requirements. Poor quantification performance may be indicative of participants not being competent in preparing reference standards. It is also possible that the extra step of reconstitution of the freeze dried urine, (not part of routine procedure), is introducing a significant error to some BS participants' analysis.

For such a large scheme, the level of participation in quantification level was low. It is accepted that this is largely a result of the mixed abilities of the participants, many being poorly-resourced laboratories. Poor uptake of quantification, however, could be as a result of it not being routine practice in some laboratories. Quantification of SM materials may not be required by the legal system. Even in the UK, purity analysis of SM is only done for large seizures. And generally with BS, analyte concentration is of greatest interest for blood samples rather than urine. It is understood that the difficulties involved in providing blood specimens in a sizable international PTS would be too great for an incentive such as ICE which has limited resources and is provided free-of-charge to participants who may also face financial difficulty. Perhaps as a compromise dried standards could be provided for reconstitution with blank blood in place of or in addition to dried urine samples.

Substance	Year/Round	Outliers	CV%	ERR%	EMa	EMb
BS						
ATS						
Amphtamine	2001_1	1	11.93	3.83	55.60	0.44
Methamphetamine	2003_2	-	16.38	5.05	54.60	0.30
Benzodiazepines						
Nordazepam	1999_1	-	30.05	10.30	44.40	0.16
Nordazepam	1999_1	-	25.89	9.06	77.80	0.18
Oxazepam	1999_1	-	31.56	8.82	12.50	0.20
Oxazepam	1999_1	-	33.02	9.96	37.50	0.18
Cannabinoids						
Cannabis	2001_1	-	19.46	7.51	62.50	0.24
Cannabis	2003_2	1	17.19	4.25	36.36	0.35
Cocaine						
Benzoylecgonine	1999_1	-	12.92	4.56	66.67	0.37
Benzoylecgonine	2001_2	-	18.29	5.11	44.44	0.33
Benzoylecgonine	2003_2	-	14.77	4.23	63.64	0.36
Opiates						
Morphine (free)	2001_1	1	67.48	27.07	12.50	0.07
SM						
ATS						
MDMA	2001_2	-	3.22	0.71	87.50	1.53
MDMA	2005_1	2	5.47	1.34	88.90	0.88
MDMA	2005_2	-	15.14	3.35	71.40	0.33
MDMA	2005_2	-	9.99	2.24	75.00	0.50
Amphetamine (base)	2001_1	1	31.45	6.80	47.06	0.15
Amphetamine (base)	2007_1	1	9.88	3.03	75.00	0.48
Amphetamine (base)	2009_1	-	6.99	1.82	86.67	0.71
Methamphetamine	2003_2	-	2.33	0.55	85.00	2.04
Methamphetamine	2008_1	1	8.59	2.55	90.91	0.59
Cocaine						
Benzoylecgonine	2001_2	-	5.66	1.26	77.27	0.84
Benzoylecgonine	2003_1	-	12.76	3.30	75.00	0.38
Benzoylecgonine	2003_2	2	10.00	2.24	75.00	0.70
Benzoylecgonine	2005_1	4	5.46	1.20	85.00	0.93
Benzoylecgonine	2005_2	-	6.14	1.30	81.82	0.82
Benzoylecgonine	2007_1	3	6.92	1.88	76.92	0.74
Benzoylecgonine	2009_1	-	5.48	1.43	86.67	0.90
Opiates						
Heroin	2001_2	-	4.00	0.91	100.00	1.23
Heroin	2003_1	-	5.84	1.60	73.33	0.81
Heroin	2005_1	1	4.52	1.05	94.44	1.12
Heroin	2009_1	-	4.32	1.13	100.00	1.14

Table 3-11: Performance Statistics

3.5.6.1 Biological Specimens

BS quantification was best for the analyte which had been most frequently included in test samples, benzoylecgonine, perhaps indicating that good performance is as a result of familiarity with the application of the analytical method. This could be misguided, however, as the poorest performance was for morphine which had been included in test samples almost as often. 86% of laboratories reported opiate identifications (only 79% for cocaine) but much fewer performed quantification (only one round contained enough appropriate data for review). Laboratories may be struggling with the quantification of opiates, although this cannot be fully explored without reference values for the total morphine content which was often the preferred measurement.

There were three outlying values amongst the BS dataset; two were produced by LC-MS methods (cannabis and morphine) and the other by GC-MS (amphetamine). Almost all of the BS laboratories used GCMS for quantification, a small number used HPLC and one used GC-NPD (GC-ECD for benzodiazepines). The instrumental method used was not observed to influence the z-score apart from the quantification of benzodiazepines in which the use of HPLC resulted in higher scores relative to GC-MS. Scores for benzodiazepine quantification by GC-NPD/GC-ECD were outside of the acceptable range but this could not be attributed to unsuitability of the instrumentation as this laboratory also scored highly with substances analysed by GC-MS. It seems more likely that this participant required some help with internal quality standards.

3.5.6.2 Seized Materials

SM performance contrasted BS performance, as quantification of opiates was the most successful SM analysis performed, with highest ratings of accuracy and satisfactory scores. No trends could be observed although it was apparent that improvement was required with ATS. Although the most frequently tested SM substance group, it contained the largest number of analytes with each individual drug being tested infrequently.

Similar to BS the majority of SM laboratories used GC-MS for quantification. A small number used GC-FID, FTIR, HPLC or GC-IRD (gas chromatography with vapour phase infrared spectrophotometric detection). (Data only available up to

2005 and assumed the same methods until 2009.) All instrumentation performed equally well. Often the poorest scores related to GC-MS, which had the greatest number of unacceptable results. Thus, the scores for performance of other methods were within that of GC-MS. There were 13 outliers: 10 of these were produced by GC-MS (2 MDMA, 1 amphetamine, 1 methamphetamine and 6 cocaine), 2 by FTIR (cocaine) and 1 by GC-FID (cocaine). None of the 21 SM laboratories that had quantified in at least six datasets were found to have poor performance, although one laboratory did appear to struggle slightly with MDMA and another with cocaine.

3.5.7 Variation in Results

With a large dispersion and the presence of outliers it was considered that ICE data was suited to further review of scoring methods recommended by Rosario *et al.*[110] It was determined how many laboratories achieved acceptable scores by each of the methods previously outlined, in addition to perceived variation of 20% of the true value for BS samples. (This was deemed fit-for-purpose variation for forensic toxicology laboratories by the Panel of Experts for the LGC QUARTZ Forensic Blood Toxicology Proficiency Testing Scheme, described in Chapter 4 page 78). This is reported in Table 3-12 (BS) and Table 3-13 (SM). The data for participant dependent scoring methods were not separated into peer groups as a large majority of laboratories used GC-MS for quantification, leaving too few for sub-grouping.

With the SM data there was not a great deal of difference in the number of acceptable scores achieved by each of the various methods of calculating the performance scores, as there was no significant variance in the SM laboratory determinations, bar some outliers. The BS laboratories, however, did not perform as well (the increased difficulty with such analyses is recognised in the increased acceptable variance in the LGC QUARTZ scheme). The effects of using different scoring systems can be more readily detected with the BS results.

In agreement with Rossario *et al*, classical statistics produced the highest number of acceptable z-scores, with 4 exceptions in which robust statistics had resulted in more acceptable z-scores due to the broadening of the acceptable range by inclusion of outliers. This confirms previous findings that robust

methods which do not reject outliers produce an unrealistically high percentage of satisfactory z-scores.[110,122,124] Comparison of the known reference value with the laboratory determinations demonstrated that both classical and robust statistics generated a falsely elevated number of acceptable z-scores. EM_a was not influenced by use of mean compared to true value.

The large variation in participant determinations influenced participant-dependent scoring systems such that a true reflection of participant performance was only obtained with perceived variation. This was evident when EM_a by the perceived method was much lower than with other methods. For example, for morphine in Round 2001_1 (Table 3-14) only one laboratory, with a determination close to the true value, received an acceptable z-score within 5% perceived. Yet by the other methods, 7 laboratories with determinations far from the true value have been awarded very good performance scores. This is clearly unhelpful as participants appear to have no cause for concern although one laboratory has detected less than one tenth of the morphine present and may be in need of corrective actions.

Z-Scoring Method	ATS		Benzodiazepines				Cannabinoids		Cocaine			Opiates
	amphetamine	Methamphetamine	Nordazepam		Oxazepam		Delta-9-THC-COOH		Benzoylecgonine			Morphine
	2001_1	2003_2	1999_1	1999_1	1999_1	1999_1	2001_1	2003_2	1999_1	2001_2	2003_2	2001_1
Classical, participant mean and standard deviation (2-Sided Grubbs at 0.05 sig level)	88.9	100.0	100.0	100.0	100.0	100.0	100.0	90.9	100.0	100.0	100.0	87.5
Classical, known true value and standard deviation (2-Sided Grubbs at 0.05 sig level)	88.9	100.0	100.0	100.0	100.0	100.0	100.0	90.9	100.0	100.0	100.0	87.5
Perceived variation of 5% of true value	55.6	54.5	44.4	44.4	12.5	37.5	62.5	36.6	66.7	44.4	63.6	12.5
Perceived variation of 10% of true value	55.6	63.6	66.7	55.6	62.5	62.5	75.0	90.9	77.8	88.9	81.8	37.5
Perceived variation of 20% of true value	56.0	73.0	89.0	89.0	100.0	88.0	75.0	91.0	89.0	100.0	100.0	50.0
Robust, known true value and MADe	55.6	63.6	100.0	100.0	100.0	87.5	75.0	90.9	77.8	88.9	100.0	87.5
Robust, participant median and MADe	55.6	72.7	100.0	88.9	100.0	100.0	75.0	90.9	77.8	100.0	100.0	87.5
ISO robust	55.6	72.7	100.0	88.9	100.0	100.0	87.5	90.9	77.8	100.0	100.0	87.5

Table 3-12: BS laboratories acceptable scores

Z-Scoring Method	ATS									Cocaine							Opiates			
	MDMA				Amphetamine			Methamphetamine									Heroin			
	2001_2	2005_1	2005_2	2005_2	2001_1	2007_1	2009_1	2003_2	2008_1	2001_2	2003_1	2003_2	2005_1	2005_2	2007_1	2009_1	2001_2	2003_1	2005_1	2009_1
Classical, participant mean and standard deviation (2-Sided Grubbs at 0.05 sig level)	100.0	88.9	100.0	100.0	94.1	91.7	100.0	100.0	90.9	100.0	100.0	90.0	80.0	100.0	76.9	100.0	100.0	100.0	94.4	100.0
Classical, known true value and standard deviation (2-Sided Grubbs at 0.05 sig level)	100.0	88.9	100.0	100.0	94.1	91.7	100.0	100.0	90.9	100.0	100.0	90.0	80.0	100.0	76.9	100.0	100.0	100.0	94.4	100.0
Prescribed variation of 5% of true value	87.5	88.9	71.4	75.0	47.1	75.0	86.7	85.0	90.9	77.3	75.0	75.0	85.0	81.8	76.9	86.7	100.0	73.3	94.4	100.0
Prescribed variation of 10% of true value	100.0	88.9	90.5	85.0	58.8	91.7	100.0	95.0	90.9	95.5	93.7	85.0	95.0	90.9	92.3	100.0	100.0	86.7	94.4	100.0
Robust, known true value and MADe	87.5	88.9	90.5	85.0	94.1	91.7	93.3	80.0	90.9	81.8	93.7	80.0	85.0	81.8	84.6	86.7	85.0	86.7	94.4	100.0
Robust, participant median and MADe	87.5	83.3	90.5	85.0	94.1	91.7	93.3	70.0	90.9	81.8	93.7	80.0	85.0	81.8	84.6	86.7	90.0	86.7	94.4	100.0
ISO robust	87.5	88.9	90.5	80.0	94.1	91.7	93.3	70.0	90.9	86.4	93.8	80.0	85.0	81.8	84.6	86.7	95.0	86.7	94.4	100.0

Table 3-13: SM laboratories acceptable scores

Morphine 2001_1 (290ng/ml)	Laboratory Determinations (ng/ml)	Z-score Calculation Method Classic	Robust	ISO	Perceived (5%)	Perceived (20%)
Participant 'true' value		264.8	329	347.4	290	290
Acceptable Deviation		192.6	222	232.1	14.5	58
Laboratory ID						
529	27.8	-1.2	-1.4	-1.4	-18.1	-4.5
526	195.5	-0.4	-0.6	-0.7	-6.5	-1.6
751	220	-0.2	-0.5	-0.5	-4.8	-1.2
757	318	0.3	0	-0.1	1.9	0.5
328	340	0.4	0	0	3.4	0.9
754	495	1.2	0.7	0.6	14.1	3.5
752	521	1.3	0.9	0.7	15.9	4
527 (outlier)	1678	7.3	6.1	5.7	95.7	23.9
Ema (%)		87.5	87.5	87.5	12.5	50

Table 3-14: Laboratory performance scores by various calculation methods for BS Morphine 2001 round 1

Delta-9-THC-COOH 2003_2 (140ng/ml)	Laboratory Determinations (ng/ml)	Z-score Calculation Method Classic	Robust	ISO	Perceived (5%)	Perceived (20%)
Participant 'true' value		121	114.7	122.4	140	140
Acceptable Deviation		18.7	19.7	20.4	7	28
Laboratory ID						
758	100.2	-1.1	-0.7	-1.1	-5.7	-1.4
526	104	-0.9	-0.5	-0.9	-5.1	-1.3
759	107.1	-0.7	-0.4	-0.7	-4.7	-1.2
525	108	-0.7	-0.3	-0.7	-4.6	-1.1
757	111.4	-0.5	-0.2	-0.5	-4.1	-1
752	114.7	-0.3	0	-0.4	-3.6	-0.9
529	128	0.4	0.7	0.3	-1.7	-0.4
751	140	1	1.3	0.9	0	0
754	144.9	1.3	1.5	1.1	0.7	0.2
328	152	1.7	1.9	1.5	1.7	0.4
527	247 (Outlier)	6.7	6.7	6.1	15.3	3.8
EMa		90.9	90.9	90.9	36.6	90.9

Table 3-15: Laboratory performance scores by various calculation methods for BS Delt-9-THC-COOH 2003 round 2

Relying on participants' determinations to define 'acceptable' can, therefore, be very misguided. This was well demonstrated by THC-COOH in Round 2003_2 (Table 3-15), even though variation within this dataset was less than in many others, which it illustrated the effect of many inaccurate measurements on a few accurate measurements. One laboratory (751) returned a result equal to the true value but a z-score of 0 was only awarded by a perceived protocol.

Whilst the scores determined by other scoring methods were good acceptable scores, more than half the other participants, who measured further from the true value, appear to have performed better than they actually did.

Perceived variation scoring is, therefore, more appropriate and worthwhile for a PTS to ensure participants are operating without significant error. The fixed percentage used, however, must be given careful consideration. At 20% some acceptable scores were questionable. This was well demonstrated with oxazepam in round 1999_1 (Table 3-16) in which all 20% perceived scores were satisfactory, however, at the extremes one laboratory quantified 50% over the true value and another 50% less than was present. An analytical result of less than half the concentration of an analyte present in a good quality sample is not fit for the purpose of legally defensible forensic toxicology.

Oxazepam 1999_1 (1150ng/ml)	Laboratory Determinations (ng/ml)	Z-score Calculation Method				
		Classic	Robust	ISO	Perceived (5%)	Perceived (20%)
Participant 'true' value		1051.4	909	1017.4	1150	1150
Acceptable Deviation		385.3	286.9	364.7	57.5	230
Laboratory ID						
527	560	-1.3	-1.2	-1.3	-10.3	-2.6
525	793	-0.7	-0.4	-0.6	-6.2	-1.6
526	870	-0.5	-0.1	-0.4	-4.9	-1.2
328	878	-0.4	-0.1	-0.4	-4.7	-1.2
759	940	-0.3	0.1	-0.2	-3.7	-0.9
754	1180	0.3	0.9	0.4	0.5	0.1
752	1440	1	1.9	1.2	5	1.3
751	1750	1.8	2.9	2	10.4	2.6
EMa (%)		100	100	100	12.5	100

Table 3-16: Laboratory performance scores by various calculation methods for BS Oxazepam 1999 round 1

It is recommended for consensus scoring systems that there should be at least 20-30 participants.[48,92] There were no datasets in the BS group which satisfied this and, although variance was much less extreme within the SM group, the dataset with the most determinations was SM cocaine 2001 round 2 (Table 3-17), which contained 22 quantifications. Unlike the studies by Ventura *et al.* there were no outliers in this data and dispersion was reasonably good (CV 5.6%).

There was nothing to suggest that unrealistic scores would be achieved by participant-dependent scoring.

Two laboratories (allowing for small unknown uncertainty) measured the true percentage of cocaine in the sample (506 and 113). They did not achieve cautionary scores by any method, however it was only perceived scoring which recognised the accuracy of these participants (ranking them 1st/2nd compared with 12th/13th by classic, robust and ISO robust methods). By the classical approach there were no unacceptable values and only two approached cautionary scores. Quite a different interpretation is given by perceived scoring, where there were five participants with unsatisfactory scores. By the ISO recommended robust method there were only three unacceptable values, in contrast to the scores that would have been awarded by a perceived method. Even with a larger data set, the use of scoring systems which rely upon participant data should be avoided.

Cocaine 2001_2 **(65%)**	Laboratory Determinations (ng/ml)	Z-score Calculation Method			
		Classic	Robust	ISO	Perceived (5%)
Participant 'true' value		68.5	68.1	67.8	65
Acceptable Deviation		7.2	3.9	4.3	3.3
Laboratory ID					
521	51.6	-2.3	-4.3	-3.8	-4.1
139	61.1	-1	-1.8	-1.5	-1.2
703	62.2	-0.9	-1.5	-1.3	-0.9
140	62.8	-0.8	-1.4	-1.2	-0.7
527	64.5	-0.5	-0.9	-0.8	-0.2
506	65	-0.5	-0.8	-0.6	0
113	65.3	-0.4	-0.7	-0.6	0.1
713	66.3	-0.3	-0.5	-0.3	0.4
110	66.8	-0.2	-0.3	-0.2	0.6
101	67.2	-0.2	-0.2	-0.1	0.7
213	68	-0.1	0	0.1	0.9
206	68.1	-0.1	0	0.1	1
141	68.6	0	0.1	0.2	1.1
215	68.6	0	0.1	0.2	1.1
526	68.7	0	0.2	0.2	1.1
108	69	0.1	0.2	0.3	1.2
187	69.5	0.1	0.4	0.4	1.4
530	70.5	0.3	0.6	0.6	1.7
135	76	1	2.1	1.9	3.4
149	80	1.6	3.1	2.8	4.6
208	81.5	1.8	3.5	3.2	5.1
151	85	2.3	4.4	4	6.2
EMa (%)		100	81.8	86.4	77.3

Table 3-17: Laboratory performance scores by various calculation methods for SM Cocaine 2001 round 2

3.6 CONCLUSIONS

Long term review of ICE was impeded by issues related to degree of participation and data retrieval. The nature of the data available did not facilitate elucidation of year-by-year improvements in participants' performance. However, some limited trend analysis was possible, as summarised below.

3.6.1 Qualitative Summary

SM laboratories performed much better in providing identifications than BS laboratories. The substance group which was most successfully analysed was cannabinoids (BS and SM). The amphetamine-type stimulants group proved most challenging to SM laboratories whilst the benzodiazepine group posed the greatest challenge to BS laboratories. Difficulties experienced with benzodiazepines were due to the large number of laboratories using TLC.

3.6.2 Methods

For BS screening, immunoassay techniques performed better than TLC or colorimetric reactions, which should not be relied upon.

TLC is unsuitable for identification with both BS and SM samples. GC-MS was the most used identification method and no problems identified with the use of this method for either BS or SM. LC-MS(/MS) performed most reliably for identification of BS and SM, however, it was not tested to the same extend as GC-MS.

3.6.3 Quantitative Summary

An extremely small proportion of laboratories showed progression from identification to quantification for at least one substance. The vast majority of those performing quantifications had done so from their first round of testing.

Quantification was much better amongst the SM laboratories than BS. No general problems of quantification could be identified. BS laboratories were

most proficient in quantifying benzoylecgonine but may have struggled with free morphine which was rarely quantified (total morphine which could not be evaluated). Conversely SM quantification was best for opiates. SM laboratories had difficulty with ATS.

The truest reflection of performance was achieved using a defined true value and a perceived acceptable variation (percentage of the defined true value). The perceived value should be greater than 5% for BS laboratories as many z-scores were unacceptable at 5%. Perceived acceptable variation need not represent best performance but should reflect acceptable performance such that it is a realistically achievable target.[86,104] It should be clear to participants, however, that pass or fail is not the outcome of such a developmental educational scheme. Successful participation means learning from the experience. How a laboratory responds to an unsatisfactory result can infer better dedication to quality than a laboratory that consistently achieves satisfactory results.[94]

One BS participant was observed to be in need of assistance in strengthening their internal quality procedures but no SM laboratories were identified with general quality issues, although one laboratory did appear to struggle with MDMA and another with cocaine.

3.6.4 ICE Review Summary

There was some limited support from BS z-score plots to suggest that participants' accuracy had improved slightly over time. The laboratories profiled were consistent ICE participants and this small improvement could be related.

The review has highlighted that assistance should be directed toward BS laboratory performance.

3.6.5 VARIATION

Scoring PT results with participant-dependent mechanisms, including the ISO recommended robust method, can be most misleading (even with 'well-behaved' data). Performance indicators determined from participant results are

unrealistic. Only a perceived variation approach provides a true reflection of performance, although careful consideration is required as to how large the permitted variance should be.

3.7 Future Work

In order that the next review be a more effective evaluation of the ICE programme, efforts should be directed towards laboratory participation, not to enrol more laboratories, but to ensure that laboratories participate regularly. From the limited data available it appears that few laboratories have committed to the scheme with participation over these first years being very erratic. Recoding and addressing reasons for non-participation/not returning results could help to resolve this issue.

It should also be stressed to participants that ICE is an assessment of their routine procedure and this is what they must apply to test materials. Observation of sporadic quantification suggests that many participants do not consistently follow routine. Some participants only provided identification for a substance in Rounds which were flanked by others in which quantification had been performed for that same substance.

Lack of clear progression amongst ICE participants could be related to insufficient feed-back. Participants were unaware of how they performed relative to others and had no means of gauging whether or not their performance was acceptable. As ISO identified, participants in PTS should know where they are within the range of capabilities, in order to aim higher. It is therefore important that efforts to provide quantification are recognised and rewarded by scoring quantification and anonymously reporting all scores to all participants. It is understood that this recommendation has been actioned by UNODC and reports for each ICE Round are now published on-line to provide anonymous feedback of all participants' performance at all levels including z-scores.[128] The reports also now provide methods analysis and are introducing relevant emerging drugs to maintain the challenge to participants. Ketamine was included in BS samples in 2008 (performance unknown) and again in 2011 where performance appeared to be generally good. Participant laboratories are officially invited to report any new substances encountered every six months.[129]

It was clear that BS assessments were more difficult than SM. In order to accelerate improvements in laboratories observed to be struggling, when consecutive unacceptable results have been returned, the participant could be asked to submit additional data for review such as calibration results and calculations, to allow the experts at UNODC to assist with corrective actions.

It is also suggested that, in order to maintain the educational, challenging ethos of ICE, participants should be encouraged to submit an uncertainty measurement with their result (in line with ISO recommendations [56]). This would also require UNODC to report the standard uncertainty associated with the 'true' concentration of each analyte. This information could be useful to all participants as, when the uncertainty is large in relation to acceptable deviation, some participants may wish to consider their action and warning limits relative to this.[96]

Chapter 4: LGC Quartz Forensic Blood Toxicology Proficiency Testing Scheme; Long Term Review

4.1 Introduction

LGC Standards is an international provider of proficiency testing schemes with many years of experience in various fields of testing (chemical, microbial and physical measurements). The scheme which is generally applicable to laboratories analysing blood samples for toxicology within a forensic setting (post mortem and other samples) is the UKAS-accredited Forensic Blood Toxicology Proficiency Testing Scheme which has been in operation for over ten years. This scheme is intended to provide laboratories with a means of safeguarding against undetected errors, enable laboratories to compare their performance with that of their peers and, additionally, provide information to participants on technical issues and methodologies via their scheme-specific advisory group.[130] All aspects of the scheme are overseen by an LGC Standards Proficiency Testing Coordinator.

4.1.1 Scheme Details

In order to ensure confidentiality, participants are allocated a unique laboratory reference number against which results are reported. Reference numbers are assigned and changed from round to round. There are four rounds offered each year and participants must apply for each round in which they wish to participate. All rounds contain forensic toxicology samples 1 and 2, whilst every other round (two per annum) contains a third forensic toxicology sample, sample 3, and a driving impairment sample, sample 4, as detailed below:

Sample 1 Identification: 10 ml blood sample
Qualitative results only. The sample will always contain one Group A substance plus up to three other drugs from either Group A or B (Table 4-1).

Sample 2 Quantification: 10 ml blood sample

Participants are provided the identity or generic classification of the drugs present and should measure their concentrations in the sample. The results should be reported together with an interpretation relating to case information provided with the sample. There will always be at least one Group A substance present.

Samples 1 and 2 are accompanied by a case study providing details that would generally be submitted with a sample to a toxicology laboratory, stating the circumstances which necessitate toxicological testing and providing information such as the clinical symptoms, post mortem observations, cause of death and history of drug use.

Sample 3 Standard Solution

A prepared drug solution to assess instrumental analysis.

Sample 4 Alcohol: 10 ml blood

A blood sample with a known amount of ethanol present. Participants are requested to measure and report the alcohol content.

Participants specify upon application which samples they wish to receive, forensic toxicology (1, 2, and 3), driving impairment (4), or both. It is not necessary to participate in the analysis of all samples provided. Test samples may contain up to four analytes from a list provided which is regularly updated; Table 4-1 is the list from 2012.[131]

Participants are asked to process the test samples as they would a routine sample and are, therefore, free to select any method of analysis which they believe to be technically appropriate. The analytical method should be reported with the result.

Before samples are provided to participants, ten randomly selected aliquots are analysed to ensure that there were no drugs present in the blood before spiking

and that the concentration of the spiked analytes is homogeneous throughout the blood samples.

Group A				Group B	
Anesthetic	Ketamine	Hypnotic	Midazolam	Anti-arrhythmic	Propafenone
Analgesic	Paracetamol		Temazepam		Flecainide
	Salicylate		Zolpidem	Anticonvulsant	Clonazepam
Anticholinergic	Procyclidine		Zopiclone		Gabapentin
Anticonvulsant	Carbamazepine	NSAID	Diclofenac	Antidepressant	Trazodone
	Lamotrigine		Ibuprofen	Antidiabetic	Metformin
	Phenytoin	Opioid Analgesic	Buprenorphine	Bronchodilator	Theophylline
Antidepressant	Amitriptyline		Codiene	Calcium antagonist	Amlodipine
	Citalopram		Dihydrocodeine	Central Stimulant	Benzylpiperazine
	Clomipramine		Fentanyl		Methylphenidate
	Dosulepin		Methadone	Hypnotic	Amobarbital
	Fluoxetine		Morphine		Butobarbital
	Imipramine		6MAM (MACM)		Clomethiazole
	Mitrazapine		Oxycodone		Gamma hydroxy butyrate
	Paroxetine		Pethidine		Lormetazepam
	Sertraline		Propoxyphene		Pentobarbital
	Venlafaxine		Tramadol		Secobarbital
Antihistamine	Cyclizine	Psychotropic	THC		Zaleplon
	Diphenhydramine		THC-COOH	Opioid analgesic	Dextromoramide
	Promethazine	Tranquilliser	Chlordiazepoxide		Dipipanone
Antipsychotic	Amisulpride		Diazepam	Opioid antagonist	Naltrexone
	Chlorpromazine		Desmethyldiazepam	Tranquilliser	Clobazam
	Clozapine		Oxazepam		Loprazolam
	Olanzapine				Lorazepam
	Quetiapine				Phenazepam
	Risperidone			Vasodilator	Sildenafil
Central Stimulant	Amfetamine			B-Blocker	Atenolol
	Cocaine				Propranolol
	Benzylecgonine			Other Substances	Carboxyhaemoglobin
	MDA				
	MDMA				
	Methamfetamine				
	Mephedrone				

Metabolites of Group A and B substances can also be included

Table 4-1: Drugs included in LGC QUARTZ Forensic Blood Toxicology Scheme. (NSAID = Non-steroidal anti-inflammatory drugs)

4.1.2 Returning Results

Participants log their results through an electronic reporting software, PORTAL, before the given deadline (around six weeks post-dispatch). Ten working days following round closure all participants' results are available for secure on-line viewing. Results not received before the deadline will not be included in the report but these laboratories can still access the report, and can calculate their own z-score using the programme available on PORTAL. Laboratories are permitted to submit more than one result (up to thirteen) in order to compare

the results obtained by different analysts or methods. Only three of these, however, can be nominated as official results to be included in the statistical analysis and report.

Additional test materials can be provided after each test round to allow participants to repeat testing as necessary, for example, to check the effectiveness of corrective actions.

4.1.3 Calculating Proficiency

The qualitative data (sample 1) is simply scored by correctly determining presence/absence of analytes. False positive identifications are noted but do not influence scoring. Correctly identifying the analyte present is considered a satisfactory performance, with each analyte scored individually. For quantitative samples, where a performance z-score is calculated for each analyte, values which are clearly in error are omitted from the overall statistics but are scored and reported as transcription to the final report is considered to be part of the proficiency assessment. A reported concentration of zero is regarded as unsatisfactory and is not accepted (the concentration should be reported as less than the lower limit of quantification). In common with other schemes a z score equal to or greater than three (or equal to or less than minus three) is considered unsatisfactory and a z-score greater than two (or less than minus two) may require attention. Scoring methods can vary depending on the nature and concentrations of the analytes.

4.1.3.1 True Value

The true value, termed the “assigned value”, is derived by one of three methods. The simplest is a formulation value, where the true value is known from sample preparation. An expert value, measured by an agreed reference laboratory could be used, otherwise a robust mean of the participants results will be used (from all methods unless the measurement is known to be method-dependent).

4.1.3.2 Acceptable Variation

Referred to in QUARTZ schemes as the 'standard deviation for proficiency assessment' (SDPA), it can take the form of a fixed value (which would be stated), a percentage of the assigned value, or robust standard deviation of the participant results.

4.1.3.3 Scoring Samples 2 and 3

Originally the SDPA was the robust standard deviation. From round 34, however, scoring was largely performed with SDPA as a percentage of the formulation value. The percentage used was dependent on analyte concentration, <0.1 mg/L = 25%, 1 to 10 mg/L = 20%, and >10 mg/L = 15%. These percentages were derived through a combination of in-house assessment to establish participant capability and expert consultation to determine fitness for purpose. Fixed fit-for-purpose SDPA with formulation value are preferred as it allows scores to be compared across rounds, and enables traceability and estimation of uncertainty. Where necessary the standard uncertainty of the assigned value will be incorporated in the scoring system, i.e. when the standard uncertainty of the assigned value is greater than 0.3 x SDPA. In these instances a z′ (z prime) score is calculated by Equation 4-1 and is interpreted following the same convention as used for z-scores. Uncertainties of the participant measurements are not considered by the scoring system but participants are provided with enough information to allow calculation of their performance using one of the metrological methods described by ISO [96] if they wish.

$$z' = (x-X) / \sqrt{(SDPA^2 + u_x^2)}$$

Equation 4-1: z prime score
Where: x = measured value, X = assigned value, and U_x = uncertainty of assigned value

If the analyte of interest is not found to have satisfactory homogeneity throughout the test material, this would be reflected by a more lenient SDPA. Acceptable homogeneity is defined as a sample variance, from duplicate measurements of ten randomly selected aliquots, which is less than a calculated critical value. This is based on IUPAC International Harmonized Protocol for the proficiency testing of analytical chemistry laboratories and meets the requirements of ISO/IEC 17043.[130]

4.1.4 Efficacy Measures

There must be eight or more participant results before analyte statistics and efficacy scores will be officially reported. These provide the percentage of satisfactory, questionable and unsatisfactory scores for each analyte and indications of the spread of the results (standard deviation, robust standard deviation, mean, median, etc.). In a well-behaved system there would be few outliers and the results would all be close to the assigned value (i.e. the robust standard deviation should be similar to acceptable variation), with no more than 5% unsatisfactory scores.

It is important to note that the second efficacy measure only conveys how closely the participants' scores were related. It is not an indication of trueness. If a number of laboratories operate similar methods which incorporate similar bias the results could be closely clustered around a value which is not the assigned value, but the variation in these would be close to that permitted and a good efficacy score would be achieved.

4.1.5 Review

At the end of each round of testing a review is conducted and any requirements for subsequent rounds are identified. Throughout its many years of operation this practice has progressed the refinement and enhancement of the Forensic Blood Toxicology PTS to produce the effective and popular accredited scheme now offered. Perhaps as a consequence of the evolutionary nature of the scheme, where assessment methods have been adapted as necessary in an attempt to maintain the truest reflection of actual laboratory performance, formal long term review has not been performed. Although participants are encouraged to monitor their own long term performance, it was considered that a long term review of the scheme would be beneficial, "to bring lessons learnt to the widest possible audience", in accordance with VAM Recommendation (f) 13.[48]

4.2 Aims of the study

The freely available unaccredited UNODC ICE scheme had been reviewed. It was unknown, however, how this performance compared to an accredited scheme. The aims were to assess the effectiveness of the Quartz PTS and compare this to ICE to determine whether ICE could develop to an accredited PTS scheme and what an appropriate scoring system might be.

4.3 Methods

In order to ensure participant anonymity, it was not possible to receive data in its raw format. Rather, data was extracted from the formal summary reports dispatched to participants following each round. Data from rounds 30 in 2007 to 48 in 2012 were collated in Excel spreadsheets.

As full processing of standard solutions (sample three) only commenced in round 39 there were only four rounds of results available, which was insufficient for long term review. This was also true of the blood alcohol tests which only have results from five rounds of testing. The review was therefore focused on performance for samples 1 and 2 with standard solution samples and blood alcohol tests addressed briefly in isolation.

As the z-score calculation for sample 2 had not always been performed using a percentage of the assigned value, scores could not be compared directly between rounds. Where possible they were recalculated with acceptable variation as a fixed percentage of the formulation value, with uncertainty considered when necessary. The percentage SDPA was unsuitable for samples which contained analytes at very low concentrations. As the most appropriate scoring could not be applied to these datasets the review was exclusive of these and also some datasets which had unknown, but increased, uncertainty.

Participant measurements had been plotted in the summary reports to demonstrate normal distribution, allowing application of z-score performance assessment. It was noted, however, that with less than twenty participants in each round deviations from a normal distribution would be difficult to identify.[132] Where the data was not normally distributed around the assigned

value it could be suggested that there was a problem with the assigned value and these datasets were excluded (Table 4-2 details excluded data and Table 4-3 and Table 4-4 provide two examples of why data was excluded). Table 4-5 lists the sample 2 datasets considered by the review with details of the scoring originally applied to each analyte in the QUARTZ report and the changes made to allow round-to-round comparison.

Substance (mg/l)	Round	Assigned Value	SDPA	Homogenous	Comments
Zopiclone 0.12	35	Median -0.063	25%	Pass	Possible degradation
Thioridazine 0.7	40	Median -0.54	20%	Pass	Possible degradation
Metformin 115.4	43	FV	15%	Pass	Small dataset
Phenytoin 15	44	Median (12.35)	15% of median + uncertainty	Pass	Possible degradation
Phenazepam 0.46	48	0.434	20% of 0.434	Pass	Possible degradation. It is unclear why the FV was changed to 0.434. There is no clear explanation for this
THC-COOH 0.005	39	FV	Robust SD	Fail	Low Level
Methamfetamine 0.00487	47	FV	Robust SD + uncertainty	Pass	Low level
6-MAM 0.05	42	Median -0.02	Rounded from Robust SD 0.007 to 0.01	Pass	Low Level: Rounding of the SDPA to 0.01 has an impact on the number of participants with satisfactory, questionable and unsatisfactory results
Norfluoxetine 0.05	47	FV	Robust SD	Pass	Low level

Table 4-2: Datasets not considered in the long term review

Formulation Value		0.12 mg/l
Median		0.063 mg/l
Participant Values	Z-score with SDPA 25% median	Z-score with SDPA 25% FV
0.014	-3.27	-3.53
0.03	-2.2	-3
0.04	-1.53	-2.67
0.047	-1.07	-2.43
0.047	-1.07	-2.43
0.05	-0.87	-2.33
0.06	-0.2	-2
0.06	-0.2	-2
0.06	-0.2	-2
0.07	0.47	-1.67
0.091	1.87	-0.97
0.1	2.47	-0.67
0.15	>4	1

Table 4-3: Round 35, Zopiclone results. (Cautionary scores, Unacceptable Scores.) **Formulation value should have been 0.12 mg/l but quantifications from a majority of participants were lower. The advisory group commented that some degradation may have taken place. The extent of the degradation, and whether it was equal across all samples, was unknown. The robust mean was, therefore, used as the assigned value. Assigned variation was 25% of this, an extra 5% to acknowledge the increased uncertainty associated with the assigned value. Some participants were allocated unacceptable or cautionary performance when they actually had measurements closer to the intended formulation value than those participants who achieved the best scores. The participants who appear to perform best actually measure the analyte of interest at half the level originally present. It is not known whether what is being assessed is the laboratory measurement performance or degree of degradation across the laboratories.**

Formulation Value		0.7 mg/l
Median		0.054 mg/l
Participant Values	Z-score with SDPA 20% median	Z-score with SDPA 25% FV
0.31	-2.13	-2.23
0.33	-1.94	-2.11
0.39	-1.39	-1.77
0.4	-1.3	-1.71
0.5	-0.37	-1.14
0.52	-0.19	-1.03
0.56	0.19	-0.8
0.6	0.56	-0.57
0.63	0.83	-0.4
0.66	1.11	-0.23
0.99	4.17	1.66
1.05	4.72	2

Table 4-4: Round 40, Thioridazine. (Cautionary scores, Unacceptable Scores.) **Almost all participant results were lower than the formulation value. Degradation may have occurred, although no comment was given to this effect. The SDPA was not increased in recognition of inflated uncertainty although the mean was used as assigned value in acknowledgement that the formulation value may not have been correct. The values listed illustrate how differently participants would have scored had the formulation value with an increased SPDA of 25% been used.**

Substance (mg/l)	Round	Assigned Value	SDPA	Homogenous	Recalculated	Comments
Chlordiazepoxide 1.1	30	FV	Robust SD	Pass	20%	
Dihydrocodeine 0.2	30	FV	Robust SD	Pass	20%	
Lamotrigine 10.13	31	FV	Robust SD	Pass	15%	
Sildenafil 5.05	32	FV	Robust SD	Pass	20%	
Citalopram 0.18	32	FV	Robust SD	Pass	20%	
MDMA 6.0	33	FV	Robust SD	Pass	20%	
MDA 0.08	33	FV	Robust SD	Pass	25%	
Temazepam 0.18	34	FV	20%	Pass		
Amfetamine 0.60	34	FV	20%	Pass		
Olanzapine 0.10	35	FV	20%	Pass		
Chlorpromazine 0.25	35	FV	20%	Pass		
Venlafaxine 0.40	36	FV	20%	Pass		
Quetiapine 0.80	36	FV	20%	Pass		
Tramadol 0.50	37	FV	20%	Pass		
Methadone 0.80	37	FV	20%	Pass		
Benzoylecgonine 0.60	38	FV	20%	Pass		
Cocaine 0.20	38	FV	20%	Pass		
Methylphenidate 0.10	38	FV	25%	Pass		
Methadone 0.85	39	FV	20%	Pass		
Amisulpride 0.3	40	Median (0.40)	20%	Pass		Acceptable - May be error with FV. All participants measured greater than the FV. The SDPA has been calculated acceptably
Oxycodone 0.2	41	FV	20%	Pass		
Sertraline 0.5	41	FV	20%	Pass		
Codeine 0.05	42	FV	20%	Pass		
Morphine 0.13	42	FV	20%	Pass		
Risperidone 1.8	43	FV	20%	Pass		
Zolpidem 0.1	44	FV	25%	Pass		
BZP 0.25	45	0.26 mg/l	20% 0f 0.26	Pass	FV and 20%	0.26 mg/l may have been selected as it's the median for the GC-MS group. At 5 of 15 this was the most prevalent method. It is unclear why the FV was not used
Mephedrone 3.8	45	FV	20%	Pass		
Clomipramine 0.228	46	FV	20%	Pass		
Flecainide 0.100	46	FV	20%	Pass		
Propanol 0.100	46	FV	25%	Pass		
Fluoxetine 0.128	47	FV	Robust SD	Pass	20%	Unclear why Robust SD was used
Cyclizine 0.098	48	FV	25%	Pass		
Propoxyphene 0.206	48	FV	20%	Pass		

Table 4-5: Result sets considered for long term review.
(FV = formulation value, Robust SD = robust standard deviation)

4.4 Results

4.4.1 Participation

Year	Round	Number of Applicants	Number of Participants	Labs returning Sample 1		Labs returning Sample 2		Labs returning Sample 3		Labs returning Sample 4	
2007	30	20	18	17	94%	17	94%	-	-	-	-
2008	31	20	17	15	88%	14	82%	-	-	-	-
2008	32	21	18	17	94%	17	94%	-	-	-	-
2008	33	20	18	15	83%	17	94%	-	-	-	-
2008	34	19	18	14	78%	18	100%	-	-	-	-
2009	35	19	16	14	88%	16	100%	9	56%	-	-
2009	36	20	18	18	100%	16	89%	-	-	-	-
2009	37	20	17	16	94%	17	100%	10	59%	-	-
2009	38	?	17	15	88%	17	100%	-	-	-	-
2010	39	?	19	13	68%	16	84%	11	58%	12	63%
2010	40	?	16	15	94%	14	88%	-	-	-	-
2010	41	?	21	19	90%	18	86%	12	57%	10	48%
2010	42	?	18	18	100%	18	100%	-	-	-	-
2011	43	?	17	15	88%	12	71%	11	65%	9	53%
2011	44	?	19	18	95%	19	100%	-	-	-	-
2011	45	?	20	18	90%	16	80%	13	65%	8	40%
2011	46	?	19	16	84%	18	95%	-	-	-	-
2012	47	?	19	17	89%	17	89%	11	58%	8	42%
2012	48	?	19	19	100%	18	95%	-	-	-	-
TOTAL			344	309	90%	315	92%	77	60%	47	49%

Table 4-6: Participation from rounds 30 to 48
The percentage of laboratories returning results for each sample is given as the proportion of laboratories returning results i.e. the number of participants. For samples 3 and 4 the percentage total is calculated only for those rounds which included these samples.

Table 4-6 details the participation rate of the QUARTZ PTS. Rounds with reduced participation may simply be the consequence of laboratories returning results after the deadline for inclusion in the summary report. This cannot be ascertained. As a consequence of laboratories not having a permanent unique laboratory reference number and in the absence of a look-up table for each round, it was impossible to monitor individual laboratory participation. The number of consistent participants could not be identified nor could it be identified when laboratories opted out of, or joined, particular rounds, although data was available for later rounds but not for the entire period of the review. A slight dip in general performance could be expected with the enrolment of new participants. This anonymised review precluded true assessment of the relationship between participation and performance.

Round	Samples Tested by Participants															
	1 Only	2 Only	1 + 2	3 Only	1 + 3	2 + 3	1 +2 + 3	4 Only	1 + 4	2 + 4	3 + 4	1 + 2 + 4	1 + 3 + 4	2 + 3 + 4	1 + 2 + 3 + 4
30	1	1	16												
31	3	2	12												
32	1	1	16												
33	1	3	14												
34		4	14												
35	1	1	6			1	8								
36	2	1	15												
37		1	6				10								
38		2	15												
39		1	1	1			5	3			1	3			4
40	2	1	13												
41	1	1	2				7	1	1			3			5
42			18												
43	1		1		1		4	2				2	1		5
44		1	18												
45			4		1		7	2				1	1		4
46	1	3	15												
47	1		4				6	1				2		1	4
48	1		18												

Table 4-7: Participation based on sample/analysis type
(1 = identification, 2 quantitative, 3 standard solution and 4 blood alcohol concentration).

Participation did appear to be fairly stable, however, ranging from 16 to 21 participants returning results each round (an average of 18). Quantification samples had the highest participation rate, very closely followed by qualitative samples. It was interesting to observe the distribution of participation across all four sample types (Table 4-7). Some laboratories did not participate with all sample types, which could have reflected routine practice. For example, performing the quantification sample 2 analysis but not the sample 3 standard solution.

4.4.2 Analyte frequency

In order to study performance it was necessary to identify analytes which were repeatedly tested to allow comparison of performance between each test. Table 4-8 details the frequency of analyte testing. Only six substances were tested three or more times across all sample types (amphetamine, benzoylecgonine, cocaine, codeine, methadone and morphine). Only methadone has been included for quantification more than once. The most tested analyte

group was opioid analgesics, followed by central stimulants then antidepressants. This was due to the large number of analytes in the test panel.

4.4.3 Qualitative Sample 1

In earlier rounds the summary reports listed results for each substance individually with the method given. In later rounds substances were merged and the analytical method was the focus. If a laboratory had used more than one method to identify all of the drugs present, an official result was counted against each analyte for each of the methods used. This unfairly influenced the percentage of correct results. With more than one analyte present it was often necessary for laboratories to apply more than one analytical method to the same sample, perhaps using different instrumentation. Each method would only be relevant to the detection of particular analytes although all instrumental methods applied were recorded for all analytes and it was not possible to distinguish which was the relevant method for each particular analyte. Thus, in the official LGC Quartz reports it was apparent that instrumental methods which had not been applied to analytes were recorded to have failed to detect the analyte. ‘Not detected’, rather than ‘not used’ was recorded and counted for each of these ‘irrelevant’ methods in the official round summary. Had it been a routine analysis the laboratory would have reported the correct content to their client as each analyte was detected by the relevant method. For this reason, where more than one instrumental method was recorded against the same laboratory reference number, this was counted as one result (with no “failed to detect” unless they did not detect the analyte by either method).

Analyte Group	Number of Times Tested	Analyte	Number of Times Tested	Qualitative Test	Quantitaive Tests	Standard Solution
Anaesthetic	1	Ketamine	1	1		
Analgesic	2	Paracetamol	2	1		1
Anti-arrhythmic	1	Flecainide	1		1	
Anticonvulsant	2	Lamotrigine	1		1	
		Phenytoin	1		Excluded	
Antidepressant	11	Fluoxetine	1	1		
		Mirtazepine	1	1		
		Sertraline	1	1		
		Trazodone	1	1		
		Citalopram	1		1	
		Clomipramine	2		1	1
		Fluoxetine	1		1	
		Norfluoxetine	1		1	
		Sertraline	1		1	
		Venlafaxine	1		1	
Antidiabetic	1	Metformin	1		Excluded	
Antihistamine	3	Cyclizine	1	1		
		Promethazine	1	1		
		Cyclizine	1		1	
Antipsychotic	6	Amisulpride	1		1	
		Chlorpromazine	1		1	
		Olanzapine	1		1	
		Quetiapine	1		1	
		Risperidone	1		1	
		Thioridazine	1		Excluded	
Calcium antagonist	1	Amlodipine	1	1		
Central stimulant	21	Amfetamine	4	2	1	1
		Benzoylecgonine	3	2	1	
		BZP	2	1	1	
		cocaine	4	2	1	1
		MDA	2	1	1	
		MDMA	2	1	1	
		Mephedrone	2	1	1	
		Methamfetamine	1		1	
		Methylphenidate	1		1	
Hypnotic	8	Amobarbital	1	1		
		GHB	1	1		
		Secobarbital	1	1		
		Temazepam	2	1	1	
		Zopiclone	2	1	Excluded	
		Zolpidem	1		1	
NSAID	2	Diclofenac	1	1		
		Ibuprofen	1	1		
Opiod analgesic	23	Buprenorphine	2	2		
		Codeine	3	2	1	
		Dihyrocodeine	2	1	1	
		Dipipanone	1	1		
		Fentanyl	1	1		
		Methadone	3	1	1	1
		Morphine	6	3	1	2
		Oxycodone	2	1	1	
		Propoxyphene	1		1	
		Tramadol	2	1	1	
		6-MAM	1		1	
Psychotropic	1	THC-COOH	1		1	
Tranquilliser	7	Chlodiazepoxide	2	1	1	
		Demoxepam	1	1		
		Desmethyldiazepam	2	2		
		Diazepam	1	1		
		Phenazepam	1		Excluded	
Vasodialator		sildenafil	1		1	
β-Blocker		Propranolol	1		1	

Table 4-8: Analyte Testing Frequency: (substance groups not tested: Anticholinergic, Bronchodilator, Opioid Antagonist and Carboxyhaemoglobin)

A total of 44 qualitative datasets were included for review. Results from only six satisfy the first efficacy measure (EM_a) of >95% correct identifications. The number of participants in each round, however, often meant that each participant accounted for greater than 5%, so even one error caused the dataset to fail the criterion. 85% was therefore considered a more realistic EM_a aim (this would generally allow for two participants to have false negatives or an incorrect identification). Less than half the datasets (19) achieved this. Figure 7 depicts by analyte type and round, the percentage of laboratories with correct detection. There was no indication that performance improved with time. Instead there was a suggestion of a general increase in the occurrence of false negatives. In many datasets there were unacceptable numbers of participants failing to identify the analyte present. This was given further attention in this review to determine whether the failure to detect was random or if there were discernable influencing factors.

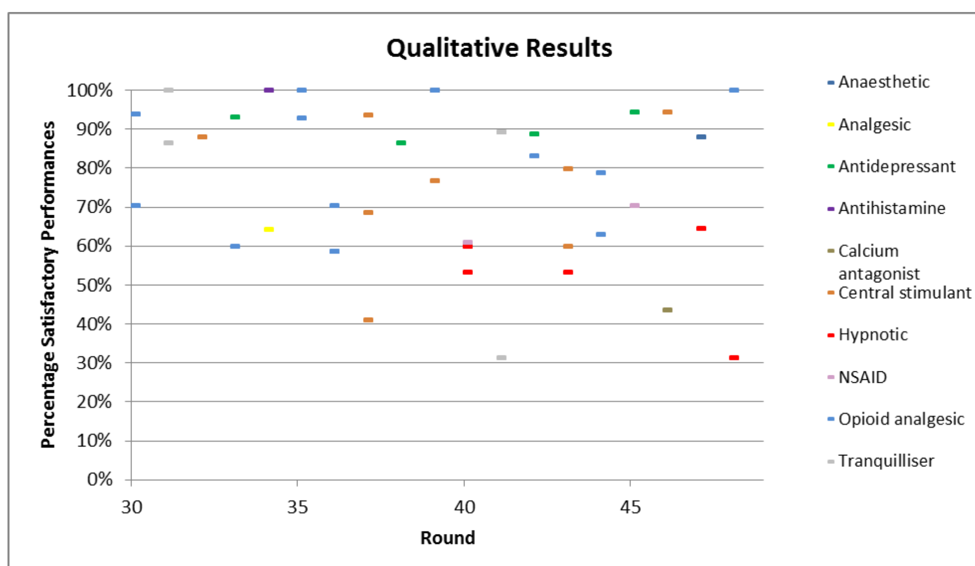


Figure 7: Percentage of participants each round that correctly identified an analyte

Of the ten analyte groups, four were considered to exhibit acceptable performance: anaesthetics, antidepressants, antihistamines and tranquilisers. There was, however, one exception in the tranquilisers group, demoxepam, 0.36 mg/l in round 4, which was detected by less than half of the participants (31.6%). Demoxepam is a metabolite of chlordiazepoxide with which laboratories performed better. It would, therefore, appear that participants did not fail to detect the metabolite but, rather, may not have performed the analysis as this analyte was not part of their routine test battery (which was true

for Forensic Medicine and Science, Glasgow University). The presence and satisfactory detection of cocaine in the samples which contained benzoylecgonine could perhaps explain the apparently poor performance with benzoylecgonine, which is a metabolite of cocaine. It is possible that benzoylecgonine was detected but only cocaine was reported. Benzoylecgonine is the main analyte in cocaine cases and is expected to form part of routine testing. This explanation could also be extended to MDA, a metabolite of MDMA. MDMA was present and satisfactorily detected in the sample for which MDA detection was poor. MDA is expected to form part of routine testing as it is a drug in its own right, as well as a metabolite of MDMA. Thus MDA may have been detected but only MDMA reported.

Of the seven analytes grouped under central stimulants cocaine and MDMA were the only two to demonstrate satisfactory detection. The detection of both BZP and mephedrone was very poor. These were relatively new drugs and it is assumed they were not detected as they did not form part of the participants' routine test battery. In addition they would not be detected by immunoassay screening and standards were not widely available. Of greater concern was the poor performance for amphetamine in round 43. It was unclear why three laboratories did not recognise amphetamine in the sample. Amphetamine was included in a subsequent sample and a good performance was recorded. Amphetamine also scored well in quantitative rounds and there did not appear to be any continuing cause for concern.

Table 4-9 outlines all of the analytes for which there were poor detection levels. Six (highlighted) require attention. There was no clear explanation as to why these drugs had poor detection. The morphine results were concerning, although there were other analytes in the opioid analgesics group with acceptable performance: dihydrocodeine, oxycodone and tramadol.

Year	Round	Analyte	Comments	Participants failing to detect
Central stimulant				
2008	32	Benzoylecgonine 0.5 mg/l	Metabolite	39.4%
2010	39	Benzoylecgonine 25 mg/l		23.1%
2009	37	BZP 0.50 mg/l	New	58.8%
2011	43	Mephedrone 0.70 mg/l	New	40.0%
2009	37	MDA 0.08 mg/l	Metabolite	31.3%
Analgesic				
2008	34	Paracetamol 6 mg/l	Possibly not reported due to low, therapeutic level	35.7%
Calcium antagonist				
2011	46	Amlodipine 0.87 mg/l		56.3%
Hypnotic				
2010	40	Amobarbital 5.72 mg/l	Barbiturates may not be routinely screened as use has declined	46.7%
2010	40	Secobarbital 5.83 mg/l		40.0%
2011	43	GHB 491.4 mg/l	Participants may not routinely screen for this drug	46.7%
2012	47	Temazepam 0.05 mg/l	Metabolite of diazepam but should have been detected	35.3%
2012	48	Zopiclone 0.20 mg/l		68.4%
NSAID				
2010	40	Diclofenac 20.0 mg/l	These drugs may not be included in routine screening	40.0%
2011	45	Ibuprofen 20 mg/l		38.9%
Opioid analgesic				
2009	36	Buprenorphine 0.05 mg/l	Detection expected to be low; it is positive that it has improved on the second testing	41.2%
2011	44	Buprenorphine 0.02 mg/l		36.8%
2009	35	Codeine 1.0 mg/l	Round 35 included other opiates and codeine might not therefore have been a target. Performance drops in Round 42 as there are no opiates present and the background directed the analysis towards antidepressants	0.0%
2010	42	Codeine 0.15 mg/l		15.8%
2009	36	Dicocal(Cyclizine/Dipipanone)		23.5% / 29.4%
2008	33	Fentanyl	Not commonly encountered in the UK, detection not expected to be good. Incidence of occurrence including this analyte served to raise awareness.	40.0%
2011	44	Methadone 0.15 mg/l		21.1%
2007	30	Morphine 0.10 mg/l		29.4%
2009	36	Morphine 0.10 mg/l		41.2%

Table 4-9: Analytes with poor detection: those highlighted require attention

There were a total of 30 false positives reported across rounds 30 to 48 (listed in Appendix 1). No particular problems were apparent from these results. Often the correct analytes were identified in addition to the false positive. It was therefore possible that these false positives were the result of a contamination event which the quality system did not signal. Alternatively, some may have been present in the blood matrix and the blank blood was not screened for every analyte. It was expected that participants had investigated any reported false positive results and the assumption was made that they were one-off random errors, as it was not possible to determine whether any participants reported multiple false positives and could have a true quality issue. However, this was possible for later rounds and there were no apparent problems.

4.4.4 Quantitative Sample 2

Performance was better for the quantification samples, perhaps because participants were advised which analytes were present. For a limited number of rounds, the organisers provided a key to link laboratory reference numbers through rounds (but did not give identification). There were four rounds in which it could be identified that a laboratory was not a regular participant (appearing to have submitted results in one or two rounds only). These laboratories had achieved very high scores, which influenced the perceived performance of dedicated participants. The results from these laboratories were therefore omitted from the review.

There were a total of 34 analyte quantifications (suitable for comparison), mapped by round for proportion of acceptable z-scores in Figure 8, and second efficacy measure (EM_b) in Figure 9. It was not possible to identify any definite improvement over time, although there was a general increase in the number of datasets for which participant results exceeded perceived fit-for-purpose performance.

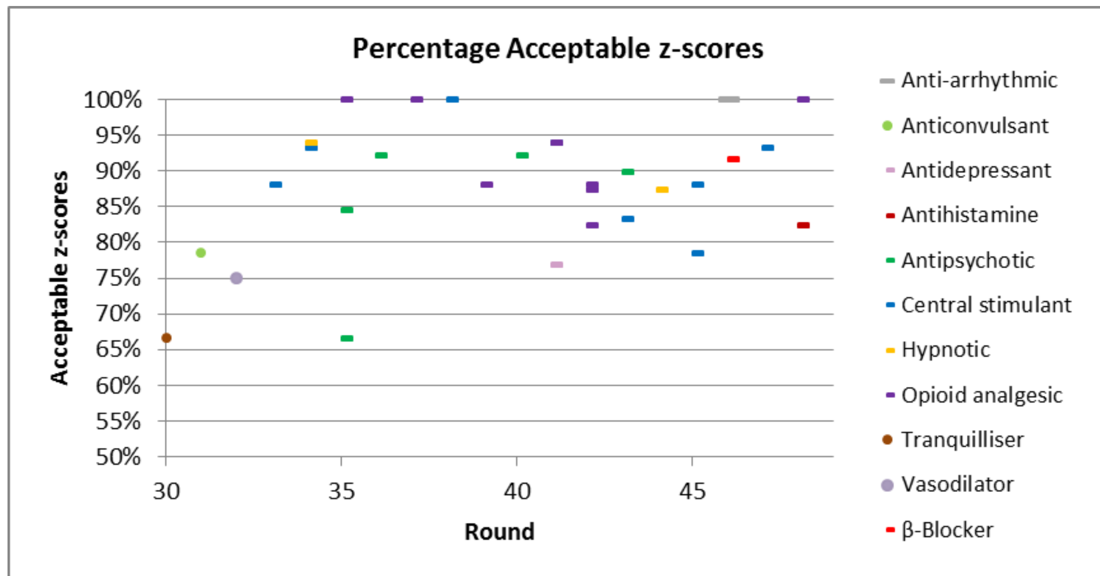


Figure 8: Percentage of acceptable z-scores for each sample 2 analyte in rounds 31 to 48

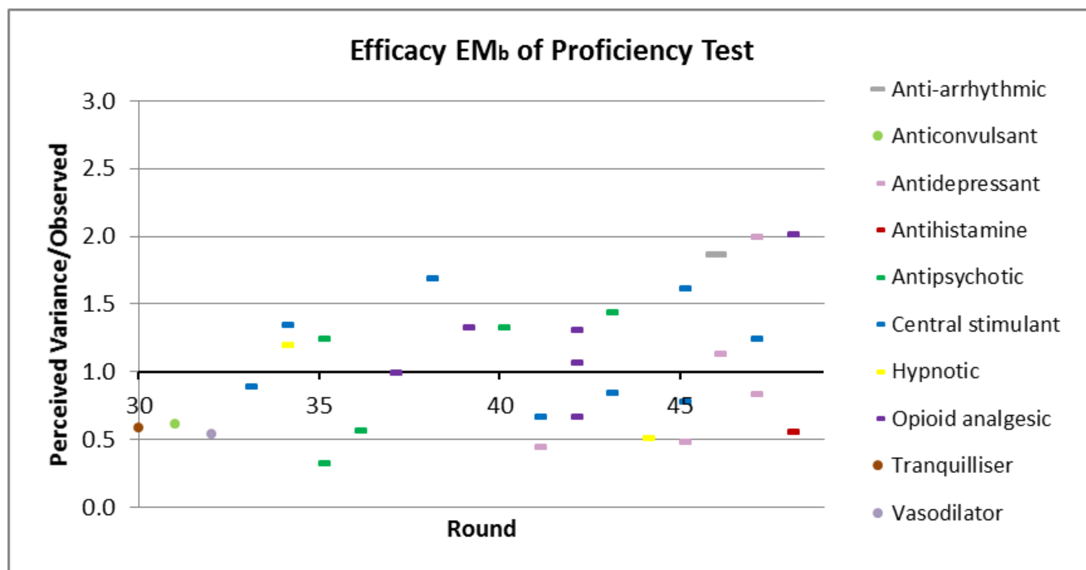


Figure 9: Efficacy values for each dataset by round.
Three values exceeded y-axis; β-Blocker Propranol (12.5 in round 46), and two opiate datasets tramadol and propoxyphene (7.14 and 6.98 in rounds 37 and 48)

28 datasets had less than 95% acceptable z-scores. Again, this appeared to be a consequence of the low number of participants, and a target EM_a of 85% was used. This is a generous figure as in a well behaved system the chances of falling outside of ±3z is remote (0.3% when standard deviation is used as acceptable variation),[87] although these figures are intended to describe large schemes. Twenty three datasets satisfied this efficacy target. Unfortunately, however, the remaining 11 exceeded this such that attention was necessary (listed in Table 4-10).

There are three analytes listed in Table 4-10 which, although they display poor quantification by participants, were not a real cause for concern. Clomipramine was at a low therapeutic level such that some participants may have considered its presence irrelevant within the circumstances of the case and therefore did not report it. Clomipramine had been included in the previous round as a sample 3 standard solution test with much better performance (with 95% of participants within acceptable limits). BZP and mephedrone were both quite new drugs to have become relevant to forensic toxicological analysis. Some participants advised this was not a drug they would routinely consider and therefore did not have validated methods for the analysis. It is, therefore, not surprising that performance was poor with these drugs.

Analyte Group	Number of times tested	Datasets with ≤15% Unsatisfactory z-scores	Datasets with ≥15% Unsatisfactory z-scores	Unsatisfactory scores			Comments
Anti-arrhythmic	1	1	0				
Anticonvulsant	1	-	1	Lamotrigine	21.40%	3 of 14	
Antidepressant				Sertraline	16.70%	3 of 18	
	5	3	2	Clomipramine	23.10%	3 of 13	Low therapeutic range
Antihistamine	1	-	1	Cyclizine	17.60%	3 of 17	
Antipsychotic	5	4	1	Olanzapine	33.30%	4 of 12	
Central stimulant				BZP	20.00%	3 of 15	New drug
	8	6	2	Mephedrone	16.70%	2 of 12	New drug
Hypnotic	2	1	1	Zolpidem	17.60%	3 of 17	
Opioid analgesic	8	7	1	Methadone	17.60%	3 of 17	
Tranquilliser	1	-	1	chlordiazepoxide	33.30%	5 of 15	
Vasodilator	1	-	1	sildenafil	25.00%	3 of 12	
β-Blocker	1	1	-				
Total	**34**	**23**	**11**				

Table 4-10: Participant performance with quantification samples

There were no clear explanations for the other analytes with poor performance. The result for the antipsychotic drug olanzapine was disappointing but most concerning was the very poor results for the benzodiazepine tranquiliser chlordiazepoxide (likely to be related to the instability of this analyte). Correct detection and quantification of this analyte is pertinent in forensic toxicological evaluation. The dataset for this analyte originated in round 30. This analyte was not included in any subsequent quantification samples and the adoption of effective corrective actions by participants, therefore, could not be explored. It

was promising, however, that when it was included as a sample 1 qualitative analyte in round 40 it was identified by 90% of participants.

No details of false positive results were reported as it was not expected that directed quantification samples would be screened for the presence of any other analytes.

4.4.5 Participant Methods

4.4.5.1 Qualitative Sample 1

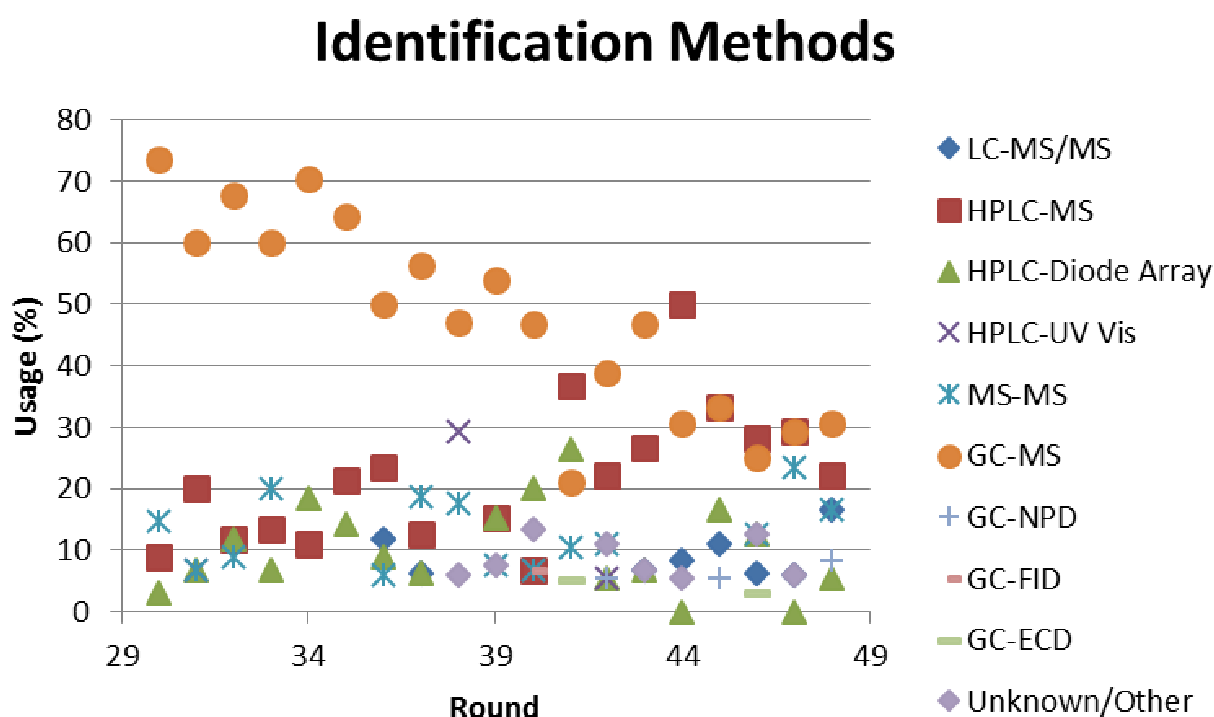


Figure 10: Methods used for the analysis of identification samples (test sample 1)

The methods used for identification are depicted in Figure 10 while a more detailed account is given in Appendix 2 indicating the percentage of unsatisfactory results for each method. It would appear that the use of GC-MS declined, although it was the most used method in almost every round. It was very difficult to study the effectiveness of particular analytical methods, as the reporting style of the indentification samples was such that it was not possible to differentiate between participants who had not attempted identification for certain analytes and those that tried but failed. For example, in round 36 the

datasets for dipipanone and cyclizine were very similar and the same laboratories failed to detect either of them, although they identified the opiates also present. This might suggest that the routine procedure for these participants did not include screening for basic drugs, although this is unlikely, or else they did not continue with further analysis having identified the presence of opiates. Of course, it is also possible that their methods were unsatifactory.

4.4.5.2 Quantitative Sample 2

A plot of the methods used in each round, Figure 11, indicates that GC-MS was the most used method overall, followed by HPLC-MS then HPLC-diode array. The general spread of methods applied in each round did not appear to have changed considerably over the review period. Participants were expected to have employed different analytical instrumentation from round to round in order to use the most suitable method for the particular analyte. The tables in Appendix 3 list the instruments used for quantification along with the number of unsatisfactory z-scores for each drug type. It was not possible to draw any conclusions as to the suitability or otherwise of particular instrumental methods.

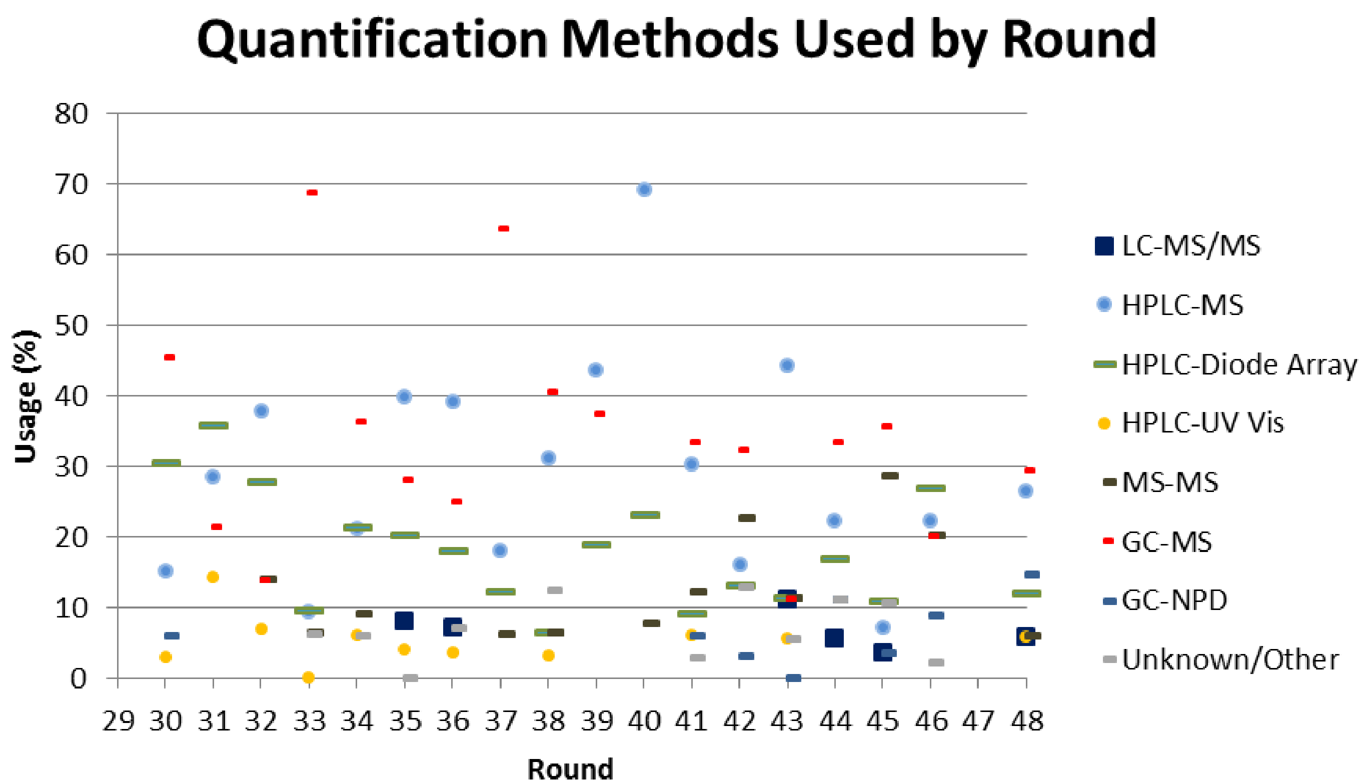


Figure 11: Methods used by participants in each round of quantitative testing

4.4.6 Standard Solution Sample 3

Standard solutions were introduced to the testing scheme at round 35. They were instigated on the advice of the advisory panel, with the intention that they could provide a diagnostic tool for the detection of the source of errors as a result of the wide range of z-scores often encountered for quantitative samples. These samples were intended to establish whether errors were introduced through pipetting, weighting, etc by reviewing the analysis sequence one step at a time. These samples were not in biological matrices in order to avoid matrix-derived variations. The first two samples provided were in methanol and did not require extraction, it was not until round 39 that analytes were provided in water and participants had to perform extractions. They were advised of a concentration range within which the analyte was present and asked to perform their routine extraction and analytical method to determine the concentration of the analyte. A summary of the participant results for the standard solution samples is outlined in Table 4-11.

Participation was relatively low for sample 3, especially when compared with the number of laboratories that quantified sample 2. Sample 3 was introduced in order to help laboratories improve their performance when quantifying. The analytes were, therefore, those most typically encountered in a forensic toxicology laboratory and the majority of laboratories should routinely analyse for these substances. It was, therefore, disappointing that this opportunity had not been fully recognised.

There were really only four rounds from which performance could be compared (shaded in Table 4-11) and, as each round contained a different analyte, varying methods may have been utilised in each, thereby hampering comparison. The general level of performance was slightly less than expected in a well-balanced system, although this again is expected to be the result of the low participation. With larger datasets these performance statistics may improve.

Substance (mg/l)	Round	Assigned Value	SDPA	Samples	Number of Sample 3 results	Number of Sample 2 Quantifications	Acceptable ($\le$-3 to $\ge$3)	EMb	Comments
Methadone 1.0	35	FV	20%	Standard in methanol, pipetting only	9	16	100.0%	0.06	
Morphine 0.2	37	FV	20%	Standard in methanol, pipetting and derivatisation	10	17	80.0%	0.57	Far outlying value (100 mg/l) removed to calculate efficacy
Morphine 1.0	39	FV	20%	Standard in water, pipetting, extraction and derivatisation	11	16	89.9%	1.33	
Amphetamine 1.0	41	FV	20%	Standard in water, pipetting, extraction and derivatisation	12	18	91.7%	1.28	
Cocaine 0.16	43	FV	20%	Standard in water, pipetting, extraction and derivatisation	11	12	81.8%	0.48	
Clomipramine 0.20	45	FV	20%	Standard in water, pipetting, extraction and derivatisation	13	16	76.9%	1.08	
Paracetamol 0.05	47	FV	RobustSD (8%)	Standard in water, pipetting, extraction and derivatisation	11	17	81.8%		This was a low level sample so robust standard deviation was used as SDPA.

Table 4-11: Performance in standard solution samples
Those highlighted are the only analytes which were suitable for comparison (prepared in water for extraction and with a perceived acceptable variation)

4.4.7 Alcohol Sample 4

The most recent version of the scheme description outlined how scoring of alcohol results was carried out.[131] For formulation values less than or equal to 100 mg of alcohol per 100 ml of blood, the SDPA is fixed at 3 mg/100ml. When the alcohol concentration is in excess of 100 mg per 100 ml of blood the SDPA is 3% of the formulation value. Previously, the alcohol SDPA was the robust standard deviation, preventing comparison between rounds. Z-scores were recalculated for the five alcohol data sets (rounds 39, 41, 43, 45 and 47) using the defined acceptable variation values rather than robust standard deviation (Table 4-12). Round 39 was unusual as this was the only round for which all participants returned results greater than the formulation value. In other rounds they were almost all lower than the formulation value, which is to be expected given the volatility of alcohol. Alcohol samples were not subjected to the same quality control measures as the quantification samples and there may have been an undetected error in the formulation value. Overall, participants completing alcohol testing performed well. The only slight regret was the rather low, and apparently declining, participation.

Round	Formulation Value (FV) mg/100ml	Range mg/100ml	SD:RSD	Number of participants						
				Total	Satisfactory		Questionable		Unsatisfacotry	
39	100	100.00 - 110.00	0.83	11	10	91%	1	9%	0	0%
41	89.9	81.00 - 90.69	1.46	10	7	70%	1	10%	2	20%
43	72	55.00 - 72.20	1.89	9	7	78%	1	11%	1	11%
45	55	48.88 - 56.10	0.89	8	8	100%	0	0%	0	0%
47	60	47.00-60.00	1.17	8	7	88%	1	12%	0	0%

Table 4-12: Results of alcohol in blood proficiency testing

4.5 Summary

There were twenty analytes which had been tested more than once across any of the three forensic toxicology sample types. The majority were included once for detection and once for quantification. Good performance across both quantification and detection was observed with only six analytes (cocaine, MDMA, codeine, dihydrocodeine, oxycodone and tramadol).

There were two analytes (chlordiazepoxide and sertraline) where detection was acceptable but quantification could have been better. Conversely, five analytes

(amphetamine, benzoylecgonine, fluoxetine, MDA and temazepam) were well quantified with poor detection. Although poor performance for detection of amphetamine was displayed in round 43, it had markedly improved in round 46 such that it was very good. Benzoylecgonine was also included for detection in two rounds. Unfortunately the improvement observed for benzoylecgonine from round 32 to 39 was only slight such that an inadequate number of participants detected the analyte. It would be useful to include this analyte in a subsequent round as there was some improvement on the second testing. It is hoped that in a third round of testing performance would reach a more acceptable level.

Other analytes presented poor performance in both qualitative and quantitative testing. The antihistamine cyclizine was the only analyte for which this could be concerning. For other analytes the difficulty had been anticipated but the analytes were included to contribute to the educational culture of the PTS, for example, BZP and mephedrone, which were new to many laboratories. These analytes, however, should now be re-tested.

Results suggested that some participants may have problems with the detection of methadone and morphine and also their quantification of methadone could be improved (Table 4-13). Sample 3 results were considered for these analytes in order to increase the information available given the concerns over performance. Over three rounds of qualitative testing the results for morphine have been varied, going from poor to acceptable to bad in round 36. This could indicate that either there is a problem with the detection of morphine, that this is a difficult analyte to detect, or that perhaps participants are unaware of when to test for this substance. Qualitative testing is based on a scenario from which participant laboratories should identify which drugs are relevant; this could be the issue rather than an analytical problem. Although, it is expected that as part of the laboratory STA samples would be subject to an immunoassay screen which should identify the possible presence of opiates.

Year/Round		Analyte	Test	Participants with Unsatisfactory Performance
2009	35	Methadone 1.0 mg/l	Standard Solution	0.0%
2009	37	Methadone 0.80 mg/l	Quantification	17.6%
2010	39	Methadone 0.85 mg/l	Quantification	12.5%
2011	44	Methadone 0.15 mg/l	Qualitative	21.05%
2007	30	Morphine 0.10 mg/l	Qualitative	29.41%
2009	35	Morphine 0.10 mg/l	Qualitative	7.14%
2009	36	Morphine 0.10 mg/l	Qualitative	41.18%
2010	42	Morphine 0.13 mg/l	Quantitative	0.0%
2009	37	Morphine 0.20 mg/l	Standard Solution	20.0%
2010	39	Morphine 1.00 mg/l	Standard Solution	9.1%

Table 4-13: Performance with methadone and morphine

The quantitative results for morphine are more promising. The first round of testing quantitatively was as a standard solution in round 37, where performance was poor, with 20% of participants having a z-score of 3 or more. In the next round this improved and on the third time of testing morphine quantitatively in round 42 the results were excellent, with all participants receiving acceptable z-scores. This is what proficiency testing aims to achieve. The difficulty for this particular review, however, is that it cannot be confirmed that the participants have improved their morphine quantification as we cannot be sure which laboratories are participating in each round. The improvement in the statistics could possibly be the results of the laboratories that had difficulty with morphine not submitting results.

Methadone and morphine are both forensically relevant analytes expected to be encountered often. These analytes should be tested frequently to ensure laboratories can safeguard these important analytical methods from error.

4.6 Discussion

4.6.1 Scheme Organisation

Given the information provided with the test samples, testing in the scheme could be considered directed. Sample 1 screening is the only undirected test. The list of drugs which could be present is known, however, and could,

therefore, prevent the reporting of false positives for drugs which are not listed. Now that participants have had some experience with this PT, the challenging, educational aspect could be maintained, and testing could be made more reflective of forensic toxicology routine practice, if the drugs to be quantified were not confirmed to participants. In addition there is an important aspect of proficiency which is not currently assessed. Participants' comments and interpretation are essential in casework. This is recognised by UKAS through interpretation being a feature which can be included in a laboratory accreditation scope. It is accredited separately from analytical methods and it should therefore form part of EQA for laboratories which are required to provide interpretation of their results. Having more participants include their interpretation for incorporation into the round summary report would also add to the educational benefits of the scheme. This would be particularly effective if participant laboratories were encouraged to involve more trainees/less senior members of staff to perform the analyses and also review the results, given that the scheme allows up to 3 sets of results to be returned. There is a wealth of information, covering a very wide range of drugs, which could benefit the education of new staff members, particularly as some of these drugs will not often be encountered through everyday casework.

There was only one recognised instance of a laboratory nominating more than one result. The ability to nominate multiple results, however, does not seem to be appropriate for the purpose of a PT which should monitor the performance that would be provided to clients by routine processing. The result which would be received by the client should be the only nominated result. Laboratories would still be able to use more methods or analysts for comparison purposes but rather than submitting these as nominated official results, they could be scored and fed back to the participant but not included in the summary report, or they could be scored and monitored by the laboratory using the on-line score calculator. Nominating multiple results does not represent the service received by customers who would only be provided one report. It is possible, however, that a client using a laboratory multiple times would receive results that have been determined by different operators each time. Submitting multiple results to the PTS allows the laboratory to monitor proficiency of staff.

Additionally it was evident from participant comments that they would often request standards or develop methods in order to perform testing of analytes that were not part of the service routinely offered. If a participant laboratory does not offer a particular analysis, there would be no expectation for them to be proficient. It is commendable and adds to the educational value of the PT that participants attempt to develop these methods and carry out the requested analyses. In such instances, however, it would be better if participants calculated their own performance score via the on-line calculator (or submit the result but not as a nominated value), rather than submitting an official result for scoring and reporting as this can complicate perception of peer performance. Laboratories routinely offering the service would lose the ability to monitor their level of proficiency amongst their peers i.e. other participants who should be proficient. When laboratories nominate results they would not routinely provide, it could falsely inflate the apparent performance of those other participants for whom the service is routine (and more importantly could have a significant influence on proficiency scores if participant results are used in the calculation).

The scheme needs to develop its reporting style to clarify whether a participant has claimed proficiency in a test for which they have provided results or whether they are attempting a non-routine analysis. Perhaps participants could be requested to indicate on the test menu which of the analytes they include in their testing service to clients. Then only results for these analytes would count as official results from the participant and be scored and included in the summary statistics.

It is also recommended that participants be requested to stipulate whether or not they have attempted an analysis. In some instances a ‘not detected’ may be recorded when in fact the correct determination is that the test material was ‘not analysed’ for the particular analyte. Again this influences the summary statistics and can provide unrealistic peer comparison. There are various reasons why an analysis might not be performed: one example could be that the sample volume did not permit the analysis. The extensive range of analytes potentially present requires a large number of different analytical methods to be employed in order to test for all possible analytes (although the testing should

be directed by the accompanying scenario). With only 10 ml of blood the number of these methods which can be utilised is restricted. In light of this, the scheme organisers may wish to consider reducing the number of potential analytes. It could be brought in line with the list of analytes with which the Home Office (through the College of Policing) request their laboratories to be proficient.

It was disappointing that some participants who would have been capable did not take part in sample 3 analyses. These participants missed the opportunity to refine practices essential for good quantification. Provided quantification scores are acceptable, however, it is understandable that laboratories under pressure of high workloads chose not to participate. Yet a larger peer group for comparison would benefit those wishing to monitor/improve these skills. There were also some participants who, in the same round, performed sample three quantifications but not sample two. This could, however, simply result from the sample two analyte not being one which the laboratory routinely measures.

4.6.2 Proficiency with Benzylpiperazine

It was good to observe the inclusion of the emerging psychoactive substances BZP and mephedrone, although performance was poor. BZP was included as both a qualitative and quantitative test. In round 37 participants were given a scenario which involved a fatality outside of a nightclub and advised that seizures were experienced by the subject. As this was a qualitative sample this dataset could reflect the number of participants who would be aware of these being symptoms of BZP and who were looking for BZP in routine casework. Unfortunately only seven of sixteen participants identified the presence of BZP. In round 45 BZP was included as a quantification sample and participants were, therefore, alerted to its presence and laboratories who would not routinely consider this analyte attempted the analysis. As the number of participants in this round was not known it was not possible to determine how many laboratories intended to participate but did not complete the analysis. The comments from the participants were interesting and it was difficult to know whether many of these laboratories would still have proceeded with BZP quantification if it had not been directed as part of a PT. One laboratory

reported that they would not quantify BZP as the concentration was too difficult to interpret due to the lack of available data. This is understandable, at the same time, however, it is by performing casework analyses that interpretation data can be collated (and perhaps then shared via PT schemes such as this).

Thus methods for the detection of new designer drugs should be implemented in forensic toxicology laboratories. UNODC, of course, aims to provide these methods as well as PT samples. Participants should develop, optimise and validate methods for the detection of BZP and mephedrone which are increasing in popularity.[133] Forensic toxicologists should familiarise themselves with the potential effects of these drugs in order that their possible use can be recognised from casework information. Including these drugs in PTSs which include a scenario describing typical circumstances of their use is, therefore, an excellent means of education to help increase familiarity with new drugs. BZP and mephedrone are not detected by routine immunoassay screening and their possible presence must be considered by the toxicologist to signal the requirement for analysis.

4.6.3 Performance Scoring

Perceived fitness-for-purpose was the most used method of calculating acceptable variation. The SDPA percentages were slighter higher than those generally suggested for analytical chemistry laboratories (5%).[99] This is in keeping with the recommendation that the expert steering committee prescribe a fit-for-purpose value for acceptable variation within the particular sector. Given poor sample quality and low analyte concentrations in forensic toxicology it was not surprising that higher percentages were required. The percentages used, however were determined by reference laboratories' performance in earlier rounds of the PTS. These percentages are perhaps too high. The lowest at 15% deems a measurement 30% from the true value as satisfactory and only those greater than 45% from the true value to be unacceptable. This does not appear to be appropriate for concentrations which could be given in evidence and should, therefore, be beyond reasonable doubt (with known uncertainty). At validation 15% from the true value is considered reasonable. This is only of analytes present at significant concentrations, however, as it is understood that

higher percentages will be necessary for some very low concentration analytes. In forensic toxicology some concentrations can be so low that a high percentage difference can be such a small value that it is a negligible difference. For example, at low concentrations there may be no significant toxicological difference between the true concentration of the drug and the values equal to this concentration plus or minus 50% of it (i.e. 0.05 mg/L, 0.1 mg/L and 0.15 mg/L may all be therapeutic concentrations).

Despite the high percentages in the Quartz scheme some analyte concentrations were still too low for application of a percentage SDPA method, and for some others, doubts that the formulation values were correct prevented use of a percentage SDPA. Robust calculations from participants' results were used in these instances. Whilst this is an approved method for performance assessment this data could not be included in the long term review as the datasets were not large enough to counter uncertainties or biases in the values: greater than thirty participants would be required.[102,132]

The necessity to exclude certain datasets for which the formulation values could not be used, highlighted the merit in stability testing. Although homogeneity testing was conducted, stability testing was not. Ideally the stability of test materials should be determined and the deadline for return of results fixed to the period for which the analyte remains stable. Where this has not been possible, and degradation is suspected, a test sample retained by LGC could be re-analysed to establish the degree of degradation which may have occurred. This could then be incorporated within the uncertainty for the formulation value and a score calculated which does not depend upon participant data. Proficiency scores calculated in such uncertain circumstances cannot benefit the participants. Where this increased uncertainty cannot be avoided or measured, it would be expedient to draw participants' attention to the unreliable nature of the assigned scores, and that a different scoring mechanism has been applied. Whilst it is always possible to determine the method used to calculate the proficiency score from the information contained within the summary report, this requires some participant endeavour. It would be helpful to participants if scoring mechanisms for each analyte could be more readily identified, to make

the significance of a performance score immediately apparent and ease monitoring of performance across rounds.

4.6.4 General Performance

Some allowances were made for poor quantification performances if the analyte was not one which would have immediate significance given the circumstances of the case or there was a low therapeutic level present which might not have been considered forensically relevant. It must be stressed to participants, however, that all analytes present are to be reported whether they are at therapeutic levels or not. This is especially important if there could have been a delay in recovering a sample and a large portion of the analyte has been lost through metabolism.

As PT scores reflect proficiency for substances experienced daily in the laboratory, the infrequently used analytes included in Quartz samples are a possible explanation for many of the poor detections. De Zeeuw [1] found that PT reflects proficiency for substances experienced daily in the laboratory; inferior quality is expected from less frequently performed methods. The analytes highlighted as problematic for detection should, however, be repeated in future rounds. Fentanyl should also have been repeated in a subsequent round to ensure participants had headed the advice of the advisory panel and were paying attention to increased occurrence of this drug. The UNODC World Drug Report also recently issued a caution with respect to the increasing popularity of fentanyl [133] and the 2011 joint report from EMCDDA and EUROPOL reported the presence of fentanyl in Scotland.[134]

There was a suggestion in participant performance that the proportion of laboratories failing to detect analytes had slightly increased over the period of the review. This could simply be a consequence of increasing numbers of analytes, some of which were emerging drugs for which participants did not have standards, validated methods, or a good general awareness.

There were not enough analytes repeatedly tested to provide an indication of the influence of participation in the scheme nor was there enough comparable

data to establish whether participating in this scheme improved participant performance. Review of the few analytes which have had repeated testing under the same sample type, however, does point toward repeated testing improving participant performance.

It is expected that an effective PTS will improve performance through time, which has been demonstrated elsewhere.[135,136] In this review, performance could not be assessed against time as constant participation could not be established; individual participation could not be monitored. Additionally, as test analytes were so varied, the usual pattern of performance being poorest in early rounds and improving in subsequent rounds before reaching a plateau was not expected to fit this scheme. This in itself, however, is a good indication of the effective educational function of the scheme. Having a large selection of test analytes reflects the true possibility of routine casework samples and also maintains a challenging aspect to the scheme. It would appear that satisfactory performance scores are not the target of this PTS, rather, the aim is for participants to satisfactorily take part and learn from the experience. After all, how a laboratory responds to an unsatisfactory result can infer better dedication to quality than a laboratory that consistently achieves satisfactory results.[94]

In order to fulfil the function of a PTS, and to be fit-for-purpose, repetitive testing is essential. The scheme offered in the rounds considered in this review, lacks the repetition necessary to guard against bias in the everyday measurements. It is not possible for such a varied scheme to effectively monitor routine performance unless the laboratory wishes to demonstrate proficiency in the ability to detect a wide range of drugs. Rather, a representative cross-section of analyses performed should be continuously monitored. As a PTS is intended to satisfy the requirement of external quality assessment, it should focus on comprehensive monitoring of the most common tests performed i.e. those with greatest volume of casework.

As these analytes are so frequently encountered a biased method will result in a large number of incorrect results being released to clients. It is, therefore, suggested that the sample three standard solution, which had poor uptake, be replaced with a test for standard analytes in blood. This should include the

analytes morphine, codeine, diazepam, cannabis, amphetamine, cocaine and methadone (over two rounds to ensure sufficient sample volume for all of the analyses, such that each analyte could be quantified at least once every year). With participants frequently testing for these analytes there would be no imposition. Participation in this part of the scheme should be as easy as simply adding the test sample to a routine batch for analysis. Repeatedly testing the same analytes would allow for a scoring system to be developed for any analyte expected to be present at very low level, with a fixed value for acceptable variation below a specified concentration (in-line with the alcohol scoring system). This would ensure continuous comparability and monitoring between rounds.

Scheme coordinators advised that participants were reluctant to seek advice with any of the testing conducted to date, although clearly some participants were in need of assistance. If repetitive testing is employed it will be clearly identified which analytes pose problems for participants (or the organisers may be able to highlight participants who struggle with particular analytes) and offer help in the form of a workshop which is offered to all participants rather than singling out and targeting laboratories. Monitoring of efficacy measures is redundant without repetitive testing. Effectiveness of the scheme cannot be assessed, other than establishing that the scheme is ineffective at the task of ensuring laboratories reach and maintain a defined target of performance. It would appear that the LGC Quartz forensic blood toxicology PTS format has only delivered one of its three aims promised to clients. Participants can compare their performance with peers, however there is no safeguarding against undetected errors, and there does not appear to be any evidence of participants being provided information on technical issues and methodologies.

4.7 Conclusion

A PTS should have clearly defined aims. The current aim of the QUARTZ Forensic Blood Toxicology PTS has education as a key factor, intending to motivate and support participants to improve their services. To provide laboratories a more effective external quality control service, which monitors method performance, the extent of the analytes tested could be reduced, and

testing should be more repetitive. As performance of participants could not be traced through time any participant specific difficulties or requirements could not be identified. It is therefore essential that participants continue to monitor their own performance over time (paying attention to scoring mechanism) and seek advice through the Technical Scheme Coordinator where they may have concerns.

Chapter 5: Comparison of ICE and QUARTZ EQA Schemes

5.1 Review

Critique of the schemes was restricted by features of both: ICE by incomplete data retrieval and inconsistent participation, and Quartz by inability to identify participants between rounds. In the main, it was not possible to assess individual participant performance. Reviews were, therefore, primarily focused on overall performance, laboratory proficiency in general, and competence with particular analytes (those which had been repeatedly tested by at least eight laboratories).

The nature of participation was a noticeable difference between the two schemes. It is understood from discussion with the Quartz co-ordinators that they have a regular group of participants with small fluctuations. The average number of laboratories to participate in Quartz was 18 (range 16-21), mostly from the UK and Europe. In contrast the ICE scheme is much larger, but participation much less regular, averaging 33 laboratories (range 22-65) from around the world. The limited anonymised laboratory identification data for later rounds of Quartz made it possible to trace some laboratories who were not regular participants (having submitted results in one or two rounds only). These laboratories had achieved very high z-scores which influenced the perceived performance of dedicated participants (and were omitted from the review, just as infrequent participants were not considered in the ICE review). This highlights the importance of being able to carefully monitor participation in association with performance statistics. Large amounts of random participation could significantly influence performance statistics, which is unfair to committed participants.

Quartz did have regular participation, yet it still did not satisfy the error detection function of a PTS. ICE testing was less frequent, but with more

specimens available for quantitative analysis in each round and a smaller test menu, more analytes were tested every year. Quartz had an extensive test menu and included fewer analytes tested every year, thereby accruing less analyte repetition. Analytes were not tested often enough for participants to effectively monitor their performance over time. Quartz round-to-round performance was randomly good or bad. The inclusion of analytes in test samples which were not frequently encountered in casework samples, and possibly did not feature in laboratory STA procedures was thought to explain the large numbers of Quartz data sets presenting less than acceptable identification. The inclusion of newly emerging drugs for which few participants would have validated methods and for which reference standards were not readily available also contributed.

A common problem in both schemes was participants not following a routine method. Laboratories must only participate in testing of analytes for which they offer a service to clients; if they do not offer an analysis they are not expected to be proficient. Submitting a result skews the statistical summary for a round and does not allow peers to know their performance relative to others who offer this service. The standard ICE results form has an option of ‘analysis not performed’ to make clear that there was no attempt to complete the test rather than the result being erroneously recorded as a false negative. The need for this clarity may have been readily perceived by the ICE coordinators because, as an international scheme, they were aware that they would encounter different laboratory protocols, and laboratories with different drugs of relevance as drug use trends and legislation differ between member states. The ‘not analysed’ option is something that other PTSs, including Quartz, should consider.

The impression given from the long term Quartz review was that laboratories were not achieving the target acceptable performance (although overall the Quartz participants displayed better performance than ICE). This was an unfair representation and was likely due to frequent inclusion of analytes that were not part of routine testing, rather than participants not having good day-to-day proficiency. The most important, high throughput analyses, were not being adequately monitored. Participating in the Quartz scheme did not allow for error detection in the most frequently used methods and, therefore, should not

be considered by the laboratory, or their clients, as demonstrating suitably reliable results. To provide a fit-for-purpose proficiency test to forensic toxicology laboratories repetitive testing of the most relevant analytes (e.g. morphine, codeine, diazepam, cannabis, amphetamine, cocaine and methadone) is essential. Only this can ensure laboratories safeguard these important analytical methods from error. As a result of the long-term review QUARTZ are now implementing such repetitive testing.

Only a small proportion of ICE laboratories returned quantifications, whilst quantification was the most popular sample type amongst Quartz participants. Regular ICE participants were shown to be more likely to perform quantification. Achieving more regular participants could help to improve the ICE quantification rate. Quantification may also be improved by ICE now returning a performance score to participants which previously it did not. Participants may not have believed it worthwhile to make the extra effort to quantify when they were not being scored. The resources available to many Ice participants makes quantification very difficult and in many jurisdictions it is not required, particularly as urine concentrations are not very meaningful. Scoring was prevented by the difficulty of assigning acceptable variation with participants of such varied abilities. For review it was set at 5% of the true value, based on validation bias being unacceptable greater than ±15%. This may have been too stringent. Acceptable variation with Quartz was much more lenient with participants determinations allowed to vary by up to 75% of the true value before they became unacceptable. This acceptable variation does not seem appropriate for measurements which could have serious legal consequences and, therefore, not fit for purpose for forensic toxicology laboratories.

The fact that ICE specimens were urine could also influence whether or not quantification was performed. In general terms, analyte concentrations in urine are not particularly useful. Blood is the preferred matrix for quantification and laboratories may not routinely quantify urine specimens. Laboratories regularly performing blood quantifications that do not have access to any other reliable form of EQA could perhaps reconstitute the urine specimens with blank blood, however, this could introduce more errors to their method and the analytes and their concentrations may not be relevant. It would be preferable if UNODC

could provide simply dried analytes at known concentration for blood reconstitution and analysis. The provision of blood samples would be impractical for such a large international scheme.

As Quartz provided blood samples and had regular participants this could explain greater uptake of quantification, however, it could also partly be due to the analytes being made known to participants; ICE testing was undirected. A PTS aims to bring participants towards a target level of performance, or to maintain this if already achieved. A plot of performance by round could, therefore, be predicted to appear as a fairly straight horizontal line within close proximity to what has been deemed acceptable. Yet, even with almost all laboratories regularly performing quantification, this prediction was not observed with Quartz quantifications. This is undoubtedly a result of the changing analytes across time. In contrast, repetitive testing in ICE enabled quantification with specific analytes, and for individual laboratories to be traced which highlighted various issues such as inappropriate methods, analytes of particular difficulty and laboratories consistently producing poor performance in need of assistance with their IQC. It is absolutely not a failing of a PTS to have laboratories that do not perform to the target standard; quite the reverse. It is reassuring to have these issues identified as UNODC can now provide technical assistance required by these laboratories for improvement. This is the true function of a PTS.

Whilst repetitive testing is essential in a PTS for error detection and to provide a measure by which corrective actions can be monitored, it can enhance educational aspects to occasionally feature some less frequently encountered analytes; particularly emerging drugs of abuse. Quartz was very strong on this educational feature. Laboratories that establish proficiency in their accredited methods elsewhere could benefit from the varied nature of this scheme to gain experience with analytes which are not encountered often or which are emerging substances of abuse. Rather than including these analytes in test specimens, ICE coordinators invite participants to report any new substances encountered at six months intervals and this information is readily shared.[129]

Whilst limitations of these schemes meant that it was not possible to determine whether participating in a PTS can improve performance over time, this has

been demonstrated elsewhere.[135,136] An aspect of PT which ensures improvement of performance is sharing of information and experiences amongst participants. Unfortunately it seems that this does not occur. It has been explained by the Quartz co-ordinators that commercial laboratories value confidentiality and, therefore, would not approach another laboratory for help as this could reveal them to be a laboratory which has performed poorly in a round. Ultimately many participants will be competitors and there is not the same motivation for developing together. A PTS introduced to allow 11 participants across Europe to collaborate on an oral fluid (OF) research project, where participants were simultaneously developing and validating their OF methods, provided very positive results. Participant performance improved with each round of testing and there was good CV% across the majority of the analytes tested.[136] This could be due to participants assisting each other with method development and problem solving in order that they all might work together to make good progress with their research. If laboratories will not communicate, it is the responsibility of PTS coordinators to interact effectively with participants to target those in need and those who perform well, and develop means to convey relevant information by a fashion which does not identify specific participants. ICE has been very good at approaching laboratories to address education of participants by developing and promoting the ICE Portal as a “global reference point and early warning advisory on new psychoactive substances”. Information such as trending observations, analytical methods, reference documents and mass spectra are shared.[137]

It was interesting that participants in both schemes, at some level, displayed difficulty with morphine analysis. Difficulty with morphine detection was also recognised in a published OF PTS review in which coefficient of variation of participants measurements reached 161%,[138] however, a subsequent PTS demonstrated laboratories to have very good proficiency with morphine.[136] The second scheme had fewer participants and they had been selected for a collaborative project presumably on the basis that they were laboratories with excellent reputations for quality. Morphine is an analyte of great importance and efforts should be made to more closely monitor this testing and assistance should be provided to those participants displaying difficulty.

ICE is not a true PTS and is not accredited as a PTS but was more effective as a means of external quality assurance. Still, it may not be recognised as a PTS for participant accreditation purposes. This was also true of the OF PTS set up with the clear and specific aim of allowing collaborative studies in the DRUID project.[136] It was not an accredited PTS but proved extremely successful in improving participant performance, such that they achieved target proficiency. The Quartz Forensic Blood Toxicology PTS is primarily an educational resource. It is a UKAS accredited PTS, yet does not provide participants an effective means of monitoring routine laboratory performance. As many of the participant laboratories are ISO 17025 accredited, it is expected that they have a supplementary source of EQA as this scheme is not appropriate to demonstrate the quality that should be attained by accredited laboratories. The QMS of an accredited laboratory should have served to highlight the failings of this scheme.

King *et al* [139] reviewed the influence of accreditation in PT performance. Only 48% of the participants accredited by the National Measurement Accreditation Service (NAMAS, now known as UKAS), CPA or ISO had a written procedure for unsatisfactory performance. They found no evidence to support accredited laboratories having better overall performance than non-accredited laboratories, although that study was carried out in 1999 before the introduction of the ISO17025 standard, which has greater analytical requirements. This is not surprising as accreditation ensures that a validation has been performed and has been demonstrated to suit the needs of the customer. It is not, however, concerned with the specific data and results generated. The requirement of accredited laboratories to have meticulous QMS did have some bearing, however, as recovery from a poor performance was generally faster amongst accredited laboratories. Laboratories without accreditation showed the highest long-term level of unsatisfactory results.

5.2 Conclusion

It is unlikely that a laboratory can find a PTS to mirror exactly all of its casework analyses, for example, there is no PTS which replicates drug driving samples with the specific road traffic vials. However, the requirements of the analysis should be as closely reproduced in a PTS as possible.

Accreditation does not provide evidence in support of the quality of a laboratory or PTS and should not direct which PTS a laboratory selects for EQA. It is the responsibility of laboratory management to ensure their EQA is appropriate for the purpose for which they are participating (safeguard against bias, familiarity with emerging substances, maintain competence with infrequently encountered analytes, etc.). This can be assessed with reference to the PTS protocol, for example, the frequency of analyte testing, the analytes included, participation, how the true value is defined and the acceptable variation. An unaccredited scheme which is more relevant to the laboratory caseload but which uses sound practices, is undoubtedly better than an accredited scheme which does not fulfil the laboratory's particular needs.

Although there are relatively few forensic toxicology PTSs, all laboratories have access to an effective means of bias detection for analytes of greatest forensic relevance, in the form of the UNODC ICE, although this is limited to analysis of urine.

Long term review ensures the effectiveness of a PTS scheme and highlights areas in which laboratories require improvement and guidance. This also requires that PT providers ensure that they employ a scoring system which realistically reflects performance. It is hoped that more providers are encouraged to more closely reflect the complex testing often required by 'forensic' laboratories; particularly testing of less traditional matrices and poor quality specimens.

For PTSs to operate effectively providers should equip themselves for long-term review by archiving the appropriate data, for example to allow round-by-round comparison and by being able to monitor individual participant performance in an anonymous way. And to get the full benefits of a PTS laboratories must be dedicated, participate regularly, and monitor their results to ensure the PTS is fulfilling the intended purpose.

It is not correct to assume good quality is present in all analyses, based on a good PT record.[48,91] It has been suggested that as most laboratories perform analyses for a range of substances and the PT cannot be all-encompassing, the selection of analytes incorporated should be as wide as possible to achieve a

representative cross-section of analyses performed.[48] The evidence from Quartz is that PTS should have a focused selection of analytes, at least some of the test samples should provide regular testing of core analytes, and, if necessary, the laboratory should participate in a selection of PTS to ensure enough of the work performed is represented. As new substances are routinely encountered these too should be incorporated by EQA. One such group of substances are the piperazines, which are the subject of the next chapter.

Chapter 6: Validation of a Method to Detect Piperazines in Blood

6.1 Piperazines

Piperazines have been available for many years and gained widespread familiarity from 2004 when they were aggressively marketed as safer, cleaner, less addictive alternatives to illegal dance drugs and introduced as 'legal highs'. Piperazine 'party pills', or piperazine-based social tonics (PBSTs), were portrayed as harmless natural highs said to elevate mood, increase mental capacity, alertness and energy. They became popular as 'dance' drugs but also appealed to shift workers, drivers and students.[140,141] At this time piperazines were legally available, although claims that piperazines were herbal, natural or harmless were erroneous.

The term 'piperazines', when used to describe a class of drugs, covers a range of piperazine-derived drugs. Those commonly encountered are depicted in figure 1. They occur in a host of shapes and forms, usually tablets of various colours, sometimes with impressed logos to look like ecstasy and less often as powders or liquids.[142,143,134]

Benzylpiperazine (BZP) given the street name 'A2', first emerged as a drug of abuse in the United States (1996) and Sweden (1999) [144-146] before seizures were progressively recorded worldwide,[140,147,148] including throughout the UK.[147] The former Forensic Science Service (FSS) in England received around four hundred piperazine cases per quarter (2008).[149]

The piperazine 'experience' has been likened to different ATS, though with lower potency.[150-153] It is understood that the optimum experience is achieved by combining different piperazines as this more closely mimics the effects of MDMA.[152-154] At the height of their popularity there were many blends available commercially, for example legal E, Legal X rapture, Frenzy, and Charge',

marketed on-line and in head shops as ecstasy alternatives, having stimulant and/or hallucinogenic effects.[155]

Originally party pills most often contained BZP, usually in combination with other piperazines, however, new varieties minus BZP became popular as BZP regulations were introduced. In many countries, including the UK until 2009, the ready availability of these substances, legally purchased rather than being 'scored', may have fostered a false sense of safety.[142,155] Legality caused users to wrongly assume that the safety of these substances had been tested and that they were quality assured with accurately labelled ingredients.[142] Naïve users, who might never have experimented with drugs were introduced to psychoactive substances simply because they were legally available.[156] It was feared that the use of these substances created a relaxed attitude towards pill use; they represented a gateway to experimentation with illegal drugs. [157,142,158] There were those who were in favour of legal party pills, however, believing that their availability and affordability discouraged the use of ecstasy with severe potential for harm.

The vast majority of piperazines (if not all) arrive in the UK from overseas factories, usually Asian (China/Hong Kong/Japan).[159] Although BZP can be synthesised from benzyl chloride and piperazine mono-hydrochloride, both cheap and readily available, it is easier to simply purchase it.

Piperazine	Reference	Structure
Benzylpiperazine	BZP	
Trifluoromethylphenylpiperazine	TFMPP	
1-(2-Methylphenylpiperazine)	*o*MPP	
1-(3-Methylphenylpiperazine)	*m*MPP	
1-(4-Methylphenylpiperazine)	*p*MPP	
1-(3-Chlorophenylpiperazine)	*m*CPP	
1-(4-Fluorophenylpiperazine)	*p*FPP	
1-(2-methoxyphenylpiperazine)	*o*MeOPP	
1-(3-Methoxyphenylpiperazine)	*p*MeOPP	
1-(3,4-methylenedioxybenzyl)piperazine	MDBP	
1,4-dibenzylpiperazine	DBZP	
4-bromo-2,5-dimethoxybenzylpiperazine	2C-B-BZP	

Table 6-1: Structures of popular piperazines

6.2 Current Piperazine Trends

Piperazines really came to fruition when efforts to control MDMA production in the 2000s meant manufacturers faced a shortage of precursors.[133] This resulted in some users experiencing difficulty in obtaining ATS due to reduced availability and increased cost. This may have created the demand for alternatives, although the demand for these substances, originally introduced as ‘legal highs’, was augmented by their appeal to others.

Ecstasy tablets which remained in circulation were found to be heavily adulterated, or with MDMA completely replaced with piperazines.[160] The FSS reported MDMA and amphetamine were often encountered in combination with piperazines and, as the number of MDMA cases steadily declined, they witnessed a related rise in the number of piperazine cases.[149] Piperazines were the most popular ‘legal high’ at this time. In recent years, however, the market for such ‘mimic drugs’ has experienced unprecedented growth.

These mimetics have many descriptors; ‘legal highs’, ‘herbal highs’, ‘research chemicals’, and ‘party pills’ have been some of the most popular. The preferred term amongst the scientific community is New Psychoactive Substances (NPS), which reflects that some are now illegal and that they may not necessarily have a stimulant effect.[159] Whilst piperazines have been misused for a number of years and are illegal in the UK they are still classified as NPS;[137] this term is defined by EMCDDA as drugs not listed under the 1971 United Nations Convention on Psychotropic Substances but which have effects similar to those which are listed. EMCDDA list piperazines as one of the main groups of NPS. The rate of production of NPS poses great difficulties for regulatory authorities and forensic analysts in attempting to identify and control them.[161] In 2011 the UK Government introduced the Home Office-funded Forensic Early Warning System that receives information from UK forensic providers which is forwarded to the EMCDDA. By 2011 eight piperazines had already been notified to the EMCDDA early warning system.[134] The UK also has a Drugs Early Warning System (DEWS), with information from police, prisons, NHS health-boards and more. Whilst the number of newly identified piperazines has declined in recent years, those which

have already been reported are continually encountered, often as mimic ecstasy which has experienced a resurgence.[137]

In 2013 piperazine seizures in the UK were found to have been greater in number than MDMA seizures, [162] yet global ATS seizures were at the highest level ever recorded.[137] A large proportion of seized ‘ecstasy’ contained substances other than MDMA. In 2011, ecstasy tablets which did not contain MDMA mostly contained diazepam (569,000) closely followed by piperazines (411,000).[163] Davies *et al* (2012) observed significant seizures of piperazines (large numbers were piperazine blends) with TFMPP most frequently encountered followed by BZP.[163] This was mirrored by Kuleya and coworkers’ (2014) analysis of ATS, all of which contained TFMPP, with BZP also common.[164] In 2013 at least 20% of UK MDMA contained mCPP; either alone or in combination with other substances.[137] Thus piperazines remain very much of relevance to UK forensic toxicologists. Quite recently a number of deaths in Scotland were associated with ‘green Rolex’ ecstasy tablets. Many of these tablets were seized and examined by the Scottish Police Authority. Whilst the main constituent was most often found to be para-methoxyamphetamine various other constituents were noted, including BZP.[165] Despite the wealth of NPS now available, piperazines endure.

6.2.1 Legal status

Under the UK Medicines Act of 1968 the sale of BZP products for human use has been illegal since March 2007.[166] In 2008 the Advisory Council on the Misuse of Drugs (ACMD) provided the UK Home Office with advice and recommendations for the control of piperazines following the European Council decision to subject BZP to ‘control measures and criminal provisions’.[149] In December 2009,

> *“1-benzylpiperazine and any compound structurally derived from1-benzylpiperazine or 1-phenylpiperazine by substitution in the aromatic ring to any extent with alkyl, alkoxy, alkylenedioxy, halide or haloalkyl substituent, whether or not substituted at the second nitrogen atom of the piperazine ring with alkly, benzyl, haloalkyl, or phenyl substitution”,*

were legislated as Class C under schedule 2, part III of the Misuse of Drugs Act (1971) and placed under Schedule 1 of the Misuse of Drugs Regulations (2001)

(*m*CPP is an exception which is under Schedule 4).[149,167,168] The ACMD report explained that classification of piperazines was based on potential for dependence and social harms, however, the impression created by the literature (or lack of it), is that the risk posed from piperazines has yet to be fully assessed.

The UK and Australia are the only countries to have banned all piperazine analogues.[169] Elsewhere some piperazines can be supplied and possessed without consequence. BZP has been controlled in all EU member states since 2009,[170] in the United states since 2002 [147,157] and in New Zealand since 2008.[171] New Zealand originally favoured regulation over illegalisation, in an endeavour to alleviate high rates of amphetamine use.[140,141,146] As popularity of the legal drugs soared annual sales were estimated at £24 million, around five million pills (2004).[171] The use of ecstasy versus party pills, however, increased.[146] Party pills were consumed in addition to habitual drug use,[142] and piperazines were illegalised.

International illegalisation of BZP cultivated demand for non-scheduled piperazines and BZP-free party pills. By 2007 mCPP had become the most common piperazine analogue in Europe,[127] and the most widely encountered NPS since monitoring began in 1997.[172] Consequently mCPP was also added to drugs legislation in many European countries.[148,149]

6.2.2 Relationship to Amphetamines

In the 1970s BZP was recognised as exhibiting stimulant properties in rats. With prospects as an anti-depressant, human clinical trials were conducted and its potential for abuse investigated. Volunteers experienced in amphetamine use who could fluently communicate effects and provide an 'educated' view of abuse appeal were recruited. It was during this trial that Campbell *et al* derived the much quoted 10:1 potency ratio.[151] Administration of 10 mg of dexamphetamine and 100 mg of BZP produced almost equal measured effects or 'scores'. BZP was actually rated more amphetamine-like than dexamphetamine indicating significant abuse potential. Campbell *et al* abandoned the trials and recommended statutory control of BZP.

There have been few subsequent studies of piperazines. As substitutes for MDMA piperazines were expected to have similar physiological action. BZP, however, was the only piperazine to exhibit clear stimulant behaviour.[154] Similar to MDMA, BZP increased heart rate, blood pressure, and auditory vigilance,[151,173,174] whilst other piperazines exerted only mild changes in these functions.[148,175] BZP has been established chiefly as a dopamine and norephedrine [168] releasing agent but also increases serotonin levels by inhibiting synaptic re-uptake, similar to amphetamine (and cocaine).[144,153,176] TFMPP, mCPP and FPP were recognised to release endogenous stores of serotonin from neurons, similar to substituted amphetamines (e.g.MDMA),[148,152,175,177] although there was no significant release of dopamine as observed with MDMA. Thus no single piperazine was found to replicate the pharmacodynamics of MDMA. Co-administration of BZP/TFMPP, however, was different.

BZP with TFMPP resulted in dopamine production in excess of the sum of both drugs taken individually. It displayed similar neurological action to MDMA, and demonstrated the existence of a piperazine interaction.[152,153] This effect was further investigated by Antia *et al* [178] with the conclusion that each of these drugs inhibit the metabolism of the other, leading to elevated levels of both drugs when co-administered. This synergistic effect could be life-threatening.[140,152]

6.2.3 Toxicity

6.2.3.1 Acute

Whilst studies of toxicity associated with piperazines and piperazine blends are few, it is documented that severe effects can occur unpredictably and at relatively low doses.[179-181] User experiences combined with controlled studies have identified adverse effects of piperazines to include: nausea, vomiting, tachycardia, hypertension, confusion, agitation, panic attacks, hallucination, dissociative symptoms, dryness of the mouth, and problems with urine retention, as well as more serious complications such as renal failure, life-threatening seizures, multi-organ failure, respiratory depression and psychosis. [141,151,155,168,173,176,181-185] Symptoms can persist for up to 24 hours and users

frequently recount experiencing a severe hangover following piperazine use.[168,181,183]

There are relatively small numbers of reported piperazine poisonings, however, these are expected to be underreported. [186,187] The inability of most routine toxicology screening to detect piperazines could contribute to this, as could misdiagnosis due to more familiar expectation of amphetamine intoxication which presents similar symptoms.[150,186,188] Patients may contribute to misdiagnosis by wrongly claiming to have consumed ATS, although they have unwittingly purchased and consumed piperazines sold in place of an ATS (e.g. mimic ecstasy).[134,162,184,186,188] In four confirmed cases of piperazine poisoning in the UK, patients believed they had purchased MDMA but tests determined they had been sold piperazines.[173,184] Although there have been no deaths attributed to piperazine use, BZP and/or TFMPP featured on 25 UK death certificates from 2008 to 2012.[163]

As piperazine use is considered to be widespread it must be assumed that the majority of users do not suffer serious adverse effects. In an assessment of the effects of BZP and TFMPP (individually and combined), however, the Medical Research Institute of New Zealand abandoned their trials after only 35 subjects had completed testing. This was due to “the frequency and nature of severe reactions” observed.[180] This partial trial was enough to demonstrate that at recommended doses BZP/TFMPP party pills “result in a high rate of severe adverse reactions,” indicating a narrow safety margin in some users, especially in blends and at high doses. During this trial in which subjects were observed to suffer severe negative effects, only the recommended dose had been administered. With piperazines promoted as safer alternatives to ATS, the perception has been that they are good quality pills but weaker, and users admit disregarding safety information and ‘hyper-dosing’.[142,174] User surveys revealed that when taking piperazines for the first time over 30% of respondents consumed more than the labelled dose (usually 2 pills), and others habitually consumed 10 or more pills, with occasional greatest consumption around 12 pills.[142,189,190]

6.2.3.2 Risk of overdose

Accidental 'overdose' could result from users' difficulty in assessing exactly what they have taken. There can be a number of reasons for this including incorrect concentration of active ingredient labelled,[155,191] variability of active ingredients across brands,[189,192] changes in constituents of brands through time, and poor labelling which is difficult to interpret.[145,193] Labels generally advise taking one or two pills and waiting only one hour to assess effects before taking more.[194] It can take more than one hour for effects to manifest so effects may be increased beyond that intended (peak subjective effects around two hours).[150,168,175] Contrasting this with MDMA, which exerts its effects relatively quickly, regular MDMA users accustomed to rapid onset may believe that party pills have not had any effect and take more, or resort to MDMA. Many users claim to purposely combine piperazines with ATS to enhance and prolong the effects.[142,146] Party pills warn against their use with other substances,[142] although the dangers of doing so are not explicit. mCPP has also been detected in cocaine and ketamine;[148,155] the pharmacokinetics of mCPP following nasal insufflation (which increases potency) are unknown.[195]

Slow onset of effects could be responsible for observed intravenous administration of piperazines, particularly BZP. [146,155,181,183,196] This could also be a result of their having similar effects to ATS but better availability.[156] Intravenous use results in a much higher bioavailability and increased risk of serious adverse reactions. Piperazines are basic making injection painful, so it is therefore unlikely that piperazines would become established for intravenous use.[156]

6.2.3.3 Chronic effects

Long term effects of piperazine use have not been established. With similar mode of action and physiological effects to ATS,[144,155,173] similar consequences of long term use may be a reasonable postulation.[187,191] MDMA has been shown to have potential to cause neurotoxicity and psychobiological problems such as depression, obsessive-compulsive disorder and paranoia.[197] With such unknown potential for long term harm the ACMD assignment of 'relatively low risk'[155] may be an underestimate. The efficacy and potency of piperazines may be lower

than ATS but the trend to consume them in large doses and in blends[198] means that they are potentially as harmful. There has also been evidence from a small number of studies to suggest that regular use of some piperazines/blends could lead to dependence.[141,151,191,199,200]

6.2.4 Metabolism

Only four groups have addressed metabolism of piperazines in human subjects. Three of these suggest that there is no significant metabolism as large concentrations of parent drug (BZP, TFMPP and mCPP) are detected in urine.[37,176,201] However, Antia *et al,* the only group to dose humans, identified that BZP is extensively metabolised.[202] The parent drug and hydroxy metabolites (3-OH-BZP slightly in excess of 4-OH-BZP) excreted in urine account for only *ca.* 12.5 % of the total administered dose. This may indicate low bioavailability of BZP, another route of excretion (e.g. billiary), or strong protein binding. Hydroxylated metabolites of BZP (3-OH-BZP and 4-OH-BZP)[202] and TFMPP (4-OH TFMPP) [203] were also detected in plasma. BZP metabolites were at extremely low levels in comparison to the parent drug (*meta* 5% and *para* 3%), but 4-OH-TFMPP was detectable in plasma in relatively high abundance, reaching a peak concentration almost as high as that of the parent drug. The metabolic interaction of BZP with TFMPP following co-ingestion was demonstrated once again, as the metabolites of individual ingestion differed to those of co-ingestion, with each losing a hydroxylated metabolite (3-OH BZP and 4-OH TFMPP respectively).[204]

The metabolic pathways of piperazines in human tissue have yet to be confirmed, although metabolic dependence of the CYP450 enzyme system is expected, with CYP2D6 believed to be the major enzyme responsible for catalysing the metabolism of piperazines.[178,205] Reliance upon CYP450 metabolism explains the susceptibility of piperazines to drug-drug interaction. Co-ingested piperazines significantly inhibit metabolism, resulting in greater toxicity. [183,186,204] It is also expected that co-ingesting a piperazine with another CYP2D6, CYP1A2 or CYP3A4 metabolised drug, e.g. MDMA and cocaine, will increase toxicity of that drug as the piperazine competes for metabolism. This

could be very dangerous if the competing drug is a medication as this may no longer function as it should.

6.3 Aim

Following the G8-Roma-Lyon expert group meeting in April 2013,[206] the UK committed to developing methods for analysis and identification of NPS to allow collection and sharing of toxicology data. It is essential that forensic toxicology laboratories develop validated quantification methods. UNODC have produced a guide to the analysis of piperazines in seized materials (2013)[207] but not yet for biological samples which are more complex.

Piperazines clearly present a means of driver impairment and pose a significant toxicological threat - they are included on the UK and Ireland Association of Forensic Toxicologists recommended drugs of abuse panel to be routinely tested in forensic toxicology laboratories.[208] A method which identifies their use and which can quantify the levels present in body fluids (particularly blood) is required.

The aim of this project was to validate a method for the detection and quantification in blood of as many 'relevant' piperazines as possible.

6.3.1 Relevant piperazines

Widespread use of BZP and TFMPP [133] make them most relevant, together with mCPP and MeOPP which are given greatest interest in the literature. 'Party pill' investigations through time have shown that they contain all four of these piperazines in addition to MBZP (methylbenzylpiperazine), DBZP (1,4-dibenzylpiperazine), pFPP and oMPP [19,193,209]; these are therefore, also of importance. UNODC reports the top five piperazines to be mCPP, BZP, TFMPP, pFPP, followed by MBZP. Piperazines most frequently encountered by the FSS in the UK were mCPP, BZP, TFMPP, and DBZP.[149] These piperazines and mCPCPP (1-(3-chlorophenyl)-4-(3-chloropropyl) piperazine) have been reported to the EMCDDA/EUROPOL Early Warning System.[134] Two further piperazines are known to have been in circulation as they have been reported to the UNODC: 2C-B-BZP (4-bromo-2,5-dimethoxybenzylpiperazine) and MePP (methylphenylpiperazine).

EWS data may not be a true representation of the piperazines in current circulation, rather, it may be a reflection of detection methods in place in drugs laboratories. It is likely that some substances have been missed. The piperazines which have been identified, however, are a good set of target analytes for a piperazines forensic toxicological method because if these are known to be present in drugs seizures they have the potential to be present in clinical and post mortem blood samples. Those piperazines identified in seizures are summarised in Table 6-2.

Piperazine	Abbreviation	Included in Method
Benzylpiperazine	BZP	✓
4-bromo-2,5-dimethoxybenzylpiperazine	2C-B-BZP	No; lack of reference standard
1-(3-Chlorophenyl)-4-(3-chloropropyl) piperazine	*m*CPCPP	No; lack of reference standard
1-(3-chlorophenyl)piperazine	*m*CPP	✓
1-(4-chlorophenyl)piperazine	*p*CPP	✓
1-(4-Dibenzyl)piperazine	DBZP	No; impurity of BZP generally at trace levels *
1-(4-fluorophenyl)piperazine	*p*FPP	✓
Methoxyphenylpiperazine	MeOPP	✓
methylbenzylpiperazine	MBZP	No; avoided by users, extremely mild effects **
Methylphenylpiperazine	MPP	✓
Trifluoromethylphenylpiperazine	TFMPP	✓

Table 6-2: Relevant piperazines for toxicological detection * [149] ** [210]

6.4 Experimental Section

6.4.1 Introduction

The piperazine structure suggests similar chemical properties to amphetamines. It was, therefore, hoped that the piperazines could be added to the Forensic Medicine and Science (FMS) validated and UKAS-accredited amphetamines GC-MS method. This method, however, proved to be unsuitable and required some changes for application to piperazines. The GC oven temperature programme, injection conditions, and derivatisation were investigated. Careful consideration of high volatility of analytes, their basic nature and high polarity was necessary. A method such as LC-MS/MS which does not operate at high temperatures, requires fewer evaporation steps, and can provide greater specificity could be more suited to piperazine analysis. This was explored through comparison of the validation parameters for GC-MS and LC-MS/MS methods.

6.4.2 Reagents and Standards

All chemicals used, listed in Table 6-1, were of at least analytical reagent grade. Deionised water was obtained from the in-house Millipore system.

Reagent	Supplier
1-(2-methylphenyl)piperazine dihyrdrochloride	ACROS Organics
1-(4-methylphenyl)piperazine dihydrochloride	ACROS Organics
1-(3-chlorophenyl)piperazine monohydrochloride	Alfa Aesar
1-benzylpiperazine	Fluka
1-(4-Chlorophenyl)piperazine	Sigma Aldrich
1-(2-methoxyphenyl)piperazine	Sigma Aldrich
1-(3-Chlorophenyl)-piperazine-d8	Sigma Aldrich
1-(3-trifluoromethylphenyl)piperazine hydrochloride	Sigma Aldrich
1-(4-fluorophenyl)piperazine	Sigma Aldrich
1-(4-methoxyphenyl)piperazine	Sigma Aldrich
1,4-Dibenzylpiperazine	Sigma Aldrich
1-methyl-3-phenyl-piperazine	Sigma Aldrich
Benzylpiperazine-d8 dihydrochloride	Sigma Aldrich
Heptafluorobutyric anhydride	Sigma Aldrich
disodium hydrogen orthophosphate anhydrous	VWR
sodium dihydrogen orthophosphate monohydrate	VWR
Ammonium Acetate	VWR
Sodium chloride	BDH Chemical Ltd
Ammonia	Fisher Scientific
Methanol	VWR
Dichloromethane	VWR
Ethyl acetate	VWR
Propan-2-ol	VWR
Acetic Acid, Glacial	Fisher Scientific
Cleanscreen DAU 200mg/10mL	Presearch

Table 6-3: Materials and suppliers

6.4.3 Preparation of Blank Blood

Validation calibration samples were prepared from expired red blood cell pouches supplied with ethical approval from the blood bank at the Western Infirmary, Glasgow. These had been frozen upon receipt. Before use they were fully defrosted and diluted 1:1 with isotonic saline solution (9.5g of sodium chloride dissolved in 1L deionised water). Whole blood was also used, which was received frozen and allowed to reach room temperature before use.

6.4.4 Preparation of Working Standards and Quality Controls

A mixed internal standard (I.S.) working solution was prepared containing 20 µg/ml of both BZP-D8 and mCPP-D8. This was prepared by adding 100 µl of 1 mg/ml BZP-D8 stock solution and 1 ml of 0.1 mg/ml mCPP-d8 to a 5 ml volumetric flask and making up to the mark with methanol. Limited resources meant that some experiments employed only one I.S. For these a separate 20 µg/ml BZP-D8 solution was prepared by adding 200 µl of BZP-D8 stock solution to a 10 ml volumetric flask and filling to the mark with methanol.

Each reference standard was prepared as a 1 mg/ml methanol stock solution, most being prepared from bulk powders. 2.5 ml of each of the nine stock solutions were combined in a 50 ml volumetric flask and filled to the line with methanol to produce a 50 µg/ml mixed reference standard intermediate solution. From this intermediate solution mixed reference standards were individually prepared (not by serial dilution). Preparation of the calibration standards is outlined in Table 6-4.

Table 6-4: Preparation of calibrators

Blood Concentration ng/ml (50µl stnd sol to 1 ml blood)	Concentration of standard solution µg/ml	ml of 50 µg/ml solution to make 10ml working solution
5	0.1	0.02
20	0.4	0.08
60	1.2	0.24
100	2.0	0.40
300	6.0	1.20
500	10.0	2.00
1000	20.0	4.00
1500	30.0	6.00
2000	40.0	8.00

Quality control samples were prepared by the same means but from different 1 mg/ml stock solutions. Practical restrictions meant that for some runs quality control samples were from the calibration batch.

6.4.5 Sample Preparation

Calibration standards were prepared by adding 50 µl of piperazines working solution and 50 µl of 20 µg/ml internal standard solution to a test-tube before adding 5 ml of mixed water:pH6 phosphate buffer (2:1 v:v) and lastly 1 ml of blank blood. This was vortex mixed then centrifuged at 2500 rpm for 10 min.

Cleanscreen® DAU solid phase extraction cartridges were used for extraction as they had previously provided good recovery for amphetamines [30,183,211-213] and piperazines [187] (although the recovery was less for mCPP).[81] Cartridges were conditioned with 3 ml methanol, 3 ml deionised water, then 1 ml pH 6 phosphate buffer, prior to sample loading. Cartridges were then washed with 3ml deionised water, 1 ml 0.1M acetic acid, then 3ml methanol and dried thoroughly on full vacuum. The analytes were eluted with 3 ml freshly prepared dichloromethane/isopropanol/concentrated ammonia (78:20:2 v:v:v). 600 µl of the extract was removed and dried at room temperature before reconstitution with 100 µl of 2 mM ammonium acetate in 5 % methanol for LC-MS/MS. The remaining extract was dried at room temperature before derivatisation with 40 µl of HFBA at 60°C for 30 min. Excess HFBA was evaporated before reconstitution in 40 µl ethyl acetate.

6.4.6 Optimisation of GC-MS

6.4.6.1 Derivatisation

The majority of the piperazine analytes were not suited to GC-MS analysis without derivatisation. Derivatisation of amine and hydroxyl moieties is necessary to prevent tailing of peaks and improve sensitivity. Amines are usually derivatised by acylation, using acetic or propionic anhydrides,[19] to replace active hydrogens which can interact with the stationary phase to negatively affect reproducibility and peak shape.[214,215] The acetyl group stabilises the piperazine ring, producing more characteristic mass spectra. Structural isomers may still have the same fragmentation, however,derivatisation improves chromatographic resolution, aiding in differentiation of isomers by retention time.[145]

Pentafluoropropionic anhydride (PFPA) is commonly cited as a derivatising reagent for amphetamines (optimum conditions 15 minutes at 50°C [216]). Given structural similarity of piperazines to amphetamines it is surprising that the application of PFPA to piperazine methods has not previously been reported. Three PFPA preparations in ethyl acetate were trialled (1:1 v:v), (2:1 v:v), and (3:2 v:v), with 50 µl of these solutions added and allowed to derivatise at a range of temperatures for various times. Excess derivatising reagent was removed by evaporation (without heat). Unfortunately this was not successful across all of the piperazines. Some analytes displayed multiple peaks due to varying sites of derivatisation; also, the derivatisation was not reproducible.

Other derivatising reagents were also trialled: acetic anhydride with pyridine, used by Stack *et al* in their piperazine metabolite investigations [29,186,188,217-220]; trifluoroacetic anhydride (TFAA), popular for derivatisation of amino and hydroxyl groups in forensic and clinical toxicology and used in published piperazine methods [144,145,221,222], not always successfully (Tsutsumi *et al* [221] problems of reproducibility); N,O-bis(trimethylsilyl)trifluoro-acetamide (BSTFA) and heptafluorobutyric anhydride (HFBA). HFBA produced the most acceptable spectra. Only o/p-MeOPP had two derivative peaks due to derivatisation occurring at two locations, but these were only apparent at high concentrations. HFBA is expensive and it is preferable for routine methods to use as little as possible. Dilution of HFBA with ethyl acetate was trialled but was found to be ineffective. Sufficient reproducibility of derivatisation by addition of 50 µl of undiluted HFBA at 60 °C for 30 minutes was demonstrated by viewing the TIC of seven HFBA derivatised aliquots of each of the individual piperazines.

6.4.6.2 Internal Standard

Prior to instrumental analysis, internal standards are added at the earliest stage to correct for any analyte loss during preparation and analysis.[16,38,74] For reliable compensation the I.S. must parallel the behaviour of the analytes.[38,141,198] This also applies to LC-MS/MS analysis in which deuteron-labelled standards can compensate for matrix effects.

Piperazine research methods have generally used ephedrine, therapeutic drugs, or other piperazines as internal standards.[145,195,223,224] These are not suitable for casework methods, although some published casework methods make use of them.[81,176,187,225] A deuterated piperazine would be preferred, however, these can be expensive and difficult to source. These constraints may render the limited number which are available unsuitable for high volume routine laboratory analysis. To overcome this, molecules of similar weight and chemistry such as deuterated amphetamine [144] or mescaline [226] have been used. These were part of methods with few targets, however, and if analytes are eluting over a considerable time, more than one internal standard is necessary.

Deuterated versions of both BZP and mCPP were obtained for this study. It would be ideal to have a deuterated version of each of the analytes but this was not possible. The best alternative is to have at least two internal standards, one which elutes close to the first eluting analyte and another which elutes close to the last. BZP-d8 and mCPP-d8 were quite well placed in the elution to represent most of the analytes as shown in Figure 12 for GC-MS method and Figure 13 for LC/MS/MS method.

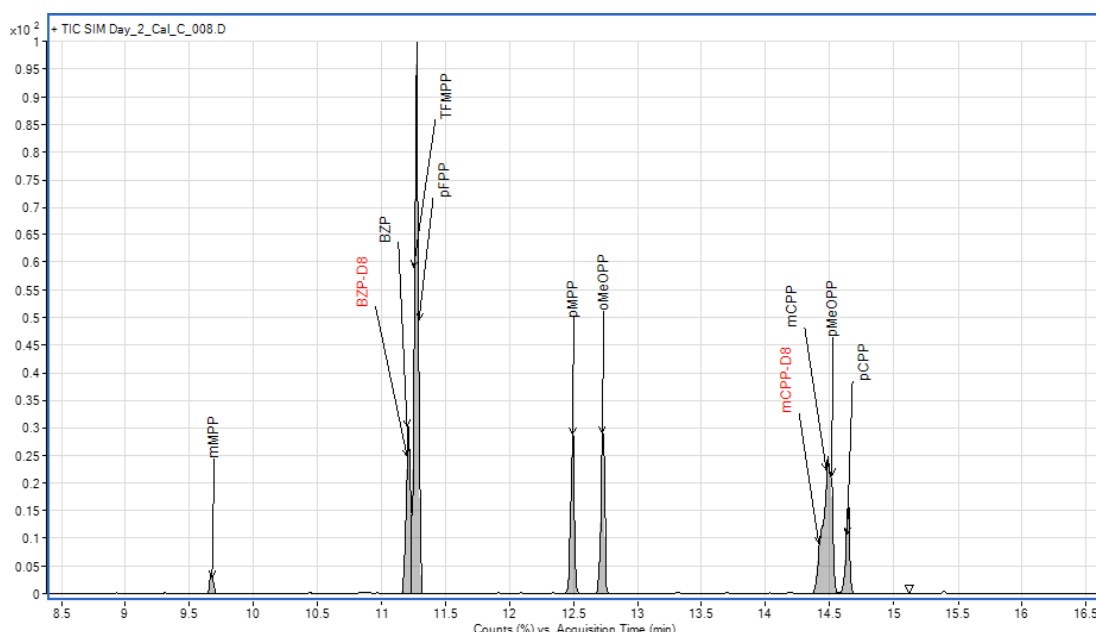


Figure 12: Placement of internal standards in GC-MS method. BZP-D8 used for mMPP, BZP, pFPP and TFMPP, mCPP-D8 used for pMPP, oMeOPP, pCPP, pMeOPP and mCPP.

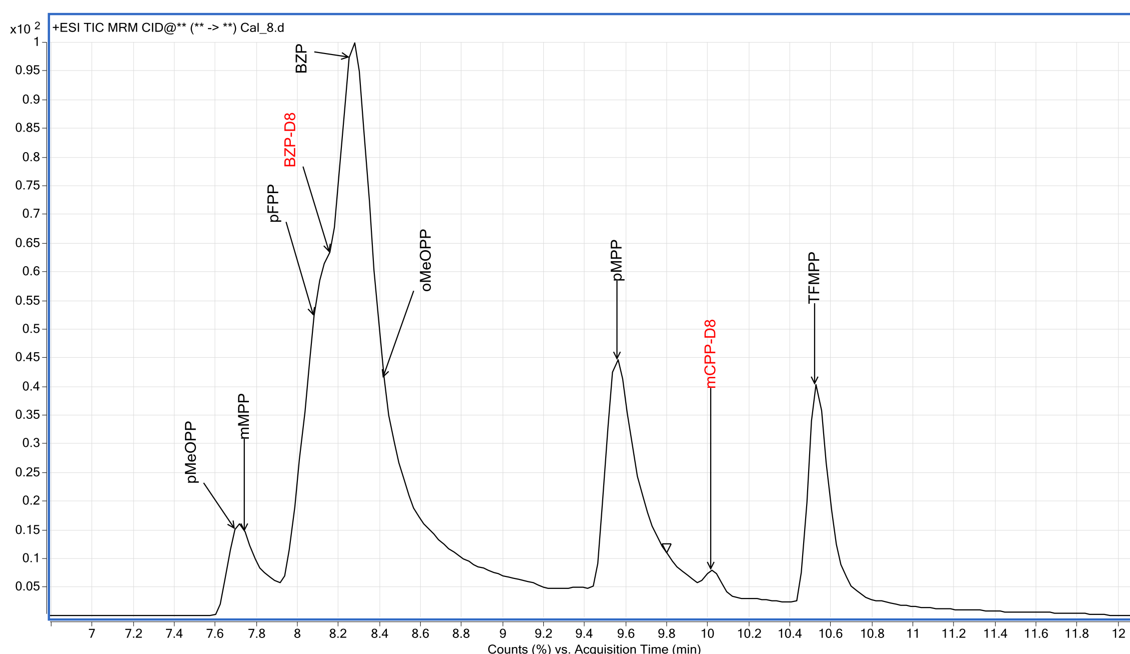


Figure 13: Placement of internal standards in LC-MS-MS method. BZP-D8 used for pMeOPP, pFPP, mMPP, oMeOPP and BZP; mCPP-D8 used for pMPP and TFMPP.

6.4.6.3 Injection Conditions

Sample introduction is a critical step in gas chromatography.[227] Splitless injection is conventionally held as the best method for analytes present in low concentration. Programmed temperature vaporisation (PTV) is associated with large volume injection (LVI) for pre-column focusing, however, its use with regular injection volumes could perhaps improve chromatography, as it is intended to remove unwanted substances. PTV could possibly remove excess derivatising reagent and solvent to provide larger on-column analyte concentration.[228] Highly volatile compounds such as piperazines, however, could also be purged. Some preliminary investigation of the potential for PTV for a piperazines method was carried out. The signs were promising with improved peak shape and reduced background noise (using PFPA derivatives), however, the GC-MS instrument with the required injection port was not available for the entire period of the study.

6.4.6.4 Oven Temperature

Temperature programmes reported in the literature (Staack *et al* [188,195,219,229] Kenyon [230] De Boer [145]) and the FMS laboratory routine amphetamine method were compared. Resolution of the piperazines was slightly better with the de Boer method and this was the focus of investigations to provide further

improvement. The final temperature programme used an initial temperature of 90° for 1 min, rising to 180 °C at 15 °C/min, then to 240 °C at 5 °/min and to 300°C at 20 °C/min, which was held for 5 minutes. The total run time was 27 minutes.

6.4.7 GC-MS Method

1 µl of sample was injected to an Agilent Technologies GC System 7890A, fitted with a DB5 column (30 m length, 0.25 mm i.d., 0.25 µm film thickness). Splitless isothermal injection at 250°C was employed with a valve off time of 0.7 min. The carrier gas was helium at a constant flow rate of 1.2 ml/min.

The GC was linked to an Agilent Technologies 5975C inert XLMSD with Triple-Axis Detection. The ion source temperature was 200ºC and the interface temperature 250ºC. EI+ mode was used with ionisation energy of 70 eV. Mass spectrometry was in the selected ion monitoring mode, using the ions detailed in Table 6-5. The relationship of these ions to the analyte structure is given in Appendix 4: generally the main fragment ions are formed by loss of functional groups from the benzene ring and splitting of the piperazine ring. Peak areas were recorded with a ChemStation data system and translated for interpretation by the Mass Hunter Quantitative Analysis programme (both supplied by Agilent Technologies).

HFBA Derivatised Analyte		Retention Index	Quantifier Ion m/z	Qualifier Ions m/z
1-(3-methylphenyl)piperazine	mMPP	1632	372	175, 104
Benzylpiperazine-d8	BZP-d8	1736	380	289, 183
1-(2-methylphenyl)piperazine	oMPP	1737	118	**
Benzylpiperazine	BZP	1739	281	372, 175
1-(4-fluorophenyl)piperazine	pFPP	1742	376	179, 123
Trifluoromethylphenylpiperazine	TFMPP	1747	426	200, 229
1-(4-methylphenyl)piperazine	pMPP	1804	372	175, 119
1-(2-methoxyphenyl)piperazine	oMeOPP	1831	388	191, 120
1-(3-chlorophenyl)piperazine-d8	mCPP-d8	1974	400	203, 171
1-(3-chlorophenyl)piperazine	mCPP	1976	392	195, 166
1-(4-methoxyphenyl)piperazine	pMeOPP	1978	388	191, 120
1-(4-chlorophenyl)piperazine	pCPP	2005	392	195, 166

Table 6-5: Ions for SIM method

6.4.8 LC-MS/MS Method Development

There were practical restrictions on the method development. In order to facilitate maximum usage of the LC-MS/MS, a standard column and mobile phase had to be used. The advantage of such restrictions, however, is that the resultant method would be readily incorporated to a busy casework laboratory.

Each piperazine analyte was 'characterised' by direct infusion to the mass spectrometer of individual piperazine solutions prepared in 2 mM 50 % methanol 0.1 % formic acid at a concentration of 1 µg/ml. The fragmentor voltage was optimised to give maximum fragmentation of the precursor ion, and the collision energy was optimised for each product ion to give greatest abundance. The ions observed and their relative intensities were recorded. Each piperazine was then introduced to the mass spectrometer in mobile phase via the column to ensure the ion intensities recorded at infusion were correct. The transition with the most intense signal was selected for quantification and the next most intense for confirmation as detailed in Table 6-6. The relationships of these ions to the piperazine structures are detailed in Appendix 5: they all tended to follow the same pattern, with fragmentation of the piperazine ring.

Analyte		Precursor Ion m/z	Fragmentor Voltage	Prodcuct Ion m/z	CE (V)	Prodcuct Ion m/z	CE (V)
Trifluoromethylphenylpiperazine	TFMPP	231.02	160	188.00	21	118.00	44
1-(3-chlorophenyl)piperazine-d8	mCPP-D8	205.00	150	158.00	20	123.00	28
1-(4-chlorophenyl)piperazine	pCPP	197.02	135	154.05	18	119.10	25
1-(3-chlorophenyl)piperazine	mCPP	197.00	150	153.97	18	118.00	37
1-(2-methoxyphenyl)piperazine	oMeOPP	193.00	130	150.00	18	119.95	37
1-(4-methoxyphenyl)piperazine	pMeOPP	193.00	135	150.00	17	176.00	15
Benzylpiperazine-d8	BZP-D8	185.10	120	91.00	22	65.00	50
1-(4-fluorophenyl)piperazine	pFPP	181.00	145	138.00	18	91.00	30
Benzylpiperazine	BZP	177.05	110	91.00	22	85.00	13
1-(3-methylphenyl)piperazine	mMpp	177.00	130	134.00	14	160.00	12
1-(2-methylphenyl)piperazine	oMPP	177.00	135	134.00	18	118.00	38
1-(4-methylphenyl)piperazine	pMPP	177.00	135	134.00	18	118.00	40

Table 6-6: Piperazine analyte identification transitions for LC-MS/MS

The analytes were then combined for injection to the LC-MS/MS with the mass spectrometer programmed to detect only the first transition for each analyte. The highest calibrator concentration 2 µg/ml of blood (4 µg/ml standard for

injection) was used to develop a gradient. Due to the presence of isomers and the similarity of each analyte an elaborate gradient was necessary to achieve the best separation possible whilst maintaining acceptable peak shapes. There is a long period (7 minutes) before any analytes elute, which is required to separate o/pMeOPP. The stability of peak retention times was assessed to ensure enough time was allowed for the column to return to equilibrium at the start of the gradient (5% solution of 2 mM acetonitrile in methanol with 95% of 2 mM acetonitrile in water).

The second transition for each analyte was then added to the method and sensitivity evaluated with non-extracted standards, then with extracted low concentration standards. Dwell time was optimised at 100 ms per step and injection volume at 20 µl.

With the final method determined each analyte was run individually to assess interferences. When each was injected individually there did not appear to be any problems of interference. When combined, however, it was clear that the analytes which co-eluted with pFPP were causing interference as a high background noise and a large shoulder to the pFPP peak at high concentration. All ion transitions identified for pFPP at infusion were trialled but none were free of this noise. Manual integration was necessary to remove the contribution from co-eluting analytes.

When interference was first assessed an aged sample of BZP-d8 was used. This produced a signal for both BZP-D8 and BZP. Fresh BZP-D8 did not produce any signal for BZP. BZP-D8 is expected to degrade to BZP. It is, therefore, essential to monitor the zero calibrator for any BZP signal.

6.4.9 LC-MS/MS Method

A 20 µl aliquot of the extract was injected on a Gemini C18 (5µm 150 x 2.0 mm) column fitted with a Gemini C18 (4 x 2.0 mm) guard column maintained at 40ºC. The mobile phase consisted of 2 mM ammonium acetate in deionised water (A) and 2 mM ammonium acetate in methanol (B) with gradient programmed to give 5 % A for 4 min, increasing to 30 % A between 4 and 5 min, held until 5.5 min

then increased to 70 % by 7 min and held for 0.25 min before increasing to 90 % A then reduced to 5 % A for 4.5 min (total run time of 16 min). Flow rate was 0.3 ml/min.

An Agilent Technologies 6420 Triple Quadrupole mass spectrometer with ESI was used. The mobile phase was introduced directly to the ion source without splitting. Analysis was performed in positive ion mode with capillary voltage of 4 KV, gas temperature 350ºC and nitrogen flow 11 l/Min. The mass spectrometer was operated in multiple reaction monitoring (MRM) mode with two transition ions, one for quantification and one for confirmation as shown in Table 6-6.

6.4.10 Validation

The GC-MS and LC-MS/MS methods were both subject to validation. Calibration response function was assessed over a wide range but restricted to 5 to 2000 ng/ml of blood for validation. The peak area ratio of the analyte to an internal standard was calculated and plotted against concentration to produce seven calibration curves over five non-consecutive days. A low (60 ng/ml), medium (1000 ng/ml) and high (2000 ng/ml) quality control (QC) sample was quantitated by each of the calibration curves. On one of these days, 5 samples at each QC concentration were analysed to investigate precision and accuracy of the method. Accuracy was measured by taking account of how different the measured value for the spiked blood was to the expected value. Instances of outliers were removed. Whilst it is desirable to have five values, three is acceptable as a minimum;[231,232] there were never less than four. A percentage ratio of the difference was reported. Precision was measured as percentage coefficient of variation (CV %). Repeatability was assessed between results from each of the five days. Acceptability for both accuracy and precision was considered to be within 15% (20% at low concentrations).

Originally LOD and LOQ were estimated by preparation of a special low level regression, around the expected range of LOD. Calculation of the standard error of the regression line of x on y provided an estimate of the uncertainty in the regression arising from the background noise. LOD was estimated to be 3 times this and LOQ ten times. This method is only suitable for linear relationships and

could not be applied to the GC-MS method which produced quadratic responses for all analytes, even at low levels. LOQ is defined as the value at which measurements have satisfactory precision (20%) and accuracy (±20%). Thus LOD and LOQ were estimated through replicate analyses at various concentrations. This was also performed for the LC-MS/MS as the estimated values by low level regression were not satisfactory in practice.

Matrix effects (Equation 6-1), recovery (Equation 6-2) and process efficiency (Equation 6-3) were calculated using three types of specially prepared standards at low, medium and high concentrations in whole blood (as described by Matuszewski *et al* [233]). The standards prepared were, A) non-extracted, B) extracted before spiking analytes into the extract, and C) analytes spiked in blood prior to extraction. Five of each type at each concentration were analysed. Only one I.S. was used, BZP-D8.

$$\text{Matrix Effect (\%)} = B/A \times 100$$

Equation 6-1: Calculation of Matrix effect, 85 to 115 % is desirable

$$\text{Recovery (\%)} = C/B \times 100$$

Equation 6-2: Calculation of Recovery, $\geq$ 60 % desirable

$$\text{Relative Process Efficiency (\%)} = C/A \times 100$$

Equation 6-3: Calculation of process efficiency, a method with no errors = 100 %

A study of short-term stability was performed to ensure the analytes remained stable over the analysis period. Low, medium and high concentration spiked blood samples were prepared by adding 100 µg/ml mixed piperazine solution to 50 ml volumetric flasks, adding a little water to dilute the methanol, then filling to the line with whole blood, as outlined in Table 6-7. These were aliquoted and divided between the bench top (room temperature 22 ± 2 ^{o}C), fridge (4 ± 2 ^{o}C) and freezer (- 20 ± 2 ^{o}C) for storage. Three aliquots of each concentration were extracted, derivatised and analysed immediately while a further 9 at each concentration were extracted and placed on the autosampler for three of each to be analysed after 24, 48, and 72 hours to assess stability of the derivatised

analytes. Three of each concentration from the bench-top and fridge were also analysed after 24, 48 and 72 hours. Three of each concentration were left in the fridge and analysed after one week. The freezer samples were analysed after each of three freeze/thaw cycles (3 at each concentration each time).

Blood Concentration	Volume of 100 µg/ml standard added
60 ng/ml	30 µl
1000 ng/ml	500 µl
2000 ng/ml	1000 µl

Table 6-7: Preparation of stability standards

6.5 Results and Discussion

6.5.1 GC-MS Validation

6.5.1.1 Specificity

The selectivity of the method was demonstrated by comparison of the chromatogram of blank blood with that of blank blood spiked with each of the piperazine analytes. Initial full scan analysis verified that the reference standards were of sufficient purity for validation and that there were not significant impurities present that might interfere with the method. The quantifier and qualifier ions had to be carefully selected, however, as combining piperazines in one method meant that they could interfere with the quantification of each other when they could not be fully resolved. All of the ions in the spectra of each co-eluting piperazine were noted. Ions shared by piperazines were only used if they had different retention times. BZP and o-MPP could not be distinguished (for quantification or reliable identification) as it was not possible to select unique qualifier ions for o-MPP as there were no ions other than m/z 118 which were not shared. As BZP is the more commonly encountered analyte o-MPP was omitted from the validation method as having it present in the piperazine mixed standard would have caused interference to the BZP response. If a real case sample contained o-MPP and no BZP, this could be identified from full scan mass spectra by looking at retention index 1737 for the ions 372 (M^+), 175, 146,132 and 118. Quantification of o-MPP would require further validation. If it was recognised that these analytes were often present in

combination, additional validation work could be performed to determine quantification parameters for the joint presence of BZP and o-MPP.

6.5.1.2 Calibration Response Function

All analytes gave quadratic rather than linear calibration curves but with acceptable correlations (R^2 > 0.99). Calibrations and average correlation coefficients are shown in Figure 14 to Figure 22. There was no carry-over between runs at the highest calibrator concentration.

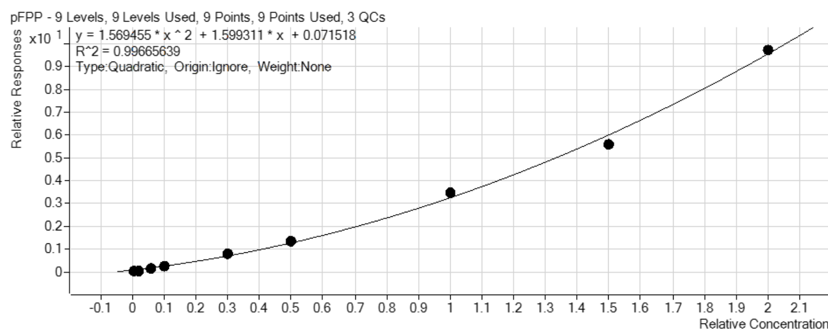


Figure 14: Calibration curve for p-FPP (mean R^2 0.9994)

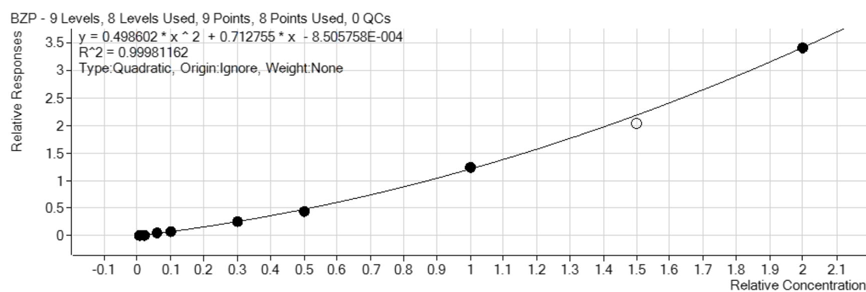


Figure 15: Calibration curve for BZP (mean R^2 0.9993)

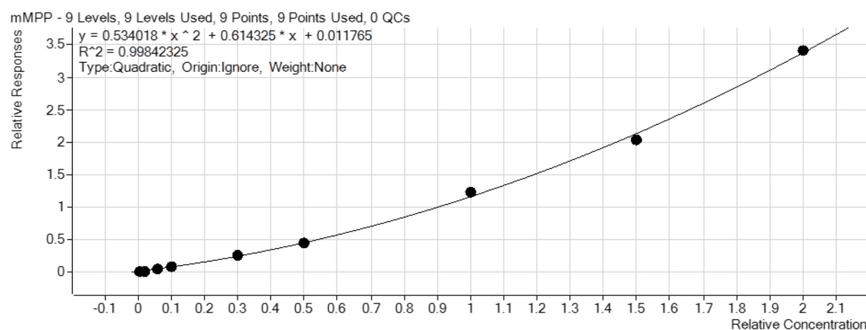


Figure 16: Calibration curve for m-MPP (mean R^2 0.9958)

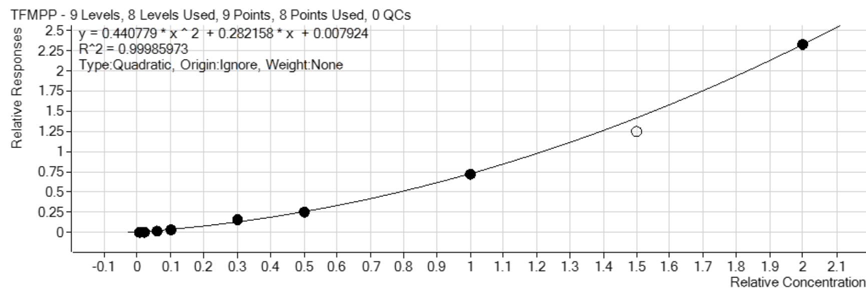


Figure 17: Calibration curve for TFMPP (mean R^2 0.9991)

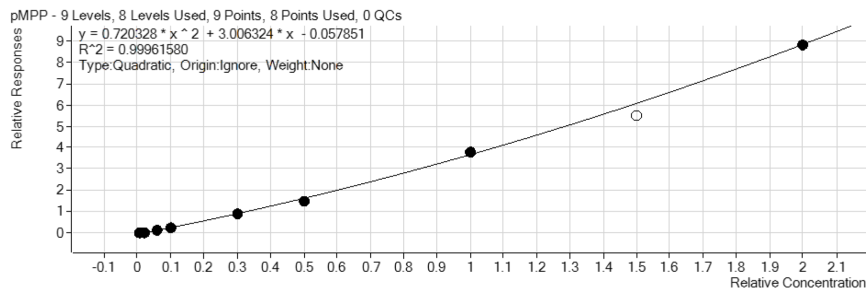


Figure 18: Calibration curve for p-MPP (mean R^2 0.9990)

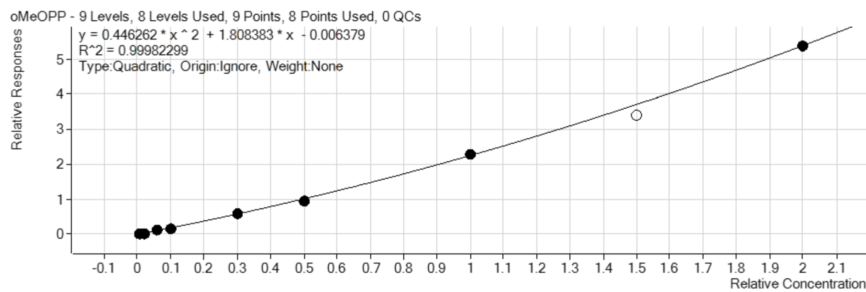


Figure 19: Calibration curve for o-MeOPP (mean R^2 0.9986)

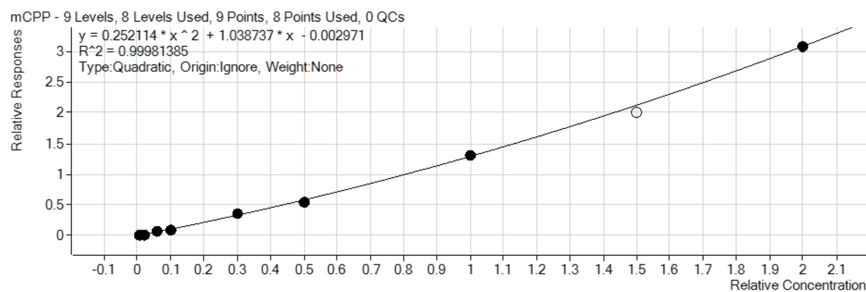


Figure 20: Calibration curve for m-CPP (mean R^2 0.9988)

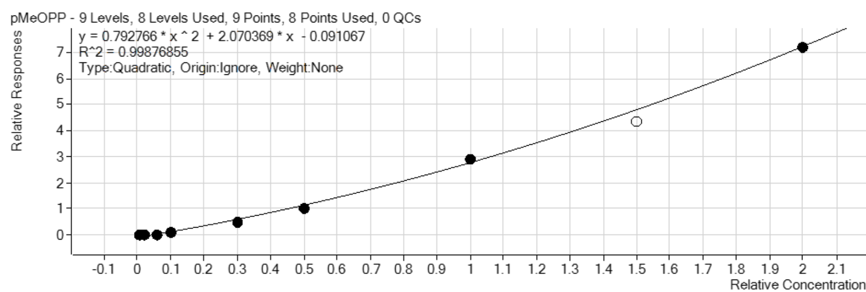


Figure 21: Calibration curve for p-MeOPP (mean R^2 0.9989)

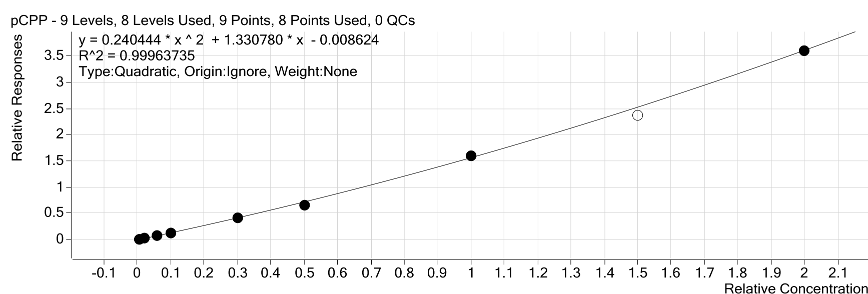


Figure 22: Calibration curve for p-CPP (mean R^2 0.9990)

6.5.1.3 Accuracy

Measurements of accuracy (bias and precision) for each analyte at low, medium and high concentrations are given in Table 6-8. These are within the acceptable limits for all analytes other than m-MPP (not shown). Despite producing calibration curves which appeared to have good concentration/ response correlations, the relationships (response functions) were different each time and the response produced by QCs often equated to concentrations significantly different to those expected. Thus m-MPP is suitable only for detection, and not quantification, by this method. The measured bias are estimates only as only p-CPP was measured with a certified reference standard.

Concentration		pCPP	pMeOPP	mCPP	oMeOPP	pMPP	TFMPP	BZP	pFPP
Low	Bias	-18.5%	15.0%	-12.4%	-17.9%	-11.2%	13.9%	-2.0%	6.1%
Medium	Bias	4.0%	-0.8%	7.0%	3.1%	-0.2%	-9.3%	6.1%	-3.3%
High	Bias	3.4%	-0.1%	4.7%	4.8%	5.7%	-7.9%	7.7%	-6.3%
Low	Precision	9.8%	14.7%	2.1%	16.9%	6.4%	1.9%	3.0%	12.0%
Medium	Precision	2.9%	13.3%	1.7%	3.6%	5.9%	11.9%	4.9%	8.8%
High	Precision	2.2%	4.3%	2.3%	5.7%	3.7%	9.8%	3.9%	8.0%
Low	Between-day-precision	17.0%	12.8%	10.6%	14.1%	10.9%	20.8%	10.1%	15.2%
Medium	Between-day-precision	7.9%	9.9%	7.5%	10.5%	6.0%	6.2%	11.6%	10.8%
High	Between-day-precision	4.6%	5.0%	4.8%	5.5%	5.2%	5.5%	5.5%	6.3%

Table 6-8: GC-MS Accuracy at low, medium and high concentrations

6.5.1.4 Limits of Detection and Quantification

Limits of detection calculated by statistical means which take account of the standard error on ‘y’ (the response, and thus signal to noise) were all very much overestimates of the detection abilities of this method. This is expected to be as a result of the non-linear calibrations. All analytes were reliably detected (acceptable quantifier to qualifier ion ratios) at the lowest calibration standard. Thus the limit of detection was set at this concentration, 5 ng/ml of blood, for

all analytes. It was necessary to determine the limit of quantification empirically using four replicates at various concentrations to find the concentration at which bias and precision were within acceptable limits (20 %). The values estimated for limit of quantification are detailed in Table 6-9.

	pFPP	BZP	TFMPP	pMPP	oMeOPP	mCPP	pMeOPP	pCPP
LLOQ ng/ml	20	30	30	30	30	20	30	20

Table 6-9: Limits of quantification for piperazines

6.5.2 LC-MS/MS Validation

6.5.2.1 Specificity

The specificity of the method was demonstrated by comparison of the chromatogram of blank blood with that of blank blood spiked with each of the piperazine analytes. It was not possible to resolve p-CPP and m-CPP. The method could detect and quantify either of these if present in isolation (using the parameters outlined in Table 6-6) but as they were both included in the calibration solutions they had to be omitted from the LC-MS/MS method validation.

All of the isomers of MPP were well resolved. Again, due to use of the same validation standards for both GC-MS and LC-MS/MS, o-MPP was omitted from the validation. Inclusion would have caused interference in the GC-MS method validation. o-MPP could be included as an analyte of the LC-MS/MS piperazines method with further validation.

LC-MS/MS performance could have been improved by optimisation of the chromatographic method. It would have been desirable to achieve greater separation of the analytes in order to include both *m*- and *p*-CPP, and to remove some of the increased background around p-FPP. It is possible that a different column and mobile phase could have achieved this.

An SCX (strong cation exchange) column allows for an organic rich mobile phase which is advantageous in achieving higher sensitivity in ESI-MS,[221,234] and is beneficial in improving separation of amines.[30,221,234,235] Such columns, however,

are expensive and cannot be applied to many methods across the laboratory. C_{18} columns are applicable to many methods and are, therefore, more economical for busy laboratories. There are nine groups that have published LC-MS/MS methods which have included piperazines. Only two have used an SCX column,[221,222,236] three used polar RP columns,[37,218,237] and the majority (four) C_{18} columns.[203,238-240] These published methods only incorporated up to three different piperazines which made separation much easier. Separation on a C_{18} column can be assisted by addition of an acid to the mobile phase.[30,221,241] Reconstituting the piperazine extracts with an acidic mobile phase would mean that they were ionised prior to being injected on the LC-MS/MS and could increase efficiency of ionisation, although non-volatile acids can cause problems for mass spectrometric detection and, where possible, should be avoided.[28]

6.5.2.2 Calibration Response Function

The mean correlation coefficients (R^2) of the calibration curves prepared from spiked blood had acceptable linearity in the range 5 to 2000 ng/ml of blood for all the piperazine analytes (> 0.99). The best relationship, however, was as a quadratic function as demonstrated in Figure 23 to Figure 29 with mean correlation coefficients provided. There was no carry-over at the highest calibrator concentration (although carry-over can be unpredictable with LC-MS/MS).

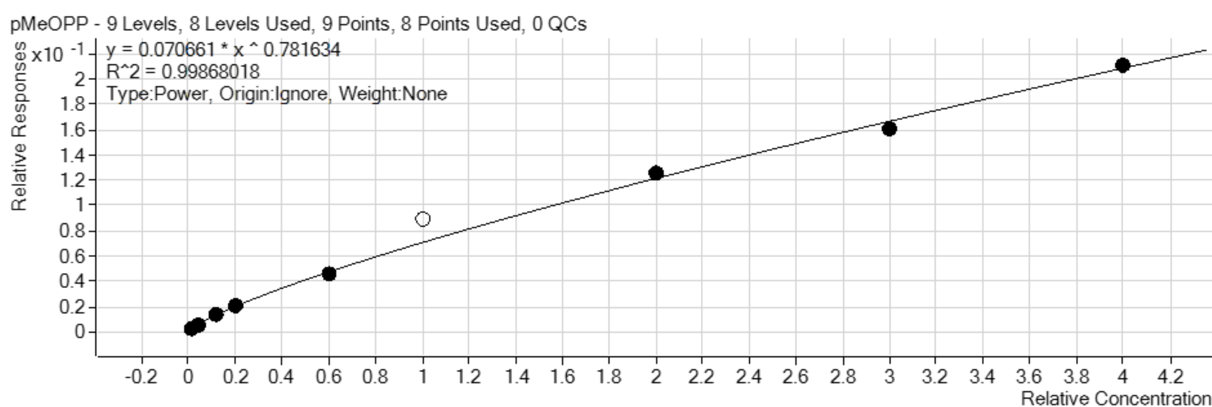


Figure 23: Calibration curve for p-MeOPP (mean R^2 0.9962)

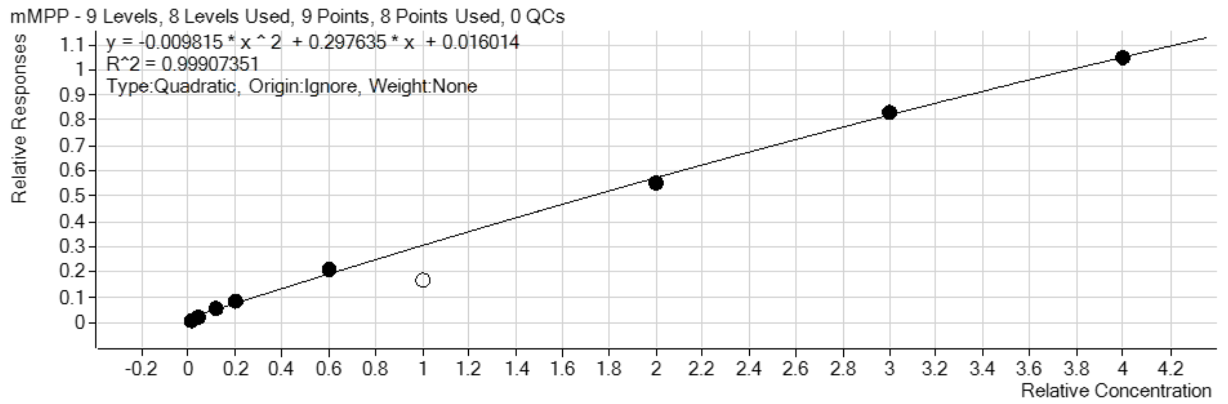


Figure 24: Calibration curve for m-MPP mean (R^2 0.9971)

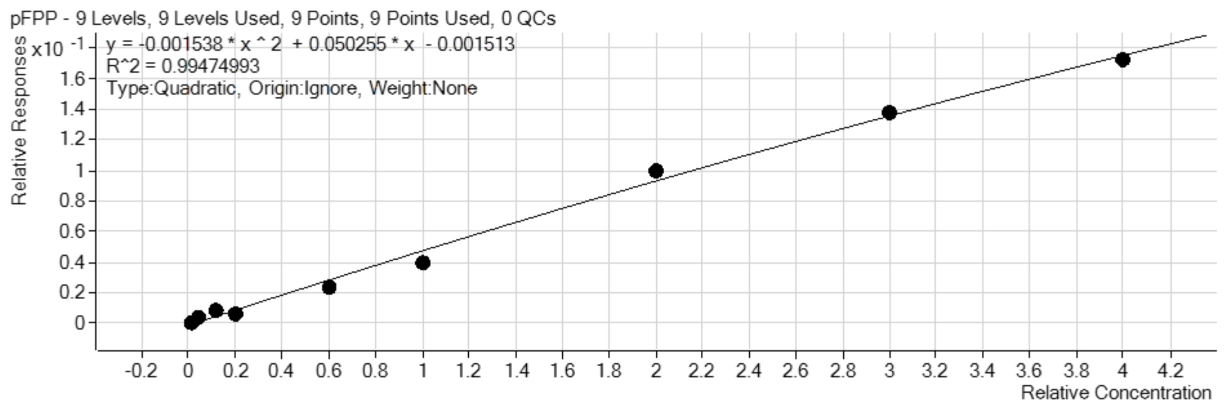


Figure 25: Calibration curve for p-FPP (mean R^2 0.9970)

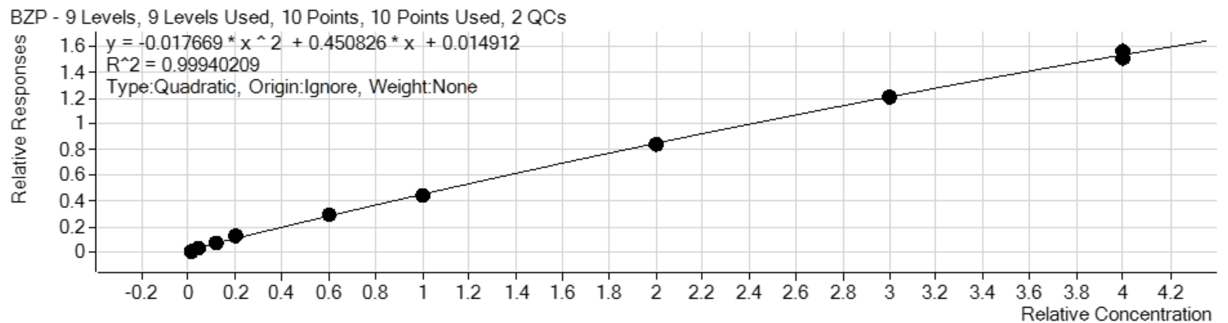


Figure 26: Calibration curve for BZP (mean R^2 0.9972)

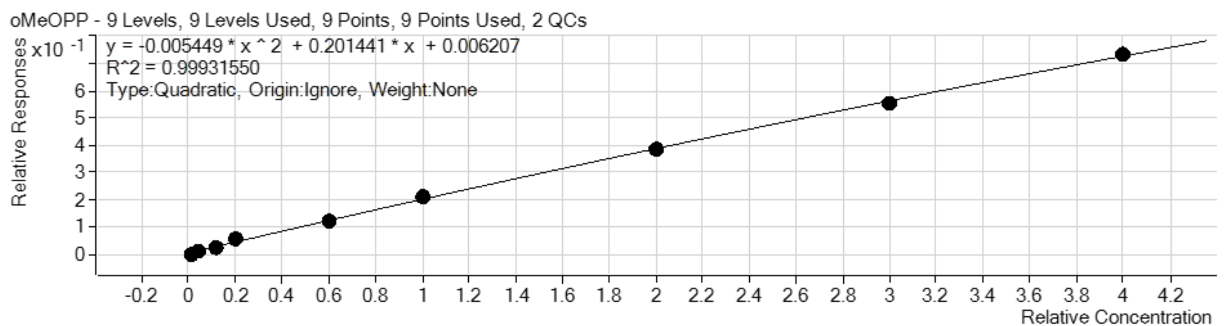


Figure 27: Calibration curve for o-MeOPP (mean R^2 0.9972)

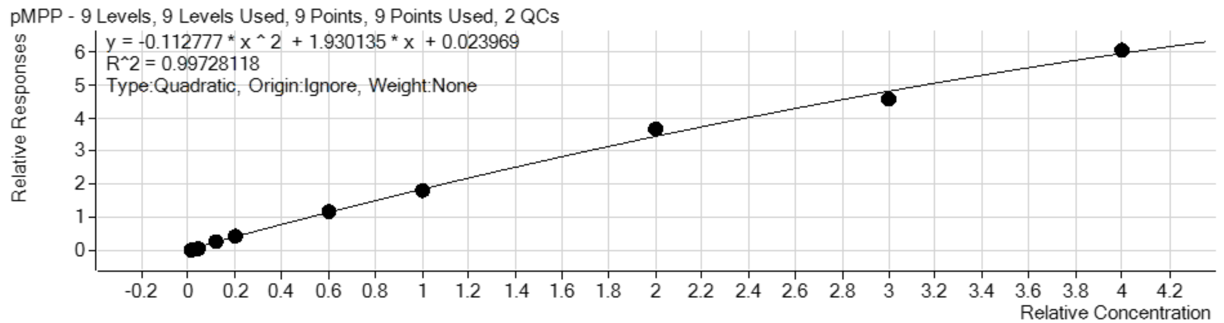


Figure 28: Calibration curve for p-MPP (mean R^2 0.9972)

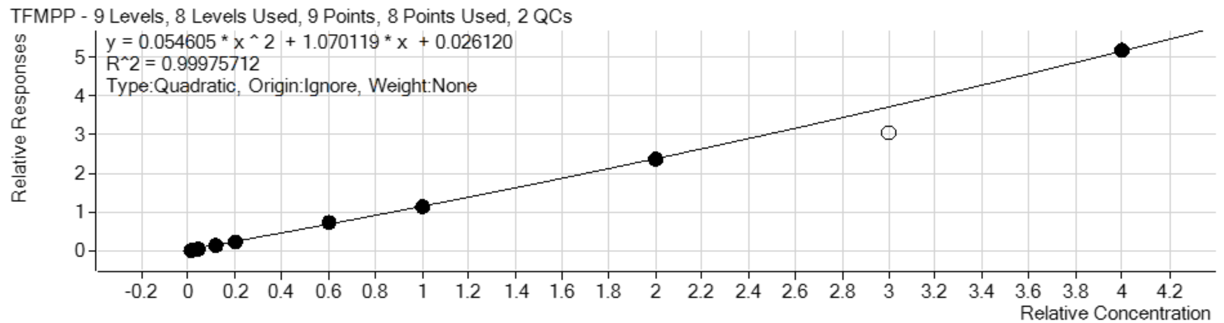


Figure 29: Calibration curve for TFMPP (mean R^2 0.9976)

6.5.2.3 Accuracy

Measurements of accuracy (bias and precision) for each analyte at low, medium and high concentrations are given in Table 6-10. These are within the acceptable limits for all analytes. m-MPP could not be quantified above 1000 ng/ml of blood as, although there appeared to be good correlation, the calculated concentrations were too far from the true values.

It was unusual that for both the GC-MS and LC-MS/MS methods m-MPP did not have a reliable response:concentration relationship. It is difficult to determine why this might have occurred. One possibility is that the calibrator and QC samples were not prepared to the correct concentrations. MPP is very hydroscopic. The bulk powder used to prepare samples had been opened for some time (years) and it may have had a high water content affecting the final concentrations of the prepared samples. The powder should be thoroughly dried to a constant weight, or freshly obtained, and the accuracy experiment repeated to determine whether this might be the cause.

Concentration		pMeOPP	mMpp	pFPP	BZP	oMeOPP	pMPP	TFMPP
Low	Bias	-5.4%	-5.9%	-7.7%	11.9%	7.0%	-7.7%	-9.6%
Medium	Bias	-0.3%	-2.1%	-10.3%	-7.0%	4.1%	-6.2%	-9.6%
High	Bias	-7.0%		3.2%	-3.4%	2.6%	2.8%	1.6%
Low	Precision	18.4%	10.9%	3.7%	12.9%	15.7%	6.3%	7.9%
Medium	Precision	10.2%	9.2%	7.5%	7.1%	7.7%	4.4%	6.0%
High	Precision	9.0%		7.0%	4.4%	13.7%	6.4%	3.6%
Low	Between-day-precision	7.0%	6.7%	8.9%	11.2%	2.0%	12.7%	12.4%
Medium	Between-day-precision	4.3%	9.8%	13.8%	8.8%	11.1%	8.8%	8.8%
High	Between-day-precision	2.8%		3.4%	6.4%	3.6%	3.3%	4.0%

Table 6-10: LC-MS/MS Accuracy at low, medium and high concentrations

Although accuracy and precision were within acceptable limits there really should be more internal standards for an LC-MS/MS method. There should be a co-eluting internal standard for every analyte (or at least within short “retention windows”).[44] This would make the method more robust, particularly with respect to any matrix effects not yet detected but which may occur in real samples (and also to signal drift which is a possibility with LC-MS/MS methods [242]). It would also help to identify any invalid results, for example, if an error in a run caused an analyte which is present not to be detected, it would also be expected that the corresponding internal standard would not be detected and signal that an error had occurred.

6.5.2.4 Limits of Detection/Quantification

Statistically estimated LLOD and LLOQ (Table 6-11) were over-estimates for all the piperazine analytes. The limits for p-FPP were particularly high due to the presence of other analytes causing high background interference. If p-FPP were present in a sample without these other piperazines it is expected that the detection levels would be reduced. p-MPP had a higher estimated LLOQ as, out of all the piperazines, this gave the weakest signal. It was determined by experimentation that at 50 ng/ml of blood, acceptable limits of bias and precision for LLOQ were achieved for all analytes other than p-MeOPP and p-MPP which had LLOQs of 60 ng/ml blood. As all analytes could be readily detected at the lowest calibration standard of 5 ng/ml of blood this was determined to be the LLOD of all the analytes.

(ng/ml)	mMPP	pMeOPP	pFPP	oMeOPP	BZP	pMPP	TFMPP
LLOQ	18	27	37	22	12	51	32
LOD	6	9	12	7	4	17	10

Table 6-11: Statistically-calculated LLOD and LLOQ for LC-MS/MS method

The values of LLOD/LLOQ by this method are slightly better than those attained by Moreno *et al* and Vorce *et al* (100 ng/ml) although their methods were intended for urine where analytes are in greater concentration.[37,223] Other methods have achieved lower limits,[81,239,240] but with fewer analytes and/or a smaller concentration range. In general terms the volume of specimen extracted and the volume of the extract injected to the column will influence the detection levels. The volume on-column was investigated for published methods and lowest levels typically ranged from 0.17 ng [236] to 1 ng on-column.[37,239] LLOD of 5 ng/ml in the validated method was equivalent to 0.2 ng on-column. This is at the lower end of the range of the published methods, thus the LLOD/LLOQ could be reduced by altering the sample volume. A full 1 ml was extracted but only around 200 µl of this was given to the LC-MS/MS method.

6.5.3 Method Performance

The results for extraction recovery, matrix effects (ME) and process efficiency (PE) by GC/MS are presented in Table 6-12 and by the LC-MS/MS method in Table 6-13.

	mMPP	mCPP	pMEOPP	pCPP	oMeOPP	pMPP	pFPP	TFMPP	BZP
Recovery									
Low	105.0%	123.6%	94.7%	140.6%	117.4%	141.4%	119.4%	135.5%	105.0%
Medium	102.5%	98.8%	100.7%	108.2%	93.6%	106.4%	110.0%	112.8%	102.5%
High	101.4%	118.3%	119.4%	87.8%	84.4%	78.6%	92.7%	83.7%	103.7%
Matrix Effects									
Low	96.0%	55.6%	19.0%	98.5%	66.5%	37.3%	70.7%	110.6%	96.0%
Medium	102.7%	65.0%	52.8%	69.9%	73.2%	72.3%	90.2%	105.1%	102.7%
High	117.6%	82.9%	80.4%	87.0%	91.6%	98.8%	136.1%	140.7%	94.6%
Process Efficiency									
Low	100.8%	72.8%	18.0%	138.5%	78.1%	52.7%	84.5%	149.9%	100.8%
Medium	105.2%	64.2%	53.2%	75.6%	68.5%	76.9%	99.2%	118.5%	105.2%
High	119.3%	98.8%	96.0%	76.4%	77.3%	77.6%	126.1%	117.8%	98.1%

Table 6-12: Recovery, matrix effects and process efficiency by GC-MS method.

	BZP	pMeOPP	oMeOPP	pFPP	pMPP	TFMPP
Recovery						
low	72.2%	70.3%	106.6%	73.1%	69.7%	58.3%
medium	90.7%	74.8%	79.4%	77.1%	72.0%	89.5%
high	77.7%	106.5%	127.5%	80.7%	93.9%	70.1%
Matrix Effects						
low	88.8%	105.0%	112.5%	79.6%	109.5%	119.9%
medium	95.0%	116.2%	107.7%	102.5%	107.1%	78.9%
high	116.1%	112.1%	98.5%	108.4%	115.1%	115.0%
Process Efficiency						
low	64.1%	73.9%	120.0%	58.2%	76.3%	69.9%
medium	86.2%	87.0%	109.8%	79.1%	77.1%	70.6%
high	90.2%	119.4%	125.7%	108.5%	108.1%	80.6%

Table 6-13: Recovery, matrix effects and process efficiency by LC-MS/MS method.

6.5.3.1 Extraction Recovery (ER)

Recovery by both methods was >60 % and, therefore, acceptable. It was notable, however, that although the same extracts were used for both methods they produced different values of recovery. The LC-MS/MS values were typically less than the GC-MS values which were often in excess of 100 %. As derivatisation and evaporation were necessary for GC-MS recovery cannot be reliably determined. Evaporation would also have influenced the LC-MS/MS results, although without derivatisation, perhaps to a lesser extent. The high values could also be a result of the co-elution by the GC-MS method, although ions were carefully selected. Acceptable bias is indicative of adequate recovery for the GC-MS method.

6.5.3.2 Apparent Matrix Effects (ME)

The majority of the LC-MS/MS samples were within the desired range of 85 to 115 % indicating no significant ion enhancement or suppression. Matrix effects are not usually present with GC-MS methods, although many of the GC-MS values were outside of the desired range indicating the presences of a (generally suppressing) apparent matrix effect. This could be the result of co-eluting siloxane peaks coming from the GC-column stationary phase suppressing ionisation but is likely to arise from discrepancies between extracted and unextracted standards caused by analyte loss, derivatisation problems, etc. A

greater concentration of analytes was detected in the unextracted standard than the extracted standard. The extracted standard required a greater degree of evaporation (extract volume was 3 ml of elution solvent compared with 50 µl of methanol containing the unextracted standard) which could be responsible for this observation. Matrix effect is not an important parameter for GC-MS validation but if it was necessary to investigate further it could be better addressed through comparison of the concentrations obtained from a spiked blood extract with that of a spiked water extract. Alternatively, matrix effect could be examined with the standard spiked into the blank extract after it had been evaporated, then the volume would be 50 µL of methanol, the same as the unextracted standard, and require the same degree of evaporation. GC-MS recoveries were observed to be greater than 100 %. This is possibly due to the lack of derivatised standards, however, it could also be due to both sample types compared having the same volume of solvent (3 ml).

Five analytes were generally unaffected by apparent ME during GC-MS analysis; m-MPP, p-CPP, TFMPP, p-FPP and BZP. p-MPP and o-MeOPP had problems with apparent ME at low concentrations. m-CPP and, in particular, p-MeOPP require attention.

6.5.3.3 Process Efficiency (PE)

Process efficiencies for both methods were, on the whole, reasonably good. The LC-MS/MS method appeared to have been better for p-MeOPP, o-MeOPP and p-MPP, and GC-MS/MS better for m-MPP, p-FPP, TFMPP and BZP. (p/m-CPP could not be compared.) Where process efficiency is low (p-MeOPP, p- and m-CPP in GC-MS method) it is, again, possible that this was a factor of loss during evaporation where a greater concentration of analyte was detected in a non-extracted sample than a spiked blood extraction, which would have been subject to drying. In order to have a true determination of process efficiency for GC-MS access to derivatised standards of the piperazine analytes would be required, as the non-extracted analytes did not represent 100% efficiency as intended.

It is also possible that the internal standard was unsuitable for the analytes with low process efficiency and did not compensate well for losses during

evaporation. It is, therefore, possible that the true process efficiency for these analytes within the validated GC-MS method is improved due to the more appropriate internal standard. These measures of method performance are generally not required for GC-MS validation, provided accuracy, precision and LLOD/LLOQ are suitable, ER, ME and PE are of little significance in a GC-MS method validation.[57]

With LC-MS/MS methods which do not require evaporation ER, ME and PE are generally measured without use of an internal standard as the internal standard could compensate for errors intended to be measured. In order to account for errors occurring at the auto-sampler, and other instrument variables (pressure, flow, etc), however, an internal standard is necessary. Thus, ideally, internal standard is added immediately prior to injection. This was not possible for the validation of these piperazine methods. The latest opportunity at which the internal standard could be added was after extraction, prior to evaporation.

6.5.4 Stability

The piperazine analytes appear to have short-term stability under all of the conditions tested. Some variation in the measurements is to be expected in accordance with the trueness of the method (precision and accuracy). Samples subjected to three freeze/thaw cycles and those stored in the fridge for up to one week were found to be stable with no significant deviations (greater than 15%) from the peak area ratios of the freshly prepared and extracted standards, Table 6-14. As samples stored in the fridge were stable, these were used to assess the stability of 'in-process', autosampler and bench-top samples, Table 6-15. There were no apparent short term stability problems.

Concentration	Storage	pFPP	BZP	pMPP	TFMPP	oMeOPP	mCPP	pMeOPP	pCPP
Low	Fridge 1 week	8.1%	4.4%	-1.9%	6.6%	-12.9%	0.9%	-4.8%	-3.5%
	Freeze/thaw x3	16.8%	1.8%	5.4%	-13.8%	-10.7%	7.3%	2.5%	-9.8%
Med	Fridge 1 week	-11.0%	-9.4%	11.6%	-12.7%	11.1%	11.9%	8.6%	14.4%
	Freeze/thaw x3	-7.1%	-7.2%	-12.5%	3.0%	-13.4%	7.0%	5.2%	-3.2%
High	Fridge 1 week	3.7%	-7.5%	12.7%	-3.3%	10.6%	13.4%	14.7%	12.3%
	Freeze/thaw x3	-13.0%	-14.6%	-12.7%	-9.0%	-8.9%	-8.5%	-5.7%	-8.2%

Table 6-14: Percentage deviation of peak height area ratios for fridge and freezer stored samples to freshly prepared and extracted samples.

Concentration	Storage	pFPP	BZP	pMPP	TFMPP	oMeOPP	mCPP	pMeOPP	pCPP
	autosampler LC-MS/MS 72 hours	5.3%	1.5%	13.7%	9.7%	-0.1%		12.1%	
Low	Autosampler GC-MS 72 hours	18.4%	9.2%	5.1%	18.6%	7.6%	20.9%	-10.8%	6.1%
	Bench 72 hours	7.5%	0.6%	8.1%	4.5%	-1.6%	5.1%	16.6%	5.8%
	autosampler LC-MS/MS 72 hours	13.2%	0.1%	2.5%	13.5%	4.1%		13.9%	
Medium	Autosampler GC-MS 72 hours	-10.5%	-12.7%	-0.8%	6.3%	6.8%	3.9%	3.7%	8.7%
	Bench 72 hours	-6.9%	-6.2%	7.1%	-1.5%	11.7%	13.0%	11.3%	12.7%
	autosampler LC-MS/MS 72 hours	-12.9%	-9.0%	5.1%	2.5%	13.6%		-9.5%	
High	Autosampler GC-MS 72 hours	-1.6%	1.3%	14.7%	-10.9%	14.4%	10.0%	-6.7%	4.2%
	Bench 72 hours	12.9%	10.5%	8.4%	13.6%	13.0%	12.5%	12.0%	11.4%

Table 6-15: Percentage deviation of peak height area ratios for in-process autosampler and bench-top samples compared to refrigerated samples.

Johnson and Botch-Jones,[224] demonstrated that BZP and TFMPP could remain stable in whole blood for at least fourteen days in the fridge (4°C). At room temperature, however, Botch-Jones found a noticeable decrease in BZP concentration after 4 days (although it was stable for up to 14 days), and after 14 days TFMPP concentration had reduced by 35%. Thus TFMPP is not expected to be stable in blood beyond the 72 hours observed in the current study. Blood thought to contain piperazines should be placed in the freezer as soon as possible or could be stored in the fridge at 4 °C for at least 1 week.

6.5.5 Concentration Range

There are few publications which detail methods used for detection of piperazines in forensic toxicology samples (listed in Table 6-16). The majority of these are LC-MS/MS urine screen methods. None contain more than four piperazines (typically BZP, TFMPP, m-CPP and p-MeOPP). This is understandable with screening methods which incorporate many drugs and have to select those which are perceived to be most relevant. It is, however, strange that two screening methods have omitted BZP, generally considered the most relevant and prevalent of the group.[201,240] It would be advantageous for a dedicated piperazines method to detect more than four analytes, and it is disappointing that two methods only detected BZP which is rarely expected to be found in isolation; BZP and TFMPP would appear to be a minimum requirement. The methods developed and validated here are superior as they include more analytes.

	Method	Matrix	Piperazines	Relationship	Weighting	Calibration Range	LLOD	LLOQ
	LC-MS/MS Screening							
Swortwood *et al*	Designer drugs screen	serum	BZP	Linear	1/x	10 - 250 ng/ml	10 - 100 pg/ml	10 ng/ml
			m CPP	Linear	1/x	10 - 250 ng/ml	10 - 100 pg/ml	10 ng/ml
			TFMPP	Linear	1/x	10 - 250 ng/ml	10 - 100 pg/ml	10 ng/ml
Wolhfarth *et al*	Designer amphetamines, tryptamines and piperazines screen	serum	BZP				5 ng/ml	
			MeOPP				2.5 ng/ml	
			m CPP				1 ng/ml	
			TFMPP				5 ng/ml	
Montesano *et al*	Aphetamines and piperazines screen	urine	FPP	Linear		LOQ - 1000 ng/ml	2 ng/ml	6 ng/ml
			MeOPP	Linear		LOQ - 1000 ng/ml	0.3 - 1 ng/ml	1 - 3 ng/ml
			p CPP	Linear		LOQ - 1000 ng/ml	2 ng/ml	6 ng/ml
			TFMPP	Linear		LOQ - 1000 ng/ml	0.2 ng/ml	0.6 ng/ml
Pichini *et al*	Hallucinogenic designer drugs screen	urine	*m* CPP	Linear		LOQ - 4000 ng/ml	16 ng/ml	53 ng/ml
Nordgren *et al*	Novel substances screen	urine	BZP				8.8 ng/ml	
Strano-Rossi *et al*	Desinger drugs screen	Oral fluid	BZP	quadratic		2 - 1000 ng/ml	2ng/ml	
	GC/MS Screen							
Peters *et al*	Amphetamines and piperazines screen	plasma	BZP	Linear	$1/x^2$	5 - 1000 ng/ml		5 ng/ml
			TFMPP	Linear	$1/x^2$	5 - 1000 ng/ml		5 ng/ml
			m CPP	Linear	$1/x^2$	5 - 1000 ng/ml		5 ng/ml
Ishida *et al*	Drugs of abuse screen	urine	BZP	Linear		100 - 5000 ng/ml		
			TFMPP	Linear		50 - 5000 ng/ml		
			m CPP	Linear		50 - 5000 ng/ml		
			p MeOPP	Linear		50 - 5000 ng/ml		
	LC-MS/MS Confirmation							
Elliot and Smith	Piperazines	Blood	BZP	Linear		312 to 10000 ng/ml	312 ng/ml	500 ng/ml
			TFMPP	Linear		31 -10000 ng/ml	20 ng/ml	31 ng/ml
Moreno *et al*	Piperazines (HPLC-DAD)	urine	BZP	Linear	$1/x^2$	100 - 5000 ng/ml		100 ng/ml
			TFMPP	Linear	1/x	100 - 5000 ng/ml		100 ng/ml
			m CPP	Linear	1/y	100 - 5000 ng/ml		100 ng/ml
			p MeOPP	Linear	$1/x^2$	100 - 5000 ng/ml		100 ng/ml
Vorce *et al*	Piperazines	urine	BZP			100 - 5000 ng/ml	100 ng/ml	100 ng/ml
			TFMPP			100 - 10000 ng/ml	100 ng/ml	100 ng/ml
	GC/MS Confirmation							
Dickson *et al*	Piperazines	urine	BZP	Linear		25 - 2000 ng/ml		
			TFMPP	Linear		25 - 2000 ng/ml		
			m CPP	Linear		25 - 2000 ng/ml		
Wikstrom *et al*	Benzylpiperazine	Blood	BZP	Linear		20 - 2000 ng/ml		
	Hair methods							
Bassindale and Berezowski	Piperazines (LC-MS/MS)	Hair	BZP	Linear				
Barroso *et al*	Piperazines (GC-MS)	Hair	TFMPP	Linear	$1/x^2$	0.05 - 4 ng/mg		0.05 ng/mg
			mCPP	Linear	$1/x^2$	0.05 - 4 ng/mg		0.05 ng/mg
			p MeOPP	Linear	1/x	0.05 - 4 ng/mg		0.05 ng/mg

Table 6-16: Published forensic toxicology piperazine methods

The best response relationship was quadratic for both the GC-MS and LC-MS/MS methods. While in some instances, it can be possible to overcome a quadratic relationship by weighted regression, this was not possible for these methods (although only one published method reports a quadratic relationship, [238] calibration model is likely to be instrument dependant). It is acknowledged in the United Kingdom and Ireland Association of Forensic Toxicologists Forensic toxicology laboratory guidelines (2010),[19] however, that "some assays are inherently non-linear and that the use of quadratic or other mathematical models may be necessary". It is quite common for LC-MS/MS methods to

produce quadratic regressions.[243] The calibrations were reproducible, with sufficient accuracy and the fact that they are not a linear relationship does not, therefore, prevent their application to casework, although careful choice of calibrators is necessary, including LLOQ and ULOQ in every run.[243]

The lack of controlled experimentation related to the toxicity of the piperazines made it difficult to assess relevant concentration range. There was some information available for the more common piperazines. For example, subjects were administered 'recommended' doses (from packaging instructions) of BZP dihydrochloride and TFMPP dihydrochloride and average peak blood concentrations were 589 ng/ml and 41 ng/ml respectively.[180] Such controlled administration is expected to underestimate peak concentrations found in regular users who may consume up to five times as much.[189] It is better to have the calibration range extend slightly higher, particularly as specimen volume may not permit a second, diluted analysis, as occurred with the method developed and applied by Swortwood *et al* which had an upper limit for BZP of 1000 ng/ml.[239] It is not expected that there would be any problem with extending the calibration range for either the GC-MS or LC-MS/MS methods. During pre-validation both were assessed with a calibration range up to 5 µg/ml. A good relationship existed for both although they were quadratic.

It is also true, however that much lower concentrations will be encountered as the result of metabolism or ingestion of a piperazine as the adulterant of another drug, where it would be in a much smaller dose. More realistic expectations of concentrations can be gathered from casework analysis where clinical BZP concentrations have ranged from 20 to 1,200 ng/ml [144,184] and post mortem BZP and TFMPP concentrations ranged from 390 to 1,700 ng/ml, [176,244] and 50 to 150 ng/mL.[176] These post mortem concentrations do not represent fatal piperazine levels as the deaths were not attributed to piperazines, other drugs were present.

These figures suggest that the GC-MS validated calibration range is ideal for BZP as, whilst the LLOQ does not extend to the lower range observed, the majority of samples are expected to fall within the centre of the range, where there is less uncertainty associated with the result. There is only one published GC-MS

method for BZP in blood and this has a very similar calibration range.[144] The validated calibration range for TFMPP is suitable, although it would be preferable to extend a little lower (current LLOQ 30 ng/ml). Many specimens may produce results in the range of increased uncertainty. The highest concentration observed and reported in the literature is a relatively low concentration (150 ng/ml [176]), which if it were measured by this validated method, accuracy and bias would be expected to be greater than 15 %. There are no comparable GC-MS methods for the detection of TFMPP in blood, although Peters *et al* have achieved a better calibration range with plasma, LLOQ 5 ng/ml, but with only three analytes.[81] The validated method favours a bigger picture of identifying what was present in the blood with limited quantification in preference to targeting only a few analytes and quantifying to low concentrations. It is possible that PTV could improve LLOD/LLOQ for piperazines present in lower concentrations such as TFMPP.

The LLOQ for LC-MS/MS is a little too high for TFMPP, however, the method is still useful for detection of lower concentration samples as the LLOD was sufficient. There is only one published LC-MS/MS method for piperazines in blood, by Elliot and Smith.[176] This method achieved an LLOQ for BZP of 500 ng/ml, which is perhaps a little too high, although allows for an extended upper limit of quantification. Elliot and Smith's calibration range for TFMPP, however, was more suitable for application to casework samples than the method developed and validated in this study as it had a lower LLOQ of 31 ng/ml. Again it would appear that the calibration ranges could be more readily tailored to the analytes as these were the only two included in the method. Swortwood *et al* also have a more appropriate TFMPP calibration range for their serum method, however, in this instance BZP is compromised as the calibration range for this analyte is much too low.[239] As previously explained, the calibration range for TFMPP could be improved by loading more analyte on-column by extracting a larger sample volume or using a larger injection volume. The injection volume, however, was optimised to give the best result for the combination of analytes present.

With fewer analytes, method performance parameters such as LLOD/LLOQ are expected to improve. With GC-MS some of the most abundant analyte ions were

shared with other substances which co-eluted and, therefore, could not be used. With LC-MS/MS fewer transitions would increase sensitivity. The common practice of consuming piperazines as blends, however, directs that a piperazine method should contain as many piperazines as possible. It would be undesirable to remove any of the analytes included in the method to increase sensitivity. If use in routine casework demonstrated that the calibration performance required improvement at the lower end (or more analytes were required to be included) it might be necessary to split the target analytes and have two runs. One run could be for more abundant analytes such as BZP and another run for those present at lower concentrations, such as TFMPP (with greater sample and injection volume), although this would require more time and expense. Use of the current LC-MS/MS method as a screen to determine what is present prior to quantification by a more developed method could be a more economical solution.

6.5.6 Specificity

At least eight different sources of blank blood were used across pre-validation and validation (with and without internal standards). No interferences from blood were observed for any of the analytes. Spiked blood, however, is not an accurate representation of blood taken from an individual who has used the piperazine drugs. For example there are no metabolites (or other drugs) present which could potentially cause interference. BZP metabolites are expected to be present at very low concentrations and should not be a significant problem. TFMPP metabolite 4-OH-TFMPP, however, could be almost as concentrated in blood as the parent drug TFMPP and it would be sensible to characterise this analyte. As it is not present if TFMPP is co-ingested with a competing substance, it should not be included in the detection method, however, simply assessed for potential to interfere. The unavailability of piperazine metabolite standards prevented this assessment. It is not expected to be a problem for the LC-MS/MS method which was more specific than the GC-MS method in which many substances co-eluted and shared similar fragmentation patterns. When metabolites become available they should be characterised and added to the method to allow information gathering from casework samples to aid interpretation (by indicating the degree of metabolism).

DBZP was not included in the validated method but was characterised to ensure it would not affect the quantification of any of the piperazines of interest, as it could be present in a specimen as an impurity found with BZP. This analyte elutes much later than the others in both the GC-MS and LC-MS/MS methods.

Specificity is complicated for piperazines as TFMPP, BZP, and most often, m-CPP and o-MeOPP have all been detected as metabolites of therapeutic drugs (usually antidepressants).[29,145,211,217-220,245-247] In these circumstances detection of the parent antidepressant drugs and their other metabolites enables differentiation.[217] It would, therefore, be prudent to check for other drugs (e.g. antidepressants, minor tranquilisers) in samples presenting only one piperazine and no other drugs of abuse. Consumption of a piperazine can only be concluded upon demonstration of the absence of antidepressant drugs.[145]

In order to complete the specificity study it is necessary to test the method with post-mortem blood due to the possibility of there being putrefaction bases which could interfere with the analysis of stimulants.[90] This was not possible as in Scotland post mortem specimens cannot be used for research purposes. When the method is applied to casework this issue can be monitored and the range of blood samples tested will increase confidence in the specificity of the methods. Validation will not be fully complete until the method has been applied to real samples.

6.6 Discussion of Further Work

Knowledge of how piperazines are used, and previous testing of amphetamine-positive specimens,[226,248] indicates that all suspected (and confirmed) amphetamine cases should be tested for the presence of piperazines. It would be desirable to retrospectively apply the validated methods to casework samples (having confirmed LLOD/LLOQ) to identify whether any piperazines had been present. Unfortunately this was not possible as permission has yet to be granted by the Procurator Fiscal. The methods could, however, be applied to any new cases received into the laboratory (following some further validation to ensure the methods transfer to the routine laboratory instruments).

Addition of another round of tests to routine testing does of course incur greater expense. The popularity of these drugs, however, demonstrates that this is essential. In order to keep costs to a minimum it is possible that the LC-MS/MS method could be used as a screening method (generally piperazines are not detected by immunoassay). This would require using a small amount of the extract. With no requirement to perform an independent extraction and derivatisation, it is a simple additional step in an amphetamines method to incorporate a relatively large range of piperazines into routine practice. The GC-MS method would be required for any confirmation, however, due to the lack of internal standards relatively high LLOQs for LC-MS/MS analysis of some analytes. Re-extraction and derivatisation would be time consuming and expensive, however, worthwhile as the screen would already have demonstrated the presence of a piperazine.

Whilst the sensitivity and process efficiencies achieved by the GC-MS method appear, at the moment, to be satisfactory, gathering of case data may suggest that greater sensitivity was required for some of the piperazines and there would be two means by which this could be explored. One would be to try reducing analyte loss through evaporation by addition of an acid “keeper” prior to evaporation as used for amphetamines.[12,216,249] The analytes are retained by forming the corresponding (non-volatile) salts. Addition of methanolic hydrochloric acid is recommended [12,215] but acetic [249] and tartaric [216] acids have also been effective. The advantage may be outweighed however by the expectation that the salts will build-up on the GC inlet or column, reducing efficiency and increasing the frequency with which instrument maintenance is required. A second, perhaps more promising, option is to ensure that excess HFBA and derivatisation by-product heptafluorobutyric acid, which can result in high background and degradation of the column, are removed completely. It is impossible to do this by evaporation without some loss of derivatives.[81] It is recommended for derivatives to be extracted into hexane and excess HFBA removed by phosphate washing. This could be investigated, however, at the moment there is no apparent need and unnecessary processing should be avoided to maintain reproducibility and accuracy of results.

Upon application of the method to casework, information on the analytes present and their concentrations should be collated. It would be beneficial to share this information (perhaps through TIAFT, UKIAFT or the Scottish Poisons Information Bureau database TOXBASE). There are systems in place to notify the forensic community what new substances have been encountered but what is lacking is data to allow interpretation. In a UK based PTS respondents cited the reason for not performing a BZP analysis was that they would be unable to interpret the results (this is detailed in Chapter 4). It would seem that it is necessary to develop the methods and perform testing before this information can become available. Even if information is not shared, routinely performing piperazines testing and monitoring results would provide an internal database to aid interpretation.

6.7 Conclusion

The validated methods are recommended to be applied to all routine amphetamine casework; the LC-MS/MS method as a simple screen and the GC-MS method for confirmation and quantification. Results from such tests should be collated in an interpretation database.

Chapter 7: Further Validation of The (Preliminary) Field Impairment Test

7.1 Introduction

7.1.1 Driving Under the Influence of Drugs (DUID)

Psychoactive substances (most illicit and some therapeutic drugs) affect the central nervous system to cause impairment of attention and information processing: skills required to drive safely.[250] The 'type' of impairment experienced differs across various drug classifications, for example, stimulants increase confidence and risk taking while depressants slow responses. Numerous studies have been conducted to try to establish the nature and extent of impairment in relation to specific drug classifications. This extensive literature has been reviewed elsewhere.[250-263] Conclusions can be inconsistent. Much of the ambiguity arises from the use of different investigative methods. Each of the methods applied to studies of drug impairment have had unavoidable faults. For example in experimental studies outcomes are influenced by factors such as sample size, 'safe' therapeutic dosing, simulated rather than 'real' driving, and testing of individual skills, rather than driving as a whole. With epidemiological research and surveys, bias can be introduced by the conditions under which data is recorded. Other problems can include having no, or an inappropriate control group; method of case selection (e.g. requiring voluntary participation may not return a representative sample of the driving population, especially not those guilty of DUID, or those experiencing the lifestyle of problem drug use); not analysing for all drugs of abuse and alcohol or considering drug interactions; not quantifying drugs detected or analysing the correct matrix to indicate recent drug use; or, failing to take account of relevant factors in the study population.[252,264] There are many studies where relevant factors are not considered, examples include the reason an accident was serious/fatal (e.g. driver not wearing a seatbelt, driver inexperience, road and weather

conditions), driver's history of drug use (tolerance), time interval between accident and taking of a blood sample, or a drug user's risk-taking lifestyle contributing to traffic accidents. The Scottish Executive found a relationship between 'sensation-seeking' and willingness to drive under the influence of drugs.[265] Many of the respondents who admitted DUID, had also been caught breaking the speed limit, had endorsements on their driving licence or had been in an accident in the previous five years.

As a result, the application of these observations to the driving population in 'real' situations is uncertain.[266] The exact characteristics of impairment related to each drug type are unknown and it is difficult to estimate trends in drug driving.

7.1.1.1 Drugs Detected

Drugs commonly detected in impaired drivers vary between countries and with time.[251,267] Studies across Europe have shown cannabis, the most used illegal drug,[268] and benzodiazepines, the most widely used class of drugs (abused and therapeutic), [269] to be most commonly detected in DUID cases.[270,269,271-273] Generally cannabis is most prevalent (in Australia [251,267], Greece [274], Denmark [275], Switzerland [275,276], Norway [277,278], Belgium [279] and Canada[280]). A factor which may have contributed to this, however, is the long detection window in chronic users, and the reliance upon detection of the metabolite 11-nor-9-carboxy-THC (THC-COOH), rather than active THC which was common place until relatively recently. Thus, it may not be that cannabis is the drug taken most often by drivers,, it could just be that cannabis is the drug most commonly detected.

After cannabis, cocaine is on the whole the most detected illegal drug in Europe.[273] Cocaine use did not feature extensively in older DUID studies, however, widespread recreational use and increased popularity of crack cocaine resulted in studies with high occurrences (Switzerland[281], England[282], and Spain[283]). Cocaine was the most detected drug in DUID cases in the Metropolitan area of England.[284] Similar to THC, however, chronic cocaine users sequester cocaine in deep body stores and small amounts can leach back into the

bloodstream for days after the drug was last used.[285] It may, therefore, be easier to detect than other DUID substances.

THC was only the third most common drug detected in Scottish drivers. Benzodiazepines, detected in 88% of DUID cases,[286] were most prevalent followed by opiates (including methadone).[258,287,288] The high rate of benzodiazepine detection could be a factor in the reliance upon impairment as the UK detection method. It has been found that through observation alone, benzodiazepine impairment is easier to detect than other drugs.[289]

Across all studies multiple drug (polydrug) use by drivers was evidenced. Around 10 % of European instances of DUID involved multiple drug use, although prevalence amongst drivers in general was relatively low, 0.4%.[290] This was not a good indication of the true occurrence in individual countries. Some had a much higher frequency (Norway,[278] Australia,[251] Belgium,[291] and Switzerland [275,276] all >50 %). This was true of Scotland where polydrug use was detected in 63 % of positive samples; benzodiazepines with opioids was most popular (59 % of positive samples).[287] Combined drug use amongst drivers was less in England though, 16%.[292] DUI of combined drugs (including alcohol) poses a serious threat to drivers and other road users.[273] In the majority of accidents involving drivers where drugs were present, a combination of psychoactive substances were detected.[290]

7.1.1.2 Prevalence of Drug Driving in the UK

The average rate of DUID in Europe was estimated at 1.9% of drivers. Again, this was encompassed in a large range (0.2 to 8.2%), and did not reflect the situation within any particular country.[273] The UK did not participate in the driving under the influence of drugs, alcohol and medicines (DRUID) extensive integrated study of European drivers, however, and statistics for DUID in the UK are lacking. It can be misleading to compare crime figures on a year to year basis as there can be many latent influences, for examples, changes in legislation which reclassify offences, or police/government campaigns which target and raise awareness of particular offences. The statistics available do suggest an increased rate of DUID; a tenfold rise from 1985 to 2003,[293] and a doubling from 2004 to 2005.[294]

Yet it is still a very low rate and more recently the number of DUID proceedings has significantly declined.[295] This is likely to be a result of difficulty in obtaining evidence for a DUID prosecution rather than a decline in DUID. When compared to alcohol impairment proceedings (of which there are far more) DUID was less likely to result in a finding of guilt.[286]

A 2008 survey suggested there were in the region of 1.7 million regular drug users in the UK [296] (around four hundred thousand problem drug users,[297,298] and a much higher number of recreational drug users, at least double [298]). It is, therefore, expected that this large drug using population will form part of the 44.8 million UK driving population.[299] An assumption supported by the RAC foundation survey in which over 45 % of young drivers admitted to having driven under the influence of illegal drugs; with one in five doing so every day.[300]

DUID has been found to be more socially acceptable than DUI alcohol.[273] 6% of 17-39 year olds reported that they had driven under the influence of drugs.[265] 68% of RAC survey respondents had travelled in a car driven by a drug-driver,[294] and 12% of respondents in the THINK! road safety survey believed it was acceptable to drive after taking Class A drugs (and many more under the influence of cannabis).[301] Drug use did not alter intent to drive.[302] These figures depict an alarming attitude and rate of DUID, yet they are expected to be underestimated. Elliot *et al* tried to characterise DUID in a six year study of road traffic fatalities in England and Wales.[244] They reported 32% of driver fatalities positive for drugs. Whilst this is a disturbing number it must be considered that of these drug positive drivers the most common drugs detected were cannabis (it is not specified whether THC or THC-COOH) with a long detection window, and “other drugs”, a category which included therapeutics such as diabetes medications and paracetamol. This very high incidence of drugs in fatalities did not correspond to government accident statistics where accidents were only recorded if a drug was a contributing factor to the cause of the accident and not simply present (Table 7-1).

Contributing Factor	2008		2010		2011		2012	
	All road traffic accidents							
		%		%		%		%
Alcohol	6,758	5.1	5,293	4.4	5,384	4.5	4,963	4.3
Drugs (illicit or medicinal)	687	0.5	565	0.5	644	0.5	622	0.5
	Percentage fatal accidents							
Alcohol		11		7		9		8
Drugs (illicit or medicinal)		3		2		3		2
Total number of accidents	131,582		120,827		118,403		114,696	

Table 7-1: Road traffic accidents for which drugs or alcohol were a contributing factor.

These figures are not expected to be a close resemblance to the true extent of drug driving, however, as they only relate to accidents which the police have attended, alcohol has not been detected and the police officer has requested blood samples to be analysed. Even for road traffic fatalities drug analysis is not routine due to cost.[292] Typically only traffic police officers will receive training for the practices to detect drugs (the field impairment test, FIT), although general duty officers often attend the scene of an accident (or stop a dangerous driver). Without drug recognition and FIT training these officers may not recognise drug behaviours and, therefore, it may not occur to them to request toxicology samples. A decline in laboratory submissions has been reported.[288] This does not necessarily mean that DUID has declined, rather it is likely to be the result of police officers who lack confidence and training failing to request samples.

The difficulty in measuring the rate of DUID in the UK stems largely from the legislation; impairment by both drugs and alcohol is covered by the same Section 4 offence. There are two consequences of this. The first is that instances of DUID cannot be counted as there is no specific charge for this (unlike a Section 5 alcohol specific offence). The second is that drugs are only considered where a driver is alcohol negative but impairment is still suspected. This is primarily due to the convenience and reliability of the alcohol breath test but it is also more economical in terms of police officers' time and public funds. The preference for the breath test is evidenced by its frequency of use in comparison to FITs, Table 7-2. It has been shown that drugs are frequently present in combination with alcohol.[303] Testing only for alcohol means substantial DUID is not detected.

Many believe the UK should have better monitoring and recording of DUID.[286] This claim has been rejected,[304] understood simply as a result of a lack of belief that the significant increase in the cost to road policing would be returned by increased road safety. This is impossible to evaluate without knowledge of its frequency of occurrence and consequences.

ACPO (Figures for England and Wales)		2009		2010		2011		2012		2013	
	Total breath tests administered	223423		170552		156569		175831			
	Total breath tests positive/refused	7638	3.4%	6662	3.9%	7124	4.6%	7123	4.0%		
Christmas Campaign	Total number of FIT conducted	489		396		540		360			
	Number of FITs that resulted in S.4 arrest	87	17.8%	82	20.1%	91	16.9%	77	21.4%		
	Number of S.s arrests without FIT	502		613		818		799			
	Total breath tests administered			100853		88071		83224		100892	
	Total breath tests positive/refused			5652	5.6%	5296	6.1%	4857	5.8%	5170	5.10%
Summer Campaign	Total number of FIT conducted			259		328		284		291	
	Number of FITs that resulted in S.4 arrest			58	22.4%	121	37.0%	63	22.0%	59	20.0%
	Number of S.s arrests without FIT			316		448		432		432	

Table 7-2: Alcohol breath tests and FITs conducted during police summer and Christmas campaigns

7.1.2 Portable Detection Devices

Policing DUID is very difficult. Efforts are being made to simplify this task. It is believed that the greatest aid would be a device which could reliably detect a range of drugs in oral fluid (OF) by the roadside.[257,305] OF testing devices have undergone substantial testing. They lacked necessary reliability (DRUID had a sensitivity and specificity target of 80%, ROSITA had target sensitivity 90% and accuracy in excess of 95%), were expensive and time consuming; none were found suitable for random roadside testing.[306]

Many countries are already employing these, possibly inadequate, screening devices across parts of Europe, the US and Australia,[307] although they are insufficient to be accepted in the UK. In 2012 the Centre for Applied Science and Technology, CAST, produced a Guide to Type Approval Procedures for Preliminary Drug Testing Devices. [308] This defined criteria for a 'road-ready' device (these can be police station based) with regards to safety considerations, calibration records, storage/operating temperatures, environmental testing procedures, target drug concentrations, and repeatability/specificity requirements.[308] The list of drugs to be detected is limited (of note for example is a lack of benzodiazepines, such as alprazolam), and type approval for a selection, or even just one of these drugs is possible.

So far, one police station based device has gained Home Office type approval; the Draeger Drug Test 5000 for cannabis (THC).[309] It is not, as yet, available in UK police stations, however, Sussex police trialled the device over the 2013 Christmas drink/drug driving campaign. The findings of this trial were not published. If OF screening is to be rolled-out across the UK it is important that the reliability of, and competence to use, such testing can be demonstrated. ISO standard ISO 2006 addresses point of care testing, stating the greatest uncertainty to be dependence upon the operator; trained police officers are expected to perform less successfully than validating laboratory staff.[68] Thus, devices should be validated by police officers within each police station and continually monitored. Proficiency testing would be expedient.

Even with such QA in place, this will only ensure that reliability of the test does not fall outside of the acceptable limits set at validation. It will not endorse OF screening as the best means of detecting DUID. DRUID determined that OF screening alone would be deficient. It could only be an additional aid, not the primary method of detection. Until drug screening devices attain the necessary sensitivity, specificity and scope, reliance upon drug detection devices must be discouraged as it could result in officers becoming less skilled in the observation of signs of impairment.[310] Drugs detected by devices will always be limited. Without identification of physical signs of impairment, a driver impaired by a drug not signalled by a device (either because it was not part of the test battery or inadequate sensitivity) could be allowed to continue driving, endangering themselves and others.

Screening with limited abilities (particularly looking for just one drug) could reverse any deterrence the device is intended to instil. Drivers may feel that they can 'get away with it' when under the influence of a substance which is not detected. Even when impairment is still suspected, it would be of increased difficulty to require a driver to provide a blood sample when they have just witnessed a negative screening result. These concerns have been recognised in Western Australia where OF screening is performed for MDMA, Methamphetamine and THC. Screening has come to be relied upon and impairment testing is not being conducted. Drivers do not perceive any risk of being detected whilst driving under the influence drugs which are not screened.[311] In light of the many NPS emerging, some taken specifically to evade detection, limiting on-site testing to a specific panel of drugs would be a major downfall (observation of clinical signs is vital to direct testing for these substances). It is essential that officers are familiar with, and confident in, the application of FIT.

7.1.3 Law Related to Driving Under the Influence of Drugs (DUID)

The legal approach to driving under the influence of drugs varies throughout the World. There are three basic strategies: impairment, zero-tolerance, and *per se* limits. Legislation could be based on one or a combination of these.

7.1.3.1 Impairment

Under impairment laws it is not illegal to drive having taken an illegal (or impairing prescription) drug. It only becomes illegal if the drug influences ability to drive. These laws are very difficult to enforce due to the complexity of demonstrating that the driver's ability to drive was affected by the drug.

Generally, these laws are administered through observation of specific factors related to drivers' co-ordination, balance and/or vital signs. The most well-known of these tests, from which others were developed, is that of the US standardised field sobriety test (SFST) which is performed at the roadside by a patrol officer. It is followed up with a drug evaluation and classification programme (DECP) performed by an officer qualified as a drug recognition expert (DRE).

7.1.3.2 Zero Tolerance

To simplify the process for conviction of DUID several European countries created zero tolerance laws.[275,278,312,313] Any amount of an illegal drug in blood is an offence; there is no requirement to demonstrate impairment. Prompt sampling is essential. Guilt is determined by LLOQ of the analytical method applied, thus, standardisation is extremely important.[277] Either one laboratory has to be responsible for all road traffic samples or analytical cut-off limits for laboratories need to be set and monitored through proficiency testing and audits.

Zero tolerance is a good solution for drugs which exhibit wide variation between blood concentrations and the impairment observed (e.g. amphetamine type stimulants), or drugs where tolerance makes limits inappropriate (e.g. opiates and benzodiazepines). A further advantage of zero-tolerance is that new substances can easily be added to the legislation (anything classed as illegal).[21]

Zero tolerance has not necessarily dispelled the complexities of defining impairment. Problems arise when drivers are found to be influenced by therapeutic drugs for which they hold a valid prescription. Interpretation of scheduled prescription drugs which could also be drugs of abuse, like methadone

or benzodiazepines, is very difficult. Expert opinion is required to state whether the concentration detected is within the therapeutic range and whether the prescription drug has been taken as directed and not abused.[314] Some countries (e.g. Belgium, Czech Republic, Finland, Germany, Latvia, Luxembourg, and Slovakia) combine zero-tolerance for illegal drugs with *per se* limits for prescribed drugs.[286] Few people are found guilty of DUID with only a therapeutic level of prescription drug present.[312] Usually the level far exceeds the therapeutic range (20-70% of patients do not take drugs as prescribed [315]) or has been combined with other substances.[264,278,312,313]

7.1.3.3 UK Legislation

Since the introduction of the Road Traffic Act in 1930, driving under the influence legislation in the UK has centred upon impairment. It is now governed by Sections 4 to 11 of the Road Traffic Act 1988 (as amended by S4 of the Road Traffic Act 1991). Amended Section 4(1) is of greatest interest:

> *"A person who, when driving or attempting to drive a motor vehicle on a road or other public place, is unfit to drive through drink or drugs shall be guilty of an offence"*

Section 11 (2) defines 'drugs' as any intoxicant other than alcohol. This could include illegal substances, prescribed and over the counter medicines, and, herbal remedies.

Section 4(5) defines 'unfit':

> *"a person shall be taken to be unfit to drive if his ability to drive properly is for the time being impaired"*

This legislation, therefore, renders impairment due to drugs illegal but it does not specify what constitutes drug impairment. Through international scientific research it has been possible for alcohol impairment to be stipulated. Section 5 outlines prescribed limits for breath, blood and urine above which a person is deemed 'unfit'. It was desirable for such definitions to be made for drug impairment and in 2012 the Crime and Courts Bill (2012) introduced Section 5A which made it an offence for a person to have a proportion of a specified

controlled drug (any substance specified in part I, II or III of Schedule 2 of the Misuse of Drugs Act 1971) present in their urine or blood which exceeds the specified limit for that drug (unless it was lawfully possessed and taken as prescribed). With the introduction of this Bill the Department for Transport formed an expert panel to define the drugs to be covered by the legislation and determine what the limits should be, and how the detection of multiple drugs (including alcohol) should be treated.[316]

Such definite limits for drugs are untenable.[255,317] Drug impairment is more complex than alcohol due to factors such as the potential for drug synergies, rapid pharmacokinetics of many impairing substances, variable effects of blood concentration depending on whether it is rising or falling, and significant intra-individual susceptibility to drug effects. Additionally the list of impairing substances is too great to be all encompassing. Thus the experts focused on drugs associated with misuse and common therapeutic drugs, and used 'real-life' situations to estimate blood concentrations believed to represent a risk to road safety (by review of epidemiological data).[318] This panel was largely composed of doctors (psychiatrists) prescribing some of these substances. There were three toxicologists on the panel and they believed the levels had been set too high.[208] The Government was also of the view that these levels could possibly be in disagreement with their zero-tolerance of illegal drug use stance; it could be an endorsement that it was alright to take illegal drugs provided "you don't have too much".[319] Zero-tolerance was, therefore, proposed but the difficulties in imposing such a law were acknowledged.[286]

To discern the best approach a 2nd expert committee was appointed (with more toxicologists) and the nation was consulted on what would be favoured out of three options: zero-tolerance, the previously recommended *per se* limits, or a compromise of lowest concentration at which a valid and reliable analytical result could be obtained (informed by toxicologists). Consultation closed and the result reported in March 2014 that the favoured option was the latter, "lowest accidental exposure limit".[320] This has been put to parliament to be written into legislation and is expected to in place by autumn 2014. It will cover 8 illicit and 8 prescription drugs (given in Table 7-3 together with recommended *per se* limits). Consultation is ongoing for amphetamine levels. Interestingly

there is a limit for LSD which currently does not feature in the UKIAFT recommended routine and driver routine testing panels.[208]

Drug	Concentration in Whole blood Proposed Alone µg/L	Recommedation when present for alcohol µg/L	Alcohol when present mg/100ml
THC	5 *	3	20
Cocaine	80**	40	20
Benzoylecgonine (BZE)	500		
Amphetamine	600	300	20
Methamphetamine	200	100	20
MDMA	300	150	20
Ketamine	200	100	20
Morphine	80	40	20
6-monoacetylmorphine	5		
Methadone	500	250	20
Diazepam	550	275	20
Oxazepam	300	150	20
Flunitrazepam	300	150	20
Clonazepam	50	25	20
Temazepam	1000	500	20
Lorazepam	100	50	20
Lysergic acid diethylamide (LSD)	1		

* Slightly higher than impairment threshold to allow for passive inhalation and THC in lipid stores. THC conc in blood declines rapidly. EMCDDA recommendation.
** Very fast elimination of cocaine, rapidly converted to BZE. BZE level higher than impairment to allow for non recent use

Table 7-3: Proposed limits and recommended *per se* limits for drugs in blood.[319,320]

The difficulty in implementing prescribed drug levels is that, unlike alcohol, drugs are not eliminated at a constant rate. The concentration at the time of driving cannot be reliably back-calculated. Attention has, therefore, been directed towards OF and sweat. Section 6C of the Road Traffic Act 1988 allows for recovery of OF or sweat at the roadside with drivers obliged to provide them (Section 7.6). The problem, however, is that roadside specimens can only be used for screening. Confirmation must be performed with blood to demonstrate the drug was having an effect on the body (and not present due to contamination), and also because it is not possible to translate an OF or sweat concentration to blood for comparison to the prescribed limits (although OF limits are expected). A further complication is the lack of approved sampling

and screening devices. Thus, for now, this legislation is largely redundant, although it does hold great potential with *per se* limits and zero-tolerance having been shown to dramatically increase the number of DUID arrests/toxicology submissions in other countries.[321-323]

Sections 6A to 6E were added to the RTA in 2003 (under the Railways and Transport Safety Act) to increase prosecution of DUID cases. Section 6B makes driver cooperation mandatory, where previously police could only act on suspicion of impairment.[257] It allows police to test for driver impairment through a roadside preliminary impairment test which meets the requirements of the Code of Practice issued by the Secretary of State.[324] This is the Field Impairment Test (FIT), described at 7.1.4. FIT does not have criteria to pass or fail but requires the trained police officer's discretion to decide whether or not a driver is impaired. Police Officers have expressed difficulty with this subjective test.[288]

Regardless of the amendments to the RTA, FIT is presently the only available method for detection of DUID and will remain essential to road policing for the foreseeable future. Thus the view of the Faculty of Forensic and Legal Medicine that;

> *"there is no need for a new offence – there is need for better application of current legislation".*[257]

This view was mirrored by Stephen Collier (who developed the UK FIT); FIT should not be replaced but evaluated to identify how it can be improved.[286]

7.1.4 The Field Impairment Test

In order to complete the task of driving safely a driver is required to combine a number of skills in close association (for example alertness, motor skills, visual acuity, reaction time, judgement and decision making). The ability to perform such skilled tasks is referred to as psychomotor performance. Impairment of psychomotor performance can be evaluated by breaking down the larger task into the best approximation of its key component skills and applying a

combination of tests to assess each of them. There are such test available, which belong to one of three groups: [325]

1. Perceptual performance tasks	Measure acuity of the senses, particularly vision and hearing; Usually an estimation of time.
2. Cognitive performance tasks	Determine intellectual functioning, concentration and mental processing; can be tested with simple arithmetic or usually a task to discriminate a specific signal from a selection.
3. Motor performance tasks	evaluate the integrity and function of motor pathways; reaction time tests (button press in response to a stimulus).

Tests of driving impairment target memory, co-ordination, balance and ability to focus and divide attention. As the nature of impairment can be different across drug groups, the assessment of impairment due to drugs requires tasks from each group to be combined. The tests which comprise FIT were originally designed to detect alcohol impairment and, therefore, test the key skills expected to be influenced by central nervous system CNS depressants. The compilation of tasks may not be broad enough to identify other effects on the CNS.

7.1.4.1 History

In the 1940s many countries began developing sobriety tests which were at that time based on clinical experience (symptoms associate with intoxication) rather than scientific principles.[326] Later, in 1981, this testing was standardised and validated as probable cause that a motorist had a blood alcohol concentration of $\geq$ 0.08 percent.[327] Laboratory validations with alcohol are not discussed here as alcohol can easily be detected in breath.[328] The Standardised Field Sobriety Test (SFST) in the USA consists of three tasks: one leg stand (OLS), walk and turn (WAT), and an eye examination for horizontal gaze nystagmus (HGN), an involuntary movement of the eye. In the 1970s, the Los Angeles Police Department, the International Association of the Chiefs of Police (IACP) and the National Highway Traffic Safety Administration joined forces and advanced the Standardised Field Sobriety Test to a Drug Evaluation and Classification Program (DECP), although the alcohol validation was yet to take place.[329-332] Selected

police officers were specially trained and certified as Drug Recognition Experts (DREs). DREs follow drug recognition training (DRT) and refer to the Drug Evaluation and Classification Program to determine whether drivers show impairment and, if so, whether this is due to alcohol or drugs, specifying up to two of seven drug categories.[329-332] DREs systematically collect and document any symptoms of impairment following the 12-step Drug Influence Evaluation (DIE), set out below in Figure 30.[329-331] A DRE exam may be requested by any police officer who witnesses a driver unsatisfactorily complete a roadside Standardised Field Sobriety Test.

1. Breath test
2. Interview
3. Preliminary exam
   - general physical appearance, use of corrective lenses
4. Eye test
   - horizontal and vertical gaze nystagmus and lack of convergence
5. Divided Attention Psychophysical tests: measure balance, motor control, ability to count and estimate time:
   - walk and turn
   - One-leg-stand
   - Romberg balance test
   - Finger-to-nose test
6. Vital signs
   - pulse rate, blood pressure, temperature
7. Dark room
   - reaction of pupils to light
8. Muscle tone
9. Skin examination for injection sites
10. Interrogation
11. Opinion
    - Officers decision of impairment and if so what drug(s) present
    - Allowed to assign two drug classes and if one is confirmed by toxicological analysis it is a vital record of impairment.
12. Toxicological analysis

Figure 30: DRE 12-step procedure

The American Drug Evaluation and Classification Program is the longest established drug detection examination protocol in the world and, therefore, the most extensively tested. The 1985 'John Hopkins Study' (Bigelow *et al* [333]) is regularly cited as the study which validates the DECP. This was not a substantial study and could not be considered a full validation as the sample size was too small and not representative of the driving population, and also, the DREs were instructors with a greater knowledge and experience than would generally be expected from most DREs. Assessments of the Drug Evaluation and Classification Program are complicated by the aim being to test whether any drug impairment was correctly classified, rather than whether the presence of a drug, any drug, was correctly asserted. This explains why the John Hopkins Study is often cited as evidence of the DECP having an accuracy of 92%, when the true ability of the

test to detect the presence of any drug was found to be much less (only 61%, although not explicitly reported).

This was a reasonable pilot study and a number of studies followed, although they were almost all field studies (reviewed elsewhere [334]) rather than rigorous laboratory validation. Field studies are not a good measure of the performance of the DECP as there are unknown influences such as driver confessions, driver being a known drug user, officer being aware of drug trends, physical evidence of drug, unknown drug dose administered, time delay to recover sample, urine rather than blood/oral fluid, and unknown time since administration. Additionally, many studies cannot provide 'miss rate' information i.e. how many impaired drivers were not detected and was it only high dose drivers detected? Eventually more laboratory studies were conducted, although none were able to validate the DECP as a reliable means of detecting drug impairment,[329,330,332] focussing instead on correct drug classification. Drugs which are more easily detected can falsely inflate this accuracy. Heishman *et al*'s [332] laboratory validation focused on drugs which are more difficult to detect, cocaine and cannabis. Only 44% of DRE observations were found to be consistent with toxicology, and only 33% of drug impairment (regardless of classification) was correctly detected; again this was not emphasised in the report.

These 'validations' were for the entire DECP omitting only the driver interview. With far fewer steps and no significant DRT, the SFST initial 'screen' is expected to be even less accurate. Its performance is unknown, however, as it has not been validated or even rigorously evaluated for detection of drugs other than alcohol. In spite of this, it was the reported 'success' of the DECP programme which led the UK to adopt a similar strategy.

In 1997 Scottish police officers travelled to California to be trained in the general principles and techniques of the DECP to devise a UK drug influence recognition training (DIRT) programme and field impairment test (FIT).[310] The FIT is performed by a specially trained police officer and if there is suspicion of impairment there will be a follow-up examination performed by a forensic medical examiner (FME). In 1999, 209 Scottish police officers and FMEs were DIRT and FIT trained.[335] The officers' course of business FITs were monitored

over a two month period. The vast majority of drivers suspected of impairment had positive toxicology (23 of 25, although for at least 11 of these there was additional information such as drugs present or a confession) and officers concluded that the training was worthwhile. The study was, therefore, extended by trained researchers from the Transport Research laboratory (TRL).[335] This was a field study which targeted individuals leaving public houses believed to be frequented by drug users. Tests were conducted in a large van which meant the full procedure could not be conducted (the 'walk and turn' task was omitted), and subjects were interviewed in relation to their drug and alcohol use, without any threat of arrest. These deviations from the FIT procedure, together with an unrepresentative population sample (only 23 'impaired' participants and the majority aged under 30 years), renders this assessment inappropriate as a FIT validation.

Researchers were instructed not to make a decision as to whether or not subjects were impaired, thus, the predictive power and utility of FIT was not determined. The researchers concluded that, without a measure of unimpaired ability, FIT was lacking. They went on, however, to recommend FIT to the Association of Chief Police Officers Traffic Committee and in 2000 a two-day 'National Drug Drive Training' course (five days for instructors) became available.[292,335,336] FIT was subsequently implemented throughout Scotland in 2002,[288] with the reliability still undetermined. In 2004 the Secretary of State for Transport issued a code of practice for preliminary impairment testing; this was simply the original FIT with the addition of a pupillary examination (outlined below).[324] It was a further two years before validation of the UK FIT procedure was reported.

This was not laboratory validation, rather, it was an extensive field study (with some of the associated complications, for example, only suspected drivers were invited to participate, some toxicology samples were from urine, and for around half the samples carboxy-THC was measured rather than THC). It is accepted, however, that this was a reasonably controlled experiment and, without the ability to administer illegal drugs, a good alternative.[287] It was concluded that FIT was a suitable screening tool but improvement was necessary. Again, the major flaw was noted to be the unavailability of a standard measure of 'not

impaired' to compare to driver's FIT performance. Comprehensive validation would require establishing 'normal' performance from unimpaired drivers.[287]

7.1.4.2 Conducting the Field Impairment Test

FIT is designed to simultaneously assess coordination and cognition. At the beginning of each of the four physical tasks comprising FIT, the officer reads specific instructions and provides a demonstration before providing the driver an opportunity to declare any disability, injury or illness that may prevent them completing the task as required. The driver's understanding of the test requirement forms part of the assessment. The driver is required to remember precise instructions, which are lengthy to delay the driver undertaking the task. The delay before starting the task is increased by the officer's demonstrations of the tasks, as delays increase opportunities to apply DIRT. The test takes 10-12 minutes to complete. Recollection whilst undertaking the task evaluates sustained attention and balance in the driver. The officer remains still and stands a short distance from the driver in order not to cause any distraction and records any signs of drug use, for example, slurred speech, drowsiness or facial itching. Officers must conduct the tests, sticking rigidly to the standard protocol throughout.

FIT is intended to be a roadside test, however the officer must be mindful of the safety of the driver, particularly as they may be under the influence of drink or drugs. An appropriate location should have a hard, level, non-slippery surface, be in a well-lit, unobstructed area and away from the public gaze.[324] If the immediate roadside is unsuitable the test can be conducted at a nearby location or at a police station, otherwise due allowance must be given to interpreting the driver's performance. Officers must also be mindful that it may not be appropriate to administer the test if the driver suffers from a medical or mental condition which could affect their performance (this includes being elderly or obese).[337] Environmental conditions such as weather, lighting conditions and shoes worn by the driver also need to be considered and recorded. All results and observations are recorded on a specific form ('FIT Form' in Scotland or MG DDF in England/Wales).

7.1.4.3 Pupillary Examination

Figure 31: Officer performing a pupillary examination

The driver's eyes are studied to assess whether their pupils are displaying constriction or dilation as this can be a symptom of drug use e.g. opiates may cause constriction whilst hallucinogens and stimulants are known to cause dilation; normal pupil diameter is given to be 3-6.5mm.[287,338] A gauge of calibrated pupil size is held to the side of the subject's face as a guide. Reddening or watering of the eyes is also recorded.

7.1.4.4 The Modified Romberg Balance Test (RBT)

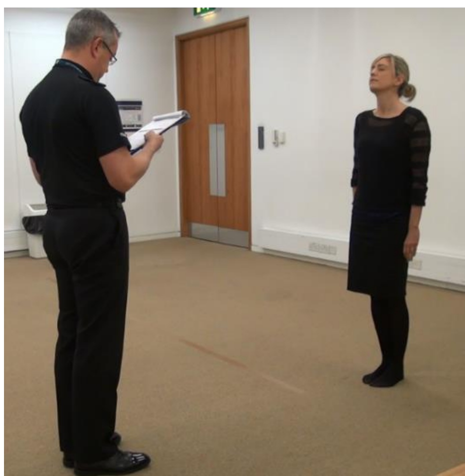


Figure 32: Officer conducting Romberg test

This task is designed to identify temporal disintegration. The driver is asked to estimate the passage of 30 seconds. An unimpaired person is expected to estimate 20 to 40 seconds; less than 20 could signal stimulant use and greater

than 40 a CNS depressant or hallucinogen. Further clues are sought by studying the body language of the subject. The driver is required to stand in a set position with their feet together, arms by their side, head tilted back and eyes closed (since vision aids balance). Any deviation from this including swaying, stepping or raising their arm(s), raising their head or opening their eyes is considered to be a possible sign of impairment.

When the driver has brought their head forward and indicated that they have finished counting, or after they have been stopped by the officer, they are asked "How long was that?". This tests whether the participant was trying to follow the instruction and perhaps made a poor estimate, or if they had not followed the instruction and, therefore not made an estimate of 30 seconds.

7.1.4.5 Walk and Turn (WAT)

Figure 33: Officer demonstrating walk and turn test

This tests the driver's motor function (gross body movement, specifically balance), ability to process instructions and divide attention. The driver is asked to assume a rather unnatural stance on the line. They must maintain this position whilst receiving instruction and demonstration. They must then keep their arms by their side and walk along a straight line taking nine heel-to-toe steps, turn in a specific manner, and take nine heel/toe steps back; counting the steps out loud. Incorrect number of steps, not counting correctly or out loud, missing heel-to-toe, turning incorrectly, raising arms, stumbling or swaying are all errors.

7.1.4.6 One Legged Stand (OLS)

Figure 34: One legged stand test

The driver is required to raise one leg, straight, 6-8 inches above the ground with toes pointing forward, and keep the supporting leg straight. They should maintain this position, count out loud in the manner “1001, 1002, 1003...”, until they are instructed to stop (30 seconds). Repeat with opposite leg. Signs of impairment include hopping, putting foot down, raising arms and not counting correctly.

7.1.4.7 Finger to Nose (FTN)

Figure 35: Finger to nose test

This is designed to assess proprioception (depth perception and balance). Proprioception is the sense that indicates whether the body is moving with required effort, as well as where the various parts of the body are located in relation to each other. This is essential in driving as it enables tasks to be

performed simultaneously, for example, feet can operate pedals, while hands steer and change gears without the driver taking their eyes off the road.

The driver is asked to close their eyes, tilt their head back and touch their nose with the tip of the index finger of each hand on command (left, right, left, right, right, left). Potential signs of impairment include not being able to touch their nose as instructed (using wrong hand or touching another part of their face), hesitation, swaying, stepping, raising arms, raising head, opening eyes.

7.1.4.8 Other Observations

Officers are instructed to note state of clothing, articulation, manner, demeanour, memory of recent events and the presence of any paraphernalia indicative of drug use on the person or in the car, as well as any other relevant observations of behaviour and mental or physical state.[339] "Relevant", however, is not defined and this can lead to extensive note-taking as all actions must be recorded since what may become relevant to the FME cannot be known in advance.

If OF screening devices can be developed with the required reliability, a positive result would negate the requirement for the FME exam. The screening device would be enough to request a blood sample. This would be helpful in reducing the delay between driving and recovery of a blood sample for analysis, however screening cannot be all-encompassing. There are many drugs for which the police FIT followed by an FME medical examination are indispensable.

7.1.5 Interpreting FIT performance

General demeanour and collective performance observed in all FIT tasks are considered by the police officer and a decision made as to whether or not the driver appears to be impaired through drugs. FIT Code of Practice states:

> *"There is no benchmark for pass or failure, nor is there any scoring system to indicate relative success."*

This has been criticised repeatedly as a severe limitation: objective definition of impairment is essential.[287,310,335,340] In the absence of a scientifically determined

definition of FIT impairment legislation relies upon impairment being adduced from:

> *"evidence of driving… such as the driver weaving on the road or even driving too slowly for the conditions" (R v McCall [1974] R.T.R 216)* [339]

The difficulty is then in demonstrating that the impairment is due to drugs, as amended Section 6B 3 of the Road Traffic Act (1988) states that the preliminary impairment test should determine:

> *"whether or not his unfitness is likely to be due to drink or drugs"*

The FME can aid this decision by assessing whether there may be a condition present other than drug use which could result in impairment. There is no nationally agreed examination protocol for FMEs, although in Scotland the results of the medical exam are recorded on an F97 Medical Examination Form and will, therefore, usually follow this format.[341,342] In England and Wales there is a similar pro-forma available. [343] The FME exam will usually consist of the same FIT tasks the police officer performed in addition to a more detailed eye examination including nystagmus and other vital signs such as blood pressure and temperature.

The FME is not legally required to give an opinion of impairment, and is advised against doing so.[286] The function of the FME has been clarified in the Court of Appeal; they should only opine as to whether or not there may be a condition present related to drug use:[344]

> "*the purpose of the medical advice is to provide a protection against the invasive requirement of a blood test when there is a clear medical explanation of the person's condition…Insofar as it might be different, his condition later at the police station is not that to which the investigation is directed…the doctor is not limited to the findings of his or her own police station examination."*

Thus, it is essential for police officers to take very detailed notes. Delay between police FIT and FME examination could result in the driver no longer displaying symptoms of drug impairment. If the officer has detailed notes which indicate possible drug impairment at the road-side, the dissipation of these

symptoms are consistent with a condition due to drug use, and the FME may still permit a blood sample to be taken.[342] Medical evidence is not considered essential as there is no requirement on the driver to consent to a medical examination. Cases have been tried on the evidence of a police officer (impairment) and toxicologist (presence of a drug) alone, for example Leetham v DPP. QB 488 1998.[310,341,345]

The FME must not be omitted from the process, however, as this examination provides independent assessment which could exclude any circumstance (e.g. fatigue), medical condition (diabetes), disease or injury as a possible cause of the impairment; a task that a police officer is not qualified to perform.[310,337,341] Often, an FME will conclude that impairment observed by a police officer was not drug-related (around 25% of examinations).[287]

Ultimately, it is the responsibility of the court, not the police officer or FME, to decide whether at the time of driving there was impairment due to drugs. As the first point of contact, however, the role of the police officer is critical. Their observational skills and ability to describe and record their findings for the court are vitally important. Detection is wholly dependent upon police officers' observational skills. It is believed, however, that many police officers do not have sufficient training in this area.[287,288] Officers doubt their ability to execute FIT effectively; they lack confidence in detecting the symptoms of drug use,[288] are unsure how to complete the required paperwork, and lack conviction in requesting toxicology samples.[292] This is apparent in the declining Section 4 toxicology submissions, which increased when FIT was first introduced.[288] It may also explain why during the Christmas drink/drug driving campaign of 2005, 11 UK police forces did not perform any FIT tests at all.[294] Officers claim a lack of regular training is partly to blame.[288]

7.1.5.1 Officer's Competence

With so many officers expressing concern over their ability to effectively conduct FIT it is clear that their ability to confidently apply FIT must be improved. This should begin with their training. UK officers' training consists of one day within a longer course. The morning covers theory and in the afternoon

the trainee officer performs two assessed FITs on colleagues. By contrast, DRE certification requires candidates to attend nine days of classroom instruction before completing a written examination and performing twelve DECP evaluations on arrested individuals with 75% of their conclusions corroborated by toxicological findings.[332] There are also criteria for decertification and reinstatement.[346] It is recognised that the DECP in the USA is more complex than FIT but they are intended to fulfil the same purpose of detecting impairment consequent to drug use. The UK training, and perhaps complexity of the test, is inadequate.

Involuntary responses, which cannot be learned or controlled, are the best diagnostic tests of drug use.[329] Symptoms strongly associated with drug use include pupil size in darkness, reaction to light, muscle tone, blood pressure and systolic blood pressure.[329,346] Better accuracy has been found with horizontal gaze nystagmus alone than has been recorded for the entire FIT battery (with alcohol). These clinical symptoms are not incorporated by FIT but ought to form part of the FME examination. [338] These are more objective measurements than the FIT tasks, with better-defined ‘normal’ response. Incorporating these symptoms to roadside FIT would increase the information on which to make an assessment and could increase police officers’ confidence with interpretation. This would also increase the relevant information available to the FME who is often significantly delayed in their evaluation of the driver. There are few FMEs and their priority is assessment of arrestees’ fitness to be detained which can mean long delay to DUID cases. Sir Peter North advised that this often deters officers processing a suspected driver, as taking them back to the police station is likely to be a waste of their time without an FME available.[286] If at least some officers were given more extensive training they could be authorised to approve a blood sample prior to the FME exam. It is not possible to omit the FME altogether as only a medical expert can properly evaluate anything which affects balance, while ensuring the health and wellbeing of the driver.[286]

With any subjective examination experience is essential. One of the key recommendations of a major study conducted by the special EU research programme, Impaired Motorists, Methods of Roadside Testing and Assessment for Licensing (IMMORTAL), was that “police officers need training in personal

contact with intoxicated motorists".[347] It would perhaps be more effective if after their initial tuition officers had a period of apprenticeship with an experienced operator before they were assessed and certified. It may be beneficial to officers if there were a minimum number of FITs to be observed following training before they could apply for certification. This would be similar to DRE training, in which the American Department of Transport requires 35 practice tests from different individuals to have been completed over six months before the officer is qualified.[348] This would acquaint officers with an expected range of driver performance. It could also act as an incentive for more FITs to be carried out, as currently FIT is not performed often enough.[349]

In Tunbridge *et al*'s [335] two month FIT evaluation, even when officers were freshly trained and at the peak of awareness, 209 officers completed a total of only 111 FITs. Almost half did not perform a single assessment. FIT is not essential to a DUID arrest so many officers chose not to do it, particularly where there is other supporting evidence.[349] This is at the root of officers lacking competence: without regularly performing FIT they lose skills, confidence and opportunities to gain experience of observing the range of abilities (both impaired and unimpaired). This is at the heart of Sir Peter North's recommendation that:[286]

> *"The number of FIT tests conducted should increase significantly, with forces making it a matter of policy to carry out the test in all cases where impaired driving is suspected, notwithstanding a negative breathalyser test."*

It is most concerning that this advice was rejected by the Secretary of State for Transport who stated FIT was simply a means of requesting a driver to exit their car in order that a police officer can observe their manner and that drug impairment would be evident without specialist training.[304] This is clearly not the case when even those officers who have received specialist training express lack of confidence in their ability to determine whether or not a driver is impaired. Perhaps it is hoped that FIT would become obsolete with the introduction of screening devices, however, lack of skills is already a significant problem which can only be expected to get worse if OF devices come to be relied upon. FITs would be conducted far less often. Systematic application of

FIT to all RTA involved drivers, and those drivers stopped for any suspicious activity, as is performed with breath alcohol, would help to increase the range of performances observed by officers, increase their confidence and familiarity with the FIT SOP.

Although FIT is accredited (see below) and should, therefore, be a standardised procedure, police officers do not always conduct FIT as specified by the SOP.[349] Sergeant Brian Poole, a certified FIT instructor responsible for the development of the FIT training programme in Scotland, advised that officers “forget” the details of the test, particularly the importance of the specific stance to be adopted with heels and toes together and requirement to use the tip of finger rather than pad in the finger to nose test. The full potential of FIT as it is currently applied is, therefore, not being realised and the relatively poor accuracy measures which were observed when it was first instigated could actually now be even lower. If a standardised system to measure impairment were to be developed and used it would be essential that FITs were truly standardised and conducted as the SOP instructs. It is only by this method that the defined ‘normal’ performance will apply. Thus not only is the QMS failing to maintain adequate quality assurance for this procedure, it has also failed to identify the skills and competencies required by police officers and how competencies should be maintained.

To maintain aptitude for FIT assessment, officers’ training should be continuous with regular refresher courses.[292] Refresher training would be useful for officers who have not had call to perform many FIT assessments, but is also necessary to update officers on recognition of emerging substances, particularly with respect to ‘legal highs’. Combining this continuing professional development with officers who are being trained for the first time would mean that the officers would benefit from shared experience. Another means of gaining experience could be for FIT tests conducted in the field to be video recorded. This could also be helpful to the FME if their examination has been delayed. In the US police routinely video their SFST. This is to provide evidence in support of their opinion in court. These would provide a good training aid, however, particularly when toxicology results have been obtained and it is known whether or not the person was impaired by drugs, allowing the officers to observe a range of

performances both impaired by drugs and unimpaired could immediately expand their experience and confidence.

The number of FIT-trained officers also needs to be addressed; there are too few to be effective. The last time the number of FIT trained officers was known was in a TRL report of 2000.[335] There was only one force where more than one third of their traffic officers had FIT training. In Strathclyde where only 5.5% of the police officers in the force were on traffic duty, only ten percent of these had FIT training. In the US an alternative to certified DRE training is the Advanced Roadside Impairment Driving Enforcement program (ARIDE), a sixteen hour training course which can be undertaken by officers who already have SFST training to gain additional skills in recognising signs and symptoms of drug use.[350] An on-line version of this course available from late 2012 makes the skills required for recognising drugged driving much more accessible to officers. Perhaps something like this would be useful in the UK to encourage more uncertified officers to consider the possibility that a driver is impaired by drugs.

7.1.6 FIT Accreditation

The FIT process is accredited to the management standard ISO 9001(2000). This accreditation standard does not address the reliability of the test result; simply that there is documentation in place to map the standardisation of the procedure. It is expected under this standard, however, that the qualification, continuous development and competence of staff be stipulated and adequately safeguarded. It is unclear how this accreditation is being monitored when officers feel so ill-equipped that they are apprehensive about performing FIT.

It is unlikely that FIT could attain accreditation to a testing standard; certainly not without validation. Such psychomotor tests lack validity as there are no databases to well-define the criteria which constitute impairment.[351,352] Johnston and Ramsey stress the urgency with which objective definition of impairment is required, stating that without such standards it is impossible to define acceptable performance versus impairment, raising fears that: [340]

> *“FIT leads to the arrest of people whose only crime is that they cannot pass FIT.”*[340]

7.2 Aim of the Present Study

In order to ascertain that there has been a change in a particular behaviour, it is essential to have first established the population distribution parameters for that behaviour.[254] It is impossible to conclude that performance is impaired, or has dropped below that which would normally be expected, without knowing what is to be expected. An important stage in validation of a test such as FIT should, therefore, be determination of 'normal' performance amongst persons not under the influence of impairing substances. This has not been done and could account for the difficulty officers have expressed with FIT.

In a survey of Scottish FMEs the majority of respondents expressed a desire for an aggregate clinical score from which to determine impairment.[337] FMEs with little experience were over-represented in this group, suggesting that a scoring system was favoured as it would increase confidence and make the process easier. It is considered that such a scoring system could also increase police officers' confidence with FIT.

In order to aid officers and improve the applicability of FIT, it was proposed that the principles of proficiency testing be applied to the robust measures of dispersion from a representative sample of the driving population. In proficiency testing schemes, unacceptable performance is defined as a 'score' more than 3 acceptable variations from the population robust average. Acceptable variation can be represented by a measure of variance within the population (expected to encompass 99.7 % of the population).

The aim was, therefore, to complete the initial stages of a FIT validation by establishing the number of FIT 'errors' displayed by persons not-impaired by drugs for development of a PTS-type scoring system for FIT interpretation.

7.3 Experimental

7.3.1 Participant Recruitment

In order to satisfy PT guidelines, when participant data is relied upon to generate values for scoring performance, at least 20 to 30 participants are

required.[102] As Dixon et al [338] identified that age is expected to impact FIT performance it was considered appropriate to separate participants into two age groups for evaluation; akin to separating laboratories into peer groups according to method in a PTS. Thus a target for recruitment was set at 60 to 80 participants across two age groups.

Prior to recruiting volunteers the Robertson Centre for Biostatistics at the University of Glasgow was consulted to determine whether this target would be expected to yield meaningful results. This required some prediction of a likely outcome. To aid this estimate a pilot study was conducted in student tutorials where students were required to learn and complete a FIT assessment as part of their coursework. Preliminary data collected from students gave a mean number of errors of 2.6. This was assumed an underestimate of the number of errors, as the test was not performed under standardised conditions, students were testing each other, and did not represent all ages of the driving population. The Robertson Centre for Biostatistics entered mean estimates from 2.6 up to ten to a purpose-written computer programme to determine that with 30 to 40 participants in each group the width of the confidence interval around the mean would be ± 2 to 3 errors; an acceptable level of certainty for performance assessment.

Ethical approval was granted by the University of Glasgow (project number 200120039), Faculty of Medicine Ethics Committee. The study was publicised by posters in university buildings and via email to the staff and some students of the College of Medicine, Veterinary and Life Sciences, the Department for Estates and Buildings and University Central Services. This allowed for a broad selection of participants from within the University, including academics, students, janitors, administrative and cleaning staff. Recruitment was also conducted within some other organisations such as solicitors' offices, where both professionals and secretarial staff accepted the invitation to participate. The only requirement for participants was that they were not under the influence of any potentially impairing substances and that they did not have any disability or illness that would prevent them from holding a normal UK driving license. By this means the sample group did not differ from the general driving population.

Before carrying out FIT, each participant was advised of the process they would undertake and those willing to continue provided written consent (Participant Information form given in Appendix 6). Each participant completed a short questionnaire (Appendix 7) related to potential sources of impairment such as injuries, disabilities, being over-tired or other conditions including use of medications, alcohol or recreational drugs. They were confirmed alcohol free through breath analysis (using either a standard UK police issue breathalyser or a Lion Intoxilyser 600, following manufacturer's guidelines). A 1 ml OF sample was recovered using a Quantisal® collection device to allow later confirmation that participants were not unwittingly under the influence of any substances which might have affected their performance in the test. Participants were provided a reference number which was used to maintain anonymity and to identify their questionnaire, OF sample and FIT form (Appendix 8).

7.3.2 Oral Fluid Samples

The Quantisal® device combined the recovered oral fluid with 3 mL of buffer and preservative in the transportation tube. Oral fluid samples were frozen on the day of recovery and allowed to come to room temperature before analysis. Before the first analysis the samples were filtered using Quantisal® filters.

7.3.2.1 Enzyme-Linked Immunosorbent Assay (ELISA)

Assay	Calibrators (ng/ml)				Cutt off (ng/ml)
	Level 1	Level 2	Level 3	Level 4	
Amphetamines	0	25	100	500	25
Benzodiazepines	0	10	60	300	10
Buprenorphine	0	5	20	100	50
Cannabinoids	0	2	10	50	10
Cocaine	0	10	60	300	10
Methadone	0	25	100	500	25
Methamphetamine	0	25	100	500	25
Opiates	0	10	60	300	10

Table 7-4: Drugs of abuse screened with cut-off concentrations

All kits and reagents were purchased from Immunalysis. The drugs of abuse which were screened are listed in Table 7-4 together with their cut-off and calibrator concentrations. Calibrators were produced by diluting methanolic stock solutions in phosphate buffer saline solution (PBSS), as outlined in Table 7-5. The calibrators were diluted 1:4 (0.5 ml:1.5 ml) in oral fluid buffer before

addition to the wells of 96-well plates to replicate the dilution of the OF samples in the Quantisal® collection devices.

Calibrators		Stock solution		Working solutions (stock solution in 10 ml PBSS)			
	Reference Standard	Volume	Conc	Level 1 µl	Level 2 µl	Level 3 µl	Level 4 µl
amphetamine, methamphetamie and methadone				0	25	100	500
100 µl	d-amphetamine 1mg/ml	10 ml	10 µg/ml				
100 µl	d-methamphetamine 1mg/ml	10 ml	10 µg/ml				
1 ml	methadone 100 µg/ml	10 ml	10 µg/ml				
Oxazepam, benzoylecgonine, morphine				0	10	60	300
1 ml	Oxazepam 1mg/ml	10 ml	100 µg/ml				
1 ml	Oxazepam 100 µg/ml (prepared above)	10 ml	10 µg/ml				
1 ml	Morphine 100 µg/ml	10 ml	10 µg/ml				
1 ml	benzoylecgonine 100 µg/ml	10 ml	10 µg/ml				
THC-COOH							
100 µl	11-nor-9-carbpxy-delta-9-tetrahydrocannabinol 100 µg/ml	10 ml	1 µg/ml	0	20	100	500
Buprenorphine							
100 µl	buprenorphine 100 µg/ml	10 ml	1 µg/ml	0	50	200	1000

Table 7-5: Preparation of ELISA calibrators

Four controls were used. One at 50 % above and one at 50 % below the cut-off concentrations. These were prepared in blank blood as described in Table 7-6; the blank blood used in preparation of these was also included as a control. The blood controls were diluted with PBS (250 µl:1000 µl control:PBS). The final control was blank OF.

Reference Standard	Methanolic working solution		Control Concentration		Working solution in 100 ml blood	
	1 ml Ref Stnd in	Concentration	-50%	50%	-50%	50%
Amphetamine (1mg/ml)	10 ml	100 µg/ml	13 ng/ml	38 ng/ml	13 µl	38 µl
Oxazepam (1mg/ml)	100 ml	10 µg/ml	5 ng/ml	15 ng/ml	50 µl	150 µl
Benzoylecgonine (1mg/ml)	100 ml	10 µg/ml	5 ng/ml	15 ng/ml	50 µl	150 µl
Morphine (1mg/ml)	100 ml	10 µg/ml	5 ng/ml	15 ng/ml	50 µl	150 µl
Methadone (1mg/ml)	10 ml	100 µg/ml	13 ng/ml	38 ng/ml	13 µl	38 µl
Methaphetamine (1mg/ml)	10 ml	100 µg/ml	13 ng/ml	38 ng/ml	13 µl	38 µl
THC-COOH (0.1 mg/ml)	10 ml	10 µg/ml	1 ng/ml	3 ng/ml	10 µl	30 µl

Table 7-6: Preparation of ELISA controls

Calibrators, controls and specimens (10 µl) were pipetted in duplicate into the wells of the microtiter plates. 100 µl of the relevant enzyme conjugate labelled with horseradish peroxidase were added and the plates were left in the dark to incubate for one hour. The wells were washed six times with 300 - 350 µL of deionised water using a microtiter plate washer, inverted and 'slapped' dry to remove residual water. Tetramethylbenzidine (TMB) Substrate (100 µl) was

added and the plates were again set aside to incubate in the dark for up to 30 minutes (monitored and stopped early if dark colouration developed). The reaction was stopped with 1M hydrochloric acid (100 µl). A negative result was indicated by yellow colouration while a positive result was colourless. The absorbance was read at 450nm, without delay, using a microtiter plate reader.

7.3.2.2 GC-MS Basic Drug Screen

The entire remaining OF specimen (*ca*. 995 µl OF in 2850 ml Buffer) was used for a basic drug screen which included methadone, amitriptyline, chlorpheniramine, chlorpromazine, citalopram, clomipramine, cyclizine, diltiazem, diphenhydramine, dothiepin, doxepin, imipramine, lignocaine, metoclopramide, mirtazapine, moclobemide, procyclidine, promethazine, propoxyphene, sertraline, tramadol, venlafaxine, and zolpidem. The specimens were extracted exactly as described for the piperazine samples (Section 6.4.5), with the exception that the 1 ml of blood was replaced with the remaining OF/buffer and the internal standard added to these was 50 µl of 10 µg/ml methadone-d9. The entire eluent was dried and reconstituted in 50 µl ethyl acetate (without derivatisation).

1 µl was analysed by splitless isothermal (250 ^{o}C) injection to the Db-5 column (30 m length, 0.25mm i.d., 0.25 µm film thickness) of the Agilent Technologies GC System 7890A coupled to an Agilent Technologies 5975C inert XLMSD with Triple-Axis Detection. The oven temperature programme has an initial temperature of 100 ^{o}C, which was raised by 8 ^{o}C/min to 200 ^{o}C and held for 10 min, then raised to 300 ^{o}C at 8 ^{o}C/min and held for 5 min, for a total run time of 40 min. The carrier gas was helium at 1 ml/min. The ion source temperature was 230 ^{o}C and the interface 250 ^{o}C. EI+ mode was used with ionisation energy of 70eV. The MS was operated in full scan mode (from m/z 40 to 550). Acquired data was processed using the Agilent MSD ChemStation E.02.01.1177 (2010) data system.

7.3.3 FIT Observations

All tests were video recorded and retained for reference. Participants were informed that these would not be held for more than five years and would be

deleted at the conclusion of the study. To minimise the influence of variables such as weather, lighting conditions, etc. testing took place indoors in a room cleared of furniture and anything which might be considered distracting (e.g. wall clocks and posters). A large number of the initial tests (around 60) were conducted by Sergeant Brian Poole of Police Scotland who brought his extensive experience of conducting FIT. Sergeant Poole is one of the very few Scotland-based Police Officers certified to instruct Drug Influence Recognition and Preliminary Impairment Testing Techniques. He authored the Police Scotland Drink & Drug Driving Standard Operating Procedure and was therefore, considered the ideal person to conduct and 'score' the FITs as he could do so in accordance with this SOP and exactly as officers are instructed to do in their course of duty.

Sergeant Poole wore his police uniform while he conducted and scored each test. The remaining tests were performed in the same fashion, with a very formal approach, by the lead researcher who had observed all of the tests conducted by Sergeant Poole. Participants were read the instructions (as scripted in the example FIT form Appendix 8) and asked to complete each of the FIT tasks. If participants could not complete a task or did not wish to attempt it they were not required to do so. Each time a participant made an error it was recorded on the FIT form. The possible errors associated with each task are listed on the FIT form and include any deviations from the specific instructions such as raising arms, not standing in the correct position, starting the test too soon, swaying, stepping out of position and any general indication of loss of balance. There was no limit to the number of errors which could be recorded.

7.3.4 Data Analysis

Any errors recorded on the FIT forms were transcribed to Microsoft Excel® 2007 worksheets to correlate the number and types of errors occurring in each task, and for FIT overall. Microsoft Excel® data analysis functions were used to apply statistical tests to the data. The tests used included those listed below.

7.3.4.1 Mann-Whitney U-test

A robust method to compare two independent sets of non-parametric data, which may contain different numbers of measurements (an alternative to the between-subjects t-test).[87,353] Mann-Whitney U-test is distribution free and most suitable to data with outliers and when there are skewed distributions.[353] The test relies on ranking the data (i.e. the lowest data value is ranked 1, then the data values are consecutively numbered in order of increasing value with equal values sharing the mean of their ranks). The entire data is ranked as a whole, not in groups. The ranks for each group are then summed and applied to Equation 7-1 as T1 and T2, where n1 and n2 are equal to the number of values in each group. Mann-Whitney U is the lower of the two values. In order to determine whether this is a statistically significant U-value it needs to be converted to a z-value by Equation 7-2. The z-value is checked against a table of z-values for significance at the relevant level. The result reported as the probability that the two groups belong to the same population.

$$\left(n1n2 + \frac{n1(n1+1)}{2} - T1\right) \text{or} \left(n1n2 + \frac{n2(n2+2)}{2} - T2\right)$$

Equation 7-1: Mann-Whitney U-test

$$z = \frac{\frac{n1n2}{2} - U}{\sqrt{\left(\frac{n1n2}{12}\right)(n1+n2+1)}}$$

Equation 7-2: Mann-Whitney U-test z-value

7.3.4.2 Kruskal-Wallis H-test

A robust method to compare more than two independent sets of non-parametric data, which may contain different numbers of measurements (an alternative to analysis of variance, ANOVA). It is important that all datasets are compared together before isolating groups for direct comparison. Comparison of smaller groups reduces the sample size and increases the likelihood of variance between groups. Thus, the data should be checked as one large set first to identify whether there is a statistically significant difference present, before comparing groups to try and identify where this difference exists.[87] The Kruskal-Wallis test

does the initial comparison using data from all of the groups. It is followed by a Kruskal-Wallis post hoc Tukey test which compares two groups at a time to identify which have significant differences.

For Kruskal-Wallis the data in each group should have the same shape of distribution (i.e. skewed in the same direction). The data is pooled and ranked. The ranks for each group are added together and the statistic 'H' is calculated. H roughly follows a chi-squared (X^2) distribution, Equation 7-3, where each group is denoted by A, B, C etc, with the number of values in each n_A, n_B, n_C (N is the sum total of these), and rank totals R_A, R_B, R_C. [87]

$$X^2 = \frac{12}{N^2 + N}\left(\frac{R_A^2}{N_A} + \frac{R_B^2}{N_B} + \frac{R_C^2}{N_C} + \cdots\right) - 3(N+1)$$

Equation 7-3: Chi-squared statistic for Kruskal-Wallis test

The significance of H is determined by comparison to tabulated values, the degrees of freedom is the number of groups minus 1. When all of the groups are identical H=0. H gains value when differences exist between the groups and the greater the difference between the groups, the greater the value of H. When H exceeds a critical value, at least one group is significantly different to at least one other.[354] A probability is reported with the H value. This is the probability that random sampling would result in a sum of ranks as far apart as those observed, if the groups contained samples from populations with identical distributions.[355] Thus, the greater the value of 'p' the more closely related are the groups.

If a large H value is produced indicative of the presence of a significantly different group, a post hoc Tukey HSD test, Equation 7-4, can be used to determine which groups have differences.[355] The mean ranks of each group are calculated. Any groups for which the difference between their mean ranks is greater than the calculated HSD value, can be said to be different.

$$HSD = \frac{q_{\alpha,k,\infty}}{\sqrt{2}}\sqrt{\frac{N(N+1)}{6n}}$$

Equation 7-4: Kruskal Wallis post hoc Tukey test.

q = value from q table (with α significance, k number of samples and ∞ degrees of freedom), N = total sample size, and n = size of the smallest sample in the analysis.

7.3.4.3 One Way Analysis of Variance (ANOVA)

Analysis of variance calculation for the comparison of more than two means to determine whether groups are significantly different. This test applies to normally distributed data and may be influenced by outliers. The null hypothesis is that all the means are the same, this is explored by dividing the total variation among all the data points into variation within each group and variation between each group.[353] It requires two values, the mean sum of squares error, MSe (equal to the within-condition variation/the total number of data points minus the number of groups), and the mean sums of squares between conditions, MSb (equal to between-condition variation/the number of groups minus one). If the null hypothesis is true the estimates of MSe and MSb should be roughly the same and their ratio should be around 1. If the differences between the groups are larger than expected MSb will be larger than MSe. The MSb/MSe ratio is denoted the 'F' value and a table of critical F values identifies whether the difference between groups is greater than would be expected if both groups come from the same population.

$$HSD = q\sqrt{MSe/n*}$$

Equation 7-5: ANOVA post hoc Tukey test
n* = the number of values used to calculate the means of interest

To identify which groups are different requires post hoc testing. This can be done by performing Students t-tests between each group,[87] or an ANOVA post hoc Tukey Analysis, Equation 7-5.[356]

7.4 Results and Discussion

7.4.1 Toxicology Screening

The results of the review of participant suitability for the study are presented in Table 7-7 and Table 7-8. The table identifies that a lot of time and effort was spent screening participants for possible presence of impairing substances, although very few participants were subsequently excluded.

One participant produced a positive alcohol breath test, although there was no alcohol declared on their questionnaire, and this was later declared as residual alcohol from drinking on the previous evening. This participant's FIT observations were not included for data analysis.

All 82 participants had negative ELISA screening for amphetamines, benzodiazepines, buprenorphine, cocaine, methadone, methamphetamine and opiates. Two participants produced positive cannabinoid ELISA screening. These presumptive positive results could not be confirmed as the entire oral fluid sample was required for the basic drugs analysis. No cannabis use was declared on these participants' questionnaires. It was possible that these individuals were not under the influence of THC, that the positive result was a consequence of previous cannabis use, as the metabolite carboxy-THC can remain present and detectable after the impairing effects of THC have dissipated. However, the FIT observations from these participants' were not used, as it could not be confirmed whether or not the participants were under the impairing influence of THC.

There was no declared use of any possibly impairing medications and no positive results from the basic drug screening which included methadone, amitriptyline, chlorpheniramine, chlorpromazine, citalopram, clomipramine, cyclizine, diltiazem, diphenhydramine, dothiepin, doxepin, imipramine, lignocaine, metoclopramide, mirtazapine, moclobemide, procyclidine, promethazine, propoxyphene, sertraline, tramadol, venlafaxine, and zolpidem. The internal standard produced a strong signal in all samples confirming that the system was functioning properly, but there were no detectable signals to indicate the presence of a basic drug.

Participant	Alcohol	Basic drugs	ELISA Screening								Clinical
			Amphetamines	Benzodiazepine	Buprenorphine	Cannabis	Cocaine	Methadone	Methamphetamine	Opiates	
1	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
2	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
4											Inner ear infection
5	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
6	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
7	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
8	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
9	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
10	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
11	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
12	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
15	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
16	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
17	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
18	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
19	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
20	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
21	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
22	negative	negative	negative	negative	negative	**Positive**	negative	negative	negative	negative	
23	negative	negative	negative	negative	negative	**Positive**	negative	negative	negative	negative	
24	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
25	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
26											Vertigo
27	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
28	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
29	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
30	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
31	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
32	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
33	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
34	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
35	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
36	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
37	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
40	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
41	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
42	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
43	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
44	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
45	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
46	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
47	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
48	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
49	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	

Table 7-7: Participant screening results (part 1)

Participant	Alcohol	Basic drugs	ELISA Screening								Clinical
			Amphetamines	Benzodiazepine	Buprenorphine	Cannabis	Cocaine	Methadone	Methamphetamine	Opiates	
50	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
51	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
52	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
53	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
54	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
55	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
56	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
57	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
58	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
59	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
60	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
61	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
62	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
63	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
64	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
65	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
66	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
67	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
68	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
69	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
70	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
71	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
72	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
73	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
74	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
75	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
76	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
77	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
79	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
82	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
84	**Positive**										
85	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
89	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
90	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
91	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
92	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
93	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
94	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
95	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	
96	negative	negative	negative	negative	negative	negative	negative	negative	negative	negative	

Table 7-8: Participant Screening Results (part 2)

7.4.2 FIT Observations

Ninety one participants completed a FIT. Five had to be discounted as they could not be considered representative of a normal unimpaired driver (3 failed toxicology screening and the others were considered to have possible impairment due to severe vertigo and an inner ear infection). A subset of results from 7 participants (4 female and 3 male) were also removed as English was not their first language and it could not be differentiated where there was an error of performance or an error of understanding due to language difficulty. The final results set consisted of observation of 79 participants, 43 females aged 18 to 65, and 36 males aged 19 to 72. In order to satisfy PTS peer group requirements when using participants' data to derive acceptable variation (>30 participants), participants were separated into two categories; participants aged under forty years (n = 41) and those aged forty and over (n = 38). Participant demographics are depicted in Figure 36.

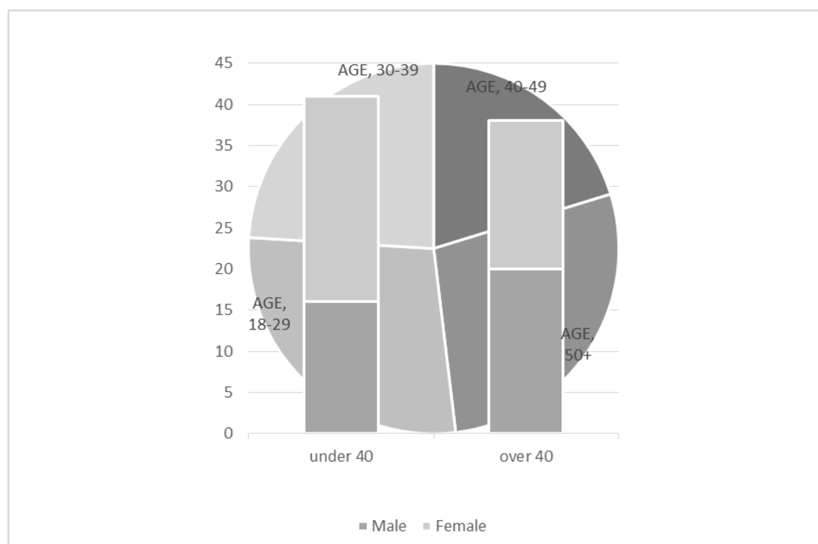


Figure 36: Participation by age and gender.

The participants were a good representation of the driving population in terms of age range and fitness abilities (although there were no unhealthy or extremely unfit participants). Participants were asked to rate their fitness from 1 (excellent) to 5 (extremely unfit). The average was 2 (the highest value was 4). 7 participants removed footwear prior to the test but 3 completed FIT with unsuitable footwear. The participant total error count for each task is listed in Table 7-9 followed by the total number of each type of error observed across all 79 participants. The total number of errors for each error type does not

represent the number of participants making the error as some participants recorded multiple occurrences of the same error type.

Number of errors Observed	Number of participants	Percentage of participants
Modified Romberg Test		
0	45	57.0%
1	26	32.9%
2	8	10.1%
Median	*0*	
Walk and Turn Test		
0	17	21.5%
1	22	27.8%
2	13	16.5%
3	9	11.4%
4	9	11.4%
5	1	1.3%
6	2	2.5%
7	5	6.3%
8	0	0.0%
9	1	1.3%
Median	*2*	
One Leg Stand		
0	15	19.0%
1	25	31.6%
2	15	19.0%
3	6	7.6%
4	3	3.8%
5	3	3.8%
6	4	5.1%
7	5	6.3%
8	2	2.5%
9	0	0.0%
10	0	0.0%
11	0	0.0%
17	1	1.3%
Median	*1*	
Finger to Nose Test		
0	23	29.1%
1	30	38.0%
2	18	22.8%
3	3	3.8%
4	4	5.1%
5	0	0.0%
6	1	1.3%
Median	*1*	

Table 7-9: Error count for each FIT task

7.4.3 Pupillary Examination

Average pupil size was around 5 mm (range 3 to 7 mm). Normal pupil diameter is given to be 3-6.5mm.[287,338] There was one participant found to be just outside of the normal range. Studies which have involved the measurement of pupil diameter have also recorded normal pupil sizes greater than 6.5 mm,[357,358] and have found pupil size to be an age-dependent variable.[357] Pupil size depends to a large extent on lighting conditions (and also other factors such as fatigue and emotion). If standard lighting cannot be applied the range of normal diameters is expected to be large.[357] Pupil diameter has previously been determined not to make a significant contribution to the FIT procedure.[287] Reaction to changing light conditions and the presence of nystagmus are better indicators of drug use.[329,332,346]

7.4.4 Modified Romberg Test

Error	No of observations	% total errors observed for each task
Modified Romberg Balance Test		
During Instruction:		
Moved feet apart	1	2.4%
Light sway	2	4.8%
Did not reply "30 seconds"	5	11.9%
Had to be stopped	1	2.4%
Closed eyes/tilted head before instructed	6	14.3%
Did not say "stop"	15	35.7%
Brought head forward	1	2.4%
Swaying:	10	23.8%
Sway	3	7.1%
Slight sway	4	9.5%
Slight sway throughout	3	7.1%
Raised arms	1	2.4%
total errors	42	

Table 7-10: Total errors observed in the modified Romberg balance task (At section 7.4.10 consideration was given to what should be a significant error as a more lenient scoring method was required. These errors were not counted in the re-evaluation at section 7.4.10)

This appeared to be the easiest of the tasks as the majority of participants did not err, giving a median of 0. Only 43% of participants made at least one error and none made more than 2. The most common error was the participant not

saying “stop” after 30 seconds but using some other phrase to signal the end of the 30 second period (15 people, 18%). Next most frequent error was swaying. It was difficult to evaluate what was a ‘normal’ level of sway with some instances recorded as a slight sway and others simply as sway. This was further complicated by the recording of sway “throughout” rather than having a count of how many times the subject was observed to sway. Other errors regularly observed were to tilt the head back and/or close eyes before being instructed to do so (6 people, 7.4%) and not to reply “30 seconds” when asked how long the participant had been counting (5 people 6.2%).

The average estimate of 30 seconds was 30 seconds with a range from 18 to 44 seconds. This does not include one participant who was still counting after one minute and had to be asked to stop. The guideline on this task is that an unimpaired person’s estimate should fall between 20 to 40 seconds. One subject estimated faster than this range and three subjects slower (not including the individual who did not stop). Thus, five participants were outside of the specified normal range.

7.4.5 Walk and Turn Test

Error	No of observations	% total errors observed for each task
Walk and Turn Test		
During Instruction:		
Did not maintian start position	2	1.2%
Asked for clarification/confirmation	4	2.5%
Stumbled and started again	1	0.6%
Sway	9	5.6%
Slight sway	2	1.2%
Stepped off line	8	5.0%
Raised arams	10	6.2%
Raised arams throughout	1	0.6%
Did not count out loud	4	2.5%
Counting Error	1	0.6%
Took more/less than nine steps	3	1.9%
Stumble	1	0.6%
Sway	3	1.9%
Slight sway	2	1.2%
Raised arms (total)	54	33.5%
Raised for balance	27	16.8%
slightly throughout	8	5.0%
throughout	19	11.8%
Missed heel/toe	31	19.3%
Stepped off	17	10.6%
Turn:	19	11.8%
Not as demonstrated	11	6.8%
Raised arms	3	1.9%
Slight stumble	3	1.9%
Stumble	1	0.6%
Swayed	1	0.6%
total errors	161	

Table 7-11: Total errors observed in the walk and turn task
(These errors were not counted in the re-evaluation at section 7.3.5.10)

The robust average number of participant errors on this task was 2, although most people made just one error. The highest was 9. The majority (around 89%) could not complete this test without error. Missing heel to toe contact on steps and raising arms were most frequently observed errors.

7.4.6 One Leg Stand

Error	No of observations	% total errors observed for each task
One Leg Stand Test		
During Instruction:		
Asked for clarification/confirmation	4	2.1%
Raised arms	3	1.6%
Stumbled	1	0.5%
Counting error	8	4.2%
Footdown and restarted	1	0.5%
Swaying total	48	25.0%
Sway	34	17.7%
left	21	10.9%
right	13	6.8%
Slight sway	9	4.7%
left	5	2.6%
right	4	2.1%
Swayed throught	5	2.6%
left	3	1.6%
right	2	1.0%
Raised arms to balance	15	7.8%
left	9	4.7%
right	6	3.1%
Raised arms throughout	26	13.5%
left	13	6.8%
right	13	6.8%
Hop	4	2.1%
left	2	1.0%
right	2	1.0%
Put foot down	27	14.1%
left	17	8.9%
right	10	5.2%
Not looking at foot	47	24.5%
left	27	14.1%
right	20	10.4%
Could not complete	5	2.6%
left	2	1.0%
right	3	1.6%
Did not point toes/moved foot	3	1.6%
left	0	0.0%
right	3	1.6%
total errors	192	

Table 7-12: Total errors observed in the one legged stand task
(These errors were not counted in the re-evaluation at section 7.3.5.10)

The robust average number of errors on this task was 1 although the range was large, from 19% of participants completing the task without error to one outlier with 17 errors. With the highest number of errors this appeared to be the most difficult task

7.4.7 Finger to Nose Test

Error	No of observations	% total errors observed for each task
Finger to nose test		
Swayed during instruction	1	1.1%
Asked for clarification/confirmation	2	2.1%
Swayed during task	1	1.1%
Tilted head back/closed eyes before asked	3	3.2%
Brought head forward	3	3.2%
Used wrong hand	2	2.1%
Instructed to take hand down/delay in returning	6	6.3%
Missed tip of nose	68	71.6%
Missed nose completely	2	2.1%
Used finger pads rather than tips	7	7.4%
total errors	95	

Table 7-13: Total errors observed in the finger to nose task
(These errors were not counted in the re-evaluation at section 7.4.10)

The robust average number of errors on this task was 1. Around 29% of participants completed the task without error. The maximum number of errors was six. The most frequent error by far was not making contact with the tip of the nose, however most people touched a part of their nose, there were only two occasions where there was contact with another part of the face.

7.4.8 Complete FIT Performance

Only one participant completed all of the FIT tasks without error. The greatest error count was 23. The robust average number of errors for the group as a whole was 4.

Number of errors	Number of participants by age group 18-27	28-37	38-47	48-57	57+	Under	Over 40	ALL
0	1					1		1
1	2		2	1		2	3	5
2	2	4	2			6	2	8
3	2	5		4		7	4	11
4	2	4	1			6	1	7
5	2	2	1	1	2	4	4	8
6		2	2	4		2	6	8
7	3	2		1		5	1	6
8	2	1	2	1		3	3	6
9					1		1	1
10	1			1	2	1	3	4
11	1		1	1	1	1	3	4
12		1	2			1	2	3
13	1					1		1
14					1		1	1
15			1				1	1
16				2			2	2
17		1				1		1
:								
23					1		1	1

Table 7-14: Number of errors for complete FIT by participant age

7.4.9 Influence of Age

Data was analysed primarily by robust statistics to overcome the influence of outliers. A two sided Grubbs test (95% confidence) identified two outliers among the data, one in each of the two age groups. The presence of outliers posed a difficulty in whether or not they should be included in the data analysis. It was considered that they should be included as the participants were known not to be under the influence of any impairing substances and, therefore, whilst a statistical outlier, still represented a possible performance that could be observed from an unimpaired driver. (The use of robust statistics, however, meant that these values had less influence on measures of dispersion.)

When the participants were divided into just two groups, over and under 40 years old (Figure 37), age was found to be a statistically significant factor in the number of errors incurred by participants (Mann-Whitney P=0.029, sig $\leq$0.05, 2 tailed). A non-parametric comparison method was considered most appropriate in order to include outliers.

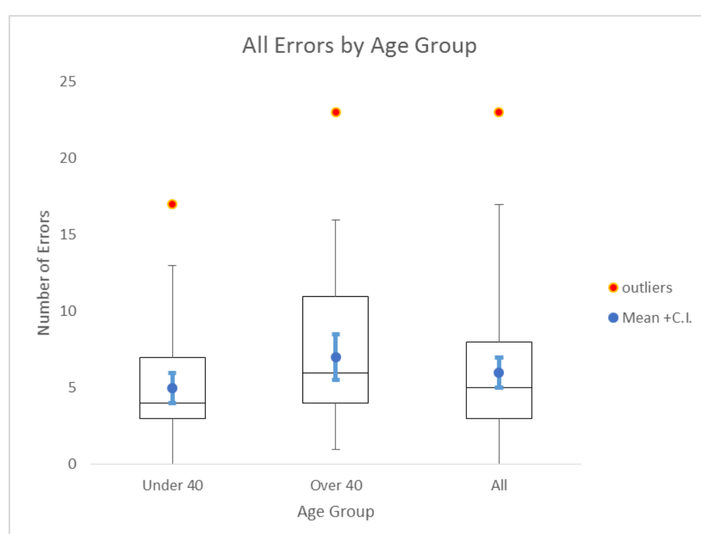


Figure 37: Dispersion of errors in the age groups over and under 40 years.

To expand upon this finding participants were separated into smaller age groups (four groups of ages 18-29, 30-39, 40-49, and 50+), a Kruskal-Wallis test was performed. This did not indicate significant differences in the mean ranks of each age group (H = 0.707 and P = 0.872). The discriminating power of non-parametric significance tests of more than 2 groups is limited, however, as information is lost in substituting data for ranks. In the interest of maximising the information which could be gained from the data, parametric analysis of variance (one way ANOVA) was applied and indicated that the mean error in each age group did differ significantly with and without outliers (F(2.727) = 3.125, p = 0.031, sig 0.05 with outliers, and F(2.73) = 3.52, p= 0.019, sig 0.05 excluding outliers). Post hoc t-tests, however, did not indicate any of the four age groups to be statistically different (the difference between the oldest and youngest group was -1.07 and the acceptable difference at 95% confidence interval was -3.69 to 1.56). When Tukey's test was applied there was an indication of a difference in performance between the age groups 30-39 and 50+ (HSD(3.16)=3.39, sig 0.05) but only when outliers were not included. This difference was slight and could be related to the small sample size of the 30-39 age group, particularly as no other differences were found. If performance in FIT was influenced by age it would be expected for the two extreme groups, those aged 18 to 29 and those aged 50 and over, to be significantly different. They were not.

Age group	No. of participants	Median Error count	Confidence Interval 95%	Inter-quartile Range	Mean error	Standard Deviation	Error range	
18-29	22	5	2 to 8	2 to 8	5	4	0 to 13	
30-39	19	4	3 to 6	3 to 6			2 to 17	
	18	4			5	2	2 to 12	*
40-49	16	6	2 to 11	2 to 9	6	4	1 to 15	
50+	22	7	5 to 11	5 to 11			3 to 23	
	21	6			8	4	3 to 16	**
Under 40	41	4	3 to 6	3 to 7			0 to 17	
	40	4			5	3.1	0 to 13	*
Over 40	38	6	5 to 10	4 to 11			1 to 23	
	37	6			7	4.4	1 to16	**
All	79	5	4 to 7	3 to 8			0 to 23	
	78	5			6	4.1	0 to 17	**

Table 7-15: Median error count and error range for by age groups (* omission of outlier 17 and ** omission of outlier 23). The robust data is that which is relevant, the parametric data is given for comparison only.

The uncertainty associated with the median of some age groups as shown by the confidence interval was large (Table 7-15), indicative of the large spread in the number of errors amongst the subjects of the groups. The large variance could be the result of a lack of data: because there is considerable variance in performance a larger sample is required in order to capture the true spread of performance. To achieve a sufficient degree of statistical power in studies of skills related to driving the European Monitoring Centre for Drugs and Drug Addiction recommend more than 18 subjects per group.[359] The group aged 40 to 49 had only 16 participants but the others had greater than 18, suggesting that the recommended number of participants should be higher.

It is possible that with increased observations the range of error rates across each group could equalise, although it is not possible to speculate. The confidence related to the medians of just two groups (over and under 40 years, each with at least 38 subjects) was more acceptable and supported the proposition that drivers over the age of at least 40 are likely to incur more errors when performing FIT. It was noted, however, that both groups had considerable overlap and the range of each was very large. The overlap between the error ranges for each age group was so great that a comparison of error count and age would be ineffective at determining whether or not an individual may be impaired. Reduction of this range had to be addressed in order to try and establish a useful baseline performance. Interpretation of an ‘error’ was

investigated by looking again at the type of errors being observed. It was found that some errors were observed more often than not. These should therefore not be considered indicative of impairment as they were not unusual behaviour for unimpaired persons. It appeared that the scoring applied was overly critical and a more conservative approach, where only unusual errors were considered significant, could make FIT observations more meaningful.

7.4.10 Omission of Common Errors

A significant proportion of UK FMEs had reservations about how FIT was scored, believing it to be too harsh.[337] Elsewhere officers are trained to be more conservative in assessment. The DECP only allows for each error type to be recorded once, even if they are observed to occur several times and allowances are made for common error types, for example, an arm raise has to be deliberate and higher than six inches.[348] It was noted during the testing that signs and symptoms could easily be missed due to the extensive note taking required. A more conservative scoring approach could mean slightly less recording on the FIT form and increased opportunity got observation of the driver.

The FIT scoring mechanism is subjective and relies upon the discretion of the observer. Some of the errors recorded occurred so frequently that they could be considered less indicative of drug impairment than others; they appear to constitute normal behaviour when balance is tested. An example of this is the observation of “sway”. It is understood that an unwarranted, sizable sway, disproportionate to that which would ‘normally’ be expected may be a valuable observation and should be recorded as an error. There is, of course, difficulty in measuring sway. There were many instances of “slight” sway recorded, however, a slight sway is to be expected. “Slight sway” was discounted as an error.

Errors which had not been counted but recorded to have occurred “throughout” a task were also removed from the second assessment as it could not be determined if these were true errors of performance or errors of understanding. As FIT tests comprehension and information processing together with balance, officers are told not to correct drivers who do not follow the instruction given.

Over the course of this research it became apparent that this was not appropriate. The level of impairment is difficult to assess if the balance tests are not performed as instructed; slight deviations from the intended position can make a task much easier to complete and it is not clear whether the deviation from instruction is because the driver has not understood/remembered the instruction, or because they are incapable of performing the task as required. If a task is not being undertaken as instructed the officer should correct the driver to alert them to this error (recording that they have done so, an error in following instruction). If the driver continues to perform the task incorrectly this can be determined a true error and the number of times the instruction is violated counted rather than "throughout" to allow a more meaningful score to be calculated. This was most applicable to the error "raised arms".

Raising arms slightly to aid balance or having them slightly raised throughout (perhaps because the instruction to keep them by the side has been forgotten) is quite different to raising them high sharply to prevent a stumble or fall, and very different to having them completely outstretched throughout because balance is extremely poor. It should be clear to an observer if arms were raised to avoid a stumble or fall as opposed to inadvertently slightly raised by the sides throughout a test. Raised arms throughout should be considered an error of not following instruction rather than a balance error if there has been no attempt to keep them by the sides. This was evident from some participants having arms raised throughout recorded on their FIT forms in addition to a count of how many times the arms were 'raised'. Instances of raised arms throughout were discounted.

Similarly the error "not as demonstrated" recorded against the turn in the WAT posed interpretive difficulty. All of the participants against whom this was recorded had attempted to perform the turn as demonstrated but made slight mistakes such as using the wrong foot or turning in the wrong direction. Downey *et al* [360] also had difficulty with this fault noting that 'improper turn' was not considered to be related to drug impairment as it was recorded most often for the placebo group. Errors of turn "not as demonstrated" were, therefore, removed. If there was no attempt to make the turn as instructed, however, then this would be countable.

The errors that were removed in the conservative assessment are identified in Table 7-13 as lighter grey text.

7.4.11 Updated performance with conservative assessment

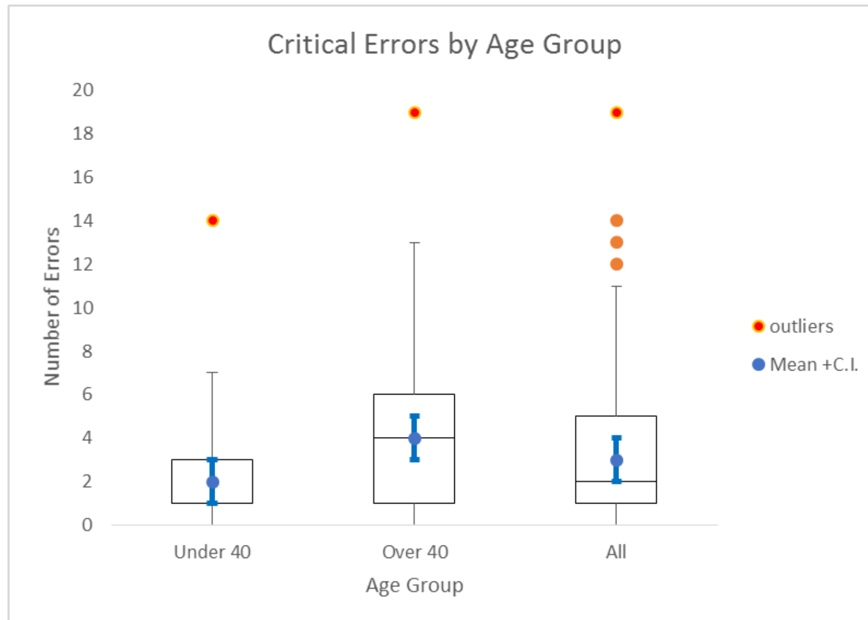


Figure 38: Dispersion of significant errors across groups aged over and under 40 years

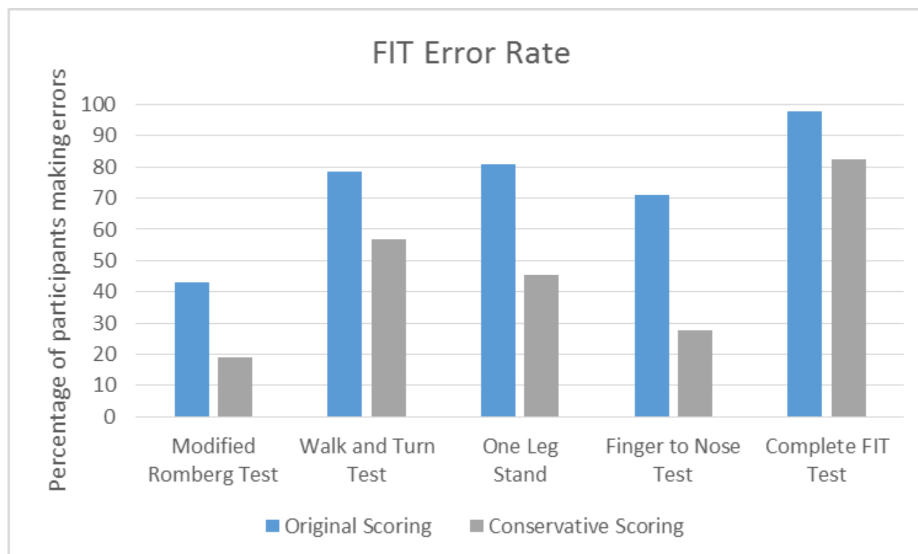


Figure 39: Comparison of the percentage of participants making at least one error by the original scoring system and with the more conservative scoring

Changing the criteria for what constitutes an error had a significant impact on perceived performance. This change in over-all FIT performance can be seen in the changing appearance of the box and whisker plot, Figure 38. The difference was most easily discerned from the new error count for each FIT task, Figure 39 and Table 7-16. There was no significant impact on the RBT. For all the other

tasks, however, the proportion of participants now making at least one error decreased (RBT from 43.0% to 19%, WAT from 78.5% to 57.0%, OLS from 81.0% to 45.6% and FTN from 70.9% to 27.8%), median values reduced by one error, and the error range decreased. The decrease in errors for the OLS was so marked that it no longer appeared to be the most difficult task. WAT was most difficult, followed by OLS. RBT was still easiest to complete. This agrees with previous studies which have largely identified OLS and WAT to be most useful FIT tasks for detection of drug impairment, and FTN least useful (after pupillary examination).[287,329,335,361,362]

The original assessment of OLS as the most difficult task correlated well with the experience of TRL that it was too sensitive, as most participants failed,[335] and Oliver *et al*'s observation that a high proportion of drug negative drivers displayed signs of impairment (even though these were composed largely of younger drivers).[287] Whilst with conservative interpretation the number of errors observed was less than those for WAT, OLS was the only task for which there was a participant that could not complete the test. It would therefore seem that OLS is actually the most difficult task to perform, although error counts for WAT could be expected to be greater as it has an increased number of requirements and increased potential for error. The majority of police surgeons disapproved of WAT and OLS as indicators of drug use but far fewer had reservations regarding RBT and FTN.[337] This could be related to the difficulty unimpaired persons have completing WAT and OLS and the much closer range of error counts for RBT and FTN (where normal performance and therefore impairment are more readily defined).

Noting only significant errors is estimated to make OLS more effective at detecting impairment, although it remains likely that unimpaired persons could experience difficulty. This highlights the importance of interpreting FIT as a whole, rather than looking at performance on individual tasks.

Number of errors Observed	Number of participants	Percentage of participants
Modified Romberg Test		
0	64	81.0%
1	12	15.2%
2	3	3.8%
Median	*0*	
Walk and Turn Test		
0	34	43.0%
1	15	19.0%
2	11	13.9%
3	6	7.6%
4	5	6.3%
5	1	1.3%
6	5	6.3%
7	1	1.3%
8	1	1.3%
9	0	0.0%
Median	*1*	
One Leg Stand		
0	43	54.4%
1	16	20.3%
2	6	7.6%
3	3	3.8%
4	4	5.1%
5	5	6.3%
6	1	1.3%
7	0	0.0%
8	0	0.0%
9	0	0.0%
10	0	0.0%
11	0	0.0%
14	1	1.3%
Median	*0*	
Finger to Nose Test		
0	57	72.2%
1	19	24.1%
2	2	2.5%
3	1	1.3%
Median	*0*	

Table 7-16: Participant error counts by task with conservative assessment

Error Count	Age Group						
	18-29	30-39	40-49	50+	<40	>40	All
0	4	6	1	3	10	4	14
1	5	6	6	2	11	8	19
2	5	2	1	2	7	3	10
3	2	1	1	1	3	2	5
4	2	1	3	2	3	5	8
5	1	2		1	3	1	4
6	2		1	6	2	7	9
7	1		1		1	1	2
8			2	1		3	3
9							
:							
:							
11				1		1	1
12				1		1	1
13				1		1	1
14		1			1		1
:							
19				1		1	1

Table 7-17: Participant total error count by age group with conservative scoring

Following the revised interpretation 14 participants (17.7%) completed FIT without error compared with just one before (1.3%), and over half completed with 2 or less errors (54.4%) compared with less than a fifth (17.7%) of participants. The error ranges for each task were reduced but performance overall still contained some relatively high error counts, Table 7-18, indicating that some people who were not under the influence of an impairing substance had difficulties in completing FIT. The largest error count had a relatively small reduction from 23 to 19. The robust average number of errors across all participants was 2, with a confidence interval 1 to 4 errors calculated by Equation 7-6 and Equation 7-7.[363]

When the results were processed as one large group the number of outliers had increased from 1 with the original scoring to 4. There was still only one outlier in each of the under 40 and 40+ age groups, however, which suggested that it was more appropriate for the data to be divided into these two groups rather than being evaluated collectively. The increased significance of the difference between the over and under 40 years groups was confirmed by statistical testing (Mann-Whitney P=0.00578, sig ≤0.05, 2 tailed). The mean number of errors for the under 40 age group was 1 (confidence interval 1 to 2), and for the 40 and

over age group it was considerably higher at 4, but with less precision (confidence interval 2 to 6).

Age group	No. of participants	Median Error	Confidence Interval	Inter-quartile Range	Mean error	Standard Deviation	Error range	
18-29	22	2	1 to 4	1 to 4	2	2	0 to 7	
30-39	19	1	0 to 2	0 to 3			0 to 14	
	18	1			2	2	0 to 5	*
40-49	16	3	1 to 6	1 to 5	3	3	0 to 8	
50+	23	6	2 to 6	2 to 6			0 to 19	
	22	6			5	3	0 to 13	**
Under 40	41	1	1 to 2	1 to 3			0 to 14	
	40	1			2	2	0 to 7	*
Over 40	38	4	2 to 6	1 to 6			0 to 19	
	37	4			4	3	0 to 13	**
All	79	2	1 to 4	1 to 5			0 to 19	
	75	2		1 to 5	3	3	0 to 11	***

Table 7-18: Number of significant errors by age group: medians and dispersion. Outliers removed: * 14, **19, ***12, 13, 14 and 19

$$\frac{n}{2} - \frac{1.96\sqrt{n}}{2}$$

Equation 7-6: Lower 95% confidence limit ranked value

$$1 + \frac{n}{2} + \frac{1.96\sqrt{n}}{2}$$

Equation 7-7: Upper 95% confidence limit ranked value

It was recognised prior to data analysis that the number of participants could only support two peer groups for PT-type FIT scoring (in order that measurement of deviation attain sufficient certainty). The large dispersion of results for the over 40 age group, and resulting substantial uncertainty associated with the robust average, could indicate that there ought to have been more than two peer groups. Whilst participation was insufficient to allow more, the possibility that further observations of unimpaired performance would have a significant contribution to the resultant scoring system was re-investigated with performance appraisal for four age ranges as before.

The Kruskal Wallis test determined at least one age group to be significantly different to the others (H = 10.251, P = 0.017). Kruskal Wallis post hoc Tukey analysis (using the values in Table 7-19) indicated that the error rates for all four age groups were statistically different to each other. Comparison of the two older age groups with the two younger groups produced the greatest differences in performance, with the older groups having significantly more errors. As information can be lost with such ranking processes, the ANOVA post hoc Tukey analysis was conducted and confirmed that the 50+ age group had statistically significantly higher error rate than the youngest group, ages 18-29 years (HSD(2.91) = 3.091, sig 0.05), and the next youngest group, ages 30-39 year (HSD(2.91) = 3.342). It was clearer with the revised, conservative scoring system that FIT performance deteriorates with age, although when the number of errors was plotted against age there was no correlation, Figure 40.

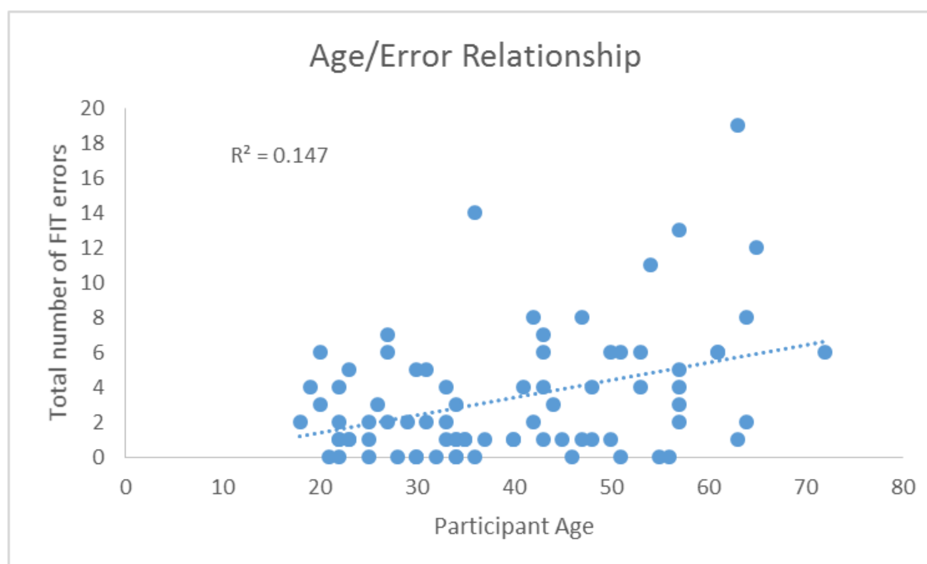


Figure 40: Relationship between participation age and FIT error count.

Deteriorating performance with age has also been noted by Dixon *et al.*[338] Performance of alcohol impaired subjects was compared with unimpaired subjects. Participants over forty years old were found to have higher levels of impairment (both when under the influence of alcohol and when in a placebo state). A substantially higher incidence of false positives was observed with subjects over 40 years of age (participants were aged 18-68; 27% false positive rate for under 40s and 67% for over 40s). This was demonstrated to be a significant problem, with false positives in the age group 41-50 years said to be greater than the number of true positives (exact figures were not given and it

was not possible to calculate the accuracy of FIT with this age group). There were also some suggestions in the work of Tunbridge et al at TRL that age could be a factor in FIT performance.[335] A greater proportion of drivers over 30 were found to fail FIT, however the toxicology results were not correlated with FIT performance and it could not be ascertained whether age or drug use impacted on performance. Additionally, the majority of subjects were under 30 and the over 30 sample size was unreliable. The analysis of FIT results by Oliver *et al* was the only study not to detect a change in performance with age. It is possible that this was because they were using the less conservative scoring method, which failed to detect an age influence in the current study, or it could be due to the majority of participants belonging to the under 40 age group.

Groups compared	Difference in mean ranks	HSD P=0.05	HSD p=0.01
18-29 with 30-39	6.85	3.09	3.89
18-29 with 40-49	5.89		
18-29 with 50+	14.93	N = 79	
30-39 with 40-49	12.74	n = 16	
30-39 with 50+	21.78	df = 60	
40-49 with 50+	9.04	k = 3	

Table 7-19: Kruskal Wallis post hoc Tukey analysis data

7.4.12 Scoring System

A scoring system which works from two different thresholds is required, one for drivers under the age of 40 and another for drivers aged 40 and over. It was hoped that acceptable variation could be calculated by robust methods. The lack of variation in performance amongst the younger age group, however, meant that most often the robust standard deviation was equal to zero. This is not a useful value for performance scoring. The alternative means of calculating acceptable variation was to remove outliers and use the conventional standard deviation. Standard deviation for each age group, together with the resultant acceptable 'unimpaired' error counts are detailed in Table 7-20. It is advised that FIT performance as a whole should be the determining factor in whether or not a driver is impaired. Under this condition it is likely that a driver is impaired if they are under 40 and make more than 7 (conservative) errors, or aged over 40 and make more than 15 (conservative) errors. By this guide one participant

under 40 (aged 36, score 8) and one person aged over 40 (age 63, score 19) known not to be under the influence of an impairing substance would be wrongly suspected. With some discretion from the officer these are not concerning. To aid with such instances the acceptable error counts for each of the individual tasks are provided. If the over-all FIT score exceeds that expected, each task could be independently assessed to determine whether there was perhaps just one task with which the driver experienced difficulties and fell outside of the normal range; this may be expected with the most difficult task, OLS. Three participants scored outside the unimpaired range for this task (2 <40 and 1 >40). For RBT no-one exceeded normal and for WAT and FTN there was one person in the <40 age group outside of the predicted range. It has been suggested that impairment should be concluded if the driver is witnessed to be impaired in two or more individual tests.[360] By this means only one person would have been suspected of impairment. Notably this was one of the participants with unsuitable footwear and if allowance was made for this she may not be deemed impaired.

The two participants with error scores beyond the unimpaired range were close to the higher end of the age groups; all other unimpaired persons (97%) would be correctly diagnosed not to be under the influence of an impairing substance. It seems that perhaps the more pressing problem could be whether or not impaired persons would exceed these definitions of “not impaired”. Two studies have demonstrated a majority of subjects being wrongly classified as impaired,[330] and a high incidence of false positive results amongst a large cohort of unimpaired subjects who were incapable of passing the test.[361]

The wide variation in performance and large range of acceptable errors observed in the current study suggest false positives are a real possibility. This phenomenon was well demonstrated by Tiplady *et al*,[364] when the mean change in performance from placebo to high dose of ethanol was less than the range of placebo performance obtained with a hand-held psychomotor testing device. Effectively an individual with good capacity for the task could remain within the range of normal performance, giving a performance similar to that of an unimpaired person with naturally poor aptitude for the task.

The DRE officers conducting the DECP are provided with a matrix which gives a broad guide to 'normal' measures, for example blood pressure, temperature, pupil diameter, the number and type of error to be expected in each physical task. The DRE officers are required to record whether the driver was normal, above or below. The allocations for normal have been questioned with only partial agreement.[329] 15 of the 29 symptoms listed in the matrix were not significantly different to a placebo 'normal' measure and one was significantly different in the opposite direction to that indicated in the DRE matrix.[329] The wide ranging inter-driver performance suggests that, even with a scoring system, police officers may still have difficulty identifying impairment from a single measure (i.e. no knowledge of how that particular person would score whilst not under the influence of an impairing substance). Thus the scoring system may lack specificity but could remove some of the subjectivity of the FIT process. It could, therefore, provide an effective aid to police officers to increase their confidence to administer FIT, seek samples for toxicology testing and arrest drivers for DUID. It could also provide the Crown with statistically justifiable evidence of possible impairment to aid management and increase prosecutions of such cases. Under these circumstances, impairment is no longer simply a personal opinion.

TASK		< 40	> 40
Complete FIT	Median	1	4
	Standard Deviation	1.96	3.45
	Outliers omitted from SD	1 (14)	1 (19)
	Cautionary error score	> 5	> 7
	Unacceptable error score	**> 7**	**> 15**
RBT	Median	0	0
	Standard Deviation	0.4	0.6
	Outliers omitted from SD		
	Cautionary error score	> 1	> 2
	Unacceptable error score	**> 2**	**> 2**
WAT	Median	0	1.5
	Standard Deviation	1.2	2.22
	Outliers omitted from SD	1 (8)	
	Cautionary error score	> 3	> 6
	Unacceptable error score	**> 4**	**>9**
OLS	Median	0	1
	Standard Deviation	0.85	1.88
	Outliers omitted from SD	2 (4 + 5)	1 (14)
	Cautionary error score	> 2	> 5
	Unacceptable error score	**> 3**	**> 7**
FTN	Median	0	0
	Standard Deviation	0.57	0.45
	Outliers omitted from SD	1 (3)	
	Cautionary error score	> 2	> 1
	Unacceptable error score	**> 2**	**> 2**

Table 7-20: Acceptable numbers of errors for unimpaired persons

A limitation to the UK FIT is that it relies on the driver having a good understanding of the English language. As comprehension is a large constituent of FIT evaluation, if a driver's first language is not English more errors are anticipated. It is also possible that the maximum error counts derived from this laboratory study were less than would be found by the road-side with English speaking drivers. Impairment can be caused by distractions such as anxiety, discomfort and sleep deprivation. Anxiety was apparent with some of the study participants, in particular one who was noted to laugh and sing throughout FIT. This could have been mistaken for the influence of a drug. Most participants displayed some signs of nervousness which may well have been exacerbated to an impairing extent had FIT been performed under usual roadside circumstances.

To minimise external influences the driver should be made as comfortable as possible. One participant arrived wearing an outdoors coat and advised that they might like to remove it as they might find it restrictive. The vast majority of participants were not wearing outdoor clothing which could perhaps have slightly biased the results as many drivers would be expected to be wearing a coat. The influence of outdoor clothing and also footwear should be further investigated, especially for drivers who may wear a restrictive uniform or protective motorcycle apparel. One participant also asked to remove their glasses as their new lenses were causing them some disorientation.

There are countless issues to be considered when assessing a person's performance in divided attention-type tasks. Some of these issues are immediately apparent such as injury or shock which may result from a road traffic accident. Fatigue is known to affect performance of functions such as reaction speed, vigilance, and coordination.[365] Studies of circadian rhythms have found performance to be poorest in the early morning and late at night.[366,367] This is mirrored in studies of driving performance, as is deterioration of driving during the mid-afternoon (1400-1500 hours) dip in wakefulness also.[259,366,368] Practical restrictions in the present study meant that testing could not be conducted late at night or in the very early hours, however testing was conducted over a range of times which included the mid-afternoon dip. Impairment caused by fatigue should be something which the FME can address.

There are some unseen factors, however, which could influence FIT performance. Police Officers have expressed concern that disorders such as dyslexia and dyspraxia negatively affect FIT performance.[337] This could also be true of attention deficit and autistic disorders. The difficulty with these is that, unlike other conditions, an FME may not identify them, particularly as many sufferers can be unaware of their condition.[337] Many adult drivers will have attended school before these conditions were widely recognised and they may never have been diagnosed.

Finally, the scoring system assumes FIT will be performed exactly as directed by the SOP. As different officers apply this to varying degrees, 'normal' scores

could be officer-specific. In order for any scoring to be effectively implemented, FIT officers must be highly competent in conducting and interpreting FIT by a standardised method. This requires intensive, regular training and monitoring.

7.4.13 Effectiveness of FIT

Some of the criticism directed towards FIT has been that the tasks involved do not directly reflect the act of driving and, therefore, cannot provide evidence that driving ability was impaired. The study of drugs and driving has greatly expanded in the time since the introduction of FIT. There is now a wealth of research to demonstrate that drugs impair driving. The law is also changing such that demonstrating impaired driving will no longer be necessary for at least eight illegal drugs. For those drugs not included in the Crime and Courts Bill and therefore without prescribed limits, the Section 4 impairment offence will remain but case law allows impaired driving to be evidenced by witnessing any act of poor driving. It is now possible, therefore, to dismiss this criticism and focus on the true requirement of FIT - to detect a condition that may be due to drugs in order that a toxicology sample can be requested following an FME excluding other medical causes. The ability of FIT to do so has been questioned.

In a study of the Standardised Field Sobriety Test, WAT and OLS were found to be particularly sensitive for cannabis use as more participants failed these tasks whilst in THC state than in placebo state (age range 21 to 35).[360] This study explored and demonstrated differences in performance from the same persons whilst in various states of intoxication. This type of study would not be possible in the field. Had the placebo and THC groups not comprised the same subjects, the difference may not have been statistically significant. Additionally, as participants completed multiple Standardised Field Sobriety Test assessments they were trained to perform the tests, effectively eliminating the comprehension and recollection aspects of the assessment. No significant predictive power was discerned by comparisons of DECP indicators between drug impaired populations and placebo group. The DECP test components which are used in FIT were determined to be ineffective.[329,330]

UK legislation states that the Secretary of State should aim to ensure that a preliminary impairment test is “designed” to indicate whether “unfitness is likely to be due to…drugs”.[369] The tasks which comprise FIT were not specifically designed for the purpose of detection of drug impairment. FIT tasks evolved from observations of alcohol intoxicated individuals and do not appear to be very good at detecting drug impairment.

7.5 Future Work

Further work is required to determine whether two sets of criteria which define impairment in those over and under forty years of age are sufficient, or if performance differences across smaller age ranges achieve statistical significance. This would require greater than 30 participants in each range. The uncertainty associated with the median for the over 40 age group was large and more observations would be desirable to increase confidence in the proposed impairment criteria for drivers over forty. The influence of time of day must also be addressed. Many drivers who commit DUID offences do so late at night or in the early hours of the morning, when impairment in even the normal population could be increased. This could be complicated, however, by the possibility that those under the influence of a stimulant would have better than baseline performance at this hour due to their fatigue compensation.

The final scoring system requires to be tested and validated. Blind testing should be performed to determine the specificity, sensitivity and accuracy in individuals who are unimpaired or under the influence of intoxicating substances. For any test of impairment to be effective and worthwhile it must be demonstrated that there is a significant difference in response between impaired and non-impaired drivers. If this can be determined, a complete FIT validation can be launched. This would require FIT to be observed and scored by a number of different officers to ensure robustness of the SOP and scoring system, similar to performing an analytical method validation under reproducible conditions i.e. applying the same method but on a different instrument or with different operators etc. Applicability of the impairment definitions must also be tested on the roadside.

7.6 Conclusion

A conservative approach to FIT interpretation of ‘error’ produced more meaningful error counts. Errors should not be recorded to have occurred “throughout”, rather a reminder should be given if it appears the instruction has not been carried out and then each instance of the fault recorded. Pupil size may not be informative as unimpaired persons were outside of the defined normal range. The defined range for time estimation was also flawed. WAT and OLS were the most difficult FIT tasks and FTN and RBT were the easiest. Performance with FIT deteriorates with age, with the decrement becoming statistically significant over the age of at least 40 years. There is large variation in the normal performance abilities of individuals, so large that establishing the range of normal performance will require many more than 30 subjects in each age group to be observed, using small age ranges. More than two age ranges is desirable.

In order to boost police officers’ confidence in performing FIT and requesting toxicology samples, a PTS-type scoring system has been devised which defines impairment for English-speaking drivers between the ages of 17 to 39 and 40 years plus.

Chapter 8: Conclusion

Consideration of the application of quality assurance to forensic toxicology services has revealed some deficiencies. It is understood that forensic toxicology is a sector where the nature of the work, analysing poor quality samples for unknown analytes possibly present at very low concentrations, means that some allowances are necessary with respect to the quality which can be achieved. Consequently, however, it is a sector within which quality should be a primary focus and implemented to the highest degree, particularly as results can have very serious legal implications. There are some areas of forensic toxicology where quality needs to be improved.

Possibly the most difficult area is in keeping pace with the emergence of new substances of abuse; the recent surge in NPS has presented a particular struggle. Each new substance necessitates development and validation of an analytical method. Review of the Quartz PTS demonstrated the difficulty laboratories have with new substances, particularly BZP which belongs to the piperazines class of NPS. Many of these continue to be forensically relevant and it is important that forensic toxicology laboratories validate methods for their detection to remain effective. In the absence of an immunoassay screening method an LC-MS/MS screening method for piperazines has been validated which requires only a small fraction of a routine basic drug extract, preserving the majority for other analyses as necessary. The lack of suitable reference standards is a problem for both the GC-MS and LC-MS/MS methods. Due to the possible presence of undetected matrix effects however, the lack of standards is particularly troublesome for LC-MS/MS and does not allow for quantification by this method. The internal standards were sufficient that the validated GC-MS method can provide quantification for a range of piperazine compounds.

Different guides from various official bodies cause confusion in knowing whether or not a method has been sufficiently validated. Validation requires laboratories to invest time and money to ensure, and demonstrate, that they have a fit-for-purpose protocol. Having expended this effort, careful application of internal

quality control is essential to confirm that the method has attained sufficient quality and that this is being safeguarded. It is vital that the suitability of IQC procedures is also proven through relevant independent external monitoring. This is particularly relevant to the analysis of NPS for which it can be difficult to obtain certified reference standards (particularly two sources for independent calibrators and controls). Unfortunately these substances are not often included in PTS.

It is difficult for forensic toxicology laboratories to obtain EQA which is entirely suited to their routine workload. EQA should primarily assist laboratories to guard against undetected errors and may additionally provide educational benefits. One scheme may not be enough and a combination may be required to achieve the necessary safeguards and maintain learning. Laboratories need to select enough schemes in which to participate to ensure a representative sample of their analyses is monitored. They must determine what they may want to achieve from EQA (e.g. bias detection for particular substances or experience with emerging substances) and rigorously review scheme protocols to ensure participation will satisfy their aims. EQA must be fit-for-purpose.

This demands laboratories understand PTS scoring systems and how these impact upon their perceived performance. To be fit for the purpose of bias detection a scheme must not use a participant-dependent scoring method, but use appropriate acceptable variation based on expert opinion. In order for a PTS to be effective in its aim of bias detection, laboratories must participate regularly (at least once per year, ideally twice) and pay close attention to their reported proficiency. Where necessary laboratories should recalculate their own proficiency score to ensure testing is to a standard appropriate for the needs of their clients. Apart from laboratory managers, less senior staff should also be coordinating PT, or at least reviewing PT reports, to gain an appreciation of the importance of quality assurance early in their career and benefit from shared experience. This is also helpful in ensuring the quality of results being released from the laboratory; it will often be the technical staff who conduct the analysis and they should be aware of their own performance.

As the selection of PTSs available to forensic toxicology laboratories was found to be very limited, long term reviews of two quite different forensic toxicology

PTSs were conducted and the strengths and weaknesses of both schemes were identified via comparison. Improvements which could be made to both schemes to further benefit their participants were highlighted.

The Quartz Forensic Blood Toxicology Proficiency Testing Scheme was primarily an educational resource. Laboratories that have established their proficiency in their most commonly-used methods can benefit from the varied nature of this scheme to gain experience with analytes which are not encountered often or which are emerging substances of abuse. In order to provide a safeguard against bias this scheme requires more repetitive testing. Without such testing this scheme does not provide participants with an effective means of monitoring routine laboratory performance. The requirement for more focused, frequent, testing of the most relevant forensic analytes was brought to the attention of the coordinators and is now being actioned.

UNODC International Collaborative Exercises is in need of more consistent participation. This is outside of the control of the coordinators. Due to the smaller array of potential analytes and, therefore, more repetitive nature of the testing ICE was more effective as a means of external quality assurance. ICE, however, is not recognised as a PTS for accreditation purposes, whereas Quartz is a UKAS accredited scheme, and also it uses urine rather than whole blood, the preferred sample type in post mortem forensic toxicology.

The reviews demonstrated that accreditation of a PTS is not an endorsement of its quality and that an unaccredited scheme can be more suitable for the intended purpose of a PTS. Laboratories should not base selection of EQA on accreditation status but on whether or not the analytes tested, the frequency of testing and the scoring system applied are appropriate for their needs. A further example of mistaken reliance on accreditation to support fitness-for-purpose was demonstrated by the ISO-accredited field impairment test.

The police FIT has never been fully validated. The first step in validation of this protocol now has been completed. Acceptable performance in the FIT for drivers over and under the age of 40 years has been defined (for drivers competent with the English language). Error counts outside of these acceptable

ranges provide police officers reason to pursue further testing with scientific justification for this which, if necessary, could be presented to the Court.

In order that police officers maintain competence with the FIT SOP a form of proficiency testing could be introduced which tests their understanding of the process and the driver errors which should be recorded. As the offence of driving with levels of prescribed drugs above prescribed limits is soon to be enforced, it is also important for fairness to drivers, that toxicology laboratories that will perform analysis of drivers' blood samples be monitored to ensure acceptable accuracy at the levels of interest. As it is anticipated that oral fluid could eventually replace blood specimens from drivers, it is promising that oral fluid PTS are being developed.

Appendix 1: Quartz False Positives in Identification Samples

Round	False Positive	False Negative
30	Clomipramine	Morphine
32	Salicylate + Clozapine	
	MDMA	Benzoylecgonine + cocaine
	Lignocane	
	Lignocane	Benzoylecgonine
35	Citalopram	
	Citalopram	
	Citalopram	
	Citalopram	
	Citalopram	
	Citalopram	
	Citalopram	
	Citalopram	Morphine
37	Amphetamine + Methamphetamine	BZP + MDA + MDMA
40	Phenobarbital	Secobarbital
	Butobarbital	
41	Promethazine	Demoxepam
	Methamphetamine	Desmethyldiazepam + Demoxepam
	Temazepam	Demoxepam
	Lignocane	Desmethyldiazepam + Demoxepam
42	MDMA	
44	Mephedrone	Buprenorphine
	Morphine	Buprenorphine
45	Venlafaxine	Trazodone
	BZP	
46	Chlodiazepoxide	Amlodipine
47	Paracetamol	Morphine
	Morphine	Ketamine + temazepam

Table Appendix I, 1: False positives observed with Quartz identification samples

Appendix 2: Quartz Participant Identification Methods

Drug Type	Opiates Group												Methadone
Round	30	30	33	35	35	36	36	39	40	42	44	Total	44
Method	Analyses/Not detected												
LC-MS/MS		1				2	2			1	2	8	1
HPLC-MS	2		2	3	3	7 /2	7 /4	2	1	4 /1	10 /4	41 /11	8 /1
HPLC-Diode Array		1	1 /1	2	2	1 /1	1	2	5	1		16 /2	
HPLC-UV Vis										1		1	
MS-MS	3 /2	2	3 /2			1 /1	1 /1	1	1	2		14 /6	
GC-MS	12 /3	13 /1	9 /3	9 /1	9	6 /3	6 /2	7	5 /2	7 /1	4 /1	87 /17	7
GC-NPD										1 /1	1 /1	2 /2	1 /1
GC-FID									1 /1			1 /1	
EIA									1 /1			1 /1	
Unknow n/Other								1	1	2	1	5	1 /1

Table Appendix 2, 1: Method used for identification of opiates

Drug Type	Basic Drugs													
Round	33	34	34	36	36	38	42	45	45	46	47	48	48	Total
Method	Analyses/Not detected													
LC-MS/MS				2 /2	2 /2			2	2	1	1	3 /2	3	16 /6
HPLC-MS	2 /1	1	2	1	1		4	6 /1	6 /3	5 /1	4	3	5	40 /6
HPLC-Diode Array	1	3	2	2 /1	2 /1		1	3	3	2		1 /1	1	21 /3
HPLC-UV Vis						5 /1	1							6 /1
MS-MS	3			1	1	3 /1	2 /1			2	4 /1	3	3	22 /3
GC-MS	9	9 /5	10	11 /1	11 /1	8 /1	7 /1	6	6 /4	4 /2	6	6 /1	5	98 /16
GC-NPD							1	1	1		1	2	1	7
Unknow n/Other						1 /1	2			2 /2	1 /1			6 /4

Table Appendix 2, 2: Method used for identification of basic drugs

Drug Type	Benzodiazepines						
Round	31	31	41	41	41	47	Total
Method	Analyses/Not detected						
LC-MS/MS						1	1
HPLC-MS	3	3	7	7 /2	7 /4	6	33 /6
HPLC-Diode Array	1	1	5	5	5 /2		17 /2
HPLC-UV Vis	1 /1	1					2 /1
MS-MS	1 /1	1	2 /1	2	2 /2	4 /1	12 /5
GC-MS	9 /1	9	4	4	4 /4	4 /3	34 /8
GC-NPD						1 /1	1 /1
GC-ECD			1	1	1 /1		3 /1
Unknow n/Other						1 /1	1 /1

Table Appendix 2, 3: Method used for identification of benzodiazepines

Drug Type	ATS					BZP	Mephedrone	GHB
Round	37	37	43	46	Total	37	43	43
Method	Analyses/Not detected					Analyses/Not detected		
LC-MS/MS	1	1	1	1	4	1	1 /1	1
HPLC-MS	2	2	4 /1	4 /2	12 /3	2 /1	4	4 /2
HPLC-Diode Array	1 /1	1	1	2	5 /1	1 /1	1	1 /1
MS-MS	3	3		2 /1	8 /1	3 /2		
GC-MS	9 /4	9 /1	7 /1	4 /3	29 /6/9	9 /5	7 /2	7 /3
GC-NPD			1		1		1 /1	1
GC-ECD				1 /1	1 /1			
Unknow n/Other			1 /1	2 /2	3 /3		1 /1	1 /1

Table Appendix 2, 4: Method used for identification of ATS, BZP, mephedrone and GHB

Drug Type	Cocaine Group				
Round	32	32	39	39	Total
Method	Analyses/Not detected				
HPLC-MS	3	1	2	2	8
HPLC-Diode Array	2 /1	2 /1	2	2	8 /2
MS-MS	1	2	1	1	5
GC-MS	11 /4	12 /1	7 /3	7	37 /8
Unknow n/Other			1	1	2

Table Appendix 2, 5: Method used for identification of cocaine

Drug Type	Barbiturates			
Round	40	40	41	Total
Method	Analyses/Not detected			
HPLC-MS	1	1		2 /6
HPLC-Diode Array	2 /1	2 /1		4 /2
MS-MS	1 /1	1 /1		2 /5
GC-MS	8 /2	8 /2		16 /8
GC-FID	1 /1	1 /1		2 /1
Unknow n/Other	2 /1	2 /1		4 /1

Table Appendix 2, 6: Method used for identification of Barbiturates

Appendix 3: Quartz Participant Quantification Methods

Drug Type	Basic Drugs																	
Round	31	32	32	36	36	35	35	38	40	41	43	44	46	46	46	48	48	Total
Method	Analyses/Unsatisfactory																	
LC-MS/MS				1	1	1	1				1	1				1	1	8
HPLC-MS	4	5	6 /1	4 /1	7 /1	4	6 /2	3	9 /1	5	5 /1	4 /2	2	4	4 /1	4 /1	5	81 /11
HPLC-Diode Array	5 /1	4	4 /1	2	3	3	2 /1		3	2	1	3	3	4 /1	4 /1	2 /1	2	47 /6
HPLC-UV Vis	2	1	1		1	1				2						1 /1	1	10 /1
MS-MS		3 /1	1 /1						1	2	2		2	4	3		2	20 /2
GC-MS	3 /2	4 /1		6	1	4 /2	3 /1	3		6		6 /1	3	4 /1	2 /1	6	4 /1	55 /10
GC-NPD										1		2	1	2	1	3	2	12
Unknown/Other				2 /1				2			1	2	1 /1					8 /2

Table Appendix 3, 1: Method used for quantification of basic drugs

Drug Type	Cocaine		
Round	38	38	Total
Method	Analyses/Unsatisfactory		
HPLC-MS	5	5	10
HPLC-Diode Array	1	1	2
HPLC-UV Vis		1 /1	1 /1
MS-MS	1	1	2
GC-MS	6 /1	7	13 /1
Unknown/Other	2	2	4

Table Appendix 3, 2: Method used for quantification of cocaine

Drug Type	Opiates					
Round	30	37	41	42	42	Total
Method	Analyses/Unsatisfactory					
HPLC-MS	2	3	5	3	5 /1	18 /1
HPLC-Diode Array	3 /1	2	1	3		9 /1
MS-MS	1	1	2	4 /1	1	9 /1
GC-MS	11 /1	10	5	5 /2	5	36 /3
GC-NPD			1	1	2	4
Other			1	2	3	6

Table Appendix 3, 3: Method used for quantification of opiates

Drug Type	Methadone		
Round	37	38	Total
Method	Analyses/Unsatisfactory		
HPLC-MS	3	7 /1	10 /1
HPLC-Diode Array	2 /2	3	5 /2
MS-MS	1		1
GC-MS	11 /1	6 /1	17 /2

Table Appendix 3, 4: Method used for quantification of methadone

Drug Type	Bezodiazepines		
Round	30	34	Total
Method	Analyses/Unsatisfactory		
HPLC-MS	3 /1	4	7 /1
HPLC-Diode Array	7	6 /1	13 /1
HPLC-UV Vis	1 /1	2	3 /1
MS-MS		2	2
GC-MS	4 /2	2	6 /2
GC-NPD	1		1
Other		1	1

Table Appendix 3, 5: Method used for quantification of benzodiazepines

Drug Type	Amphetamine Type Stimulants				Mephedrone	BZ
Round	33	33	34	Tot	45	45
Method	Analyses/Unsatisfactory					
LC-MS/MS			1	1	1	
HPLC-MS	1	2	3	6		2
HPLC-Diode Array	1 /1	2 /1	1	4 /2	2 /1	1
MS-MS	1	1		2	4	4 /1
GC-MS	11	11 /1	10 /1	43 /2	5 /1	5 /2
GC-NPD						1
Unknow n/Other	1	1	1 /1	3 /1	1	2

Table Appendix 3, 6: Method used for quantification of ATS, mephedrone and BZP

Appendix 4: GC-MS Fragmentation of HFBA Derivatised Piperazines

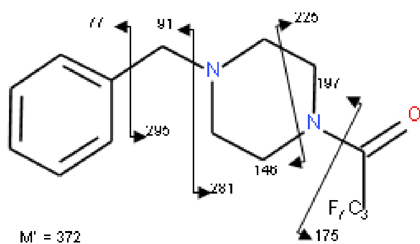


Figure Appendix 4, 1: GC-MS fragmentation of BZP-D8

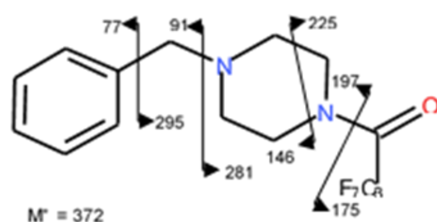


Figure Appendix 4, 2: GC-MS fragmentation of BZP

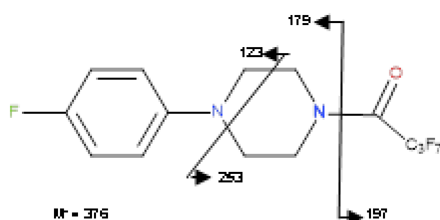


Figure Appendix 4, 3: GC-MS fragmentation of p-FPP

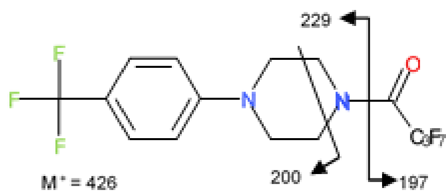


Figure Appendix 4, 4: GC-MS fragmentation of TFMPP

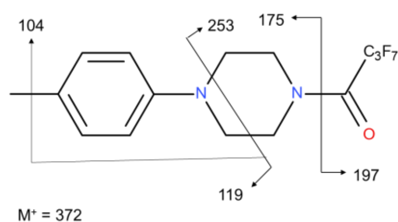


Figure Appendix 4, 5: GC-MS fragmentation of MPP

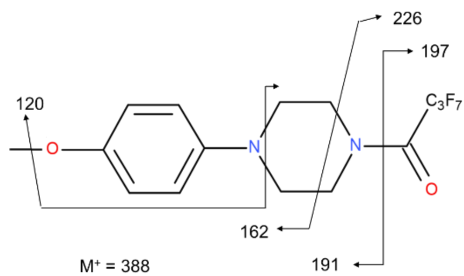


Figure Appendix 4, 6: GC-MS fragmentation of MeOPP

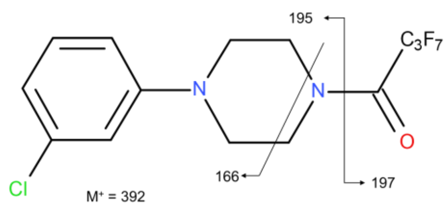


Figure Appendix 4, 7: GC-MS fragmentation of CPP

Appendix 5: LC-QqQ Fragmentation of Piperazines

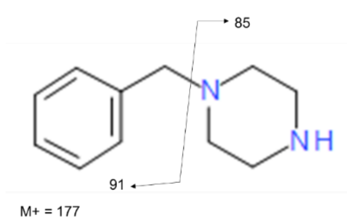


Figure 41: LC-MS/MS fragmentation of BZP

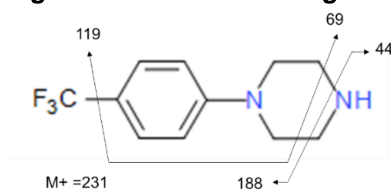


Figure 42: LC-MS/MS fragmentation of TFMPP

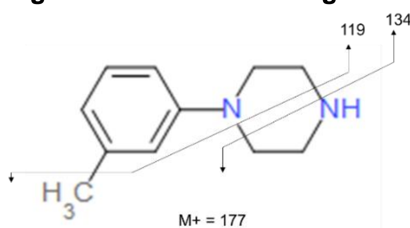


Figure 43: LC-MS/MS fragmentation of MPP

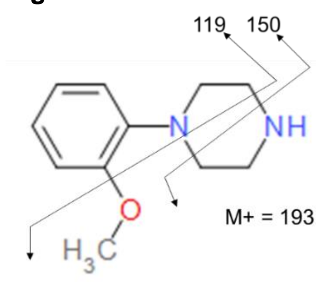


Figure 44: LC-MS/MS fragmentation of MeOPP

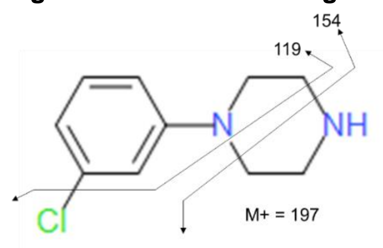


Figure 45: LC-MS/MC fragmentation of CPP

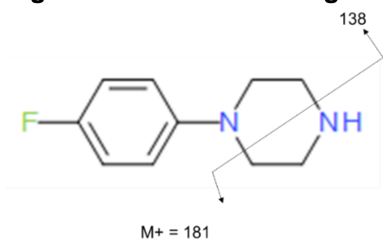


Figure 46: LC-MS/MS fragmentation of *p*-FPP

Appendix 6: Participant Information

PARTICIPANT INFORMATION SHEET

Proficiency in the 'Field Impairment Test' (FIT) amongst non drug users

You are being invited to volunteer as a participant in my research project. I am a part-time postgraduate research student in the Department of Forensic Medicine and Science at the University of Glasgow. In addition to contributing towards my PhD qualification, the research I am undertaking is intended to help me, in my role as a consultant scientist within The Forensic Institute, and others working within the Criminal Justice System. My research is funded by The Forensic Institute and the University of Glasgow.

Before you decide to contribute it is important for you to understand why this research is being undertaken and what it will involve. Please take time to read the following information carefully. If you are unclear of anything or if you would like more information then please ask.

What is the purpose of the Study?

When a police officer observes a driver to be behaving in an uncontrolled manner the driver can be stopped to allow an investigation as to the cause of the erratic driving. A well-known, commonly encountered, reason is the consumption of alcohol. This can be easily identified with a roadside breath test. It is known, however, that many drivers may be impaired due to the consumption of drugs (licit or illicit). Currently there is no roadside drug detection analysis approved for use at the roadside. Rather the police officer may ask the driver to complete a physical test known as a 'Preliminary Impairment Test' (PIT). This can take the form of the Field Impairment Test (FIT) which comprises the five specific tasks, designed to simultaneously assess coordination, balance and understanding. Instructions are read to the driver and the tasks are also demonstrated. The five tasks are:

- *Pupillary examination*
  Pupils examined for dilation/constriction
- *Rhomberg Balance*
  Feet together, eyes closed and head tilted back, estimate when 30 seconds has elapsed
- *Walk and turn*
  Take nine steps heel-to-toe in a straight line, turn, take nine steps back
- *One-leg balance*
  Raise one leg whilst keeping supporting leg straight with toes pointed forward. Hold this position until told to stop. Repeated with opposite leg.
- *Finger to nose*
  Tilt the head back and touch the nose with the index finger of each hand in turn (3 times)

The driver's response to this test is believed to indicate if they are under the influence of a drug. The result of test is admissible in a Court of Law, however, no detailed studies have taken place to establish the variation of response that may be expected from a range of individuals who are known not to have taken drugs. The proportion of non-impaired persons who could not pass this test is unknown. For this reason I want to observe non-drug impaired individuals perform these tasks to establish the general level of acceptable proficiency. This will take into account participant age. The study will be performed over a period of around six months.

What will happen if I take part?

It is up to you to decide whether or not to take part. If you do decide to take part you will be given this information sheet to keep and be asked to sign a consent form. You will also be

required to complete a questionnaire which requests details of your age, level of fitness, routine medication, drug/alcohol consumption in the past 48 hours, any physical or mental disabilities and any other factors which you feel might influence your performance in the test. **If you decide to take part you are still free to withdraw at any time and without giving a reason.**
A decision by any university student not to participate will not affect their grades in any way.

To participate in the study, you are asked to consent freely to a preliminary impairment test as described above (this may be assessed by a member of the University or a specially trained police officer). Your performance will be video recorded. The video recording will be used to gain an accurate assessment only. It will not be viewed by anyone other than the researchers and will be deleted upon completion of the study.

If there are any components of the test that you do not feel able to perform, you will not be required to do so. It is anticipated that the test will not take more than 30 minutes to complete (including time for discussion and paperwork). The tests will take place indoors, not by the roadside. Prior to performing the FIT you will be required to blow into a 'breathalyser' to test for alcohol consumption. You will also be required to provide a sample of saliva. This will be used to confirm that participants were not unwittingly under the influence of any drugs which may have affected their performance in the test. The saliva sample will be collected by placing a purpose-made cotton swab into the mouth, this is completely pain free. Saliva samples will be consumed by the drugs analysis, they will not be used for any other testing (for example they will not be tested for DNA).

At this stage your contribution will be complete and you will not be required to provide any further assistance.

Will taking part remain confidential?
All information which is collected about you during the study will be kept strictly confidential. Information collected from participants will not include their name or any other information which may allow identification. Each result and saliva sample will be stored under a case number (this will not be your matriculation number). You will be given a note of this number as it will be used to identify and destroy your saliva sample and results form should you wish to withdraw from the study at a later date.

What will happen to the results of the Study?
The results of the tests will be anonymised and securely stored within the Department of Forensic Medicine and Science for a period not longer than five years. The results of this study may be published in scientific journals but these would not contain any information by which a participant could be identified. You will not receive any results from these tests.

Who has reviewed the Study?
This project has been reviewed by the College of Medical, Veterinary and Life Sciences Ethics Committee.
If you have a concern regarding any aspect of how this research has been conducted please contact Prof Billy Martin the chair of the MVLS Ethics Committee.

Contact for Further Information
For any further information concerning this project please contact Carrie Mullen at the address given above.

Thank you for participating in the project, your cooperation and help are very much appreciate

Appendix 7: Participant Questionnaire

Participant Number:

Carrie Mullen
Forensic Medicine Department
University of Glasgow
GLASGOW
G12 8QQ
carriem@theforensicinstitute.com

Participant Questionnaire

Proficiency in the 'Field Impairment Test' (FIT) amongst non-drug users.

The following questions are intended to identify any factors which may influence your performance. All information requested is relevant but if you do not wish to complete any of the questions you do not have to participate, you may withdraw from the study. This information will be stored securely within the Forensic Medicine Department and will not be made available to any other persons or for any other research studies. This data will not be stored for longer than five years. You may request to have this form returned at any time by quoting your participant number.

Please provide truthful answers throughout, only study coordinators will have access to the answers you provide and you cannot be identified. If you are unsure of any questions please ask.

Time of Test: ______________________

Age: ______________________

Occupation: ______________________

Gender: ______________________

Height: ______________________

Weight: ______________________

1. **Have you ever completed a field impairment test before?** **YES / NO**

   How many times?______________________________

2. **How would you describe your current general health?**

Healthy 1 2 3 4 5 Poor

3. **Are you aware of any weaknesses, illnesses or injuries which might influence your ability to perform the FIT?** (e.g. headache, cold, flu, tiredness, declared disabilities, ear infections, head injuries, dyslexia or dyspraxia)

____

____

____

____

4. **How would you rate your general level of fitness?**

Fit 1 2 3 4 5 Unfit

5. **Have you taken any prescribed or over the counter medicines in the last 48 hours?** **YES /NO**

If yes, please provide details_______________________________

____

____

6. **Do you smoke?** **YES / NO**

If yes,

On average how many?_______________________________

When was your last cigarette?_______________________________

7. **Have you consumed alcohol in the previous 48 hours?** **YES / NO**

____

____

8. **How would you describe the quality of your last sleep?**__________

____

How long have you been awake?______________________________

Are you used to long periods of wakefulness? (e.g. shiftwork)

____

9. **How would you rate your current state of wakefulness?**

   Energetic 1 2 3 4 5 Exhausted

10. **Have you eaten recently?** (What did you have and how long ago?)

    ____

    ____

    ____

11. **Do you hold, or have you held, a valid driver's license?**

    ____

12. **Would you be willing to drive a motor vehicle at the present time?**

    YES / NO

Appendix 8: Scottish Field Impairment Test Form

PRELIMINARY IMPAIRMENT TEST
Section 3A/4 (and) 6B RTA 1988

INTRODUCTION AND GENERAL GUIDANCE

This form is for use by authorised police officers during the application of a Preliminary Impairment Test on a subject who has been required to co-operate. Where a test is abandoned the reasons should be recorded. If the questions are read from a card, the wording must be identical to those used in this form and the card must remain available for production at court. A record of any medical condition or disability claimed at any time during the tests, and a record or any response or gesture made to any question or at any other time, must be recorded. Any ancillary pupil gauge used must be retained for production at court if required.

RELEVANT DETAILS

BRIEF CIRCUMSTANCES OF INCIDENT/DEMEANOUR/BEHAVIOUR OF SUSPECT

NAME		**DOB:**	**MALE/FEMALE**
ADDRESS			
DAY/DATE		**TIME STARTED**	
LOCATION TEST			
WEATHER	FINE: **YES/NO**	RAIN: **YES/NO**	SNOW: **YES/NO** WIND: **YES/NO**
LIGHTING CONDITION	DAYLIGHT: **YES/NO**	TWILIGHT: **YES/NO**	DARKNESS: **YES/NO**
STREET LIGHTING	LIT **YES/NO**	UNLIT: **YES/NO**	NIL: **YES/NO**
ROAD SURFACE	WET: **YES/NO**	DRY: **YES/NO**	
FOOTWEAR	SUITABLE: **YES/NO**	UNSUITABLE: **YES/NO**	REMOVED: **YES/NO**

If unsuitable state reason:

Additional Comments:

OFFICER DEFINED LAMINATED CARD USED: **YES/NO**

STATEMENT TO SUSPECT

It is suspected that you may have committed an offence contrary to section 3A or 4 of the Road Traffic Act 1988 as amended, and I require you to co-operate in a Preliminary Impairment Test. I must warn you that failure to co-operate without reasonable excuse is an offence and may render you liable to prosecution.

Do you agree to co-operate in a preliminary test, and if not, what is your reason for refusing **YES/NO**

If no state reason:

1

PUPILLARY EXAMINATION			
I am going to examine the size of your pupils, comparing them to this gauge, which I will hold up to the side of your face. All I require you to do is look straight ahead and keep your eyes open wide.			
Do you understand	**YES/NO**		
Are you wearing contact lenses:	**YES/NO**		
Pupil size left	**CONSTRICTED/NORMAL/DILATED**	**Approx Size:**	
Pupil size right	**CONSTRICTED/NORMAL/DILATED**	**Approx Size:**	
A pupil size: 1.0 – 2.5 (inclusive) normally indicates constriction. 7.0 – 9.0 (inclusive) normally indicates dilation			

NOTE CONDITION OF EYES:	**WATERY**	**YES/NO**	**REDDENING**	**YES/NO**
Additional Comments:				

MODIFIED ROMBERG BALANCE TEST	
Stand up straight with your heels and toes together and your arms down by your sides. (**Demonstrate**) Maintain that position while I give you the remaining instructions. Do not begin until I tell you. When I tell you, tilt your head back slightly, close your eyes. (**Demonstrate – but do not close your eyes**) When you think 30 seconds has passed, bring your head forward, open your eyes and say, "**STOP**".	
Do you understand	**YES/NO**
Do you have any disability or medical condition that prevents you from participating in this test:	**YES/NO**
Tilt your head back, close your eyes – Begin	
ABLE TO BALANCE DURING INSTRUCTIONS:	**YES/NO**

IF 'NO'	STEPS:	SWAYS:	RAISE ARMS:

COMPLIES WITH INSTRUCTIONS:	**YES/NO**				
IF 'NO'	EYES OPEN	HEAD RAISED	STEPS	SWAYS	RAISED ARMS
TIME (Seconds)					

Estimates of 30 secs @: **How Long Was That REPLY:**

Additional Comments:

WALK AND TURN TEST	
(Identify a real or imaginary line. Do not use the kerb or anywhere the subject may fall)	
Place your left foot on the line. Place your right foot on the line in front of your left touching heel to toe (**Demonstrate**). Put your arms down at your sides and keep them there throughout the test. Maintain that position while I give you the remaining instructions. Do not begin until I tell you.	
Do you understand	**YES/NO**
When I tell you, you must take nine heel to toe steps along the line. On each step the heel of the foot must be placed against the toe of the other foot. (**Demonstrate**) When the ninth step has been taken, you must leave the front foot on the line and turn around using a series of small steps with the other foot. After turning you must take another nine heel to toe steps along the line. During the test you must watch your feet at all times and count each step out loud. Once you start walking do not stop until you have completed the test. (**Demonstrate complete test**)	
Do you understand	**YES/NO**
Do you have any disability or medical condition that prevents you from participating in this test:	**YES/NO**
ABLE TO BALANCE DURING INSTRUCTIONS:	**YES/NO**

IF 'NO'	STEPS:	SWAYS:	RAISED ARMS:	STARTS TO SOON:

COMPLIES WITH INSTRUCTIONS:	YES/NO	IF 'NO':

Any deviation from the instructions should be indicated below and on the diagram above.

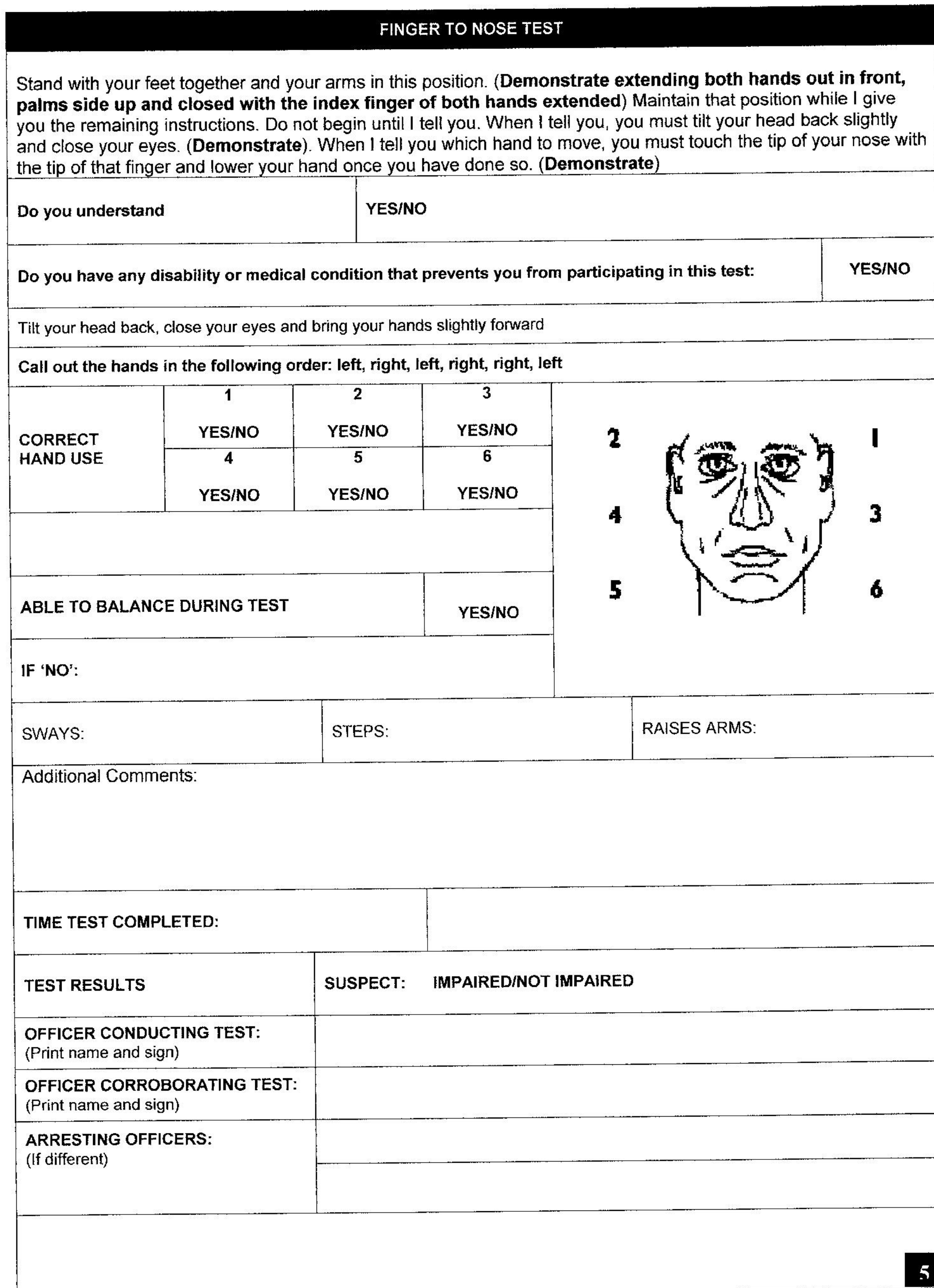
STOPS WALKING	MISS HEEL/TOE	RAISES ARMS	STEPS OFF LINE

CORRECT TURN	YES/NO
If No STATE REASON:	
COUNTS OUT LOAD	YES/NO
CORRECT STEP COUNT	YES/NO

FROM TURN:	TO TURN:
Additional Comments:	

3

ONE LEG STAND TEST				
Stand with your feet together and your arms down by your sides. (**Demonstrate**) Maintain that position while I give you the remaining instructions. Do not begin until I tell you.				
Do you understand	**YES/NO**			
When I tell you to you must raise your right foot 6 to 8 inches (or 15 to 20cms) off the ground, keeping your leg straight and your toes pointing forward with your foot parallel to the ground. (**Demonstrate**) You must keep your arms down by your sides, and keep looking at your raised foot while counting out loud in the following manner; one thousand and one, one thousand and two and so on until I tell you stop.				
Do you understand	**YES/NO**			
Do you have any disability or medical condition that prevents you from participating in this test:	**YES/NO**			
(Time 30 Seconds)				
NOTE: REPEAT FOR OTHER FOOT				
ABLE TO BALANCE DURING INSTRUCTIONS:	**YES/NO**			
IF 'NO':	STEPS:	SWAYS:	RAISES ARMS:	
COMPLIES WITH INSTRUCTIONS:	**YES/NO**			
IF 'NO':				
LEFT LEG	SWAYS	HOPS	PUTS FOOT DOWN	RAISES ARMS
TIME (seconds)				
RIGHT LEG	SWAYS	HOPS	PUTS FOOT DOWN	RAISES ARMS
TIME (Seconds)				
COUNTED CORRECTLY	**YES/NO**			
Additional Comments:				



FINGER TO NOSE TEST

Stand with your feet together and your arms in this position. (**Demonstrate extending both hands out in front, palms side up and closed with the index finger of both hands extended**) Maintain that position while I give you the remaining instructions. Do not begin until I tell you. When I tell you, you must tilt your head back slightly and close your eyes. (**Demonstrate**). When I tell you which hand to move, you must touch the tip of your nose with the tip of that finger and lower your hand once you have done so. (**Demonstrate**)

Do you understand	**YES/NO**

Do you have any disability or medical condition that prevents you from participating in this test:	**YES/NO**

Tilt your head back, close your eyes and bring your hands slightly forward

Call out the hands in the following order: left, right, left, right, right, left

	1	2	3
CORRECT HAND USE	YES/NO	YES/NO	YES/NO
	4	5	6
	YES/NO	YES/NO	YES/NO

ABLE TO BALANCE DURING TEST	YES/NO
IF 'NO':	

SWAYS:	STEPS:	RAISES ARMS:

Additional Comments:

TIME TEST COMPLETED:	
TEST RESULTS	SUSPECT: IMPAIRED/NOT IMPAIRED
OFFICER CONDUCTING TEST: (Print name and sign)	
OFFICER CORROBORATING TEST: (Print name and sign)	
ARRESTING OFFICERS: (If different)	

5

List of References

1. De Zeeuw RA. Recent developments in analytical toxicology: for better or worse. *Toxicology Letters* 1998;102:103-108.

2. Madej K. Analysis of meconium, nails and tears for determination of medicines and drugs of abuse. *Trends in Analytical Chemistry* 2010;29:246-259.

3. Beck O. Exhaled breath for drugs of abuse testing - Evaluation in criminal justice settings. *Science and Justice* 2014;54:57-60.

4. De Giovanni N, Fucci N. The current status of sweat testing for drugs of abuse: A review. *Current Medical Chemistry* 2013;20:545-561.

5. Klintz P. Proceedings of the Society of Hair Testing (SoHT) annual meeting in Chamonix (21-26 March 2011). *Forensic Science International* 2012;218.

6. Bosker WM, Huestis MA. Oral Fluid Testing for Drugs of Abuse. *Clinical Chemistry* 2009;55:1910-1931.

7. Kadehjian LJ. Specimens for Drugs-of-Abuse Testing In: Wong RC,Tse HY, eds. *Drugs of Abuse; Body Fluid Testing* New Jersey: Humana Press, 2005.

8. Langman LJ. The use of oral fluid for therapeutic drug management: Clinical and forensic toxicology. *Annals of the New York Academies of Sciences* 2007;1098:145-166.

9. Drummer O. Review: Pharmacokinetics of illicit drugs in oral fluid. *Forensic Science International* 2005;150:133-142.

10. Aps JKM, Martens LC. Review: The physiology of saliva and transfer of drugs into saliva. *Forensic Science International* 2005;150:119-131.

11. Kuijten C (Ed)Driving under the Influence of Drugs, Alcohol and Medicines, DRUID, Evaluation of oral fluid screening devices by TISPOL to harmonise European police Requirements (ESTHER) 2009.

12. United Nations International Drug Control Programme. *Recommended Methods for the Detection and Assay of Heroin, Cannabinoids, Cocaine, Amphetamine, Methamphetamine and Ring-Substituted Amphetamine Derivatives in Biological Specimens*. New York, US: United Nations, 1995.

13. Franke JP, de Zeeuw RA. Solid-phase extraction procedures in systematic toxicological analysis. *Journal of Chromatography B* 1998;713:51-59.

14. Rivier L. Criteria for the identification of compounds by liquid chromatography-mass spectrometry and liquid chromatography -multiple mass spectrometry in forensic toxicology and doping analysis. *Analytica Chimica Acta* 2003;492:69-82.

15. Maurer HH. Multi-analyte procedures for screening for and quantification of drugs in blood, plasma, or serum by liquid chromatography-single stage or tandem mass spectrometry (LC-MS or LC-MS/MS) relevant to clinical and forensic toxicology. *Clinical Biochemistry* 2005;38:310-318.

16. Maurer H. Hyphenated mass spectrometric techniques - indispensable tools in clinical and forensic toxicology and doping control. *Journal of Mass Spectrometry* 2006;41:1399-1413.

17. Bogusz MJ. Quality assurance aspects of identification with chromatographic - mass spectrometric methods. *Problems in Forensic Sciences* 2009;LXXVII:7-28.

18. Wilson JF. External quality assessment schemes for toxicology. *Forensic Science International* 2002;128:98-103.

19. Flannagan RJ, Taylor A, Watson I, et al. *Fundamentals of Analytical Toxicology*. Sussex, England: John Wiley & Sons, 2007.

20. Jackson ARW, Jackson JM. *Forensic Science*. Essex, England: Pearson Education, 2008.

21. Segura J, Ventura R, Jurado C. Derivatization procedures for gas chromatographic-mass spectrometric determination of xenobiotics in biological samples, with special attention to drugs of abuse and doping agents. *Journal of Chromatography B* 1998;713:61-90.

22. McNair HM, Miller JM. *Basic Gas Chromatography*. New York: Wiley and Sons, 1998.

23. McMaster MC. *GC/MS A Practical User's Guide*. <st1:state style="font-family: Arial, sans-serif; font-size: 12pt;" w:st="on">New York: Wiley and Sons, 2008.

24. Willoughby R, Sheehan E, Mitrovich S. *A Global View of LC/MS; How to solve your most challenging analytical problems*. 2 ed. Pittsburgh: Global View Publishing, 2002.

25. Thieme D, Sachs H. Improved screening capabilities in forensic toxicology by application of liquid chromatography-tandem mass spectrometry. *Analytical Chimica Acta* 2003;492:171-186.

26. Scott RPW. *Chrom-Ed Book Series; Liquid Chromatography*: Library4Science, 2003.

27. Pihlainen K, Sippola E, Kostiainen R. Rapid identification and quantification of compounds with forensic interest using fast liquid chromatography-ion trap mass spectrometry and library searching. *Journal of Chromatography A* 2003;994:93-102.

28. McMaster MC. *LC/MS A Practical User's Guide*. New Jersey: John Wiley & Sons, 2005.

29. Verweij AMA, Lipman PJL. Comparison of mass spectrometric ionisation techniques for the analysis of phenethylamines. *Journal of Chromatographic Science* 1996;34:379-382.

30. Katagi M, Tatsuno M, Miki A, et al. Discrimination of dimethylamphetamine and methamphetamine use: Simultaneous determination of dimethylamphetamine-N-oxide and other metabolites in urine by high-performance liquid chromatography electrospray ionisation mass spectrometry. *Journal of Analytical Toxicology* 2000;24:354-358.

31. LCQ Deca XP Plus Hardware Manual. California: ThermoFinnigan, 2002.

32. Chambers E, Wagrowski-Diehl D, Lu Z, et al. Systematic and comprehensive strategy for reducing matrix effects in LC/MS/MS analyses. *Journal of Chromatography B* 2007;852:22-34.

33. Maurer HH. Current role of liquid chromatography-mass spectrometry in clinical and forensic toxicology. *Analytical and Bioanalytical Chemistry* 2007;388:1315-1325.

34. De Zeeuw RA. Substance identification: the weak link in analytical toxicology. *Journal of Chromatography B* 2004:3-12.

35. Hartstra J, Franke JP, de Zeeuw RA. How to approach substance identification in qualitative bioanalysis. *Journal of Chromatography B* 2000;739:125-137.

36. Kidwell DA, Riggs LA. Comparing two analytical methods: minimal standards in forensic toxicology derived from information theory. *Forensic Science International* 2004;145:85-96.

37. Vorce SP, Holler JM, Levine B, et al. Detection of 1-Benzylpiperazine and 1-(3-Trifluoromethylphenyl)-piperazine in urine analysis specimens using GC-MS and LC-ESI-MS. *Journal of Analytical Toxicology* 2008;32:444-450.

38. SOFT/AAFS *Forensic Toxicology Laboratory Guidelines; Available at:* www.soft-tox.org/files/Guidelines_2006_Final.pdf. Accessed August 2008, 2006.

39. Zellner BD, Bicchi C, Dugo P, et al. Linear retention indices in gas chromatographic analysis: a review. *Flavour and Fragrance Journal* 2008;23:297-314.

40. De Zeeuw RA. Drug screening in biological fluids; The need for a systematic approach. *Journal of Chromatography B* 1997;689:71-79.

41. Fernandez MMR, Wille SMR, Samyn N, et al. High-Throughput analysis of amphetamines in blood and urine with on-line solid-phase extraction-liquid chromatography-Tandem mass spectrometry. *Journal of Analytical Toxicology* 2009;33:578-587.

42. *Principles of Forensic Toxicology*. 2 ed. Washington, US.: American Association for Clinical Chemistry Press, 2003.

43. Barroso M, Costa S, Dias M, et al. Analysis of phenylpiperazine-like stimulants in human hair as trimethylsilyl derivatives by gas chromatography-mass spectrometry. *Journal of Chromatography A* 2010;1217:6274-6280.

44. Anderson M, Gustavsson E, Stephanson N, et al. Direct injection LC-MS/MS method for identification and quantification of amphetamine, methamphetamine, 3,4-methylenedioxyamphetamine and 3,4-methylenedioxymethamphetamine in urine drug testing. *Journal of Chromatography B* 2008;861:22-28.

45. *Recommended Guidelines for Quality Assurance and Good Laboratory Practices Manual for use by National Laboratories*: UNODC, 1995.

46. European Laboratory Guidelines for Legally Defensible Workplace Drugs Testing (2002) Available at: http://www.ewdts.org/data/uploads/documents/ewdtsguidelines.pdf. Accessed April 2014.

47. Cooper G, Paterson S, Oselton MD. The United Kingdom and Ireland Association of Forensic Toxicologists Forensic toxicology laboratory guidelines (2010). *Science and Justice* 2010;50:166-176.

48. Lawn RE, Thompson M, Walker RF. *Proficiency Testing In Analytical Chemistry*. Cambridge: The Royal Society of Chemistry, 1997.

49. CITAC/EURACHEM *Guide to Quality in Analytical Chemistry - An Aid to Accreditation (2002) Available at: http://www.european-accreditation.org/publication/citac-eurachem-ta. Accessed: Sep 2010*.

50. Taverniers I, De Loose M, Van Bockstaele E. Trends in quality in the analytical laboratory: ii. Analytical method validation and quality assurance. *Trends in Analytical Chemistry* 2004;23:535-552.

51. Bailey C, Barwick V. *Laboratory Skills Training Handbook*: LGC VAM, 2007.

52. Malkoc E, Neuteboom W. The current status of forensic science laboratory accreditation in Europe. *Forensic Science International* 2007;167:121-126.

53. The Forensic Science Regulator, Government Digital Service, https://www.gov.uk/government/organisations/forensic-science-regulator/about, Accessed April 2014.

54. EURACHEM Guide - *The Fitness for Purpose of Analytical Methods; A Laboratory Guide to Method Validation and Related Topics*. Teddington, UK: LGC, 1998.

55. The International Laboratory Accreditation Cooperation; www.ilac.org; Accessed October 2009.

56. British Standard BS EN ISO/IEC 17025:2005 General Requirements for the competence of testing and calibration laboratories. Available at http://www.iso.org/iso/catalogue_detail.htm?csnumber=39883. Accessed April 2014.

57. Peters FT, Drummer OH, Musshoff F. Validation of new methods. *Forensic Science International* 2007;165:216-224.

58. Peters FT, Maurer HH. Bioanalytical method validation and its implications for forensic and clinical toxicology - A review. *Accreditation and Quality Assurance* 2002;7:441-449.

59. European network of Forensic Science Institutes; http://www.enfsi.eu/uploads/files/PT%20Providers%20outside%20ENFSI.pdf, Accessed April 2010.

60. Thompson M, Ellison SLR, Wood R. Harmonised Guidelines for Single-Laboratory Validation of Methods of Analysis (IUPAC Technical Report). *Pure and Applied Chemistry* 2002;74:835-855.

61. Ferrara SD. Importance of quality assurance in testing drugs of abuse. *Forensic Science International* 1993;63:305-309.

62. Ferrara DS, Tedeschi L, Frison G, et al. Quality control in toxicological analysis. *Journal of Chromatography B* 1998;713:227-243.

63. Report on the Consultative Meeting on Quality Assurance and Good Laboratory Practices: United Nations International Drug Control Programme, Technical Services Division, 1993.

64. EURACHEM/CITAC. *Quality Assurance for Research and Development and Non-routine Analysis*, . Teddington UK: LGC, 1998.

65. ENFSI Standing Committee for Quality and Competence; Validation and Implementation of (new) Methods, Guidance Document, 2006.

66. Linsinger TPJ. Use of recovery and bias information in analytical chemistry and estimation of its uncertainty contribution. *Trends in Analytical Chemistry* 2008;27:916-923.

67. Bressolle F, Bromet-Petit M, Audran M. Validation of liquid chromatographic and gas chromatographic methods. Applications to pharmacokinetics. *Journal of Chromatography B* 1996;686:3-10.

68. Hartmann C, Smyers-Verbeke J, Madssart DL, et al. Validation of bioanalytical chromatographic methods. *Journal of Pharmaceutical and Biomedical Analysis* 1998;17:193-218.

69. Spiking in Analytical Method Development and Validation, A guide for authors. *Journal of Pharmaceutical and Biomedical Analysis* 2005;38:v.

70. Stockl D, D'Hondt H, Thienpont LM. Method validation across the disciplines - Critical investigation of major validation criteria and associated experimental protocols. *Journal of Chromatography B* 2008;877:2180-2190.

71. Boulanger B, Chiap P, Dewe W, et al. An analysis of the SFSTP guide on validation of chromatographic bioanalytical methods: progress and limitations, 32:753-765. *Journal of Pharmaceutical and Biomedical Analysis* 2003;32:753-765.

72. *EURACHEM/CITAC Guide CG4 Quantifying Uncertainty in Analytical Measurement*. Second ed, 2000.

73. US Department of Health and Human Services Food and Drug Administration, Centre for Drug Evaluation and Research, Centre For Vetinary Medicine; Guidance for Industry, Bioanalytical Method Validation, 2001.

74. Dadgar D, Burnett PE, Choc MG, et al. Application issues in bioanalytical method validation, sample analysis and data reporting. *Journal of Pharmaceutical & Biomedical Analysis* 1995;13:89-97.

75. Shah VP, Midha KK, Dighe S, et al. Analytical Method Calibration: Bioavailability, Bioequivalence and Pharmacokinetic Studies. *Pharmaceutical Research* 1992;9:588-592.

76. Peters FT, Kraemer T, Maurer HH. Drug Testing in Blood: Validated Negative-Ion Chemical Ionization Gas Chromatographic-Mass Spectrometric Assay for Determination of Amphetamine and Methamphetamine Enantiomers and Its Application to Toxicology Cases. *Clinical Chemistry* 2002;48:1472-1485.

77. Hibbert DB. Method validation of modern analytical techniques. *Accreditation and Quality Assurance* 1999;4:352-356.

78. Barwick V. *Preparation of Calibration Curves: A Guide to Best Practice*: LGC ltd, 2003.

79. Shah VP, Midha KK, Findlay JW, et al. Bioanalytical Method Validation - A Revisit with a Decade of Progress. *Pharmaceutical Research* 2000;17:1551-1557.

80. "Text on validation of analytical procedures" (1995) International Conference on Harmonisation. Federal Register 60 .p.11260. Available at http://www.fda.gov/downloads/Drugs/Guidances/ucm073381.pdf. Accessed April 2014.

81. Peters FT, Schaefer S, Staack RF, et al. Screening for and validating quantification of amphetamines and of amphetamine- and piperazine-derived designer drugs in human blood plasma by gas-chromatography/mass spectrometry. *Journal of Mass Spectrometry* 2003;38:659-676.

82. Araujo P. Key aspects of analytical method validation and linearity evaluation. *Journal of Chromatography B* 2009;877:2224-2234.

83. Green JM. A Practical Guide to Analytical Method Validation. *Analytical Chemistry* 1996;68:205A-309A.

84. Heyden YV, Nijhuis A, Smeyers-Verbeke J, et al. Guidance for robustness/rugedness tests in method validation. *Journal of Pharmaceutical and Biomedical Analysis* 2001;24:723-753.

85. Miller WG. The role of proficiency testing in achieving standardisation and harmonisation between laboratories. *Clinical Biochemistry* 2009;42:232-235.

86. Thomson M, Ellison S, Wood R. The International Harmonised Protocol for the proficiency testing of analytical chemistry laboratories (IUPAC Technical Report). *Pure and Applied Chemistry* 2006;78:145-196.

87. Miller JC, Miller JN. *Statistics for Analytical Chemistry*. Third ed. West Sussex: Ellis Horwood Ltd, 1993.

88. EURACHEM/WELAC Accreditation for Chemical Laboratories (1993): United Kingdom Accreditation Service Publication ref: LAB 27, 1993.

89. ISO 17043: 2010 – Conformity Assessment – General Requirements for proficiency testing. Available from http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=29366, Accessed April 2014.

90. ISO Guide 43: 1997 - Proficiency Testing by interlaboratory comparisons: Part one: Development and operation of proficiency testing schemes.

91. ISO Guide 43: 1997 - Proficiency Testing by interlaboratory comparisons: Part two: Selection and use of proficiency testing schemes by laboratory accreditation bodies.

92. Araujo P, Froyland L. Hierarchical classification designs for the estimation of different sources of variability in proficiency testing experiments. *Analytica Chimica Acta* 2006;555:348-353.

93. Nogueira JMF, Cortez L. EPTIS: The new European database of proficiency testing schemes for analytical laboratories. *Trends in Analytical Chemistry* 2001;20:457-461.

94. EURACHEM Guide - Selection, use and interpretation of proficiency testing (PT) Schemes by laboratories, 2000, LGC (as part of VAM), http://www.eurachem.org/index.php/publications/pubarch/pt2000wd, Accessed May 2014.

95. European Network of Forensic Science Institutes Standing Committee for Quality and Competence; *Guidance on the conduct of proficiency tests and collaborative exercises within ENSFI*, 2005.

96. British Standard ISO 13528:2005 Statistical methods for use in proficiency testing by laboratory comparisons. Available from http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=35664. Accessed May 2014.

97. Chan KM, Cheug STC, Wong Y, et al. Proficiency test for food and herbal medicine in the Asia Pacific Region. *Trends in Analytical Chemistry* 2010;29:562-576.

98. Earnshaw A, Smith RA, Owen L. How proficiency testing can improve the quality of analytical data using vitamin analysis as an example. *Food Chemistry* 2009;113:781-783.

99. Boley NP, De Bievre PD, Taylor PDP, et al. Requirements vs acceptability in proficiency testing schemes and other interlaboratory comparisons, 6 244-251. *Accreditation and Quality Assurance* 2001;6:244-251.

100. Royal Society of Chemistry (2002) Understanding and acting on scores obtained in proficiency testing schemes, Available at: http://www.rsc.org/Membership/Networking/InterestGroups/Analytical/AMC/. *Analytical Methods Committee Technical Brief*, 2002.

101. Taylor A, Angerer J, Arnauld J, et al. Quality specifications for evaluation and comparison of performance among external assessment schemes in occupational and environmental laboratory medicine. *Accreditation and Quality Assurance* 2006;11:440-445.

102. Kuselman I, Fajgelj A. Key metrological issues in proficiency testing – response to “Metrological compatibility –a key issue in further accreditation” by K. Hedron. *Accreditation and Quality Assurance* 2011;16:99-102.

103. Belli M, Ellison SLR, Fajgelij A, et al. Implementation of proficiency testing schemes for a limited number of participants, 12 391-398. *Accreditation and Quality Assurance* 2007;12:391-398.

104. Yip Y, Tong W. Assessing laboratory performance in intercomparisons for inorganic analysis. *Trends in Analytical Chemistry* 2009;28:1276-1294.

105. Analytical Methods Committee RSoC. Robust Statistics - How not to reject outliers. *Analyst* 1989;114:1693-1696.

106. Vankeerberghen P, Vandenbosch C, Smeyers-Verbeke J, et al. Some robust statistical procedures applied to the analysis of chemical data. *Chemometrics and Intelligent Laboratory Systems* 1991;12:3-13.

107. Analytical Methods Committee RSoC. The role of proficiency testing in method validation. *Accreditation and Quality Assurance* 2010;15:73-79.

108. *Czichos H, Saito T, Smith LR, Smith L. (Eds) Springer handbook of materials measurement methods, 2006, 978(3)20786*

109. Thompson M, Ellison SLR. Fitness for purpose - the integrating theme of the revised Harmonised Protocol for Proficiency Testing in Analytical Chemistry Laboratories *Accreditation and Quality Assurance* 2006;11:373-378.

110. Rosario P, Martinez JL, Silvan JM. Comparison of different statistical methods for evaluation of proficiency test data. *Accreditation and Quality Assurance* 2008;13:493-499.

111. Hedron K. Proficiency testing in the light of a new rationale in metrology. *Accreditation and Quality Assurance* 2008;13:217-222.

112. Eurachem/CITAC Guide (2007) *Use of uncertainty information in compliance assessment*

113. UNODC. *Report of the Standing Panel on the International Collaborative Exercises (ICE) within the International Quality Assurance Programme (IQAP)*. Vienna: United Nations Office on Drugs and Crime, 2000.

114. ASCLAD LAB Accredited Proficiency Testing Schemes, http://www.a2la.org/dirsearchnew/ptproviders.cfm, Accessed March 2010.

115. College of American Pathologists, http://www.cap.org/apps/cap.portal, Accessed March 2010.

116. RTI International Quality Assurance Samples; http://www.rti.org/page.cfm?obj=477DC09E-C80B-41B1-8523300CD2F142DE#certification. Accessed March 2010.

117. Asia Pacific Laboratory Accreditation Cooperation, Proficiency Testing Frequency Benchmarks, http://www.aplac.org/documents/pt/aplac_pt_006_issue_1.pdf. Melbourne, Australia: APLAC Proficiency Testing Committee, 2009.

118. Asia Pacific Laboratory Accreditation Cooperation (APLC), Proficiency Testing Directory (2009) https://www.aplac.org/documents/pt/aplac_pt_003_issue_12.pdf, Accessed May 2014.

119. Austox Urine Proficiency Program Drugs of Abuse, www.austox.com, Accessed May 2012.

120. Proficiency Testing Organisations Schedules, United Kingdom Accreditation Service, http://www.ukas.com/about-accreditation/accredited-bodies/proficiency-testing-organisations.asp, Accessed May 2014.

121. London Toxicology Group, Version 1.0 ‘guidelines’, http://www.ltg.uk.net, Accessed March 2010, March 2001.

122. Ventura M, Ventura R, Pichini S, et al. ORALVEQ: External quality assessment scheme for drugs of abuse in oral fluid results obtained in the first round performed in 2007. *Forensic Science International* 2008;182:35-40.

123. Society of Hair Testing, www.soht.org, Accessed March 2011.

124. Ventura M, Stramesi C, Pichini S, et al. HAIRVEQ 2006: Evolution of laboratories' performance after different educational actions. *Forensic Science International* 2008;176:2-8.

125. Chaturvedi AK. The FAA's Postmortem Forensic Toxicology Self-Evaluation Proficiency Test Programme: The First Seven Years. *Journal of Forensic Sciences* 2000;45:422-428.

126. Protocol to the International Collaborative Exercises within the International Quality Assurance Programme, United Nations Officer on Drugs and Crime. (1999) Available at https://www.unodc.org/unodc/en/scientists/protocol-for-the-international-collaborative-exercises-within-the-international-quality-assurance-programme.html. Accessed March 2009: United Nations Office on Drugs and Crime.

127. International Collaborative Exercises, Drug Analysis: United Nations Office on Drugs and Crime, 2007.

128. United Nations Office on Drugs and Crime, Quality Assurance Support, https://www.unodc.org/unodc/en/scientists/quality-assurance-support.html, Accessed 2013.

129. Tracking designer drugs, legal highs and bath salts, United Nations Office on Drugs and Crime (UNODC), 2012, http://www.unodc.org/unodc/en/frontpage/2012/November/tracking-designer-drugs-legal-highs-and-bath-salts.html. Accessed Nov 2012.

130. General Protocol Proficiency Testing Schemes, Issue 6 (April.2012): LGC Standards, 2012.

131. QUARTZ Forensic Blood Toxicology Proficiency Testing Scheme; Scheme Description, Issue 10 (18.01.2012): LGC Standards, 2012.

132. Kuselman I, Fajgeli A. IUPAC/CITAC Guide: Selection and use of proficiency testing schemes for a limited number of participants – chemical analytical laboratories (IUPAC Technical Report). *Pure and Applied Chemistry* 2010;85:1099-1135.

133. *World Drugs Report 2012*. Vienna, Austria: United Nations Office on Drugs and Crime.

134. Europol 2011 Annual Report on the implementation of Council Decision 2005/287/JHA. *In accordance with Article 10 of Council Decision 2005/387/JHA*

on the information exchange, risk assessment and control of new psychoactive substances: European Monitoring Centre for Drugs and Drug Addiction, 2012.

135. Guant W, Whetton M. Regular participation in proficiency testing provides long term improvements in laboratory performance: an assessment of data over time. *Accreditation and Quality Assurance* 2009;14:449-454.

136. Pil K, Esposito FM, Verstraete A. External quality assessment of multi-analyte chromatographic methods in oral fluid. *Clinica Chemica Acta* 2010;411:1041-1045.

137. United Nations Office on Drugs and Crime, *World Drugs Report 2013* vienna: UNODC, 2013.

138. Clarke J, Wilson JF. Proficiency testing (external quality assessment) of drug detection in oral fluid. *Forensic Science International* 2005;150:161-164.

139. King B, Boley N, Kannan G. The correlation of laboratory performance in proficiency testing with other QA characteristics. *Accreditation and Quality Assurance* 1999;4:280-291.

140. Johnston AC, Lea RA, Brennan KA, et al. Benzylpiperazine: a drug of abuse? *Journal of Psycopharmacology* 2007; 21:888-894.

141. Austin H, Monasterio E. Acute psychosis following ingestion of 'Rapture'. *Australian Psychiatry* 2004;12:406-408.

142. Sheridan J, Butler R. "They're legal so they're safe, right?" What did the legal status of BZP-party Pills mean to young people in New Zealand. *International Journal of Drug Policy* 2010;21:77-81.

143. Drug Profiles, BZP and Other Piperazines, European Monitoring Centre for Drugs and Drug Addiction (EMCDDA), 2012, Available at http://www.emcdda.europa.eu/publications/drug-profiles/bzp, Accessed July 2012.

144. Wikstrom M, Holmgren P, Ahlner J. A2 (N-Benzylpiperazine) a New Drug of Abuse in Sweden. *Journal of Analytical Toxicology* 2004;28:67-70.

145. de Boer D, Bosman IJ, Hidvegi E, et al. Piperazine-like compounds: a new group of designer drugs-of-abuse on the European Market. *Forensic Science International* 2001;121:47-56.

146. United Nations Office on Drugs and Crime. Amphetamines and Ecstasy: Global ATS Assessment. New York: United Nations, 2008.

147. Europol-EMCDDA Joint Report on a new psychoactive substance: 1-benzylpiperazine (BZP), Available from: http://www.emcdda.europa.eu/html.cfm/index132204EN.html.

148. Bossong MG, Van Dijk JP, Niesink RJM. Methylone and mCPP, two new drugs of abuse? . *Addiction Biology* 2005;10:321-323.

149. Control of 1-benzylpiperazine (BZP) and related compounds, Advisory Council on Misuse of Drugs, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/119178/ACMD-BZP-Report.pdf. Accessed September 2008.

150. Bye C, Munro-Faure AD, Peck AW, et al. A comparison of the effects of 1-Benzylepiperazine and Dexamphetamine on human performance tests. *European Journal of Clinical Pharmacology* 1973;6:163-169.

151. Campbell H, Cline W, Evans M, et al. Comparison of the effects of dexamphetamine and 1-benzylpiperazine in former addicts. *European Journal of Clinical Pharmacology* 1973;6:170-176.

152. Baumann MH, Clark RD, Budzynski AG, et al. Effects of "Legal X" Piperazine Analogs on Dopamine and Serotonin Release in Rat Brain. *Annals New York Academy of Sciences* 2004;1025:189-197.

153. Baumann MH, Clark RD, Budzynski AG, et al. N-Substituted Piperazines Abused by Humans Minic the Molecular Mechanism of 3,4-Methylenedioyymethamphetamine (MDMA or 'Edstasy'). *Neuropsychopharmacology* 2005;30:550-560.

154. Yarosh HL, Katz EB, Coop A, et al. MDMA-like behavioural effects of N-substituted piperazines in the mouse. *Pharmacology Biochemistry and Behaviour* 2007;88:18-27.

155. Europol-EMCDDA Joint Report on a new psychoactive substance: 1-benzylpiperazine (BZP), Available from: http://www.emcdda.europa.eu/html.cfm/index132204EN.html. Accessed April 2008.

156. Analysis of Submissions. Proposal to Classify BZP and related substances as Class C controlled drugs under the Misuse of Drugs Act 1975. Final Report, Allen and Clarke Policy and Regulation Specialists, Wellington, New Zealand, April 2007, http://www.health.govt.nz/system/files/documents/publications/bzpanalysissubmissions.pdf, Accessed Oct 2012.

157. New Drug Under Formal Scrutiny, Council asks EMCDDA to assess risks of BZP, News Release from the EU drugs agency in Lisbon, 2007, European Monitoring Centre for Drugs and Drug Addiction, http://www.emcdda.europa.eu/index.cfm?fuseaction=public.AttachmentDownload&nNodeID=27522&sLanguageISO=EN, Accessed April 2008.

158. Green JA. Partying on? Life after BZP-based party pills. *New Zealand Medical Journal* 2008;121:35-42.

159. Annual Report on the Home Office Forensic Early Warning System (FEWS), A system to identify new psychoactive substances in the UK, Home Office, May 2012, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/225840/fews.pdf, Accessed June 2012.

160. Global SMART (Synthetics Monitoring: Analyses, reporting and Trends) Update: United Nations Office on Drugs and Crime, 2010.

161. Communication from the Commission to the European Parliament and the Council, Towards a stronger European response to drugs, European Commission, Brussels, 2011, Available at http://ec.europa.eu/justice/anti-drugs/files/com2011-6892_en.pdf, Accessed Feb 2012.

162. Advisory Council on the Misuse of Drugs, Consideration of the Novel Phsychoactive Susbtances ('Legal highs'), 2011, https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/119139/acmdnps2011.pdf. Accessed Dec 2011.

163. Davies C, English L, Stewart C, et al. The United Kingdom Focal Point on Drugs Annual Report to the European Monitoring Centre for Drugs and Drug Addiction, United Kingdom Drug Situation, Department of Health, UK, 2012, http://www.nwph.net/ukfocalpoint/writedir/userfiles/file/Report%202012/REPORT2012FINAL.pdf, Accessed June 2013.

164. Kuleya C, Hall S, Gautam L, et al. An optimised gas chromatographic-mass spectrometric method for the chemical characterisation of benzylpiperazine and 1-arylpiperazine based drugs. *Analytical Methods* 2014;6:156-163.

165. Police Scotland, Statement of Opinion Unit, *Drug Collection Scheme*, July 2013.

166. Benzylpiperazine (PEP) Pills Are Dangerous And Illegal, UK, Medical News Today, 2007, www.medicalnewstoday.com/articles/66472.php, Accessed Dec 2008.

167. Home Office News and Events, May 2009, http://drugs.homeoffice.gov.uk/news-events/latest-news/proposals-for-control-of gbl, Accessed June 2009, 21 May 2009.

168. Benzylpiperazines, Toxbase, Dec 2009, http://www.toxbase.org/Poisons-Index-A-Z/B-Products/Benzylpiperazines/?print=y, Accessed Jan 2010.

169. Murphy DL, Lesch KP, Aulakh CS, et al. III'Serotonin-selective Arylpiperazines with Neuroendocrine, Behaviour, Behavioral, Temperature, and cardiovascular Effects in Humans. *Pharmacological Reviews* 1991;34:527-551.

170. United Nations Office on Drugs and Crime. Amphetamines and Ecstacy: Global ATS Assessment. New York: United Nations, 2008.

171. Wilkins C, Sweetsur P. Difference in harm from legal BZP/TFMPP party pills between North Island and South Island users in New Zealand: A case of effective industry self-regulation? *International Journal of Drug Policy* 2010;21.

172. Risk Assessment Report of a new psychoactive substance: 1-benzypiperazine (BZP) in accordance with Article 6 of Council Decicion 2005/387/JHA on information exchange, risk assessment and control of new psychoactive substances, European Monitoring Centre for Drugs and Drug Abuse, June 2007, Http://edne-cma.emcddaeuropa.eu/assets/upload/Risk_Assessment_Report_BZP.pdf. Accessed May 2008.

173. Wood DM, Dargan PI, Button J, et al. Collapse, reported seizure - and an unexpected pill. *The Lancet* 2007;369.

174. Logan BK. Amphetamines: An update on Forensic Issues. *Journal of Analytical Toxicology* 2001;25:400-404.

175. Tancer ME, Johanson CE. The subjective effects of MDMA and mCPP in moderate MDMA users. *Drug and alcohol Dependence* 2001;65:97-101.

176. Elliot S, Smith C. Investigation of the First Deaths in the United Kingdom Involving the Detection and Quantitation of the Piperazines BZP and 3-TFMPP. *Journal of Analytical Toxicology* 2008;32:172-177.

177. Nikolova I, Danchev N. Piperazine based substances of abuse - a new party pills on Bulgarian drug market. *Biotechnology and Biotechnology* 2008;22:652-655.

178. Antia U, Tingle MD, Russel BR. Metabolic interactions with piperazine-based 'party-pill' drugs. *Journal of Pharmacy and Pharmacology* 2009;61:877-882.

179. Gee P, Fountain J. Party on? BZP party pills in New Zealand. *Journal of New Zealand Medical Council*;120.

180. Thompson I, Williams G, Aldington S, et al. *Report for the Ministry of Health: The benzylpiperazine (BZP) / trifluoromethylphenylpiperazine (TFMPP) and alcohol safety study* Medical Research Institute of New Zealand, Wellington., 2006.

181. Gee P, Gilbert M, Richardson S, et al. Toxicity from the recreational use of 1-benzylpiperazine. *Clinical Toxicology* 2008;46.

182. Alansari M, Hamilton D. Nephrotoxicity of BZP-based herbal party pills: a New Zealand case report. *The New Zealand Medical Journal* 2006;119:1959-1961.

183. New Drug Under Formal Scrutiny, Council asks EMCDDA to assess risks of BZP, News Release from the EU drugs agency in Lisbon: European Monitoring Centre for Drugs and Drug Addiction, 2007.

184. Wood DM, Button J, Lidder S, et al. Dissociative and Sympathomimetic Toxicity Associated with Recreational Use of 1-(3-Trifluoromethylphenyl)Piperazine

(TFMPP) and 1-Benzylpiperazine (BZP). *Journal of Medical Toxicology* 2008;4:254-257.

185. Drugs and Chemicals of Concern (BZP, TFMPP), United States Drug Enforcement Administration, 2001, http://www.deadiversion.usdoj.gov/drugs_concern/bzp_tmp/bzp_tmp.htm. Accessed Jun 2008.

186. Staack RF. Piperazine designer drug of abuse. *Lancet* 2007;369:1411-1413.

187. Ishida T, Kudo K, Inoue H, et al. Rapid Screening for and Simultaneous Semiquantitative Analysis of Thirty Abused Drugs in Human Urine Samples Using Gas Chromatography - Mass Spectrometry. *Journal of Analytical Toxicology* 2006;30:468-477.

188. Staack RF, Theobald DS, Paul LD, et al. In vivo metabolism of the new designer drug 1-(4-methoxypheny)piperazine MeOPP) in rat and identification of the human cytochrome P450 enzymes responsible for the major metabolic step. *Xenobiotica* 2004;34:179-192.

189. Wilkins C, Sweetsur P, Girling M. Patterns of benzylpiperazine/trifluoromethylphenylpiperazine party pill use and adverse effects in a population sample in New Zealand. *Drug and Alcohol Review* 2008;27:633-639.

190. Nicholson TC. Prevalence of use, epidemiology and toxicology of 'herbal party pills' among those presenting to the emergency department. *Emergency Medicine Australasia* 2006;18:180-184.

191. Thompson I, Williams G, Aldington S, et al. *Report for the Ministry of Health: The benzylpiperazine (BZP) / trifluoromethylphenylpiperazine (TFMPP) and alcohol safety study* Wellington: Medical Research Institute of New Zealand, 2006.

192. Intelligence Report(EXT-2574) Drugs: Risks Associated with New Psychoactive Substances. United Kingdom: Serious and Organised Crime Agency, Sep 2011.

193. Davies S, Wood DM, Smith G, et al. Purchasing ‘legal highs’ on the internet – is there consistency in what you get? *QJM* 2010;103:489-493.

194. Russell B. Party pills - how little is known? *New Zealand Family Physician* 2006;33:46-48.

195. Staack RF, Paul LD, Schmid D, et al. Proof of a 1-(3-chlorophenyl)piperazine (mCPP) intake—Use as adulterant ofcocaine resulting in drug-drug interactions? *Journal of Chromatography B* 2007;855:127-133.

196. Gee P, Fountain J. Party on? BZP party pills in New Zealand. *Journal of New Zealand Medical Council* 2007;120.

197. Parrott E, Sisk E, Turner JJD. Psychobiological problems in heavy 'ecstasy' (MDMA) polydrug users. *Drug and Alcohol Dependence* 2000;60:105-110.

198. Herbert CE, Hughes RN. A comparison of 1-benzylpiperazine and methamphetamine in their acute effects on anxiety-related behaviour of hooded rats. *Pharmacology, Biochemisry and Behavior* 2009;92.

199. Meirinne E, Kajos M, Kankaanpaa A, et al. Rewarding properties of 1-Benzylpiperazine, a new drug of abuse in rats. *Basic & Clinical Pharmacology & Toxicology* 2006;98:346-350.

200. Fantegrossi WE, Winger G, Woods JH, et al. Reinforcing and discriminative stimulus effects of 1-benzylpiperazine and trifluoromethylphenylpiperazine in rhesus monkeys. *Drug and Alcohol Dependence* 2005;77:161-168.

201. Pichini S, Pujadas M, Marchei E, et al. Liquid chromatography-atmospheric pressure ionization electrospray mass spectrometry determination of "hallucinogenic designer drugs" in urine of consumers. *Journal of Pharmaceutical and Biomedical Analysis* 2008;47:335-342.

202. Antia U, Lee HS, Kydd RR, et al. Pharmacokinetics of 'party pill' drug N-benzylpiperazine (BZP) in healthy human participants. *Forensic Science International* 2009;186:63-67.

203. Antia U, Tingle M, Russell B. Validation of an LC-MS method for the detection and quantification of BZP and TFMPP and their hydroxylated metabolites in human plasma and its application to the pharmacokinetic study of TFMPP in humans. *Journal of Forensic Sciences* 2010;55.

204. Antia U, Tingle MD, Russell BR. In vivo interaction between BZP and TFMPP (party pill drugs). *The New Zealand Medical Journal* 2009;122:29.

205. Maurer HH, Kraemer T, Springer D, et al. Chemistry, Pharmacology, Toxicology, and Hepatic Metabolism of Designer Drugs of the Amphetamine (Ecstasy), Piperazine and Pyrrolidinophenone Types, A Synopsis. *Therapuetic Drug Monitoring* 2004;26:127-131.

206. G8 Statement of Intent, Collection and Sharing of Data on NPS (June 2013) https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/208634/NPS_G8_RLG_Statement_of_Intent.pdf. Accessed May 2014.

207. Crime UNOoDa. *Recommended methods for the identification and analysis of piperazines in seized materials*, 2013.

208. Osselton D. Borkenstein Alcohol, Drugs and Driving Course. Glasgow, 2014.

209. Zuba D, Byrska B. Prevalence and co-existence of active components of 'legal highs'. *Drug Testing and Analysis* 2012.

210. Bilinski H, Holownia P, Kapka-Skrzypczak L, et al. Designer Drug (DD) abuse in Poland; a review of the psychoactive and toxic properties of substances found from seizures of illegal drug products and the legal consequences thereof. Part II - Piperazines/Piperidines, Phenylethylamines, Tryptamines, and miscellaneous 'Others'. _Annals of Agriculture and Environmental Medicine_ 2012;19:871-882.

211. Miller RL, DeVane CL. Analysis of trazodone and m-chlorophenylpiperazine in plasma and brain tissue by high performance liquid chromatography. _Journal of Chromatography_ 1986;374:388-393.

212. W R, Krause D, Bartels H, et al. LC-MS/MS analysis of pholedrine in a fatal intoxication case. _Forensic Science International_ 2003;133:101-106.

213. Gonzalez-Marino I, Quintana JB, Rodriguez I, et al. Comparison of molecularly imprinted, mixed-mode and hydrophilic balance sorbents performance in the solid-phase extraction of amphetamine drugs from wastewate samples for liquid chromatography-tandem mass spectrometry determination. _Journal of Chromatography A_ 2009;1216:8435-8441.

214. Telepchak MJ, August TF, Chaney G. _Forensic and clinical Applications of solid Phase Extraction_. New Jersey: Humana Press, 2004.

215. Sample Preparation Products, Certify Methods Manual, Varian Inc, 2006.

216. Cooper GAAC. Application of solid-phase extraction for the analysis of drugs in biological matrices. _Department of Forensic Medicine and Science_: University of Glasgow, 1999.

217. Staack RF, Maurer HH. Piperazine-Derived Designer drug 1-(3-Chlorophenyl)piperazine (mCPP): GC-MS Studies on its Metabolism and its Toxicological Detection in Rat Urine Including Analytical Differentiation from its Precursor Drugs Trazodone and Nefazodone. _Journal of Analytical Toxicology_ 2003;27:560-568.

218. Staack RF, Fritschi G, Maurer HH. New designer drug 1-(3-trimethylphenyl)piperazine (TFMPP): gas chromatography/mass spectrometry and liquid chromatography/mass spectrometry studies on its phase I and II metabolites and on its toxicological detection in rat urine. _Journal of Mass Spectrometry_ 2003;38:971-981.

219. Chen Y, Guo Z, Wang X, et al. Sample Preparation. _Journal of Chromatography A_ 2008;1184:191-219.

220. Staack RF, Maurer HH. New designer drug 1-(3,4-methylenedioxybenzyl)piperazine (MDBP): studies on its metabolism and toxicological detection in rat urine using gas chromatography/mass spectrometry. _Journal of Mass Spectrometry_ 2004;39:255-261.

221. Tsutsumi H, Katagi M, Miki A, et al. Development of simultaneous gas chromatography-mass spectrometric and liquid chromatography-electrospray

ionisation mass spectrometric determination method for the new designer drugs, N-benzylpiperazine (BZP), 1-(3-trifluoromethylphenyl)piperazine (TFMPP) and their main metabolites in urine. *Journal of Chromatography B* 2005;819:315-322.

222. Tsutsumi H, Katagi G, Miki M, et al. Metabolism and the Urinary Excretion Profile of the Recently Schedules Designer Drug N-Benzylpiperazine (BZP) in the Rat. *Journal of Analytical Toxicology* 2006;30:38-43.

223. Moreno IED, da Fonseca BM, Barroso M, et al. Determination of piperazine-type stimulants in human urine by means of microextraction in packed sorbent and high performance liquid chromatography-diode array detection. *Journal of Pharmaceutical and Biomedical Analysis* 2012;61:93-99.

224. Johnson RD, Botch-Jones SR. The stability of four designer drugs: MDPV, Mephedrone, BZP and TFMPP in three biological matrices under various storage conditions. *Journal of Analytical Toxicology* 2013;37:51-55.

225. Suckow RF. High performance liquid chromatographic method for the analysis of plasma m-chlorophenylpiperazine. *Journal of Chromatography* 1990;528:228-234.

226. Dickson AJ, Vorce SP, Holler JM, et al. Detection of 1-Denzylpiperazine, 1-(3-trifluoromethylphenyl)-piperazine, and 1-(3-chlorophenyl)-piperazine in 3,4-methylenedioxymethamphetamine-positive urine samples. *Journal of Analytical Toxicology* 2010;34:464-469.

227. Anderson K, Jalava K, Lock E, et al. Development of a harmonised method for the profiling of amphetamines III. Development of the gas chromatographic method. *Forensic Science international* 2007;169:50-63.

228. Mondello L, Tranchida PQ, Dugo P, et al. Comprehensive Two-Dimensional Gas Chromatography-Mass Spectrometry: A Review. *Mass Spectrometry Reviews* 2008;27:101-124.

229. Staack RF, Maurer HH. Toxicological detection of the new designer drug 1-(4-methoxyphenyl)piperazine and its metabolites in urine and differentiation from an intake of structurally related medicaments using gas chromatography-mass spectrometry. *Journal of Chromatography B* 2003;798:333-342.

230. Kenyon S, Button J, Ramsey J, et al. "Legal highs" analysis of tablets and capsules containing piperazines. . *LTG December Meeting*. London, UK, 2006.

231. Goldberger BA, Huestis MA, Wilkins DG. Commonly practiced quality control and quality assurance procedures for gas chromatography/mass spectrometry analysis in forensic urine drug-testing laboratories. *Forensic Science Review* 1997;9:60-80.

232. *Scientific Working Group for Forensic Toxicology (SWGTOX) Standard Practices for method Validation in Forensic Toxicology, 2013, available at* http://www.swgtox.org/documents/Validation3.pdf, Accessed Feb 2014.

233. Matuszewski BK, Constanzer ML, Chavez-Eng CM. Strategies for the assessment of matrix effect in quantitative bioanalytical methods based on HPLC-MS/MS. *Analytical Chemistry* 2003;75:3019-3030.

234. Kuwayama K, Inoue H, Kanamori T, et al. Analysis of amphetamine-type stimulants and their metabolites in plasma, urine and bile by liquid chromatography with a strong cation-exchange column-tandem mass spectrometry. *Journal of Chromatography B* 2008;867:78-83.

235. Katagi M, Tatsuno M, Miki A, et al. Simultaneous determination of selegiline-N-oxide, a new indicator for selegiline administration, and other metabolites in urine by high-performance liquid chromatography-electrospray ionization mass spectrometry *Journal of Chromatography B* 2001;759:125-133.

236. Bassindale TA, Berezowski R. Quantitative analysis of hair samples for 1-benzylpiperazine (BZP) using high performance liquid chromatography triple quadrupole mass spectrometry (LC-MS/MS) detection. *Analytical and Bioanalytical Chemistry* 2011;401:2013-2017.

237. Wohlfarth A, Weinmann W, Dresen S. LC-MS/MS screening method for designer amphetamines, tryptamines, and piperazines in serum. *Analytical and Bioanalytical Chemistry* 2010:2403-2414.

238. Strano-Rossi S, Anzillotti L, Castrignano E, et al. Ultra high performance liquid chromatography-electrospray ionization-tandem mass spectrometry screening methods for direct analysis of designer drugs, "spice" and stimulants in oral fluid. *Journal of Chromatography A* 2012;1258:37-42.

239. Swortwood MJ, Boland DM, DeCaprio AP. Determination of 32 cathinone derivatives and other designer drugs in serum by comprehensive LC-QQQ-MS/MS analysis. *Analytical and Bioanalytical Chemistry* 2013;405:1383-1397.

240. Montesano C, Sergi M, Moro M, et al. Screening of methylenedioxyamphetamine and piperazine-derived designer drugs in urine by LC-MS/MS using neutral loss and precursor ion scan. *Journal of Mass Spectrometry* 2012;48:49-59.

241. Mortier KA, Dams R, Lambert WE, et al. Determination of paramethoxyamphetamine and other amphetamine-related designer drugs by liquid chromatography/sonic spray ionisation mass spectrometry. *Rapid Communications in Mass Spectrometry* 2002;16:865-870.

242. Bischoff R, Hopfgartner G, Karner HT, et al. Summary of a recent workshop/conference report on validation and implementation of bioanalytical methods: Implications on manuscript review in the Journal of Chromatography B. *Journal of Chromatography B* 2007;860:1-3.

243. Tan A, Awaiye K, Trabelsi F. Impact of calibrator concentrations and their distribution on accuracy of quadratic regression for liquid chromatrography-mass spectrometry bioanalysis. *Analytica Chimica Acta* 2014;815:33-41.

244. Elliot S, Woolacott H, Braithwaite R. The prevalence of drugs and alcohol found in road traffic fatalities: A comparative study of victims. _Science and Justice_ 2009;49:19-23.

245. Yao M, Shah VR, Shyu WC, et al. Sensitive liquid chromatographic-mass spectrometric assay for the simultaneous quantitation of nefazone and its metabolites hydroxynefazodone m-chlorophenylpiperazine and triazole-dione in human plasma using single-ion monitoring. _Journal of Chromatography B_ 1998;18:77-85.

246. Dodd S, Buist A, Burrows G, et al. Determination of nefazodone and its pharmacologically active metabolites in human blood plasma and breast milk by high-performance liquid chromatography. _Journal of Chromatography B_ 1999;730.

247. Franc JE, Duncan GF, Farmen RH, et al. High-performance liquid chromatographic method for the determination of nefazodone and its metabolites in human plasma using laboratory robotics. _Journal of Chromatography_ 1991;570:129-138.

248. Rust KY, Baugertner MR, Dally AM, et al. Prevalence of new psychoactive substances: A retrospective study in hair. *Drug Testing and Analysis* 2012;4:402-408.

249. Chen XH, Wijsbeek J, van Veen J, et al. Solid-phase extraction for the screening of acidic, neutral and basic drugs in plasma using a single-column procedure on Bond Elute Certify. _Journal of Chromatography_ 1990;529:161-166.

250. Hindmarch I. Psychomotor Function and Psychoactive Drugs. _British Journal of Clinical Pharmacology_ 1980;10:189-209.

251. Drummer OH, Gerostamoulos J, Batziris H, et al. The incidence of drugs in drivers killed in Australian road traffic crashes. _Forensic Science International_ 2003;134:154-162.

252. Morland J. Driving under the influence of non-Alcohol Drugs. _Forensic Science Review_ 2000;12:79-105.

253. Akram G, Forsyth AJM. A literature review detailing the nature and prevalence of dance drugs and driving. _International Journal of Drug Policy_ 2000;11:265-277.

254. Literature Review on the Relation between Drug Use, Impaired Driving and Traffic Accidents (CT.97.EP.14), European Monitoring Centre for Drugs and Drug Addiction, Lisbon, 1999, file:///C:/Users/Carrie/Downloads/EMCDDA%20Literature%20review%20on%20the%20relation%20between%20drug%20use,%20impaired%20driving%20and%20traffic%20accidents.pdf, Accessed Dec 2008.

255. Publication HO. Drugs, Driving and Road Safety, Home Office Publication available at http://drugs.homeoffice.gov.uk/publication-search/communities/drivingandroadsafety. Accessed March 2009, 2006.

256. Impairment Research Update Spring 2003: Department for Transport- Road safety Division, 2003.

257. *Illegal Drugs and Driving*: International Council on Alcohol, Drugs and Traffic Safety (ICADTS) Working Group on Illegal Drugs and Driving, 2000.

258. Pompidou Group, Driving and Alcohol, drugs and medicine: concerted European Action needed, 2006, https://wcd.coe.int/ViewDoc.jsp?id=1098107&Site=COE, Accessed Dec 2008.

259. Seymour A, Oliver JS. Role of drugs and alcohol in impaired drivers and fatally injured drivers in the Strathclyde police region of Scotland, 1995-1998. *Forensic Science International* 1999;103:89-100.

260. Ramaekers JG, Kuypers KPC, Wood CM, et al. IMMORTAL Deliverable D-R4.4 Experimental studies on the effects of licit and illicit drugs on driving performance; psychomotor skills and cognitive function, 2004.

261. Walsh JM, De Grier JJ, Christopherson AS, et al. Drugs and Driving. *Traffic Injury Prevention* 2004;5:241-253.

262. Couper FJ, Logan B. Drugs and Human Performance Fact-sheet National: Highway Traffic Safety Administration US Department of Transportation, 2004.

263. Literature Review on the Relation between Drug Use, Impaired Driving and Traffic Accidents (CT.97.EP.14). Lisbon: European Monitoring Centre for Drugs and Drug Addiction, February 1999.

264. Drummer O, Gerostamoulos J, Batziris H, et al. The involvement of drugs in drivers of motor vehicles killed in Australian road traffic crashes. *Accident Analysis and Prevention* 2004;36:239-248.

265. Myant K, Hope S, McIntosh J, et al. Illicit Drugs and Driving; Scottish Executive Social Research, 2006.

266. Movig KLL, Mathijssen MPM, Nagel PHA, et al. Psychoactive substances and the risk of motor vehicle accidents. *Accident Analysis and Prevention* 2004;36:631-636.

267. Holmgren P, Holmgren A, Ahlner J. Alcohol and drugs in drivers fatally injured in traffic accidents in Sweden during the year 2000-2002. *Forensic Science International* 2005;152:11-17.

268. Moore THM, Zammit S, Lingford-Hughes A, et al. Cannabis use and risk of psychotic or affective mental health outcomes: a systematic review. *The Lancet* 2007;370:319-328.

269. Drugs and Driving: Drugnet Europe; EMCDDA, 2007, http://www.emcdda.europa.eu/html.cfm/index41242EN.html, Accessed April 2009, 2007.

270. Hughes B. *Drugs in Focus: Briefing of the European Monitoring Centre for Drugs and Drug Addiction*: EMCCDA publications office, 2009.

271. Annual Report on the State of the Drugs Problem in Europe, European Monitoring Centre for Drugs and Drug Addiction EMCDDA, Lisbon, 2011, http://www.emcdda.europa.eu/publications/annual-report/2011, Accessed Feb 2012.

272. Julien RM. A Primer of Drug action. 9th ed. New York: Worth Publishers, 2001.

273. Schumacher M (Ed), Urmeew R, Auerbach K. Driving under the Influence of Drugs, Alcohol and Medicines, DRUID, *Final Report: Work performed, main results and recommendations*, 2012.

274. Papadodima AA, Athanaselis SA, Dona AA, et al. Driving under the influence in Greece: A 7-year survey (1998-2004). *Forensic Science International* 2008;174:157-160.

275. Bernhoft IM, Steentoft A, Johansen SS, et al. Drugs in injured drivers in Denmark. *Forensic Science International* 2005;150:181-189.

276. Augsburger M, Rivier L. Drugs and alcohol among suspected impaired drivers in Canton de Vaud Switzerland. *Forensic Science International* 1997;85:95-104.

277. Christophersen AS, Morland J. Drugged driving, a review based on the experience in Norway. *Drug and Alcohol Dependence* 1997;47:125-135.

278. Skurtveit S, Abotness B, Christophersen AS. Drugged driving in Norway with benzodiazepine detections. *Forensic Science International* 2002;125:75-82.

279. Laloup M. Correlation of Δ9-tetrahydrocannabinol concentrations determined by LC-MS-MS in oral fluid and plasma from impaired drivers and evaluation of the on-site Drager Drugtest. *Forensic Science International* 2006;161:175-179.

280. Mercer GW, Jeffery WK. Alcohol, drugs, and impairment in fatal traffic accidents in British Columbia. *Accident Analysis and Prevention* 1995;27:335-345.

281. Augsburger M, Donzé N, Ménétrey A, et al. Concentration of drugs in blood of suspected impaired drivers. *Forensic Science International* 2005;153:11-15.

282. Burch HJ, Clarke EJ, Hubbard AM, et al. Concentrations of drugs determined in blood samples collected from suspected drugged drivers in England and Wales. *Journal of Forensic and Legal Medicine* 2013;20:278-289.

283. Carmen del Río M, Gómez J, Sancho M, et al. Alcohol, illicit drugs and medicinal drugs in fatally injured drivers in Spain between 1991and 2000. *Forensic Science International* 2002;127:63-70.

284. Lamping S. An Overview of Drugs and Driving in England and Wales. . 50 Years of Toxicology at the University of Glasgow 2009.

285. Cone EJ, Weddington WW. Prolonged occurrence of cocaine in human saliva and urine after chronic use. *Journal of Analytical Toxicology* 1989;13:65-68.

286. North P. Report of the Review of Drink and Drug Driving and Law. London: Department for Transport, 2010.

287. Oliver JS, Seymour A, Wylie FM, et al. University of Glasgow and Department for Transport Road Safety Research Report No63: Monitoring the Effectiveness of UK Field Impairment Tests. 2006.

288. Officer J. Trends in drug use of Scottish drivers arrested under Section 4 of the Road Traffic Act - A 10 year review *Science and Justice* 2009;49:237-241.

289. Blencowe T, Pehrsson A, Lillsunde P. (Eds) Driving under the Influence of Drugs, Alcohol and Medicines, DRUID, Analytical Evaluation of oral fluid screening devices and preceding selection procedures, 2010.

290. *Summary of Main DRUID Results*: Driving Under the Influence of Drugs, Alcohol, and Medicines DRUID, 2012.

291. Smink BE, Ruiter B, Lusthof KF, et al. Driving under the influence of alcohol and/or drugs in the Netherlands 1995-1998 in view of the German and Belgian legislation. *Forensic Science International* 2001;120:195-203.

292. Jackson PG, Hilditch CJ. A Review of Evidence Related to Drug Driving in the UK: A Report Submitted to the North Review Team: Department for Transport, 2010.

293. De Souza R. As it happens Drug and Alcohol Misuse in Edinburgh: Edinburgh Drug Action Team, 2003.

294. King E, Delany K, Rainger S, et al. Prescription to Fix Drug Driving, RAC Foundation, 2006, http://www.racfoundation.org/index.php?option=com_content&task=view&id=357&Itemid=35, Accessed Dec 2009.

295. Home Office Scientific Development Branch briefing (unpublished). Overview of drug driving technological challenges. 25 February 2010. Cited by North P (2010) Report of the Review of Drink and Drug Driving and Law. Department for Transport; London.

296. Department of Health, UK Focal Point on Drugs, United Kingdom Drug situation, 2012 Edition,

www.nwph.net/ukfocalpoint/writedir/userfiles/file/Report%202012/REPORT2012FINAL.pdf, Accessed Dec 2012.

297. Drugs Misuse Statistics Scotland, 2008, http://www.drugmisuse.isdscotland.org/publications/abstracts/isdbull.htm, Accessed April 2008.

298. Hoare J, Flatley J. Drug Misuse Declared Findings from the 2007/08 British Crime Survey, Home Office Statistical Bulletin, 2008, http://webarchive.nationalarchives.gov.uk/20110220105210/rds.homeoffice.gov.uk/rds/pdfs08/hosb1308.pdf. Accessed Oct 2009.

299. DVLA Annual Report 2011-2012, http://www.dft.gov.uk/dvla/ARA2011-12.aspx, Accessed July 2012.

300. King E, Delany K, Rainger S, et al. High risk for young drug drivers RAC Foundation, 2006, www.racfoundation.org/index.php?option=com.content&task=view&id=359&item=35, Accessed Dec 2009.

301. Think! Campaign 2011 Annual Survey, Department for Transport, http://www.dft.gov.uk/publications/think-research/, Accessed Jul 2012.

302. Furr-Holden D, Voas RB, Kelley-Baker T, et al. Drug and alcohol impaired driving among electronic music dance event attendees. *Drug and Alcohol Dependence* 2006;85:83-86.

303. Logan BK, Barnes L. Combined drug and alcohol use in drivers of vehicular assault and homicide (K27). Proceedings of the American Academy of Forensic Sciences, Seattle, WA. 2006:354-355.

304. Secretary of State for Transport *The Government Response to the Reports by Sir Peter North CBE QC and the Transport Select Committee on Drink and Drug Driving*: Her Majesty's Stationery Office, UK, 2011.

305. Verstraete, A.G. and Raes, E. (Eds) (2006) Rosita - 2 Project Final Report.

306. Kuijten C, Houwing S, Mathijssen R, et al. Driving under the Influence of Drugs, Alcohol and Medicines, DRUID, Protocol of "Workshop on drug driving detection by means of oral fluid screening" In: Blencowe T,Lillsunde P, eds, 2009.

307. Roadside Detection of Driving Under the Influence of Drugs - An International Progress Report, 2008, http://www.european-police.eu/Review/binarywriterservlet?imgUid=32e5df6e-d562-031f-e7ba-c907b988f2ee&uBasVariant=11111111-1111-1111-1111-111111111111.

308. *Preliminary Drug Testing Devices; A Guide to type Approval Procedures for Preliminary Drug Testing Devices Used for Transport Law Enforcement in Great Britain*. St Albans, UK: Home Office Centre for Applied Science and Technology.

309. Crime and Policing Newsletter, Home Office, Jan 2013, https://www.gov.uk/government/publications/crime-and-policing-newsletter-january-2013/crime-and-policing-newsletter-january-2013, Accessed Jan 2014.

310. Stark M, Tunbridge R, Rowe D, et al. Drugs, driving and sobriety tests - a review of recent developments. *Journal of Clinical Forensic Medicine* 2002;9:126-132.

311. Woolley JE, Baldock MRJ. *Review of Western Australian Drug Driving Laws*: Centre for Automotive Safety Research, 2009.

312. Jones AW, Holmgren A, Holmgren P. High concentrations of diazepam and nordiazepam in blood of impaired drivers: association with age, gender and spectrum of other drugs present. *Forensic Science International* 2004;146:1-7.

313. Skurtveit S, Christophersen AS, Morland J. Female drivers suspected for drunken or drugged driving. *Forensic Science International* 1995;75:139-148.

314. Jones AW, Holmgren A, Kugelberg FC. Concentrations of scheduled prescription drugs in blood of impaired drivers: Considerations for interpreting the results. *The Drug Monit* 2007;29:248-260.

315. Linnoila M. Tranquillizers and driving. *Accident Analysis and Prevention* 1976;8:15-19.

316. Drink and Drug Driving, Department for Transport, 2012, http://www.dft.gov.uk/topics/road-safety/drink-and-drug-driving/#drug, Accessed July 2012.

317. Reisfield GM, Goldberger BA, Gold MS, et al. The mirage of impairing drug concentration thresholds: A rational for zero tolerance per se driving under the influence of drugs laws. *Journal of Analytical Toxicology* 2012;36:353-356.

318. Wolff K, Brimblecombe R, Forfar JC, et al. Driving Under the Influence of Drugs. *Report from the Expert Panel on Drug Driving*. London: Department For Transport, Feb 2013.

319. Regulations to specify the drugs and corresponding limits for the new offence of driving with a specified controlled drug in the body above the specified limit - A Consultation Document. *Proposed drugs and corresponding limits to be specified in Regulations using the power in section 5A of the Road Traffic Act 1988 (as inserted by the Crime and courts Act 2013)*: Department for Transport, 2013.

320. *Drug Driving - Summary of Responses to the 2 Consultations: Specifying the drugs and their corresponding limits for inclusion in regulations for the new offence of driving with a specified drug in the body.* Department for Transport., March 2014.

321. Steentoft A, Simonsen KW, Linnet K. The frequency of drugs among Danish drivers before and after the introduction of fixed concentration limits. *Traffic Injury Prevention* 2010;11:329-333.

322. Ojaniemi KK, Lintonen TP, Impinen AQ, et al. Trends in driving under the influence of drugs: A register-based study of DUID suspects during 1977-2007. *Forensic Science International* 2009;41:191-196.

323. Jones AW. Driving under the influence of drugs in Sweden with zero concentration limits in blood for controlled substances. *Traffic Injury Prevention* 2005;6.

324. *Code of Practice for Preliminary Impairment Tests*: Department for Transport, 2004.

325. Hope AT, Woolman PS, Gray WM, et al. A system for psychomotor evaluation; design, implementation and practice effects in volunteers. *Anaesthesia* 1998;53:545-550.

326. Bramness JG, Skurtveit S, Morland J. Clinical impairment of benzodiazepines - relation between benzodiazepine concentrations and impairment in apprehended drivers. *Drug and Alcohol Dependence* 2002;68:131-141.

327. Stutser J, Burns M. Validation of the standardised filed sobriety test battery at BACs below 0.10 percent. *US Department of Transportation National Highway Traffic Safety Administration, California* 1998.

328. Papafotiou K. An Evaluation of the Efficiency of Sobriety Testing to Detect Blood Levels of Cannabis and Impaired Driving Abillity. Melbourne, Australia: Swinburne University of Technology, 2001.

329. Schechtman E, Shinar D. Modelling drug detection and diagnosis with the 'drug evaluation and classification program' *Accident Analysis and Prevention* 2005;37:852-861.

330. Shinar D, Schechtman E. Drug identification performance on the basis of observable signs and symptoms. *Accident Analysis and Prevention* 2005;37:843-851.

331. Farrell LJ, Kerrigan S, Logan BK. Recommendations for Toxicological Investigation of Drug Impaired Driving. *Journal of Forensic Sciences* 2007;52:1214-1218.

332. Heishman SJ, Singleton EG, Crouch DJ. Laboratory Validation Study of Drug Evaluation and Classification Program: Theanol, Cocaine and Marijuana *Journal of Analytical Toxicology* 1996;20:468-483.

333. Bigelow GE, Bickel WE, Roach JD, et al. Identifying types of drug intoxication: Laboratory evaluation of the Subject Examination Procedure. Washington, DC: National Highway Traffic Safety Administration.

334. Beirness DJ, LeCavalier J. Evaluation of the Drug Evaluation and Classification Program: A Critical Review of the Evidence. _Traffic Injury Prevention_ 2007;8:368-376.

335. Tunbridge RJ, Keigan M, James FJ. Recognising drug use and drug related impairment in drivers at the roadside. TRL Report 464. Crowthorne: Transport Research laboratory, 2000.

336. Staack RF, Fritschi G, Maurer HH. Studies on the metabolism and toxicological detection of the new designer drug N-benzylpiperazine in urine using gas chromatography-mass spectrometry. _Journal of Chromatography B_ 2002;773:35-46.

337. O'Keefe M. Drugs driving - standardized field sobriety tests: a survey of police surgeons in Strathclyde. _Journal of Clinical Forensic Medicine_ 2001;8:57-65.

338. Dixon PR, Clark T, Tiplady B. Evaluation of a roadside impairment test device using alcohol. _Accident Analysis and Prevention_ 2009;41:412-418.

339. Ledingham D. Drugs and driving: a retrospective study of the analyses of blood and urine specimens submitted to the Lothian and Borders Police Forensic Laboratory. _Journal of Clinical Forensic Medicine_ 1999;6:133-140.

340. Johnston A, Ramsey J. Drugs, driving and sobriety tests. _Journal of Clinical Forensic Medicine_ 2003;10:126-127.

341. Wall, Karch. Traffic Medicine in Clinical Forensic Medicine. _A Physicians Guide_. NewJersey: Humana Press Inc, 2005.

342. Sexton BF, Tunbridge RJ, Board A, et al. The influence of cannabis on driving; Report 477: Department of the Environment, Transport and the Regions. Transport Research Laboratory.

343. The Royal College of Physicians, The Faculty of Forensic and Legal Medicine, Section 4 RTA Assessment Pro-Forma, http://fflm.ac.uk/librarydetail/4000006.

344. Angel v Chief Constable of South Yorkshire. Judgement available on-line: http://www.bailii.org/ew/cases/EWHC/Admin/2010/883.html.

345. Fernie CGM. Drugs and driving - a comparison of observations made by police surgeons and police officers. _Journal of Clinical Forensic Medicine_ 1999;6:191.

346. Porath-Waller AJ, Bierness DJ, Beasley EE. Toward a more parsimonious approach to drug expert evaluations. _Traffic Injury Prevention_ 2009;10:513-518.

347. Klemenjak W, Braun E, Alvarez J, et al. Deliverable A3.2 Final Programme Report, IMMORTAL., 2005.

348. New Jersey DWI Defense Centre: Field Sobriety Testing Fact and Fiction, http://njdwi.com/field-sobriety-testing-1.htm, Accessed Nov 2013.

349. Poole B. Borkenstein Course (2013) Glasgow.

350. Dupont RL, Voas RB, Walsh JM, et al. The need for drugged driving per se laws: a commentary. *Traffic Injury Prevention* 2012;13:31-42.

351. Grant SA, Millar K, Kenny GNC. Blood alcohol concentration and psychomotor effects. *British Journal of Anaesthesia* 2000;85:401-406.

352. Grant SA, Murdoch J, Millar K, et al. Blood propofol concentration and psychomotor effects on driving skills. *British Journal of Anaesthesia* 2000;3:396-400.

353. Wright DB. *First Steps in Statistics* London: Sage Publications, 2002.

354. Handbook of Biological Statistics, Kruskal-Wallis and Mann-Whitney U-test. http://udel.edu/~mcdonald/statkruskalwallis.html, Accessed April 2014.

355. Post-hoc comparisons after a significant Kruskal-Wallis, university of Nevada, http://wolfweb.unr.edu/homepage/maddux/stat/cep740/transformations.second.lec.after.factorial.html. Accessed April 2014.

356. Post hoc tests for ANOVA, University of Oregon, http://pages.uoregon.edu/stevensj/posthoc.pdf, http://pages.uoregon.edu/stevensj/posthoc.pdf, Accessed April 2014.

357. Winn B, Whitaker D, Elliot DB, et al. Factors effecting light-adapted pupil size in normal human subjects. *Investigative Opthalmology and Visual Science* 1994;35:1132-1137.

358. Watson AB, Yellott J. A united formula for light-adapted pupil size. *Journal of Vision* 2014;14.

359. Drugs and Driving, European Monitoring Centre for Drugs and Drug Addiction, 2007, http://www.emcdda.europa.eu/html.cfm/index44716EN.html, Accessed Dec 2008.

360. Downey LA, King R, Papafotiou K, et al. Detecting impairment associated with cannabis with and without alcohol on the Standardised Field Sobriety Tests. *Psychopharmacology* 2012;224:581-589.

361. Papafotiou K, Carter JD, Stough C. The relationship between performance on the standardised field sobriety tests, driving performance and the level of Δ9-tetrahydrocannabinol (THC) in blood. *Forensic Science International* 2005;155:172-178.

362. Smink BE, Lusthof KJ, de Gier JJ, et al. The relation between the blood benzodiazepine concentration and performance in suspected impaired drivers, 15.p.483-488. *Journal of Forensic Legal Medicine* 2008;15:483-488.

363. Statistics and Research Methodology, non-parametric tests, confidence interval for a single median, https://epilab.ich.ucl.ac.uk/coursematerial/statistics/non_parametric/confidence_interval.html, Accessed April 2014.

364. Tiplady B, Degia A, Dixon P. Assessment of driver impairment: Evaluation of a two-choice tester using ethanol. *Transport Research Part F8* 2005:299-310.

365. Williamson AM, Feyer AM, Mattick RP, et al. Developing measures of fatigue using alcohol comparison to validate the effects of fatigue on performance. *Accident Analysis and Prevention* 2001;33:313-326.

366. Carrier J, Monk TH. Circadian Rhythms of Performance: New Trends. *Chronobiology International* 2000;17:719-732.

367. Philip P, Sagaspe P, Moore N, et al. Fatigue, sleep restriction and driving performance, 37.p.473-478. *Accident Analysis and Prevention* 2005;37:473-478.

368. Lenne MG, Triggs T, Redman JR. Time of day variations in driving performance. *Accident Analysis and Prevention* 1997;29:431-437.

369. The Road Traffic Act 1988, Available at http://www.legislation.gov.uk/ukpga/1988/52/section/6B. Accessed May 2014.