

出國報告（出國類別：考察）

澳洲國際認證組織對測試實驗室人員能力試驗樣品提供者之認證管理制度
考察報告

服務機關：行政院環境保護署環境檢驗所

姓名職稱：陳滄欽 專員

派赴國家：澳洲

出國期間：105 年 12 月 8 日至 12 月 14 日

報告日期：106 年 3 月 14 日

摘 要

本次前往澳洲國家測量研究所(National Measurement Institute, NMI)之化學與生物計量部門(Cheical and Biological Metrology Branch)中化學參考數值組(Cheical Reference Values)與參考氣體混合組(Reference Gas Mixtures)、澳洲 NATA 子機構 Proficiency Testing Australia 進行考察，透過與執行檢驗室人員能力試驗提供單位之專家針對空氣與水等基質之能力試驗樣品之 Assigned value、均勻度、穩定度測定方式與評估規範、能力試驗樣品分發後之結果統計方式、ISO/IEC 17043 對能力試驗提供單位之規範等方面進行交流討論，得到盲樣測試統計範圍過大之可能解決之道，可供國內盲樣製備與管理之參據。

目 錄

摘 要.....	III
壹、 目的.....	1
貳、 過程.....	1
參、 考察內容.....	2
一、 NMI 化學與生物計量部門中之化學參考數值組(Chemical Reference Values)	2
二、 Proficiency Testing Australia (PTA , 澳洲 NATA 子機構).....	8
三、 NMI 化學與生物計量部門中之 Reference Gas Mixtures	10
肆、 心得.....	14
伍、 建議.....	14
附 錄.....	16

壹、目的

環保法規標準制訂、環境影響評估調查、環境品質監測、公害污染防治及公害稽查管制等，均需要準確精密之檢測數據品質為依據，為確保檢測數據之公信力，提升環境檢測水準，於 79 年 1 月 10 日正式成立本所（環保署環境檢驗所）。為管理公民營環境檢驗測定機構，訂定「環境檢驗測定機構管理辦法」，當中為確保各環境檢驗測定機構之數據品質，於該管理辦法第 21 條規定「檢測機構或其檢測人員應依規定接受中央主管機關之採樣技術評鑑或盲樣測試」，並於同法第 24 條第 1 項第 5 款及第 6 款、第 24 條第 2 項第 3 款訂定相關罰責。

為落實環境檢驗測定機構之檢測品質維持確保檢測準確性之管理，本所每年皆會進行盲樣配製、購買與發放之作業。因盲樣相關作業係涉及檢測機構之檢測數據品質與其不合格後之罰責，相關流程是否無誤且與國際趨勢是否一致則需更加謹慎。國際間以 ISO 17043「Conformity assessment-General requirements for proficiency testing」（能力試驗一般的要求）規範能力試驗機構並進行認證，國內雖有財團法人全國認證基金會(TAF)推動國內各類驗證機構、檢驗機構及實驗室各領域之國際認證，惟目前經 TAF 認證 ISO 17043 之能力試驗執行機構並未有環境相關基質之試驗項目。

有鑑於此，實有必要考察他國之認證組織測試實驗室人員能力試驗與規範能力試驗機構之管理制度，從中吸取經驗，以免於閉門造車且能期能與國際制度接軌，強化環境檢測機構盲樣測試制度與提高其檢測數據品質。本次藉考察澳洲經認證之能力試驗機構之機會，就進一步瞭解其制度面與執行面之相關經驗，供本所於能力試驗規劃之參考，強化認證公信力。

貳、過程

考察期間：105 年 12 月 7 日至 14 日

（一）105 年 12 月 7 日 起程赴澳洲

桃園搭機至澳洲雪梨市

（二）105 年 12 月 8 日

抵達雪梨，準備至 NMI 考察資料

（三）105 年 12 月 9 日 至澳洲國家測量研究所(National Measurement Institute, NMI) (North Ryde 辦公區)化學與生物計量部門(Cheical and Biological Metrology Branch)中之化學參考數值組(Cheical Reference Values)考察

考察議題：考察能力試驗執行機構(水、土壤等基質)之實驗室管理

受訪人：Paul Armishaw(Manager of Chemical Reference Values)

(四) 105 年 12 月 10-11 日 假日 (整理資料)

(五) 105 年 12 月 12 日 至澳洲 NATA 子機構 Proficiency Testing Australia 考察

考察議題：考察認證機構對能力試驗執行機構之管理

受訪人：Mr Philip Briggs (General Manager)等人

(六) 105 年 12 月 13 日 至 NMI (Lindfield 辦公區)化學與生物計量部門(Cheical and Biological Metrology Branch)中之參考氣體混合組(Reference Gas Mixtures) 考察

考察議題：考察氣體基質之能力試驗執行機構其執行方式

受訪人：Dr. Damian Smeulders (Manager of Reference Gas Mixtures)等人

晚間搭機返臺

(七) 105 年 12 月 14 日 返程 (雪梨至桃園)

參、考察內容

一、NMI 化學與生物計量部門中之化學參考數值組(Cheical Reference Values)

NMI 是隸屬於澳洲工業部(Department of Industry, Innovation and Science)下的一個單位，於 2004 年成立，總部設在雪梨市 Lindfield 地區，除此外於雪梨市另於 North Ryde 及 Londonderry 設有兩個分部、於墨爾本(Melbourne)及伯斯(Perth)亦有分部。其下主要有四個主要部門：化學與生物計量部門、物理計量部門、分析服務部門、法定度量衡部門，其目前組織架構如圖 1 所示。

NMI 目前約有 370 名員工，每年之預算經費總額約 7,400 萬澳元 (約 17 億 7,600 萬新臺幣)，其中 3,100 萬澳元由澳洲政府預算提供，4,300 萬澳元需由 NMI 自行籌措 (約六成需自行籌措經費)。此次考察對象之化學參考數值組設立 North Ryde，目前連同經理人 Paul Armishaw 共計 13 名成員，每年之預算經費總額約 180 萬澳元 (約 4,300 萬新臺幣)，主要執行化學方面之能力試驗(Cheical Proficiency Testing)，其為經澳洲國家檢測協會 NATA (National Association of Testing Authorities) 認證 ISO/IEC 17043:2010 之化學能力試驗機構。ISO/IEC 17043 用於以能力試驗為目的之實驗室間比對，以判定個別的實驗室在特定試驗(或量測)上的表現，能力試驗機構依此規範進行能力試驗規畫與執行，以符合國際性要求。

其認證包括環境基質、農作物、食物和飲料、藥物、毒品等方面，認證項目如下表：



National Measurement Institute

December 2016

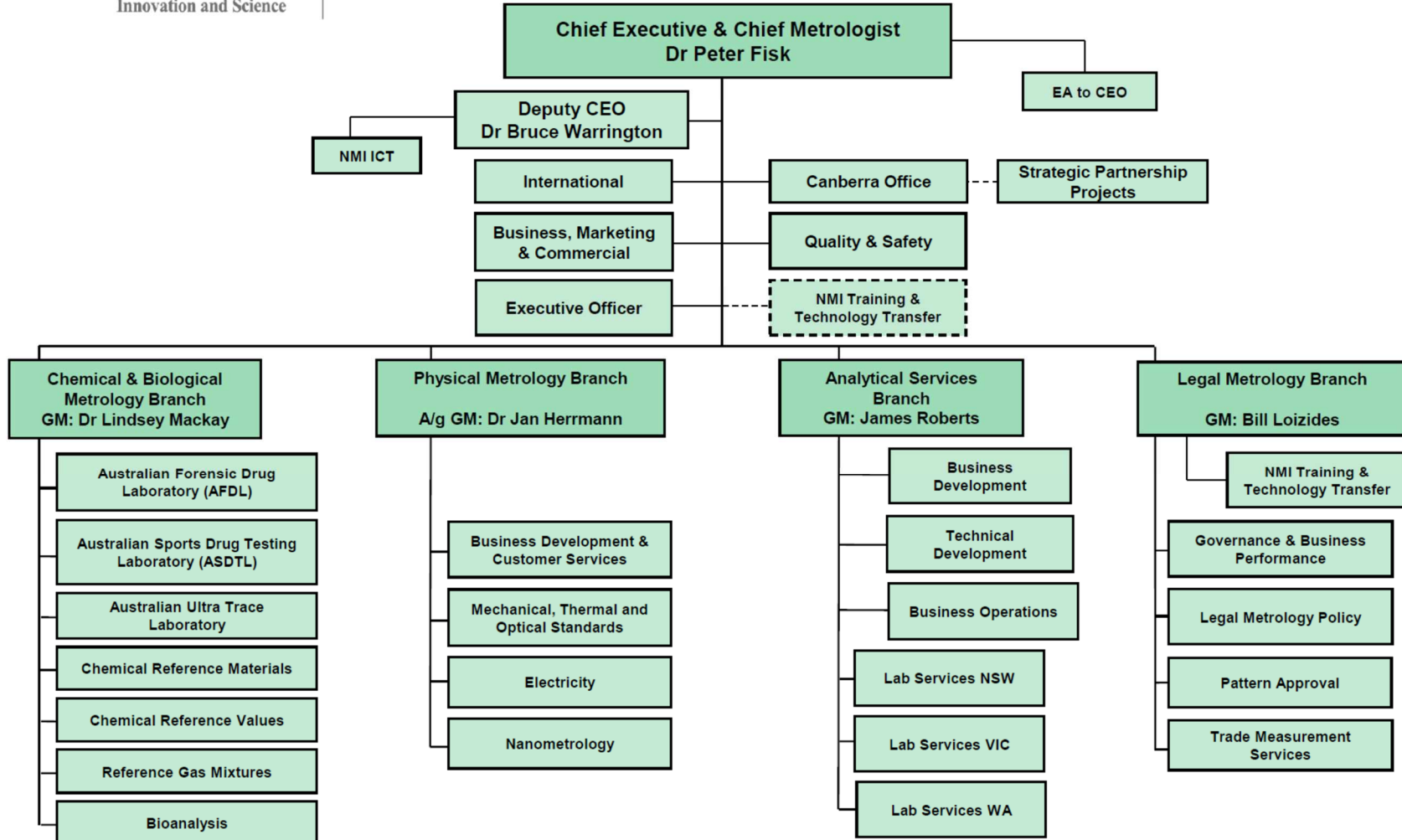


圖 1 NMI 組織架構圖

Environment	
40.01 Chemical Composition, Residues and Contaminants	
.01 Waters Pesticides Petroleum Hydrocarbons Metals Anions Dimethylsulfidepropionate Fluorinated Chemicals	.02 Soils and Sediments Pesticides Petroleum Hydrocarbons Metals Anions Fluorinated Chemicals
Agriculture, Foods and Beverages	
41.01 Chemical Composition, Residues and Contaminants	
.02 Foods and Food Products Pesticide residues Metals Nutrient elements Allergens Fluorinated Chemicals	.04 Potable Water Pesticide residues Metals Anions
41.03 Nutritional Content	
.03 Foods and Food Products Vitamins Nutrient elements	
Health and Community Services	
42.02 Pharmaceuticals	
.01 Active ingredients Metals	.02 Contaminants Metals
Legal	
43.03 Controlled Substances	
.01 Forensic Drugs Analysis of controlled drugs Clandestine laboratory investigation	

此單位所製備之能力試驗樣品皆以實際基質進行配製與添加，例如考察當日檢驗室即正執行蘋果果肉與汁液之混合，做為農藥含量檢測之添加使用；另外水質樣品會至溪流中進行採樣、毒品檢測之驗尿樣品則由收集檢驗室同仁之尿液再添加待測物等，依此配製之能力試驗樣品將能更貼近實際檢測之情形，能降低參加能力試驗之受測檢

驗室是否無法排除樣品基質影響之疑慮。此外，Mr. Paul Armishaw 指出，該單位提供之能力試驗樣品係以實際濃度樣品提供，與本所對檢測機構發放之盲樣或 ERA、RTC 等能力試驗機構所提供之能力試驗樣品係為高濃度樣品，需再自行取樣稀釋方成為能力試驗樣品不同。最大之差異點應在於該批次能力試驗樣品發放完畢後，NMI 即未再做為他用，而本所於年度績效測試發放之盲樣批次，將在環境檢測機構提出許可證申請時會進行發放，而 ERA、RTC 之能力試驗樣品（PT 樣品），將另外成為品質管制樣品（QC 樣品）販賣，而樣品於高濃度的情況下將較利於保存，因此若依據 NMI 之執行方式，其樣品之有效期限將較為縮短，影響相關作業。

NMI 做為一個經認證核可之能力試驗測試機構，其執行能力試驗皆有相關計畫書，而其相關統計方式則彙編成「Statistical Manual」（附錄 1），重點摘要如下：

1. 均勻度測試：於批次樣品分裝後，即隨機挑選至少 7 個樣品（最好 10 個）進行分析，並將每個樣品分成兩份進行重複分析。由所得之數據進行 Cochran's Test，由查表獲得臨界值，判斷 Cochran's Test 計算值是否超過臨界值，若未超過臨界值則表示未有偏離值(outlier)。之後再透過 one-way ANOVA 分析判定，若三種測試值未超過其各自之臨界值，則通過均勻度測試。
2. 設定值(Assigned value)計算：為求得樣品真值濃度之最佳可行建立方式，通常以 Robust mean 做為 Assigned value。
3. 能力試驗之標準偏差(σ)：依據 Horwitz function 之模式推算（如下），是以樣品之濃度落於不同區間而有各別之計算標準，但非為能力試驗參與者之實際數據加以計算而得。

$$\sigma = \begin{cases} 0.22 * c & \text{if } c < 1.2 * 10^{-7} \\ 0.02 * c^{0.8495} & \text{if } 1.20 * 10^{-7} \leq c \leq 0.138 \\ 0.01 * c^{0.5} & \text{if } c < 0.138 \end{cases}$$

where c = concentration, (eg. the assigned value X expressed as a dimensionless mass ratio 1ppm = 10^{-6} or % = 10^{-2})

依據 ISO 13528 (Statistical methods for use in proficiency testing by interlaboratory comparisons)，有幾種方式可求得 σ ：

- (1) 依據能力試驗的目的(目標)，由專家判斷或法規強制規定；
- (2) 依據先前的能力試驗得到的估計值或由經驗得到預期值（經驗值）
- (3) 由統計模式得到的估計值
- (4) 由精密度實驗得到
- (5) 能力試驗參與者之結果得到的標準偏差

4. 計算標準分數(z-score)：此數值為用來判定參加能力試驗單位其分析結果是否令人滿意，一般若其 z-score 大於等於 3 者，則分析結果不令人滿意，亦即結果不合格。

$$z = \frac{(\chi - X)}{\sigma}$$

where:

z = z-score

χ = individual laboratory result

X = assigned value

σ = target standard deviation.

由 NMI 之統計方法有幾個可供本所盲樣試驗之比較參考，分別為：

1. 能力試驗樣品製備後僅著重在均勻度測試上，而未考慮「準確度」。方法雖與本所略有不同，但因本所於盲樣測試結果之判定上，除由剔除偏離值之數據統計其平均值 ± 3 倍標準偏差做為合格範圍外，亦併行以配製值開立百分比做為合格範圍，因此準確度對本所配製之盲樣也很重要，故本所同時測試準確度與精確度之做法亦較國際規範嚴格。
2. 於 Assigned value 之判定，因本所配製之盲樣皆有進行準對度測試而獲得可信賴之配製值，因此亦可由配製值做為 Assigned value。
3. 能力試驗之標準偏差(σ)可參考 NMI 係以經驗模式所得之公式計算之做法，其與本所現行剔除偏離值後之數據進行標準差計算之方式可由下列例子發現最大之差別：

某試驗之原始配製濃度為 1.2 mg/L，參與試驗者之數據為

代碼	A	B	C	D	E	F	G	H	I	J	K	L
數值	1.24	1.17	1.23	2.69	1.30	0.44	0.20	0.78	1.21	1.20	1.10	1.23

因有配製濃度做為參考，以上述 12 家中之檢測數據直觀上 D、F、G 等三個檢驗室之分析數據有顯著的偏離。但以統計方式篩選偏離值時，僅 D 檢驗室之數值會被列為偏離值，而剩餘 11 家之數據平均值則為 1.0047，標準差為 0.3824。合格範圍為平均值 ± 3 倍標準差，即為-0.104 mg/L~2.123 mg/L，此範圍雖為經統計所得，但數據顯示只要不分析得到偏高之數值，即使偏低至無法分析，皆須判定合格，此與執行能力試驗去管控檢驗室人員檢測能力之目的有落差。

若同樣之數據改以 NMI 之評估方式，以經驗模式所得之公式計算標準偏差(σ)，而 Assigned value 則分別選用 12 家檢驗室數據之平均值與測試樣品之配製濃度值進行 z-score 之計算，結果為：

代碼	A	B	C	D	E	F	G	H	I	J	K	L
數值	1.24	1.17	1.23	2.69	1.30	0.44	0.20	0.78	1.21	1.20	1.10	1.23
Z-score (以平均值計)	0.53	0.14	0.47	8.61	0.86	-3.93	-5.54	-2.03	0.36	0.31	-0.25	0.47
Z-score (以配製值計)	0.21	-0.16	0.16	7.98	0.54	-4.07	-5.62	-2.25	0.05	0	-0.54	0.16

可發現此案例不論以數據平均值或配製值去做為 z-score 之計算依據時，D、F、G 三檢樣室之 z-score 皆超出 ± 3 ，而將被判定為不合格，此與直觀上判定結果較為貼近。而各別算出其合格範圍為：

- (1) 以 12 家檢驗室之平均值去計算：0.61 mg/L \sim 1.68 mg/L。
- (2) 以配製值去計算：0.64 mg/L \sim 1.76 mg/L。

進一步去探討為何本所現行執行方式，其統計方式與方法也都合於規範，但最終數值卻有不合常理之情形，可能為若檢驗室檢測數值呈現常態分佈時，則不會造成數值之不合理情形產生，但若數據群組呈現有兩個集團時（分佈不對稱或可能有雙波峰產生時），則造成標準差較大而產生合格範圍合規定卻不合理之情形產生。

此外，NMI 執行能力試驗樣品製備後，其批次樣品執行測試完畢後即無再做他用，穩定度之測試僅為於樣品發放後，再抽取三個樣品執行檢測，其數據與執行均勻度測試所得之數據進行比對，無差異則代表穩定度無疑。但因其配製後至分發樣品之時程較短，穩定度不易有所變化，因此可以此方式執行。但本所之盲樣因後續另有他用，故先前已經驗證配製樣品之穩定期限，但仍有與 Mr. Paul Armishaw 進行意見交換，其建議為若要執行多年的穩定度測試，為減少每批次分析所造成的誤差不同而影響穩定度之判斷，可於樣品配製完成後，將其放置於零下 80 度之環境（其視為內含物皆為穩定不會有所變化），而以 2 年之穩定度做為例子，將 3 個樣品於配製後即擺放於室溫，其餘置放於於零下 80 度之環境；定期（如半年）再各別取出 3 個樣品擺放於室溫下，2 年後共有 5 個時段共 15 個樣品，再同一批次進行分析檢驗。可藉由此分析結果去判定 2 年內是否穩定度有變化，若有變化是於哪一時段開始有變化，即可進行其穩定期限之規範依據。

此次至 NMI 下之化學參考數值組進行考查，獲得許多關於能力試驗機構於數據統計與相關檢測之經驗與啟發，是此行重要的收穫。



圖2 至 NMI 位於 North Ryde 之化學參考數值組考察，主要訪談對象為此組之 Manager Paul Armishaw (左圖右 1)，該單位執行能力試驗樣品之檢驗程序皆依詢 ISO 17025 之實驗室管理系統執行，於儀器旁擺放相關使用紀錄本(右圖)。

二、Proficiency Testing Australia (PTA，澳洲 NATA 子機構)

PTA 是澳洲國家檢測協會 NATA (National Association of Testing Authorities) 的子機構，但其管理與董事會與 NATA 是各自分開的。總部設在雪梨市，另於布里斯本亦設有分部。PTA 因為 NATA 之子機構，所以不能接受 NATA 認證 ISO/IEC 17043:2010，因此 PTA 自 2012 年起即由紐西蘭國際認證機構(International Accreditation New Zealand, IANZ)進行認證成為能力試驗機構。其除了認證化學方面之能力試驗，亦有食物、校正、非破壞性測試、機械、建築材料、環境(微生物)等方面之能力試驗，觸角廣範，其認證項目詳如附錄 2。

PTA 現在之總經理為 Mr. Philip Briggs，他在能力試驗領域之經驗豐富，曾任亞太實驗室認證聯盟 (Asia Pacific Laboratory Accreditation Cooperation, APLAC) 之能力試驗委員會主席，本次考察即與對於認證機構對能力試驗執行機構如何進行管理與 Mr. Philip Briggs 進行意見交換。Mr. Philip Briggs 表示，為了避免球員兼裁判的情形，國際上正在研擬如 NATA 等國際認證組織，與其有關之任何公司、組織等皆不可成為能力試驗執行機構。

至 PTA 前原以為該單位與 NMI 下之化學參考數值組或如 ERA、RTC 等能力試驗機構相同，皆為自行配製能力試驗樣品、分發與統計，結果確出乎預期。PTA 僅有辦公室而無任何實驗室，其能力試驗樣品皆由其他能力試驗機構取得，目前 PTA 之樣品主要來自於紐西蘭的 Global Proficiency Ltd (GP) 及美國的 ERA 等 2 個機構。進一步就此與 Mr. Philip Briggs 請教若化學方面之能力試驗機構受 ISO 17043 認證時若無自行配製樣品，主要重點為何? 答案是能力試驗計畫之審核，Mr. Philip Briggs 指出 ISO 17043 認證並未要求能力試驗機構一定要具備實驗室，僅針對其規畫書內容做審查，當中包括

如何取得能力試驗樣品。以 PTA 為例，就會針對其如何規劃樣品濃度、要求提供樣品之機構其品質管控等進行審查，因此 PTA 僅告知 GP 或 ERA 其所需樣品濃度、時程，其餘樣品配製、均勻度測試、穩定度測試等即由 GP 或 ERA 公司提供。PTA 執行能力試驗之做法則如「Guide to Proficiency Testing Australia」(附錄 3)，其重點如下：

1. 計算標準分數(z-score)之方式與 NMI 略為不同，其計算公式為：

$$Z = \frac{A - \text{median}(A)}{\text{normIQR}(A)}$$

其中，A 為各別檢驗室之測值；median(A)為所有檢驗室排序後之中位數；normIQR(A)為常態化四分位數距差，為所有檢驗室之數據排序後，取其位於 3/4 處之值減去位於 1/4 處之值再乘上一個校正係數而得。

2. 穩定度：部分由 GP 公司製備之測試樣品，其均勻度與穩定度同時測試，取 7 個樣品存放於冷藏環境，另外取 3 個樣品放置在 35°C 的環境下 3 天，並一起進行分析。其假設每升溫 7°C，則老化(反應)速率會加倍，因此若存放於 4°C 與 35°C 則相差 31°C，則 3 天*2^(31/7)等於約 64 天，即由製備後 64 天皆為穩定。

經詢問為何由 ERA 製備之樣品其能力試驗報告中穩定度與 GP 公司有所不同，ERA 表示部分樣品其穩定度為先前已有相關之歷史數據可供參考，因此即設定樣品之穩定期限，而不須每次皆執行。

PTA 由其他能力試驗機構取得樣品之作法與方式，與本所由第一組提出盲樣計畫，再由第三、四、五組協助配製盲樣或購買外部盲樣之執行方式相似，顯示本所之規劃與取得國際認證組織認證之能力測試機構執行模式一致。



圖 3 至 PTA 位於 Rhodes 之總部考察，其與 NATA 於同一建築物中，主要訪談對象為 Philip Briggs (General Manager, 右 1，他為前 APLAC 能力試驗委員會主席)

三、NMI 化學與生物計量部門中之 Reference Gas Mixtures

Reference Gas Mixtures 雖與 Chemical Reference Values 皆隸屬於化學與生物計量部門，但兩單位是分設於不同地區，Reference Gas Mixtures 位於 NMI 在 Lindfield 地區的總部中。本所管理之環境檢驗測定機構中，為數不少之機構皆執行排放管道氣體自動監測作業，因此本所皆會執行相關盲樣測試，而目前有執行氣體方面之能力試驗之機構極少，故透過 Chemical Reference Values 之 Mr. Paul Armishaw 介紹才得以聯絡上 Dr. Damian Smeulders 進而至 Reference Gas Mixtures 考察，機會難得。

Reference Gas Mixtures 主要的業務有兩方面：

1. 配製高純度標準氣體：除了一般經 NMI 分析具有濃度與不確定度之標準參考物質外(CRMs)，因為 NMI 具有澳洲國家一級標準件(質量)，亦以重量法配製而得之一級參考物質 (primary reference materials, PRMs)，該配製皆可追溯至該標準件。此外，也接受氣體產品之驗證作業，可提供分析證明書(COA)。
2. 執行氣體之能力試驗：目前提供之能力試驗項目多為天然氣（含硫量）、液化石油氣（丙烷、丁烷）、排放管道氣體（一氧化碳、二氧化碳、甲烷、氧氣）等商品或數%之高濃度氣體。

該單位亦經澳洲國家檢測協會 NATA (National Association of Testing Authorities) 認證，做為氣體分析檢驗與標準參考物質製造者，其認證 ISO/IEC 17025:2005 與 ISO Guide 34 (2009)。可製備之標準參考物質項目略整理如下表：

Fuel gases (Coal Mine Gas, Coal Seam Gas/Coal Seam Methane (CSG), Coke Oven Gas, Liquefied Natural Gas (LNG), Natural Gas) CO ₂ 、CO、He、H ₂ 、H ₂ S、O ₂ 、甲烷、乙烷等
Environmental gases (Automotive exhaust) CO ₂ 、CO、O ₂ 、丙烷
Mine and Workplace safety CO ₂ 、CO、H ₂ 、甲烷、丙烷
Coal Mine Safety CO ₂ 、O ₂
Speciality gases Xenon、乙烯、乙炔等

藉由 Dr. Damian Smeulders 介紹得知 Reference Gas Mixtures 於執行能力試驗時之相關流程。首先須先配製能力試驗用氣體（置於氣體鋼瓶中），而氣體配製採量測重量之方式執行（此與該單位製備 CRMs 之流程是一致的），流程略述如下：

1. 清洗氣體鋼瓶：將鋼瓶置入專用烘箱中，接上氣體閥件後，將鋼瓶中之氣體以真空泵浦抽出，再填入氮氣，反覆此步驟數次（此時皆於加熱之烘箱中執行），

最後將氣體鋼瓶中之氣體完全抽出。

2. 將清洗過後之鋼瓶擺於天平上，接上專用氣體閥件，導入已事先預估計算出配製濃度須導入之待測氣體體積量後（約略值），再藉由末重與初重差，稱重得到實際導入之待測氣體重量。再將氮氣依相同方式導入並紀錄重量差，並加以換算即可得到配製濃度。
3. 若配製混合氣體時，則先各別導入待測氣體並紀錄其各別重量後，才導入氮氣加以平衡。
4. 將配製後之鋼瓶躺放於專用滾瓶設備，將其滾動使鋼瓶內之氣體均勻混合，最終進行濃度分析確認。
5. 將鋼瓶裝箱、寄交至受測機構。
6. 受測機構於分析後，出具報告並將鋼瓶寄還 NMI（亦可另外購買該瓶做為參考物質）。

於能力試驗結果統計方面，有兩點較化學方面之能力試驗 z-score 計算方式不同，如下：

$$z = \frac{\chi - X}{\sigma}$$

Where:

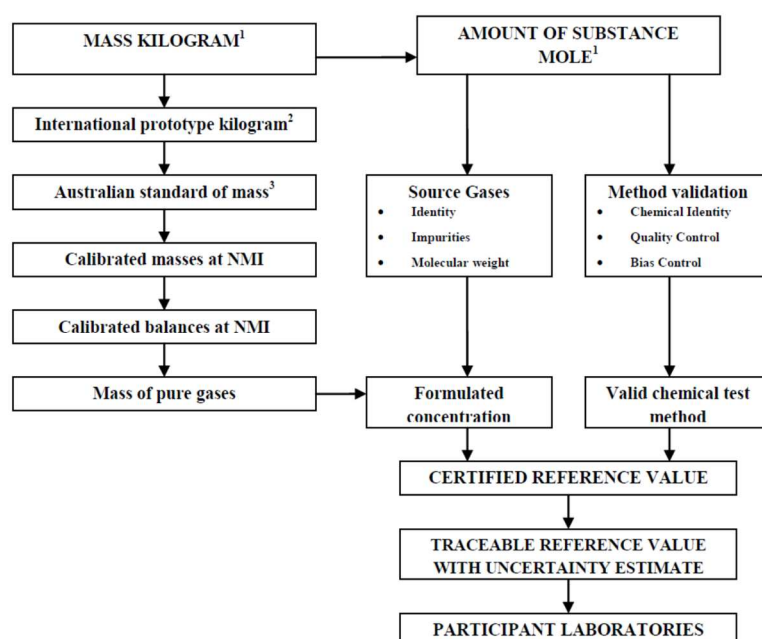
Z = z-Score

χ = Participant result

X = Reference value

σ = Target standard deviation

1. 採用參考數值去計算，非為平均值、中位數等。而參考數值之獲得方式如下圖。



2. 採用「設定」之標準偏差，非為實際數據統計或依經驗公式推算而之標準偏差，係依實際實驗與經驗所訂定出之數值，其會依濃度高低範圍不同而有不同之標準偏差設定值。據 Dr. Damian Smeulders 指出，若依能力試驗參與者之數值計算而得之標準偏差數值通常較大，範圍較為寬鬆，因此採用設定之標準偏差值做為 z-score 計算，而此方式仍符合 ISO 13528 中標準差之選則方式。

透過 Dr. Damian Smeulders 之介紹並實際參觀如何製備能力試驗之氣體樣品，對配製氣體樣品之程序有初步之了解與收獲，可做為日後執行本所環境檢測機構之檢驗室人員能力試驗測試規劃之參據。



圖 4 至 NMI Reference Gas Mixtures 考察，此為清洗能力試驗用鋼瓶之設備。左圖為烘箱，可見其內有閥件可供鋼瓶以快速接頭做連接；右圖為一真空泵浦，烘箱內管線連接至此藉以抽除鋼瓶內之殘存氣體。



圖 5 藉由左圖之閥件控制導入鋼瓶內之標準氣體(或氮氣)之體積量，再由右圖之天平進行重量量測，以獲得鋼瓶內氣體重量，並可用來計算鋼瓶氣體濃度。



圖6 多數氣體配製皆由液體標準品加以氣化而得，左圖為NMI配製氣化標準品之設備，其上有壓力錶可供換算其氣體濃度；右圖為配製後之氣體鋼瓶置放於滾瓶機上，供瓶內氣體均勻混合用。



圖7 配製後之氣體鋼瓶於未使用時多以倒放之方式存放(如左圖)，以免造成鋼瓶內待測物分層而濃度不均之情形；至NMI位於Lindfield之Reference Gas Mixtures考察，主要訪談對象為Dr. Damian Smeulders(右圖右1)。

肆、心得

- 一、 本次至澳洲工業部所屬之 NMI 所屬化學與生物計量部門下之兩個單位 (Chemical Reference Values 與 Reference Gas Mixtures) 與澳洲 NATA 子機構 Proficiency Testing Australia, 考察經國際認證組織認證之檢驗室能力試驗提供者其管理方式 (包括能力試驗樣品製備、均勻度測試、數值統計等方面), 比對現行本所依據環境檢驗測定機構管理辦法第 21 條執行之盲樣測試, 執行方式雖不盡相同, 但各有優點, 本所之盲樣測試程序與國際能力試驗提供機構相較絲毫不遜色。
- 二、 本次考察之 NMI 兩個單位, 分別提供化學與氣體之能力試驗, 其於能力試驗結果統計時其標準差分別依據模式公式計算與經驗設定固定值, 雖有差異但皆符合 ISO 13528 之規範, 可供日後本所執行盲樣測試時之參據。
- 三、 PTA 非自身配製能力試驗樣品而是由 GP 與 ERA 公司提供, 與本所現行分為盲樣使用與供給計畫相同, 表示本所之做法亦可符合 ISO 17043 之規範, 與國際能力試驗提供者之做法一致。
- 四、 能力試驗提供氣體之機構極少, NMI 之 Reference Gas Mixtures 除了執行能力試驗樣品製備外, 亦配製原級參考物質與驗證參考物質, 因此其利用其專業性(專家)經驗, 訂定計算 z-score 時之標準差數值, 而參與能力試驗之單位亦尊重此專家權威判斷, 對技術專業極尊重。

伍、建議

- 一、 NMI 之 Chemical Reference Values 於製備能力試驗樣品過程嚴謹, 且自行測試時多採用原級方法 (Primary Method) 驗證數值, 目前亦在製備底泥中戴奧辛之樣品, 可供本所日後參加國際比測時之另一選擇。
- 二、 NMI 之 Chemical Reference Values 係採用 Horwitz function 之模式推算計算能力試驗之標準偏差(σ), 經參照該算法後, 可避免本所執行之盲樣測試結果若分屬 2 個族群時, 雖剔除偏離值但所得之統計範圍不合理之情形。建議若依現行盲樣統計

方式，建議若出現合格範圍超出 NELAC 組織對於能力試驗樣品經過多年資料累積統計回歸分析獲得之合格範圍時，加採該模式進行計算判斷結果。

- 三、可透過 GP 公司提供 PTA 能力試驗樣品假設每升高 7°C 其衰退速度就會加倍之穩定性測試方法，並參考 ERA、GP 等販賣相同樣品(如水中重金屬)類似濃度之有效期限(已經驗證穩定性無虞)，延長目前本所自行配製之能力試驗樣品之有效期限，以減少過期樣品廢棄之情形產生，發揮每批次最大之效能。

附 錄

附錄	內 容	頁數
附錄 1	NMI 之 Chemical Reference Values 其 Statistical Manual (可由 NMI 網站下載)	15
附錄 2	Proficiency Testing Australia (PTA)受 IANZ 之認證內容與項目 (可由 IANZ 網站下載)	8
附錄 3	GUIDE TO PROFICIENCY TESTING AUSTRALIA (可由 PTA 網站下載)	29



Australian Government
Department of Industry,
Innovation and Science

National
Measurement
Institute

Chemical & Biological Metrology

Statistical Manual

Chemical Reference Values

Issue No.: 3.7 **Issue Date:** 25/10/2016

Approved By: Section Manager

Amendments: Refer Revision History

Control: The electronic copy on the WAN is the latest version of this document. Any paper copy is UNCONTROLLED and should be checked against the electronic copy before use.

Prepared by: Raluca Iavetz

Contents

1.	<i>INTRODUCTION</i>	3
2.	<i>SUFFICIENT HOMOGENEITY TESTING</i>	3
2.1.	Sample Selection and Measurement	3
2.2.	Statistical Analysis of Homogeneity Data	3
2.2.1.	Visual Appraisal for Data Pathologies	4
2.2.2.	Cochran's Test.....	5
2.2.3.	Estimate of Analytical and Sampling Variances	5
2.2.4.	Test for Sufficient Analytical Precision ($s_{an} < 0.5 \sigma$)	6
2.2.5.	Test for Acceptable Between Sample Variance	6
2.3.	Uncertainty Due to Inhomogeneity	7
2.4.	Alternative Homogeneity Testing Procedure used in NMI CPT	7
3.	<i>ESTABLISHING THE ASSIGNED VALUE (X)</i>	8
3.1.	Consensus of Participants' Results	8
3.2.	Measurement by a Reference Laboratory	9
3.3.	Use of a Certified Reference Material	10
3.4.	Formulation	10
4.	<i>SETTING THE TARGET STANDARD DEVIATION (σ)</i>	10
4.1.	By Perception	10
4.2.	From a Predictive Model	10
5.	<i>CALCULATION OF Z-SCORES AND E_N-SCORES</i>	10
5.1.	Introduction	10
5.2.	Invalid results or extreme outliers	11
5.3.	Calculation of z-scores	11
5.4.	Calculation of E_n-scores	11
6.	<i>SUMMARY STATISTICS AND GRAPHS</i>	12
6.1.	Summary Statistics	12
6.2.	Bar Plots	12
6.3.	Scatter Plots of z-Scores	13
6.4.	Box-and-whisker plot	13
6.5.	Kernel density plot	14
7.	<i>REFERENCES</i>	14
8.	<i>REVISION HISTORY</i>	15

1. Introduction

The Chemical Proficiency Testing (CPT) Statistical Manual outlines the statistical methods used by CPT. These methods are based on the procedures described in ISO 13528:2005 (E) “Statistical methods for use in proficiency testing by interlaboratory comparisons”¹ and “The International Harmonized Protocol for the Proficiency Testing of Analytical Chemistry Laboratories”².

The role of the CPT Statistical Manual is to set out the procedures used in assessing the homogeneity of the test materials sent to the participants’, the method of establishing the assigned value and the target standard deviation of a PT study as well as the tools used to assess and compare individual laboratory performance.

2. Sufficient Homogeneity Testing

2.1. Sample Selection and Measurement

Homogeneity testing of the prepared and packaged proficiency test samples should be conducted as soon as possible after packaging.

Select a minimum of 7 (but preferably 10) of the packaged units strictly at random from the entire batch, or by stratified random sampling throughout the fill sequence if fill trend effects are suspected. This must be done in a formal way, by assigning a sequential number to the units (either by label or by their position in a linear sequence). The selection is made by use of a random number table or computer random number generation software. It is not acceptable to select the units in any other way (eg by “shuffling” or “selection at random”).

Homogenise each selected test unit within its container, then take two appropriately sized test portions from each. Label the test portions as “1a”, “1b”, “2a”, “2b” etc. Test portions must be sufficiently large, particularly for solid samples, so as not to compromise the precision of the test results.

Sort the entire set of test portions into a random order, again using a random number table or computer random number generation software.

Analyse each test portion for each analyte of interest, maintaining this random order throughout. The testing should be performed under repeatability conditions (in as short a time as is practical, by a single analyst, preferably in a single sample batch). The analytical method selected must be sufficiently precise to allow a satisfactory estimation of between-sample variance and therefore should have a repeatability standard deviation (s_{an}) of less than half of the target standard deviation (σ) set for the study.

Include appropriate quality control samples (blanks, recoveries, control samples) with each batch of test samples.

2.2. Statistical Analysis of Homogeneity Data

The statistical procedure below follows the “The International Harmonized Protocol for the Proficiency Testing of Analytical Chemistry Laboratories”².

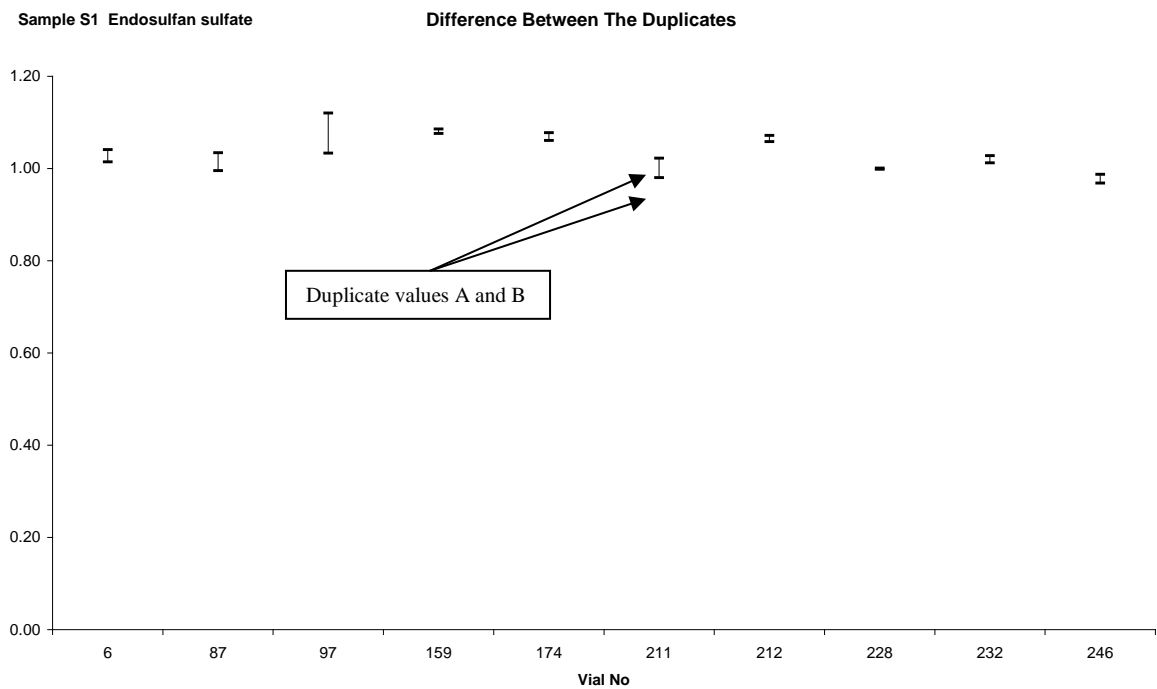
The data in the Table 1 are taken from AQA 06-02, Sample S1 Endosulfan Sulfate

Table 1 Duplicated results for ten distribution units and intermediate stages of calculation in Cochran's test

Sample	A (mg/kg)	B (mg/kg)	D = A-B	S = A+B	D ² =(A-B) ²
6	1.041	1.014	0.027	2.055	0.00070
87	1.034	0.995	0.039	2.029	0.00151
97	1.120	1.033	0.087	2.153	0.00756
159	1.076	1.086	-0.010	2.161	0.00010
174	1.078	1.061	0.017	2.139	0.00028
211	1.023	0.980	0.042	2.003	0.00178
212	1.058	1.072	-0.013	2.130	0.00018
228	1.001	0.998	0.002	1.999	0.00001
232	1.012	1.028	-0.015	2.040	0.00023
246	0.987	0.969	0.019	1.956	0.00035

2.2.1. Visual Appraisal for Data Pathologies

The data presented is inspected visually for suspect features such as discordant duplicated results, outlying samples, trends or discontinuities.



No obvious trends, outliers or discontinuities.

2.2.2. Cochran’s Test

Analytical outliers should be deleted from the data before one-way analysis of variance (ANOVA) is carried out; Cochran’s test is suitable.

Calculate the test statistic (C):

$$C = \frac{D_{\max}^2}{\sum D_i^2}$$

$$= \frac{0.00756}{0.0127}$$

$$= 0.595$$

where C = Cochran’s statistic test

D_{\max} = the largest difference between duplicates

D_i = difference of each pair of duplicates

Table 2 Critical values for the Cochran test statistic for duplicates

m^1	95%
7	0.727
8	0.680
9	0.638
10	0.602
11	0.570
12	0.541
13	0.515
14	0.492
15	0.471
16	0.452
17	0.434
18	0.418
19	0.403
20	0.389

¹ m is the number of samples that have been measured in duplicate.

The 5% critical value for ten samples from Table 2 is 0.602.

No analytical outlier was identified.

2.2.3. Estimate of Analytical and Sampling Variances

One-way ANOVA is used to estimate the analytical and sampling variance and is performed in Excel.

The output from one-way Anova is presented in the table below:

ANOVA						
Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	0.0244	9	0.00271	4.27	0.0166	3.020
Within Groups	0.00635	10	0.000635			

$$\text{So } s_{an}^2 = MS_{within} \\ = 0.0006351$$

where s_{an}^2 = the analytical variance

and

$$s_{sam}^2 = \frac{MS_{between} - MS_{within}}{2} \\ = \frac{0.00271 - 0.000635}{2} \\ = 0.00104$$

where s_{sam}^2 = the between-sample variance

2.2.4. Test for Sufficient Analytical Precision ($s_{an} < 0.5 \sigma$)

The target standard deviation (σ) is the product of the mean of all duplicate results (χ) and the between-laboratory coefficient of variation (CV) which is established by the study coordinator.

$$\sigma = \chi * CV \\ = 1.03 * 0.15 \\ = 0.155 \text{ mg/kg}$$

The analytical standard deviation (s_{an}) is the square root of the analytical variance estimated from ANOVA above.

$$s_{an} / \sigma = \frac{0.0252}{0.155} \\ = 0.163$$

This is less than the critical value of 0.5. The method is precise enough to detect significant in-homogeneity.

2.2.5. Test for Acceptable Between Sample Variance

Calculate the allowable sampling variance (σ_{all}^2) as

$$\sigma_{all}^2 = (0.3 * \sigma)^2 \\ = (0.3 * 0.155)^2 \\ = 0.00216$$

where σ = target standard deviation

The critical value is:

$$c = F_1\sigma_{all}^2 + F_2s_{an}^2$$

$$c = 1.88 * 0.00216 + 1.01 * 0.000635$$

$$= 0.00471$$

The values for factors F_1 and F_2^2 are presented in Table 2.

Table 3 Factors F_1 and F_2 for use in testing for sufficient homogeneity

m^1	20	19	18	17	16	15	14	13	12	11	10	9	8	7
F_1	1.59	1.60	1.62	1.64	1.67	1.69	1.72	1.75	1.79	1.83	1.88	1.94	2.01	2.10
F_2	0.57	0.59	0.62	0.64	0.68	0.71	0.75	0.80	0.86	0.93	1.01	1.11	1.25	1.43

¹ m is the number of samples that have been measured in duplicate.

Compare the sampling variance s_{sam}^2 with the critical value.

The sampling variance ($s_{sam}^2 = 0.00104$) is less than the critical value (0.00471). The samples are sufficiently homogeneous.

The results of the sufficient homogeneity testing is summarised in Table 4.

Table 4: Homogeneity test results

	Value	Critical	Result
Cochran	0.595	0.602	Pass
s_{an}/σ	0.16	0.5	Pass
s_{sam}^2	0.00104	0.00471	Pass

Note: even though statistically significant differences between the test samples have been detected using one-way Anova (P value < 0.02), the inhomogeneity is small enough to be of no practical consequence when compared to the expected between laboratory variability.

2.3. Uncertainty Due to Inhomogeneity

The uncertainty associated with inhomogeneity (u_{hom}) is incorporated into the uncertainty of the assigned value.

- If $F > 1$, then u_{hom} = the sampling standard deviation (s_{sam}) estimated from ANOVA
- If $F < 1$, then u_{hom} = the standard deviation of all results (s_{total}) divided by root 6.

The logic is:

If $F > 1$, sampling variance has been observed, so this can be used to estimate the uncertainty due to inhomogeneity.

If $F < 1$, then the sampling variance is smaller than the analytical variance. This means that any inhomogeneity is so small that the homogeneity testing does not have the power to detect it. The observed variation is almost all due to analytical variance. However this is not proof that the samples are perfectly homogeneous. Inhomogeneity is somewhere between zero, and the analytical variance (estimated as the standard deviation of all results, s_{total}), and it is likely to be closer to 0 than to s_{total} . This approximates a triangular distribution, hence the choice of root 6 as the divisor.

2.4. Alternative Homogeneity Testing Procedure used in NMI CPT

Sometime the above approach for homogeneity testing is not practical. For the analysis of total petroleum hydrocarbons and PFOS/PFOA in water it is necessary to use the whole sample for each analysis and so it is not possible to analyse in duplicate. An alternative is to perform single analyses on a minimum of 5 packaged units (but preferably 7 to 10). The standard deviation of replicate analysis results is an indicator of sample homogeneity. When is not possible to conduct replicate measurements, the standard deviation of the results can be used as s_{sam}^1

The proficiency testing samples may be considered to be adequately homogeneous if:

$$S_{sam} \leq 0.3 \sigma$$

3. Establishing the Assigned Value (X)

The assigned value is the “best practicable estimate of the true value of the concentration (or amount) of analyte in the test material.”³ Methods for establishing assigned value are presented below.

3.1. Consensus of Participants’ Results

The consensus of participants results is used as the assigned value when this value is the only practical method available for the proficiency test. The consensus of participants results is not traceable to any external reference, so although expressed in SI units, metrological traceability is not established.

CPT will calculate an assigned value by this method only if there is a minimum of six results to ensure a reasonable estimate.

The assigned value for the test material used in a proficiency study is the robust average of the results reported by all the participants in the round. This is a modern approach to the outlier problems in a proficiency study in which the influence of the outliers and heavy tails is down-weighted and is calculated using the procedure described in “ISO13258:2015(E), Statistical methods for use in proficiency testing by interlaboratory comparisons – Annex C”¹.

When the assigned value is derived from robust average the uncertainty is estimated as:

$$U_{\text{rob mean}} = 1.25 * S_{\text{rob mean}} / \sqrt{p}$$

where:

$U_{\text{rob mean}}$ = robust mean standard uncertainty

$S_{\text{rob mean}}$ = robust mean standard deviation

p = number of results

The expanded uncertainty ($U_{\text{rob mean}}$) is the standard uncertainty multiplied by a coverage factor $k = 2$ at approximately 95% confidence level.

A worked example is set out below in Table 5 and 6.

Table 5 Participant results AQA 08-13 methamphetamine

Lab Code	Concentration Sample S3
2	71.2
3	57.0
4	55.4
5	58.1
6	55.4
7	58.4
8	60.67
9	55.65
10	57.2
11	55.4
12	59.6
13	45.9
14	57.3
15	56.0
16	55.3
17	61
18	56.5
19	57.7
20	100
21	58.4
22	54.3

Table 6 Robust average and associated uncertainty

No. results (p)	21
Robust mean	57.4
$S_{rob\ mean}$	2.6
$u_{rob\ mean}$	0.7
k	2
$U_{rob\ mean}$	1.4

So the assigned value is $57.4 \pm 1.4\%$ methamphetamine base (m/m).

3.2. Measurement by a Reference Laboratory

An assigned value and uncertainty may be obtained by a suitably qualified measurement laboratory using a method with sufficiently small uncertainty. This is probably the closest approach to obtaining the true value for the test material but it may be very expensive. This approach is used when practical and when resources are available for certain analytes and matrices.

NMI uses primary methods such as Isotope Dilution Mass Spectrometry for which the result is traceable directly to SI and is of the smallest achievable uncertainty. When reference value is used as the assigned value, performance scores are calculated for any number of participants.

3.3. Use of a Certified Reference Material

When the material used in a proficiency testing scheme is a certified reference material (CRM) its certified reference value is used as the assigned value. The uncertainty of the assigned value is derived from the information on uncertainty provided on the certificate.

3.4. Formulation

Formulation is the addition of a known amount or concentration of analyte to a base material which is either free of the analyte or its concentration accurately known. The assigned value is then determined from the proportions of the materials used and the known concentrations added.

This method is advantageous if pure substances are available to spike the test samples, as the added amount can be measured extremely accurately by gravimetric or volumetric methods. Consequently, there is usually no difficulty in establishing the traceability of the assigned value.

The uncertainty is estimated from the uncertainties in analyte concentrations of the materials used and gravimetric and volumetric uncertainties, through moisture content or any other changes during mixing if significant. For more details to estimate standard uncertainty follow the approach described in the "Guide to the expression of uncertainty in measurement"⁵.

4. Setting the Target Standard Deviation (σ)

The target standard deviation (σ) is the product of the assigned value (X) and the between laboratory coefficient of variation (CV).

The between laboratory coefficient of variation is a measure of the between laboratory variation that in the judgement of the study coordinator would be expected from participants given the analyte concentration. It is important to note that this is not the coefficient of variation of participants results.

4.1. By Perception

The target standard deviation could be fixed arbitrarily by the study coordinator based on a perception of how laboratory should perform. The perception is based on practical experience and published models^{4, 5, 6} and varies depending on the concentration in the matrix. The values of target standard deviation for various projects are presented in the CPT Study Protocol.

4.2. From a Predictive Model

Thompson⁶ suggested a contemporary model to calculate the reproducibility standard deviation (σ) based on the Horwitz function⁴. This model predicts a standard deviation from a given concentration (c) and requires c to be dimensionless mass ratio, eg. 1ppm $\equiv 10^{-6}$ or % $\equiv 10^{-2}$.

$$\sigma = \begin{cases} 0.22 * c & \text{if } c < 1.2 * 10^{-7} \\ 0.02 * c^{0.8495} & \text{if } 1.20 * 10^{-7} \leq c \leq 0.138 \\ 0.01 * c^{0.5} & \text{if } c < 0.138 \end{cases}$$

where c = concentration, (eg. the assigned value X expressed as a dimensionless mass ratio 1ppm $\equiv 10^{-6}$ or % $\equiv 10^{-2}$)

5. Calculation of z-scores and E_n -scores

5.1. Introduction

Scoring is the method of converting a participant's raw result into a standard form that adds judgemental information about performance.

Laboratory performance is assessed by comparing reported test results to the assigned value using both z-scores and E_n -scores.

5.2. Invalid results or extreme outliers

Results are identifiably invalid if they are

- expressed in the wrong units,
- transposed
- gross errors
- extreme outliers (eg outside the range of $\pm 50\%$ of the assigned value)
- non-numerical (eg NR not reported, NT not tested, 'less than')

and excluded from statistical analysis and scoring. [1, 2]

5.3. Calculation of z-scores

z-scores are an indication of how much the reported result differs from the assigned value. The assigned value (X) and the target standard deviation (σ) have a critical influence on the calculation of z-scores and must be selected with care if they are to provide a realistic assessment of laboratory performance.

$$z = \frac{(\chi - X)}{\sigma}$$

where:

z = z-score

χ = individual laboratory result

X = assigned value

σ = target standard deviation.

z-scores are interpreted as follows:

- $|z| \leq 2$ satisfactory.
- $2 < |z| < 3$ questionable
- $|z| \geq 3$ unsatisfactory

5.4. Calculation of E_n -scores

E_n -scores (more properly called E_n numbers) are an alternative to z-scores. They provide a measure of how closely a reported laboratory result agrees with the assigned value, taking account of uncertainties in both the result and assigned value. Where a laboratory does not report an uncertainty estimate, an uncertainty of zero (0) is used to calculate the E_n -score.

The E_n -score is an objective measure of whether or not an individual result is consistent with the assigned value. Unlike z-scores, E_n -scores do not require the setting of a target standard deviation.

$$E_n = \frac{(\chi - X)}{\sqrt{U_\chi^2 + U_X^2}}$$

where:

E_n = E_n -score

χ = individual laboratory result

U_χ = expanded uncertainty of the individual laboratory result

X = assigned value

U_X = expanded uncertainty of the assigned value

E_n scores are interpreted as follows:

- $|E_n| \leq 1$ satisfactory.
- $|E_n| > 1$ questionable.

6. Summary Statistics and Graphs

6.1. Summary Statistics

Summary statistics: mean, median, maximum, minimum, robust standard deviation and robust coefficient of variation are calculated from the participants' results and tabulated with the participant results.

A guide to the number of significant figures for the summary statistics is given by Hibbert and Gooding⁷. The recommendation is two significant figures for uncertainty and then the result to the same order of magnitude (eg. uncertainty 0.011 M then the concentration would be expressed as 0.115 ± 0.011 M – 95% confidence interval).

6.2. Bar Plots

Bar charts of results and performance scores are included in the final report. An example chart with interpretation guide is shown in Figure 1. Included with the participant results chart is a histogram.

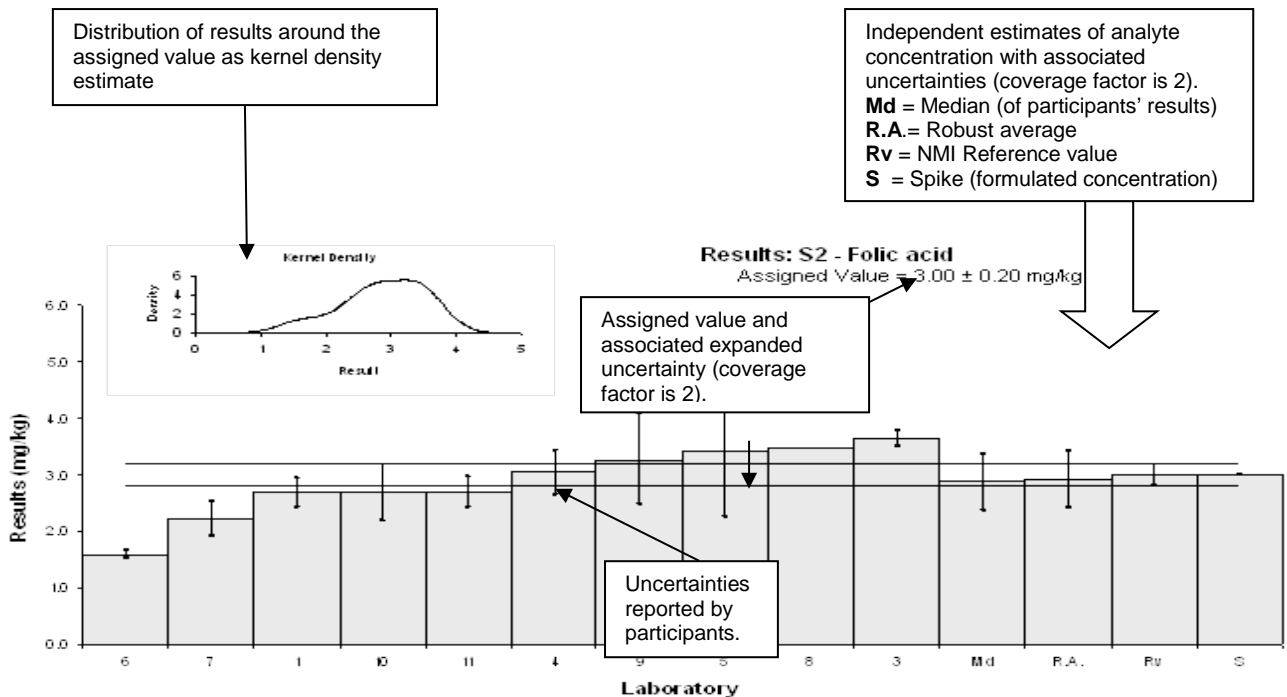


Figure 1 Guide to Presentation of Results

Z-scores and E_n -scores are plotted against the Lab Code Number. Example z-score chart is presented in Figure 2.

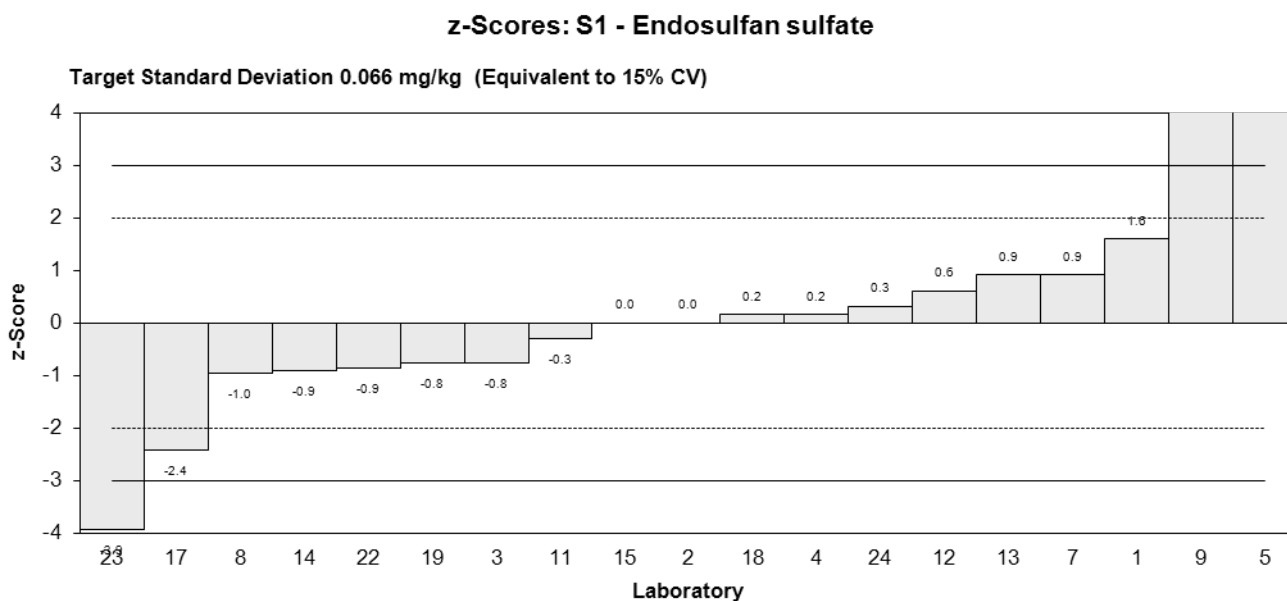


Figure 2. Bar chart z-scores

6.3. Scatter Plots of z-Scores

The z-score scatter plot is presented in Figure 3.

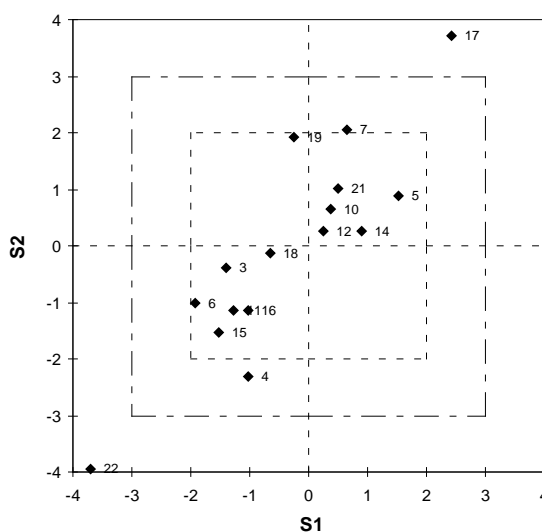


Figure 3 z-score scatter plot for sample S1 and S2

The plot has two squares, the inner square corresponding to a z-score of $|z| < 2$, the outer square corresponding to a z-score of $|z| < 3$. Laboratories falling within the centre square have z-scores with $|z| < 2$ for both samples. Laboratories falling between the inner and outer squares have z-scores with $|z|$ between 2 and 3 for at least one sample. Laboratories falling outside the outer square have at least one z-score with $|z| > 3$.

Within laboratory and between laboratory variability is indicated in the same fashion as for a conventional Youden Plot. For laboratories plotted in the upper right and lower left quadrants, between laboratory variability predominates. For laboratories plotted in the upper left and lower right quadrants, within laboratory variation predominates.

6.4. Box-and-whisker plot

Box and whisker plots⁸ are helpful in interpreting the distribution of data. The diagram shows the quartiles of the data, using these as an indication of the spread. It is made up of a "box", which lies

between the upper and lower quartiles. The median can also be indicated by dividing the box into two. The "whiskers" are straight line extending from the ends of the box to the maximum and minimum values. Example is presented in Figure 4.

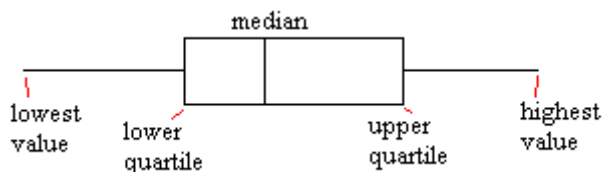


Figure 4 Box-and-whisker plot

6.5. Kernel density plot

An alternative to histograms for visualising the distribution of results is the kernel density estimate. Details about kernel density estimates are presented in AMC Technical Brief no 4. The technical brief and the software required to produce kernel density plots are found at the Royal Society of Chemistry UK.⁹

The Kernel density plot is used to identify modes in the distribution of participants' results. It is also used to identify outlying results.

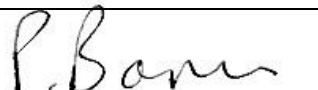
7. References

1. ISO13528:2015 (E), Statistical methods for use in proficiency testing by interlaboratory comparisons, ISO, Geneva, Switzerland.
2. Thompson, M., Ellison, S. L. and Wood, R., The International Harmonised Protocol for the Proficiency Testing of Analytical Chemistry Laboratories, *Pure Appl. Chem.*, 78 (1), 145-196, 2006.
3. Lawn, R. E., Thompson, M. and Walker, R. F., *Proficiency Testing in Analytical Chemistry*, LGC, Teddington, UK, 1997.
4. Horwitz, W., Evaluation of analytical methods used for regulations of food and drugs, *Anal. Chem.*, 54, 67A-76A.6, 1982.
5. Thompson, M., and Lowthian, P.J., A Horwitz-like function describes precision in a proficiency test, *Analyst*, 120, 271-272, 1995.
6. Thompson, M., Recent trends in inter-laboratory precision at ppb and sub-ppb concentrations in relation to fitness for purpose criteria in proficiency testing, *Analyst*, 125, 385-386, 2000.
7. Hibbert, D. B. and Gooding J. J., *Data Analysis for Chemists – An introductory guide for students and laboratory scientists*, first edition, Unversity Press, New York, 2006.
8. Stephen L. R. E., Barwick V. J. and Farrant T. J. D., *Practical Statistics for the Analytical Scientist – A bench guide*, 2nd edition, RSC Publishing, Cambridge, 2009.
9. Royal Society of Chemistry UK, <http://www.rsc.org/>, 2010.

8 Revision History

Date	Issue Number	Reasons for revision
April 2006	1.0	First issue after move to NSW
August 2006	1.1	Issues raised at NATA audit addressed
November 2007	1.2	Issues raised at Internal audit addressed
February 2009	2.0	Issues raised at NATA audit addressed
December 2010	3.0	Complete revision
February 2012	3.1	Small amendments to Chapter 3, 5 and 6 to reflect new requirement in ISO 17043
August 2012	3.2	Changed from NMI Pymble to NMI North Ryde
September 2012	3.3	Issue raised at Internal audit addressed
July 2013	3.4	Review minor change to example chart.
February 2014	3.5	Histogram replaced with Kernel plot in example chart
May 2016	5.2	Invalid result definition expanded
October 2016	2.4	Homogeneity for samples that cannot be analysed in duplicate updated as per ISO 13528:2015 Appendix B

Schedule to	
CERTIFICATE OF ACCREDITATION	
Proficiency Testing Australia (PTA)	Client No: 7356
PO Box 7507, Silverwater, NSW, 2128 7 Leeds Street, Rhodes, NSW, 2138 Telephone: 0061 2 9736-8397 www.pta.asn.au Fax: 0061 2 9743-6664	
Authorised Representative: Mr Philip Briggs General Manager Programme Accredited Proficiency Testing Provider Accreditation Number: 3 Initial Accreditation Date: 30 May 2012	
Conformance Standard ISO/IEC 17043:2010 Conformity assessment - General requirements for proficiency testing	
Proficiency Testing Services Summary Site 1: 7 Leeds Street, Rhodes, NSW 2138 Environmental Food Chemical Construction Materials Mechanical Non-Destructive Testing Calibration Site 2: 628 Ipswich Road, Annerley, QLD 4103 Environmental Chemical	

Authorised: General Manager 	Issue 7	Date: 01/04/16	Page 1 of 8
---	---------	----------------	-------------

Schedule to
CERTIFICATE OF ACCREDITATION

Proficiency Testing Australia (PTA)
 Accredited Proficiency Testing Provider
SCOPE OF ACCREDITATION Accreditation No 3

Site 1: 7 Leeds Street, Rhodes, NSW 2138

Environmental

Programme	Sample Types	Properties
Cryptosporidium and Giardia	Spiked water samples to represent environmental waters	Cryptosporidium and Giardia
Legionella	Lenticule discs, representing cooling tower water when rehydrated	Legionella

Food

Programme	Sample Types	Properties
Non-Pathogens	Freeze dried vials with an accompanying matrix e.g. whole milk powder	Microbiological parameters including microbial loading and indicator organisms
Pathogens	Freeze dried vials with an accompanying matrix e.g. whole milk powder	Detection of target pathogens
Food	Alternating rounds of Dairy, may include: <ul style="list-style-type: none"> • skim milk powder • whole milk powder • cheese Alternating rounds of Non-dairy, may include: <ul style="list-style-type: none"> • meat paste • fish paste • wheat flour • fruit 	Chemical properties including proximates

Schedule to
CERTIFICATE OF ACCREDITATION

Proficiency Testing Australia (PTA)
 Accredited Proficiency Testing Provider
SCOPE OF ACCREDITATION Accreditation No 3

Chemical

Programme	Sample Types	Properties
Air and Emissions	Impinger solutions, glass fibre filter or filter paper	Chemical parameters in liquid solution, filter or filter paper
Bitumen	Samples of bitumen	Physical properties including: <ul style="list-style-type: none"> • Viscosity • Density (Bottle) • Penetration @ 25°C • Flash point (COC)
Cement	OPC Cement	Chemical parameters including: <ul style="list-style-type: none"> • Chemical Composition • Loss on Ignition • Insoluble Residue • Specific Surface Area
Coal	Washed coal	Chemical and physical properties including metals
Metal Alloys	Disc of a metal alloy	Range of chemical compositional analysis including: <ul style="list-style-type: none"> • Various metals
Paint	Tins of water based paint, pre-coated panels and/or uncoated panels	Various physical parameters including (for water based paint): <ul style="list-style-type: none"> • Consistency • Density • Non-Volatiles by Mass and by Volume • Specular Gloss Various physical parameters (for paint panels) including: <ul style="list-style-type: none"> • Measurement of Specular Gloss • Methods of Colour Measurement • Determination of Pencil

Schedule to
CERTIFICATE OF ACCREDITATION

Proficiency Testing Australia (PTA)
 Accredited Proficiency Testing Provider
SCOPE OF ACCREDITATION Accreditation No 3

		Hardness of a Paint Film <ul style="list-style-type: none"> • Adhesion (crosscut) • Dry Film Thickness – Paint Inspection Gauge
Polychlorinated Biphenyls (PCBs)	Oil	Total PCBs and various Arochlors
Soils	Soil sample in sealed ampoule	Chemical composition – Pesticides and Metals
Waters (Chemical)	Potable, effluent water	A range of chemical and physical determinations including: <ul style="list-style-type: none"> • Various metals • Orthophosphate • Hardness • Total solids

Construction Materials

Programme	Sample Types	Properties
Aggregates	Aggregate sample	Various physical properties including: <ul style="list-style-type: none"> • Material Finer Than 75 µm • Flakiness Index • Particle Size Distribution • Particle Density on a Saturated-Surface-Dry Basis
Asphalt	Bucket of asphalt	Various physical properties including: <ul style="list-style-type: none"> • Bitumen Content • Maximum Density • Bulk Density • Grading measurements
Concrete	Concrete cylinders	Various physical properties

Schedule to
CERTIFICATE OF ACCREDITATION

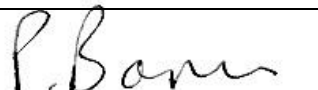
Proficiency Testing Australia (PTA)
 Accredited Proficiency Testing Provider
SCOPE OF ACCREDITATION Accreditation No 3

		including: <ul style="list-style-type: none"> • Dimensions • Mass per unit Volume • Compressive Strength • Type of Failure
Soils	Soil sample	Various physical properties including: <ul style="list-style-type: none"> • Apparent Particle Density • Moisture Content • Liquid Limit • Linear Shrinkage

Mechanical

Programme	Sample Types	Properties
Hardness Testing of Metals	Metals	A range hardness analysis including: <ul style="list-style-type: none"> • Vickers • Rockwell • Brinell
Tensile Testing of Metals	Metals	A range mechanical analysis including: <ul style="list-style-type: none"> • Thickness • Yield • Tensile Strength (Rm)
Textiles	Textiles	A range chemical and mechanical analysis including: <ul style="list-style-type: none"> • Quantitative Fibre Analysis • Breaking Load • Extension
Impact Testing	Metals	Charpy V-Notch Pendulum Impact Test

Non-Destructive Testing

Authorised: General Manager		Issue 7	Date: 01/04/16	Page 5 of 8
--------------------------------	---	---------	----------------	-------------

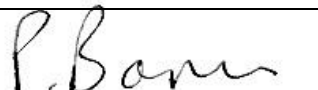
Schedule to
CERTIFICATE OF ACCREDITATION

Proficiency Testing Australia (PTA)
 Accredited Proficiency Testing Provider
SCOPE OF ACCREDITATION Accreditation No 3

Programme	Sample Types	Properties
Magnetic Particle Inspection	Pipe, tee, plate and Y test items	Various tests in accordance with AS 1171 and result reporting in accordance with AS 4037
Radiography	Pipe or plate test item	Various tests in accordance with: AS 2177:2006 (Non-destructive testing - Radiography of welded butt joints in metal), AS 2314:2006 (Radiography of metals - Image quality indicators (IQI) and recommendations for their use), AS 4041:2006 (Pressure piping), Class 1 and AS 1210: 1997 (Pressure vessels), Class 1
Ultrasonics	Pipe, Plate or Tee test items	Various tests in accordance with: AS 2207: 2007 (Non-destructive testing - Ultrasonic testing of fusion welded joints in carbon and low alloy steel)

Calibration

Programme	Sample Types	Properties
Acoustic and Vibration Electrical Gravimetric Heat and Temperature Optics and Radiometry Physical and Dimensional Metrology	PTA reference test item	As specified, compared to known values of the reference item

Authorised: General Manager 	Issue 7	Date: 01/04/16	Page 6 of 8
--	---------	----------------	-------------

Schedule to
CERTIFICATE OF ACCREDITATION

Proficiency Testing Australia (PTA)
 Accredited Proficiency Testing Provider
SCOPE OF ACCREDITATION Accreditation No 3

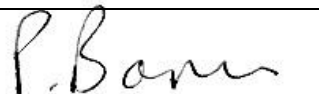
Site 2: 628 Ipswich Road, Annerley, QLD 4103

Environmental

Programme	Sample Types	Properties
Water (Biological)	Alternating potable and effluent water	Parameters including microbial loading and indicator organisms
Algae	Freshwater and seawater samples	Identify and enumerate genera and species

Chemical

Programme	Sample Types	Properties
Asbestos Identification – Building and Related Products	Bulk samples	Detection of: <ul style="list-style-type: none"> • Chrysotile Asbestos • Amosite Asbestos • Crocidolite Asbestos • Synthetic mineral fibres (SMF) • Organic Fibres
National Asbestos Programme	Slides	Fibre counting
Asbestos in Soils	Soil (or similar) samples	Detection of: <ul style="list-style-type: none"> • Chrysotile Asbestos • Amosite Asbestos • Crocidolite Asbestos • Synthetic mineral fibres (SMF) • Organic fibres
Geochemical	Metal ore	Range of elements including: <ul style="list-style-type: none"> • Various metals • Loss on ignition (LOI)
Wine	White wine	Various chemical parameters

Authorised: General Manager 	Issue 7	Date: 01/04/16	Page 7 of 8
--	---------	----------------	-------------

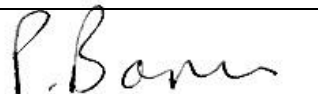
Schedule to
CERTIFICATE OF ACCREDITATION

Proficiency Testing Australia (PTA)
Accredited Proficiency Testing Provider
SCOPE OF ACCREDITATION

Accreditation No 3

	Red wine	including: <ul style="list-style-type: none">• Alcohol• Acids• Dissolved gases• Sugars• Metals
--	----------	--

Uncontrolled copy printed from the Internet

Authorised: General Manager		Issue 7	Date: 01/04/16	Page 8 of 8
--------------------------------	---	---------	----------------	-------------

PROFICIENCY TESTING AUSTRALIA

GUIDE TO PROFICIENCY TESTING AUSTRALIA



2016

© Copyright Proficiency Testing Australia

Revised May 2016

PROFICIENCY TESTING AUSTRALIA

PO Box 7507 Silverwater NSW 2128 AUSTRALIA

CONTENTS

	<i>Page</i>
1. Scope	2
2. Introduction	
2.1 Confidentiality	2
2.2 Funding	3
3. References	3
4. Quality Management of Proficiency Testing Schemes	3
5. Testing Interlaboratory Comparisons	
5.1 Introduction	4
5.2 Working Group and Program Design	4
5.3 Sample Supply and Preparation	5
5.4 Documentation	5
5.5 Packaging and Dispatch of Samples	5
5.6 Receipt of Results	6
5.7 Analysis of Data and Reporting of Results	6
5.8 Other Types of Testing Programs	6
6. Calibration Interlaboratory Comparisons	
6.1 Introduction	7
6.2 Program Design	8
6.3 Test Item Selection	8
6.4 Documentation	8
6.5 Test Item Stability	8
6.6 Evaluation of Performance	8
6.7 Reference Values	9
6.8 Measurement Uncertainty (MU)	9
6.9 Reporting	9
6.10 Measurement Audits	9
Appendix A Glossary of Terms	10
Appendix B Evaluation Procedures for Testing Programs	12
Appendix C Evaluation Procedures for Calibration Programs	24

1. Scope

The purpose of this document is to provide participants in Proficiency Testing Australia's (PTA) programs with an overview of how the various types of proficiency testing programs are conducted and an explanation of how laboratory performance is evaluated. The document does not attempt to cover each step in the proficiency testing process. These are covered in PTA's internal procedures which are in compliance with the requirements of ISO/IEC 17043¹.

The main body of this document contains general information about PTA's programs and is intended for all users of this document. The appendices contain: a glossary of terms (A); information on the evaluation procedures used for testing programs (B); and details of the evaluation of the results for calibration programs (C).

2. Introduction

The competence of laboratories is assessed by two complementary techniques. One technique is an on-site evaluation to the requirements of ISO/IEC 17025². The other technique is by proficiency testing which involves the determination of laboratory performance by means of interlaboratory comparisons, whereby the laboratory undergoes practical tests and their results are compared with those of other laboratories. The two techniques each have their own advantages which, when combined, give a high degree of confidence in the integrity and effectiveness of the assessment process. Although proficiency testing schemes may often also provide information for other purposes (e.g. method evaluation), PTA uses them specifically for the determination of laboratory performance.

PTA programs are divided into two different categories - testing interlaboratory comparisons, which involve concurrent testing of samples by two or more laboratories and calculation of consensus values from all participants' results, and calibration interlaboratory comparisons in which one test item is distributed sequentially among two or more participating laboratories and each laboratory's results are compared to reference values. A subset of interlaboratory comparisons are one-off practical tests (refer Section 5.8) and measurement audits (refer Section 6.10) where a well characterised test item is distributed to *one* laboratory and the results are compared to reference values.

Proficiency testing is carried out by PTA staff. Technical input for each program is provided by Technical Advisers. The programs are conducted using collaborators for the supply and characterisation of the samples and test items. All other activities are undertaken by PTA.

2.1 Confidentiality

All information supplied by a laboratory as part of a proficiency testing program is treated as confidential. There are, however, three exceptions. Information can be disclosed to third parties:

- with the express approval of the client(s);
- when PTA has an agreement with or requirement in writing from the Commonwealth or a State Government which requires the provision of information, and the relevant parties/clients have been informed in writing of such agreement or requirement;
- when PTA has any concerns about the conduct of any aspect of the proficiency testing process or in relation to any safety, medical or public health issues identified in the proficiency testing process.

PTA sample suppliers, distributors and Technical Advisers are required to sign confidentiality declarations at the commencement of each program round.

2.2 Funding

PTA charges a participation fee for each program. This fee varies from program to program and participants are notified accordingly, prior to a program's commencement.

3. References

1. ISO/IEC 17043:2010 *Conformity assessment: General requirements for proficiency testing*
2. ISO/IEC 17025:2005 *General requirements for the competence of testing and calibration laboratories*
3. ISO/IEC 17011:2004 *Conformity assessment: General requirements for accreditation bodies accrediting conformity assessment bodies*
4. ISO/IEC Guide 98-3:2008 *Uncertainty of measurement – Part 3: Guide to the expression of uncertainty in measurement (GUM)*
5. ISO 13528:2015 *Statistical methods for use in proficiency testing by interlaboratory comparisons*
6. APLAC PT001 (revised 2008) *Calibration interlaboratory comparisons*
7. APLAC PT002 (revised 2008) *Testing interlaboratory comparisons*

4. Quality Management of Proficiency Testing Schemes

In accordance with best international practice, PTA maintains and documents a quality system for the conduct of its proficiency testing programs. This quality system complies with the requirements specified in ISO/IEC 17043:2010¹.

5. Testing Interlaboratory Comparisons

5.1 Introduction

PTA uses collaborators for the supply and homogeneity testing of samples. All other activities are undertaken by PTA and technical input is provided by program Technical Advisers.

In the majority of interlaboratory comparisons conducted by PTA, subdivided samples (taken from a bulk sample) are distributed to participating laboratories which test these concurrently. They then return results to PTA for analysis and this includes the determination of consensus values.

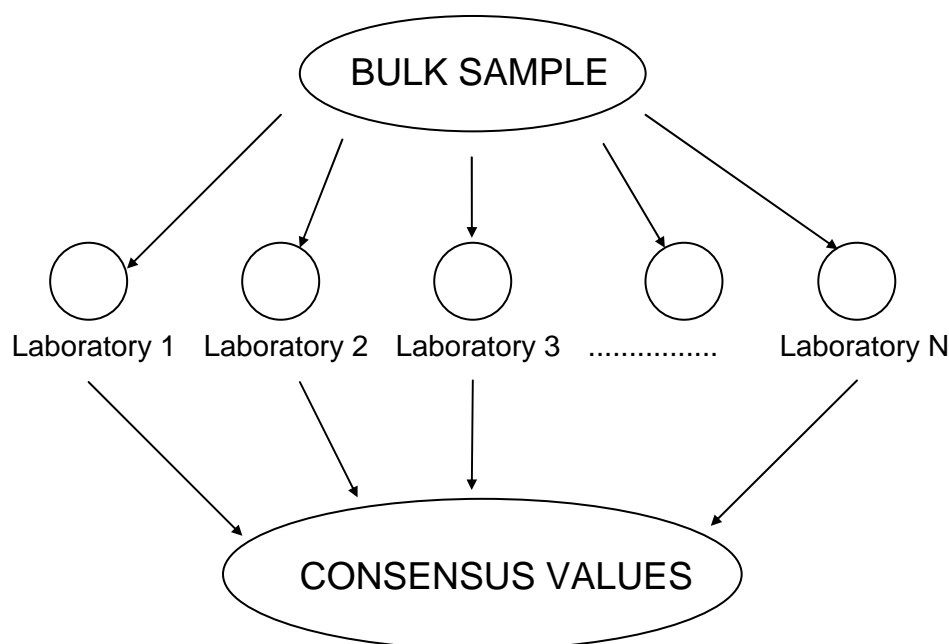


Figure 1: Typical Testing Interlaboratory Comparison

5.2 Working Group and Program Design

Once a program has been selected, a small working group is formed. This group usually comprises one or more Technical Advisers, and the PTA Scientific Officer who will act as the Program Coordinator.

It is most important that at least one, but preferably two, technical experts are included in the planning of the program and in the evaluation of the results. Their input is needed in at least the following areas:

- nomination of tests to be conducted, range of values to be included, test methods to be used and number/design of samples required;
- preparation of paperwork (instructions and results sheet) particularly with reference to reporting formats, number of decimal places to which results should be reported and correct units for reporting;
- identification and resolution of any difficulties expected in the preparation and maintenance of homogeneous proficiency test items, or in the provision of a stable assigned value for a proficiency test item;
- technical commentary in the final report and, in some cases, answer questions from participants.

An appropriate statistical design is essential and therefore must be established during the preliminary stages of the program (see Appendix B for further details).

5.3 Sample Supply and Preparation

The Program Coordinator is responsible for organising the supply and preparation of the samples. It is often the case that one of the Technical Advisers will also act as the program's sample supplier. In any case, the organisation preparing the test items is always one that is considered by PTA to have demonstrable competence to do so.

Sample preparation procedures are designed to ensure that the samples used are as homogeneous and stable as possible, while still being similar to samples routinely tested by laboratories. A number of each type of sample are selected at random and tested, to ensure that they are sufficiently homogeneous for use in the proficiency program. Whenever possible, this is done prior to samples being distributed to participants. The results of this homogeneity testing are analysed statistically and may be included in the final report.

5.4 Documentation

The main documents associated with the initial phase of a proficiency program are:

(a) *Letter of Intent*

This is sent to prospective participants to advise that the program will be conducted and provides information on the type of samples and tests which will be included, the schedule and participation fees.

(b) *Instructions to Participants*

These are carefully designed for each individual program and participants are always asked to adhere closely to them.

(c) *Results Sheet*

For most programs a pro-forma results sheet is supplied to enable consistency in the statistical treatment of results.

Instructions and Results Sheets may be issued with, or prior to, the dispatch of samples.

5.5 Packaging and Dispatch of Samples

The packaging and method of transport of the samples are considered carefully to ensure that they are adequate and able to protect the stability and characteristics of the samples. In some cases, samples are packaged and dispatched from the organisation supplying them, in other cases they are shipped to PTA for this distribution. It is also ensured that certain restrictions on transport such as dangerous goods regulations or customs requirements are complied with.

5.6 Receipt of Results

Results from participating laboratories for PTA testing programs are required to be sent to either our Sydney office or Brisbane office. A 'due date' for return of results is set for each program, usually allowing laboratories two to three weeks to test the samples. If any results are outstanding after the due date, reminders are issued, however, as late results delay the data analysis, these may not be included. Laboratories are requested to submit all results on time.

5.7 Analysis of Data and Reporting of Results

Results are usually analysed together (with necessary distinctions made for method variation) to give consensus values for the entire group. The results received from participating laboratories are entered and analysed as soon as practicable so that the final report can be issued to participants within six weeks of the due date for results.

The evaluation of the results is by calculation of robust z-scores, which are used to identify any outliers. Summary statistics and charts of the data are also produced, to assist with interpretation of the results. A detailed account of the procedures used to analyse results appears in Appendix B.

Participants are issued with an individual laboratory summary sheet (refer Appendix B) which indicates which, if any, of their results were identified as outlier results. Where appropriate, it also includes other relevant comments (e.g. reporting logistics, method selection).

A final report is produced at the completion of a program and includes data on the distribution of results from all laboratories, together with an indication of each participant's performance. This report typically contains the following information:

- (a) introduction;
- (b) features of the program - number of participants, sample description, tests to be carried out;
- (c) results from participants;
- (d) statistical analysis, including graphical displays and data summaries (outlined in Appendix B);
- (e) a table summarising the outlier[†] results;
- (f) PTA and Technical Adviser's comments (on possible causes of outliers, variation between methods, overall performance etc.);
- (g) sample preparation and homogeneity testing information; and
- (h) a copy of the instructions to participants and results sheet.

Note: [†] *Outlier results are the results which are judged inconsistent with the consensus values (refer Appendix A for definition).*

The final program report is released on the PTA website, and participants are notified of its availability via email.

5.8 Other Types of Testing Programs

PTA conducts some proficiency testing activities which do not exactly fit the model outlined in Section 5.1. These include known-value programs where samples with well established reference values are distributed (e.g. slides for asbestos fibre counting).

Further examples are one-off practical tests where material of known composition (e.g. certified reference material) is presented to one laboratory. This type of activity is also extensively used in the calibration area (refer Section 6.10, Measurement Audits). These activities do not, or by their nature cannot, use the usual consensus values as the basis for the evaluation of performance.

Some of PTA's testing interlaboratory comparisons do not produce quantitative results - i.e. qualitative programs where the presence or absence of a particular parameter is to be determined (e.g. pathogens in food). By their nature the results must also be treated differently from the procedures outlined in Appendix B.

6. Calibration Interlaboratory Comparisons

6.1 Introduction

PTA uses collaborators for the supply and calibration of test items. All other activities are undertaken by PTA and technical input is provided by program Technical Advisers. Each calibration laboratory has its capability uniquely expressed both in terms of its ranges of measurements and the least measurement uncertainty (or best accuracy) applicable in each range. Because calibration laboratories are generally working to different levels of accuracy, it is not normally practicable to compare results on a group basis such as in interlaboratory *testing* programs. For calibration programs, we need to determine each individual laboratory's ability to achieve the level of accuracy for which they have nominated (their *least measurement uncertainties*).

The assigned (reference) values for a calibration program are not derived from a statistical analysis of the group's results. Instead they are provided by a Reference Laboratory which must have a higher accuracy than that of the participating laboratories. For PTA interlaboratory comparisons, the Reference Laboratory is usually Australia's National Measurement Institute (NMI), which maintains Australia's primary standards of measurement.

Another difference between calibration and testing programs is that there is usually only one test item (also known as an artefact) which has to be distributed sequentially around the participating laboratories, making these programs substantially longer to run. Consequently, great care has to be taken to ensure the measurement stability of the test item.

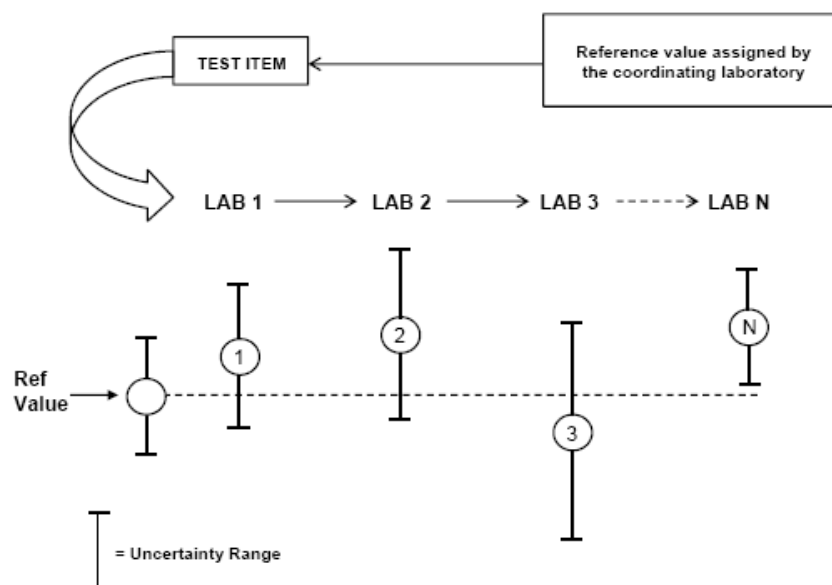


Figure 2: Typical Calibration Interlaboratory Comparison

In Figure 2, LAB 3 has a larger uncertainty range than LAB 1. This means that LAB 1 has the capability to calibrate higher accuracy instruments. This situation, where laboratories are working to different levels of accuracy, is valid provided that each laboratory works within their capabilities and that their nominated level of accuracy (measurement uncertainty) is suitable for the instrument being calibrated.

6.2 Program Design

Once a program has been selected, a small working group is formed. This group usually comprises one or more Technical Advisers and a PTA Scientific Officer who will act as the Program Coordinator. The group decides on the measurements to be conducted, how often the test item will need to be recalibrated and the range of values to be measured. They also formulate instructions and results sheets. PTA programs are designed so that it will normally take no more than eight hours for each participant to complete the measurements.

6.3 Test Item Selection

Because there can often be a substantial difference in the nominated measurement uncertainties of the participating laboratories, the test item must be carefully chosen. For example, it would be inappropriate to send a 3½ digit multimeter to a laboratory that had a nominated measurement uncertainty of 5 parts per million (0.0005%) because the resolution, repeatability and stability of such a test item would limit the measurement uncertainty the laboratory could report to no better than 0.05%. What is necessary is a test item with high resolution, good repeatability, good stability and an error that is large enough to be a meaningful test for all participants.

In some intercomparisons (especially international ones), the purpose may not only be to determine how well laboratories can measure specific points but also to highlight differences in methodology and interpretation.

6.4 Documentation

A *Letter of Intent* is sent to all potential participants to advise that the program will be conducted and to provide as much information as possible.

Instructions to Participants are carefully designed for each individual program and it is essential to the success of the program that the participating laboratories adhere closely to them. For most programs a pro-forma *Results Sheet* is used, to ensure that laboratories supply all the necessary information in a readily accessible format.

6.5 Test Item Stability

The test item is distributed sequentially around the participating laboratories. To ensure its stability, it is usually calibrated at least at the start and at the end of the circulation. For test items whose values may drift during the course of the program (e.g. resistors, electronic devices, etc.) more frequent calibrations and checks are necessary.

6.6 Evaluation of Performance

As stated in Section 6.1, calibration laboratories are generally working to different levels of accuracy. Consequently, their performance is *not* judged by comparing their results with those of the other laboratories in an interlaboratory comparison. Instead, their results are compared only to the Reference Laboratory's results and their ability to achieve the accuracy for which they have nominated is evaluated by calculating the E_n number. For further details please refer to Appendix C.

6.7 Reference Values

Australia's National Measurement Institute (NMI) provides most of the reference values for PTA's Calibration interlaboratory comparisons. The majority of the participating laboratories' reference equipment is also calibrated by NMI.

As stated previously, it is important to select test items with high resolution, good repeatability and good stability. This is to ensure that these factors do not contribute significantly to the reference value uncertainty. Likewise, the Reference Laboratory must have the capability to assign measurement uncertainties that are better than the participating laboratories. Otherwise it will be more difficult to evaluate each laboratory's performance.

Where a test item has exhibited drift, the reference values will usually be derived from the mean of the Reference Laboratory calibrations carried out before and after the measurements made by the participating laboratories. Where a step change is suspected, then the reference values will be derived from the most appropriate Reference Laboratory calibration.

6.8 Measurement Uncertainty (MU)

To be able to adequately compare laboratories they must report their uncertainties with the same confidence level. A confidence level of 95% is the most commonly used internationally. Laboratories should also use the same procedures to estimate their uncertainties as given in the ISO Guide⁴.

Laboratories should not report uncertainties smaller than their nominated measurement uncertainty.

6.9 Reporting

An individual summary sheet is sent to laboratories to give them feedback on their performance. The summary sheet states the E_n values for each measurement based on the preliminary reference values and usually does not contain any technical commentary.

A *Final Report* is issued on the PTA website (www.pta.asn.au) at the conclusion of the program. This typically contains more information than is provided in the summary sheet - including all participant's results and uncertainties, final E_n numbers, technical commentary and graphical displays.

6.10 Measurement Audits

The term *measurement audit* is used by PTA to describe a practical test whereby a well characterised and calibrated test item (or artefact) is sent to a single laboratory and the results are compared with a reference value (usually supplied by NMI).

Procedures are the same as for a normal interlaboratory comparison except that usually only a simple report is generated.

APPENDIX A

GLOSSARY OF TERMS

GLOSSARY OF TERMS

Further details about many of these terms may be found in either Appendix B (testing programs) or Appendix C (calibration programs). A number of these are also defined in ISO/IEC 17043¹.

assigned value	value attributed to a particular property of a proficiency test item
consensus value	an assigned value obtained from the results submitted by participants (e.g. for most testing programs the median [†] is used as the assigned value)
E_n number	stands for error normalised and is the internationally accepted quantitative measure of laboratory performance for calibration programs (see formula in Appendix C)
false negative	failing to report the presence of a parameter (e.g. analyte, organism) which is present in the sample
false positive	erroneously reporting the presence of a parameter (e.g. analyte, organism) which is absent from the sample
interlaboratory comparison	organisation, performance and evaluation of measurements or tests on the same or similar items by two or more laboratories in accordance with predetermined conditions
measurement uncertainty (MU)	non-negative parameter characterising the dispersion of the quantity values being attributed to a measurand, based on the information used
outlier	observation in a set of data that appears to be inconsistent with the remainder of that set, e.g. absolute z-score greater than or equal to three (i.e. 3.0) for testing programs
reference value	an assigned value which is provided by a Reference Laboratory
robust statistics	statistical method insensitive to small departures from underlying assumptions surrounding an underlying probabilistic model
z-score (Z)	a normalised value which assigns a “score” to the result(s), relative to the other numbers in the group - e.g. $(\text{result} - \text{median}^{\dagger}) \div \text{normalised IQR}^{\dagger}$

NOTE: [†] the median, normalised interquartile range (IQR) and other summary statistics are defined in Appendix B.

APPENDIX B

EVALUATION PROCEDURES FOR TESTING PROGRAMS

	<i>Page</i>
B.1 Introduction	13
B.2 Statistical Design	13
B.3 Data Preparation	14
B.4 Summary Statistics	15
B.5 Robust Z-scores and Outliers	17
B.6 Graphical Displays	18
B.7 Laboratory Summary Sheets	21

B.1 Introduction

This appendix outlines the procedures PTA uses to analyse the results of its proficiency testing programs. It is important to note that these procedures are applied only to *testing* programs, not *calibration* programs (which are covered in Appendix C). In testing programs the evaluation of results is based on comparison to assigned values which are usually obtained from all participants' results (i.e. consensus values).

The statistical procedures described in this appendix have been chosen so that they can be applied to a wide range of testing programs and, whenever practicable, programs are designed so that these 'standard' procedures can be used to analyse the results. In some cases, however, a program is run where the 'standard' statistical analyses cannot be applied - in these cases other, more appropriate, statistical procedures may be used.

For all programs the statistical analysis is only one part of the evaluation of the results. If a result is identified as an outlier, this means that statistically it is significantly different from the others in the group, however, from the point of view of the specific science involved (e.g. chemistry), there may be nothing "wrong" with this result. This is why the assessment of the results is always a combination of the statistical analysis and input by Technical Advisers (who are experts in the field). In most cases the Technical Adviser's assessment matches the statistical assessment.

B.2 Statistical Design

In order to assess the testing performance of laboratories in a program, a robust statistical approach, using z-scores, is used. Z-scores give a measure of how far a result is from the assigned value, and give a "score" to each result relative to the other results in the group. Section B.5 describes the method used by PTA for calculating z-scores.

For most testing programs, simple robust z-scores are calculated for each sample. Occasionally, the samples in a program may be paired and robust z-scores can be calculated for the sample pair. If paired samples are used they may be identical ("blind duplicates") or slightly different (i.e. the properties to be tested are at different levels). The pairs of results which are subsequently obtained fall into two categories: uniform pairs, where the results are expected to be the same (i.e. the samples are identical or the same sample has been tested twice); and split pairs, where the results should be slightly different. The pairing of samples allows the assessment of both between-laboratories and within-laboratory variation in a program.

One of the main statistical considerations made during the planning of a program is that the analysis used is based on the assumption that the results will be approximately normally distributed. This means that the results roughly follow a normal distribution, which is the most common type of statistical distribution (see Figure 3).

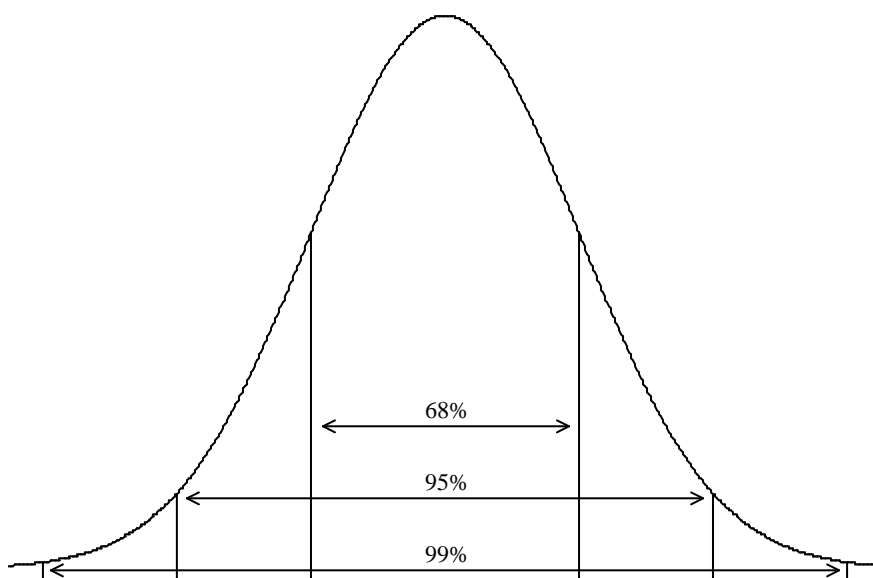


Figure 3: The Normal Distribution

The normal distribution is a “bell-shaped” curve, which is continuous and symmetric, and is defined such that about 68% of the values lie within one standard deviation of the mean, 95% are within two standard deviations and 99% are within three. To ensure that the results for a program will be approximately normal the working group (in particular the Technical Adviser) must think carefully about the results which might be obtained for the samples which are to be used.

For example, for the results to be continuous, careful consideration must be given to the units and number of decimal places requested - otherwise the data may contain a large number of repeated values. Another problem which should be avoided is when the properties to be tested are at very low levels - in this case the results are often not symmetric (i.e. skewed towards zero).

B.3 Data Preparation

Prior to commencing the statistical analysis, a number of steps are undertaken to ensure that the data collected is accurate and appropriate for analysis.

As the results are submitted to PTA, care is taken to ensure that all of the results are entered correctly. Once all of the results have been received (or the deadline for submission has passed), the entered results are carefully double-checked. It is during this checking phase that gross errors and potential problems with the data in general may be identified.

In some cases the results are then transformed - for example, for microbiological count data the statistical analysis is usually carried out on the \log_{10} of the results, rather than the raw counts. When all of the results have been entered and checked (and transformed if necessary) histograms of the data - which indicate the distribution of the results - are generated to check the assumption of normality.

These histograms are examined to see whether the results are continuous and symmetric. If this is not the case the statistical analysis may not be valid. One problem which may arise is that there are two distinct groups of results on the histogram (i.e. a bi-modal distribution). This is most commonly due to two test methods giving different results, and in this case it may be possible to separate the results for the two methods and then perform the statistical analysis on each group.

B.4 Summary Statistics

Once the data preparation is complete, summary statistics are calculated to describe the data. PTA uses eight summary statistics - number of results, median, uncertainty of the median, normalised interquartile range (IQR), robust coefficient of variation (CV), minimum, maximum and range. All of these are described in detail below.

The most important statistics used are the median and the normalised IQR - these are measures of the centre and spread of the data (respectively), similar to the mean and standard deviation. The median and normalised IQR are used because they are robust statistics, which means that they are not influenced by the presence of outliers in the data.

The no. of results is simply the total number of results received for a particular test/sample, and is denoted by N. Most of the other statistics are calculated from the sorted results, i.e. from lowest to highest, and in this appendix X[i] will be used to denote the ith sorted data value (e.g. X[1] is the lowest value and X[N] is the highest).

The median is the middle value of the group, i.e. half of the results are higher than it and half are lower. If N is an odd number the median is the single central value, i.e. X[(N+1)/2]. If N is even, the median is the average of the two central values, i.e. (X[N/2] + X[(N/2)+1])/2. For example if N is 9 the median is the 5th sorted value and if N is 10 the median is the average of the 5th and 6th values.

The normalised IQR is a measure of the variability of the results. It is equal to the interquartile range (IQR) multiplied by a correction factor[†], which makes it comparable to a standard deviation. The interquartile range is the difference between the lower and upper quartiles. The lower quartile (Q1) is the value below which, as near as possible, a quarter of the results lie. Similarly the upper quartile (Q3) is the value above which a quarter of the results lie. In most cases Q1 and Q3 are obtained by interpolating between the data values. The IQR = Q3 – Q1 and the normalised IQR = IQR × correction factor.

Since the median is a consensus value, it has an uncertainty originating from the testing conditions of the laboratories that participated in the program and other factors. The (standard) uncertainty of the median is calculated as:

$$\text{uncertainty}(\text{median}) \approx \sqrt{\frac{\pi}{2}} \times \frac{\text{normalised IQR}}{\sqrt{N}}$$

where N = no. of results.

The robust CV is a coefficient of variation (which allows for the variability in different samples/tests to be compared) and is equal to the normalised IQR divided by the median, expressed as a percentage - i.e. robust CV = 100 × normalised IQR ÷ median.

The minimum is the lowest value (i.e. X[1]), the maximum is the highest value (X[N]) and the range is the difference between them (X[N]–X[1]).

On page 17 is an example of the summary statistics as they appear in a final report.

NOTE: [†] The interquartile range of normally distributed data is not equivalent to the familiar ±1 SD interval. To convert an IQR into a ±1 SD range, it must be scaled by a correction factor. The correction factor is calculated by using expected normal scores of order statistics and depends on the number of results reported for the test/sample.

Example: Data Set and Summary Statistics

Waters (Chemical) Results for PTA Sample 1 - Total Solids, Total Suspended Solids and Total Dissolved Solids

Lab Code	PTA Sample 1						Total Solids Robust Z-Scores	Total Suspended Solids Robust Z-Scores	Total Dissolved Solids Robust Z-Scores
	Total Solids		Total Suspended Solids		Total Dissolved Solids				
	Result ± MU mg/L		Result ± MU mg/L		Result ± MU mg/L				
1	584	25	200	6	389	25	-0.91	-0.51	-1.64
2	600	60	204	#	405	40	-0.29	-0.26	-0.10
3	572	15	195	20	406	20	-1.37	-0.84	0.00
4	624	#	216	#	431	#	0.64	0.51	2.41
5	575	#	192	10	444	#	-1.25	-1.03	3.66 §
6	631	113	209	#	410	#	0.91	0.06	0.39
7	640	64	176	#	351	8.6	1.25	-2.06	-5.30 §
8	600	1	180	#	360	36	-0.29	-1.80	-4.43 §
9	581	58.1	185	7.6	410	41	-1.02	-1.48	0.39
10	592	#	190	34	432	1	-0.60	-1.16	2.51
11	567.5	#	230	23	395	39.5	-1.54	1.41	-1.06
12	621	13	222	1	410	#	0.52	0.90	0.39
13	602	#	181	18.1	370	#	-0.21	-1.73	-3.47 §
14	625	63	182	#	426	10	0.67	-1.67	1.93
15	620	8.37	195	#	368	#	0.48	-0.84	-3.66 §
16	611	16.74	223	1.7	413	41	0.13	0.96	0.67
17	586	#	226	#	407	7.53	-0.83	1.16	0.10
18	627	30	201	20	402	4.21	0.75	-0.45	-0.39
19	619	40	213	10.58	396	#	0.44	0.32	-0.96
20	700	#	214	5.79	408	20	3.57 §	0.39	0.19
21	600	6.28	178	#	398	60	-0.29	-1.93	-0.77
22	624	64.90	207	15	409	6.13	0.64	-0.06	0.29
23	588	#	209	15	406	28.42	-0.75	0.06	0.00
24	619	31.7	211	21	405	#	0.44	0.19	-0.10
25	634	15	203	3.02	410	20.5	1.02	-0.32	0.39
26	624	10	218	27.47	390	59	0.64	0.64	-1.54
27	604	72	226	32.3	396	47.5	-0.13	1.16	-0.96
28	578	58	182	#	411	15	-1.14	-1.67	0.48
29	601	60	213	6	404	8.8%	-0.25	0.32	-0.19
30	<500	40	216	15.1	419	10	0.98	0.51	1.25

NOTES: § denotes an outlier, i.e. |z-score|≥3.0.
 # indicates that no results were submitted.
 "N/A" indicates not applicable.

**TOTAL SOLIDS, TOTAL SUSPENDED SOLIDS AND TOTAL DISSOLVED SOLIDS -
SUMMARY STATISTICS (mg/L)**

Statistic	Total Solids	Total Suspended Solids	Total Dissolved Solids
No. of Results	30	30	30
Median	607.5	205.5	406.0
Normalised IQR	25.9	18.5	10.4
Uncertainty (Median)	5.9	4.2	2.4
Robust CV	4.3%	9.0%	2.6%
Minimum	567.5	176	351
Maximum	700	230	444
Range	132.5	54	93

B.5 Robust Z-scores and Outliers

To statistically evaluate the participants' results, PTA uses z-scores based on robust summary statistics (the median and normalised IQR).

If a sample in a testing program is labelled A, then the robust z-score (denoted by Z) for a laboratory's sample A result would be:

$$Z = \frac{A - \text{median}(A)}{\text{normIQR}(A)}$$

where the median and normalised IQR of all the sample A results are denoted by median(A) and normIQR(A), respectively.

The calculated z-scores are tabulated in the report for a program, alongside the corresponding results and the results are assessed based on their z-scores. The interpretation of z-scores is as below:

$ Z \leq 2.0$	indicates a "satisfactory" performance
$2.0 < Z < 3.0$	indicates a "questionable" performance
$ Z \geq 3.0$	indicates an "unsatisfactory" performance

where $|Z|$ denotes the absolute value of the z-score.

An outlier is defined as any result with an absolute z-score greater than or equal to three, i.e. $Z \geq 3.0$ or $Z \leq -3.0$. Outliers are identified in the tabulated results in a report by a marker (§) beside the z-score. When an outlier is identified the sign of the z-score indicates whether the result is too high (positive z-score) or too low (negative z-score). Laboratories that obtain outliers or questionable results in a program are encouraged to review their results.

In the example on page 16, laboratory 5 has a positive outlier for Total Dissolved Solids and laboratory 20 has a positive outlier for Total Solids. Laboratories 7, 8, 13 and 15 have negative outliers for Total Dissolved Solids.

In some circumstances it may not be possible to calculate a robust z-score using the formula above. This occurs when the normalised IQR is equal to zero (which could occur if more than 50% of the results submitted by participants were identical and equal to the median). In other circumstances it may be possible to calculate a robust z-score using the formula above, but the spread of results (as measured by the normalised IQR) might be so small that even a slight deviation from the median will result in an outlier. In yet other circumstances the spread of results (as measured by the normalised IQR) might be so large that it is extremely unlikely that any result would ever be classified as an outlier.

If the normalised IQR is equal to zero, or if the spread of results is too large or too small, in the opinion of the Technical Adviser, then a target coefficient of variation (CV) is used to calculate z-scores. These z-scores are calculated by:

$$Z = \frac{A - \text{median}(A)}{\text{target CV} \times \text{median}(A)}$$

where the target CV is expressed as a decimal.

The actual value used as the target CV to calculate such z-scores is chosen in consultation with the Technical Adviser and usually takes into account historical data (most likely obtained from previous rounds of the program, or similar interlaboratory testing programs).

When pairs of results have been obtained in a program, two z-scores may be calculated - a between-laboratories z-score and a within-laboratory z-score. These are based on the sum and difference of the pair of results, respectively.

Suppose the pair of results are from two samples labelled A and B. The standardised sum (denoted by S) and standardised difference (D) for the pair of results are:

$$S = (A + B) / \sqrt{2} \text{ and } D = \begin{cases} (B - A) / \sqrt{2}, & \text{if } \text{median}(A) < \text{median}(B) \\ (A - B) / \sqrt{2}, & \text{otherwise.} \end{cases}$$

Each laboratory's standardised sum and difference are calculated, followed by the median and normalised IQR of all the S's and all the D's - i.e. median(S), normIQR(D), etc.

The between-laboratories z-score (denoted by ZB) is then calculated as the robust z-score for S and the within-laboratory z-score (ZW) is the robust z-score for D, i.e.

$$ZB = \frac{S - \text{median}(S)}{\text{normIQR}(S)} \quad \text{and} \quad ZW = \frac{D - \text{median}(D)}{\text{normIQR}(D)}$$

B.6 Graphical Displays

In addition to tables of the results and z-scores, and summary statistics, a number of graphical displays of the data are included in the report for a program. The two most commonly used graphs are the ordered z-score bar-chart and the Youden diagram - both of which are described in detail below.

These charts are to assist the Program Coordinator and Technical Advisers with the interpretation of the results and are very useful to participants - especially those participants with outliers because they can see how their results differ from those submitted by other laboratories.

Ordered Z-score Chart

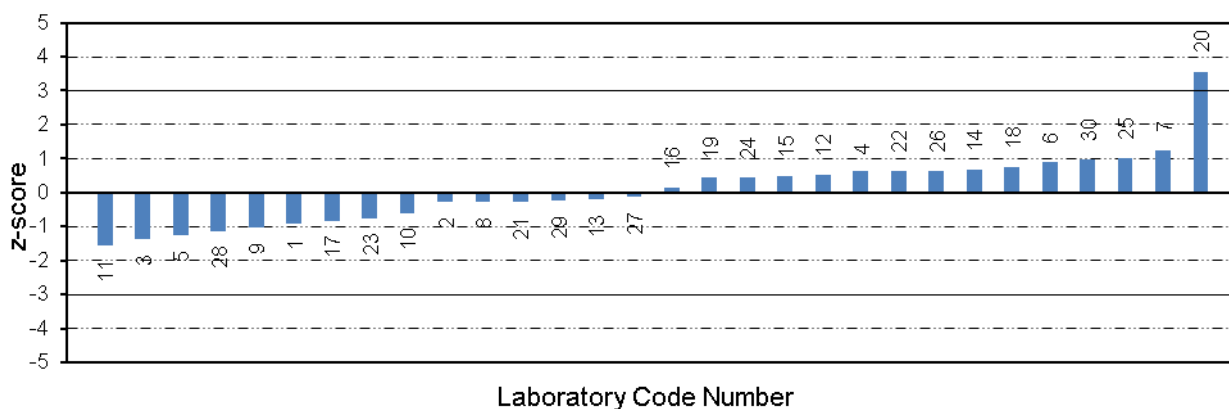
An ordered z-score chart is generated for the z-scores calculated for each test. An example is included below. On these charts each laboratory's z-score is shown, in order of magnitude, and is marked with its code number. From this each laboratory can readily compare its performance relative to the other laboratories.

These charts contain solid lines at +3.0 and -3.0, so the outliers are clearly identifiable as the laboratories whose "bar" extends beyond these cut-off lines. The y-axis is usually limited, so in some cases very large or small (negative) z-scores appear as extending beyond the limit of the chart - for example, laboratory 7 for the Total Dissolved Solids z-score bar-chart on page 20.

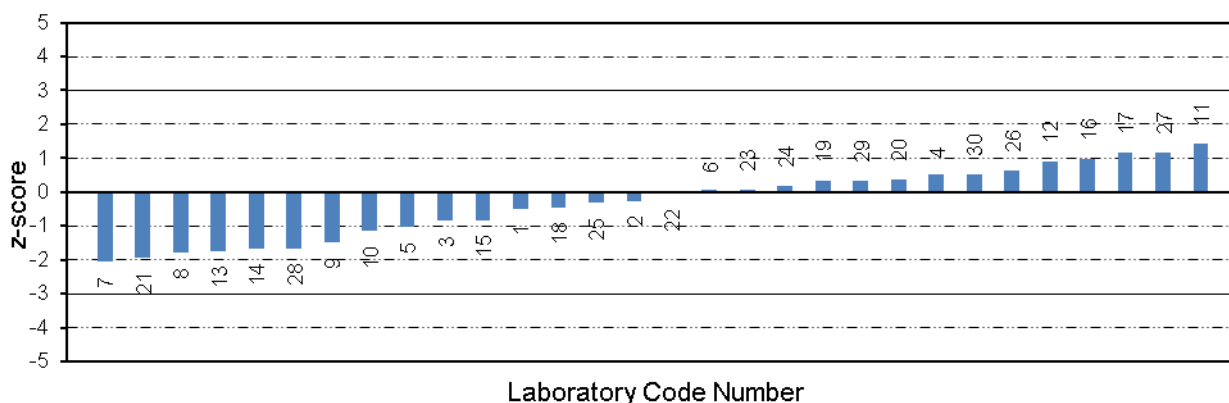
The advantages of these charts are that each laboratory is identified and the outliers are clearly indicated, however, unlike the Youden diagrams, they are not graphs of the actual results.

Examples: Ordered Z-Score Charts

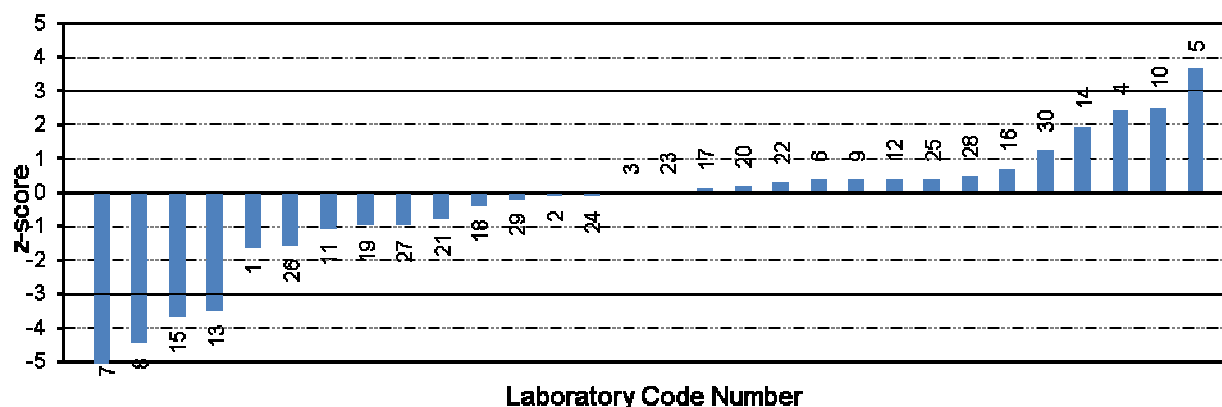
Total Solids - Sample PTA 1 - Robust Z-Scores



Total Suspended Solids - Sample PTA 1 - Robust Z-Scores



Total Dissolved Solids - Sample PTA 1 - Robust Z-Scores



Youden Diagrams

These charts are generated for pairs of results. Youden diagrams are produced for biological program reports where results have been log transformed, for duplicate samples, and for duplicate results requested from the same sample. Youden two-sample diagrams are presented to highlight laboratory systematic differences. They are based on a plot of each laboratory's pair of results, represented by a black spot •.

These diagrams also feature an approximate 95% confidence ellipse for the bivariate analysis of the results, and dashed lines which mark the median value for each of the samples. The ellipse is estimated by re-scaling an approximate 95% confidence region (which is a circle) in the bivariate z-scores space back to the original data space.

All points which lie outside the ellipse are labelled with the corresponding laboratory's code number. Note, however, that these points may not correspond with those identified as outliers. This is because the outlier criterion ($|Z| \geq 3.0$) has a confidence level of approximately 99%, whereas the ellipse is an approximate 95% confidence region.

This means that, if there are no outliers in the data, it can be expected that about 5% (i.e. one in twenty) of the results will lie outside the ellipse, however, as proficiency testing data usually contains some outliers, more than 5% of points will be outside the ellipse in most cases. The points outside the ellipse on the Youden diagram will roughly correspond to those with absolute z-scores greater than 2.0. Laboratories with results outside the ellipse which have not been identified as outliers (those which have $2.0 < |Z| < 3.0$) are encouraged to review their results.

An example of a Youden diagram is included below. All of the laboratories with outliers, i.e. $|Z| \geq 3.0$, and those with $2.0 < |Z| < 3.0$ lie outside the ellipse.

The advantages of these diagrams are that they are plots of the actual data - so the laboratories with results outside the ellipse can see *how* their results differ from the others - and results with an absolute z-score greater than 2.0 are highlighted.

As a guide to the interpretation of the Youden diagrams:

- (i) laboratories with significant systematic error components (i.e. between-laboratories variation) will be outside the ellipse in either the upper right hand quadrant (as formed by the median lines) or the lower left hand quadrant, i.e. inordinately high or low results for *both* samples;

and

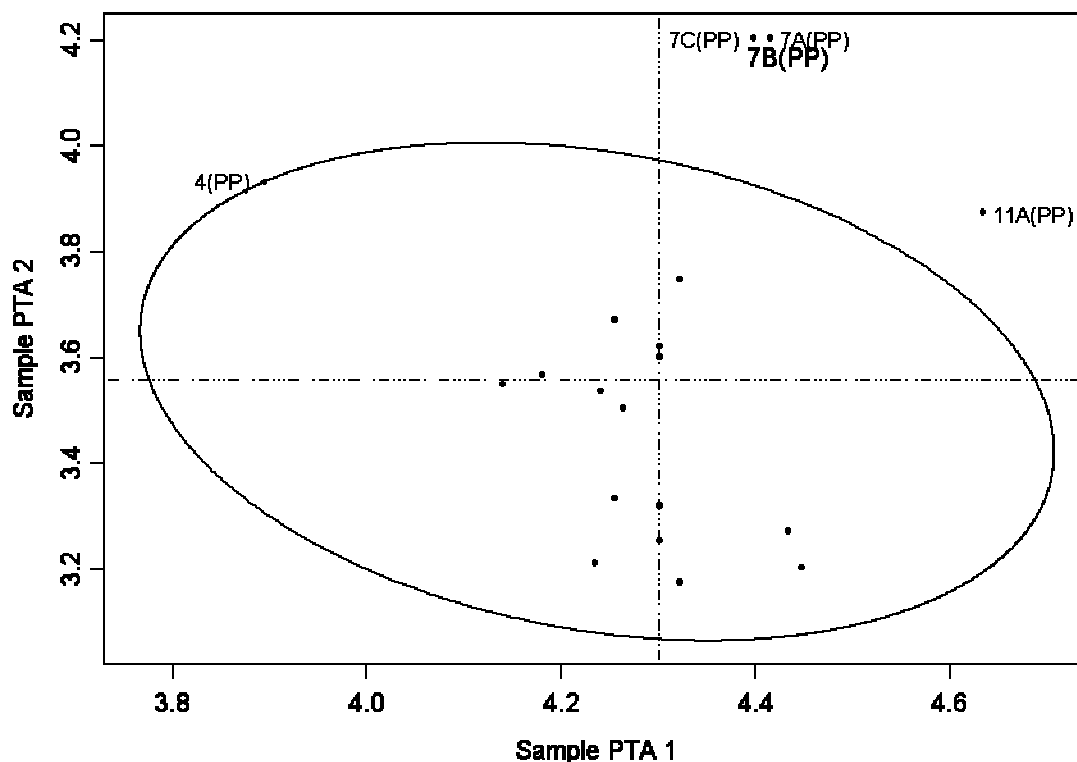
- (ii) laboratories with random error components (i.e. within-laboratory variation) significantly greater than other participants will be outside the ellipse and (usually) in either the upper

left or lower right quadrants, i.e. an inordinately high result for one sample and low for the other.

It is important to note, however, that Youden diagrams are an illustration of the data only, and are *not* used to assess the results (this is done by the z-scores).

Example: Youden diagram

Aerobic Plate Count, All Methods Pooled - log(cfu/g)



B.7 Laboratory Summary Sheets

In addition to the final report, which contains complete details of the statistical analysis, an individual summary sheet is prepared for each participant. This laboratory summary sheet contains all of the participant's results, alongside the statistics for that test/sample and the associated z-scores. Comments about the program in general and specific to the laboratory (if necessary) are also included.

An example summary sheet appears on page 23. At the top of the page is the title of the program and the identity of the laboratory. The main part of this summary sheet consists of: the test and sample identity; the laboratory's result including its MU (where required); the number of results; median and normalised IQR for each test/sample; and the z-scores (or two z-scores for a sample pair) for each test.

Any outliers are again marked with a § next to the z-score. At the bottom of the page is a section for notes and comments. In this case there are no special laboratory-specific remarks. From this summary sheet we can see quickly and easily that:

- (1) this laboratory submitted results for all of the tests;
- (2) the laboratory has reported one outlier; and
- (3) the laboratory has reported one questionable result.

Seeing all of a laboratory's z-scores together can be very useful, even if no outliers were reported. For example, where a pair of samples is tested, if all of the between-laboratories z-scores are negative (or positive) this may be indicative of a laboratory bias - i.e. all of its results are lower (or higher) than the consensus values.



Proficiency Testing Australia

LABORATORY SUMMARY SHEET

Proficiency Testing - Waters (Chemical) Round [###]
- Total Solids, Total Suspended Solids, Total Dissolved Solids -

Report No. [###]

Date of summary sheet issue: [Date]

Lab Name: [name of Laboratory/company, including Site name]

Laboratory Code: [##]

Location: [state/country]

Analyte	Sample	Laboratory result \pm MU (mg/L) ¹	Median ²	Norm. IQR ³	Robust CV ⁴	No. of results	Robust z-score ⁵
Total Solids (TS)	PTA 1	640 \pm 64	607.5	25.9	4.3%	30	1.25
Total Suspended Solids (TSS)	PTA 1	176 \pm *	205.5	18.5	9.0%	30	-2.06 ?
Total Dissolved Solids (TDS)	PTA 1	351 \pm 9	406.0	10.4	2.6%	30	-5.30 §

No. of outlier results is: 1

¹ A "." indicates that no result was returned for this sample/test.

² The *median* is the middle result. It is a measure of the centre of the data set.

³ The *normalised IQR* is a measure of the spread of the results. It is calculated by multiplying the *interquartile range (IQR)* by a factor which converts the *IQR* to an estimate of the *standard deviation*. The *IQR* is the difference between the *upper* and *lower quartiles* (i.e. the values above and below which a quarter of the results lie, respectively).

⁴ The *robust coefficient of variation (robust CV)* is calculated by dividing the *normalised IQR* by the *median* and expressed as a percentage. The *robust CV* allows for the variability in different samples/tests to be compared.

⁵ Each z-score marked with a "§" is an *outlier* (i.e. $|z\text{-score}| \geq 3.0$). Laboratories are also encouraged to review results which have an absolute z-score value between two and three (i.e. $2.0 < |z\text{-score}| < 3.0$), these have been marked with a "?".

⁶ For the purposes of consistency in reporting, summary sheet results and MU values have been rounded to zero decimal places for all analytes.

This summary sheet should be read in conjunction with the final report found at www.pta.asn.au. The above results are from one proficiency program only and may not be fully representative of a laboratory's overall performance. Therefore, this summary sheet should not be used solely to evaluate laboratory competence.

APPENDIX C

EVALUATION PROCEDURES FOR CALIBRATION PROGRAMS

	<i>Page</i>
C.1 Introduction	25
C.2 Calibration Programs	25
C.3 Graphical Displays for Calibration Programs	26
C.4 Measurement Audit Programs	26
C.5 Measurement Uncertainty (MU)	27

C.1 Introduction

This appendix outlines the procedures PTA uses to evaluate the results of its *calibration* programs and *measurement audit programs* (refer to Appendix B for procedures applicable to *testing* programs). The procedures used by PTA are consistent with those used for international calibration programs run by the European Cooperation for Accreditation (EA) and Asia Pacific Laboratory Accreditation Cooperation (APLAC).

C.2 Calibration Program

As stated in Section 6.6, PTA uses the E_n number to evaluate each individual result from a laboratory. E_n stands for **Error normalised** and is defined as:-

$$E_n = \frac{LAB - REF}{\sqrt{U^2_{LAB} + U^2_{REF}}}$$

where: *LAB* is the participating laboratory's result
REF is the Reference Laboratory's result
 U_{LAB} is the participating laboratory's reported uncertainty
 U_{REF} is the Reference Laboratory's reported uncertainty

For a result to be acceptable the E_n number should be between -1.0 and +1.0 i.e. $|E_n| \leq 1.0$. (The closer to zero the better.)

In *testing* interlaboratory comparisons a laboratory's z-score gives an indication of how close the laboratory's measurement is to the assigned value, however, in *calibration* interlaboratory comparisons the E_n numbers indicate whether laboratories are within their particular measurement uncertainty of the reference value (assigned value).

The E_n numbers do not necessarily indicate which laboratory's result is closest to the reference value. Consequently, calibration laboratories reporting small uncertainties may have a similar E_n number to laboratories working to a much lower level of accuracy (i.e. larger uncertainties).

In a series of similar measurements a normal distribution of E_n numbers would be expected. So when considering the significance of any results with $|E_n|$ marginally greater than 1.0, all the results from that laboratory are evaluated to see if there is a systematic bias e.g. consistently positive or consistently negative values of E_n .

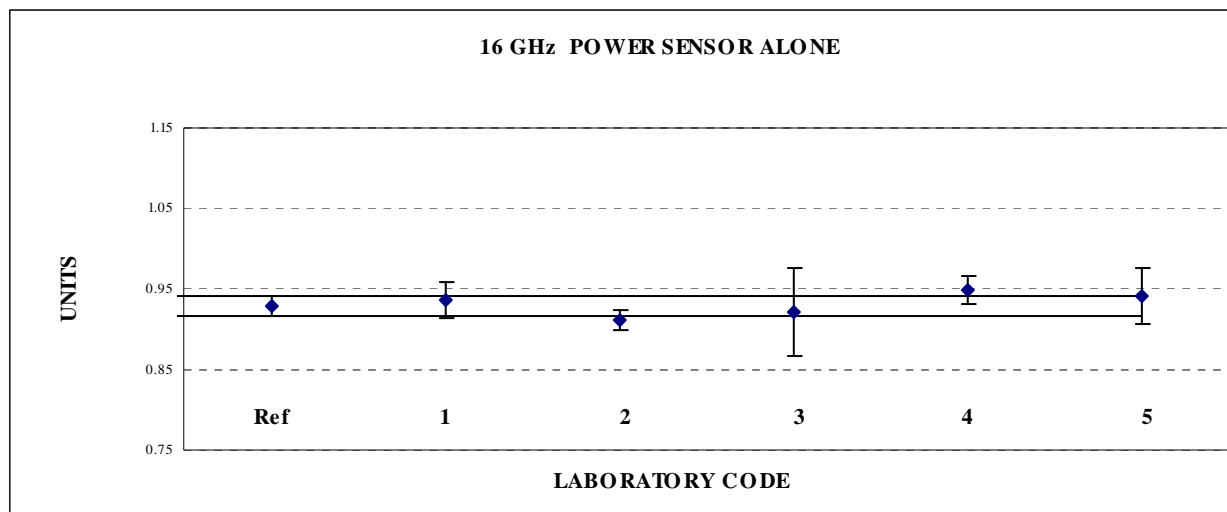
A sample of results from a radio frequency power interlaboratory comparison, their corresponding reported uncertainties and E_n numbers are tabulated below. The result for laboratory 2 is considered unsatisfactory.

16 GHz Power Sensor Alone

Lab Code	Results	U_{95}	E_n
REF	0.929	0.011	
1	0.936	0.022	0.28
2	0.911	0.012	-1.09
3	0.921	0.054	-0.14
4	0.949	0.018	0.94
5	0.942	0.035	0.35

C.3 Graphical Displays for Calibration Program

Graphs of reported results and their associated uncertainties are included in final reports for *calibration* programs. The example graph below shows a plot of the results tabulated in Section C.2. Each laboratory's result is represented by a ♦ mark. The bars protruding above and below the ♦ mark represent that laboratory's reported measurement uncertainty, that is, the region in which the laboratory has statistically calculated (with a 95% confidence level) that the "true value" may lie, or in other words, their estimate of how accurately they can measure.



It is important to note however that the graphs are an illustration of the data only and allow a broad comparison of all participants' results/uncertainties. They do not represent an assessment of results (this is done by the E_n numbers).

C.4 Measurement Audit Programs

A sample of results from a pressure transducer *measurement audit*, the laboratory's corresponding reported uncertainties and E_n numbers are tabulated below. The results for decreasing applied pressures at 9.9999 MPa, 7.5000 MPa and 5.0000 MPa are considered unsatisfactory.

10 MPa Pressure Transducer

APPLIED PRESSURE	REF VALUE MPa	REF U_{95} MPa	LAB MEAN MPa	LAB U_{95} MPa	E_n NO.
5.0000	4.8983	0.0014	4.8982	0.002	-0.03
7.5000	7.3478	0.0014	7.3466	0.002	-0.46
9.9999	9.7973	0.0019	9.7970	0.004	-0.08
9.9999	9.8133	0.0025	9.7972	0.004	-3.72
7.5000	7.3605	0.0031	7.3462	0.002	-3.88
5.0000	4.9074	0.0025	4.8971	0.002	-3.51

Graphs of reported results and their associated uncertainties are provided for *measurement audit* programs when necessary.

C.5 Measurement Uncertainty (MU)

The measurement uncertainty reported by the laboratory is used in the E_n number. The test items used in these programs usually have sufficient resolution, repeatability and stability to allow the laboratory to report an uncertainty equal to their claimed "*best measurement capability*".

End of Document