EVALUATING MEASUREMENT UNCERTAINTY IN AMINO ACID RACEMIZATION ANALYSIS: TOWARDS A NEW CHRONOLOGY

Joanne Powell

PhD

2012

Evaluating Measurement Uncertainty in Amino Acid Racemization Analysis: Towards a New Chronology

Joanne Powell

PhD

University of York

Department of Archaeology

October 2012

Abstract

Unlike other Quaternary dating methods, amino acid racemisation (AAR) geochronology has the potential to provide age estimates that span the entire Quaternary period, a crucial period for understanding past climate change and human evolution. It has become a critical technique for Quaternary Science and uses the time/temperature dependent kinetics of protein decomposition to provide relative age estimates of fossil samples. The accuracy of age estimates relies heavily on the accuracy of analytical data and accurate determinations of uncertainty estimates.

This thesis takes internationally established principles of measurement uncertainty determination and applies them to AAR. Analytical uncertainty is considered in the context of intra- and inter-laboratory measurement results. A retrospective evaluation of intra-laboratory precision using ANOVA is given, and results from an inter-laboratory proficiency study, evaluated as estimates of bias, are summarised (paper submitted). The final sections look at uncertainty from existing archaeological site data, including sampling effects. A model is proposed that utilises decomposition correlations between amino acids to provide *a priori* uncertainty estimates. These are then used to update observed site data using a Bayesian approach to derive posterior uncertainty estimates and D/L values. A further model is tentatively presented which could potentially be used to derive quantitative age estimates once uncertainty within the kinetic and temperature models have been characterised and accounted for.

Table of Contents

List of Figures

List of Tables

List of Appendices provided on CD

CHAPTER 4

Chpt 4: Appendix 1_scatter charts.docx

Chpt 4: Appendix 2_replicates.docx

Chpt 4: Appendix 3_box plots.docx

Chpt 4: Appendix 4_histograms.docx

Chpt 4: Appendix 5_box plots minus outliers.docx

Chpt 4: Appendix 6_histograms minus outliers.docx

Chpt 4: Appendix 7_kernal densities.docx

Chpt 4: Appendix 8_confidence intervals.docx

CHAPTER 5

Chpt 5: Appendix 1_draft paper Part 2.pdf

Chpt 5: Appendix 2_Report PT Mollusc A bias(NPC).pdf

Chpt 5: Appendix 3_ Report PT Mollusc B bias(NPC).pdf

Chpt 5: Appendix 4_ Report PT OES A bias(NPC).pdf

Chpt 5: Appendix 5_ Report PT OES B bias(NPC).pdf

Chpt 5: Appendix 6_ Report PT opercula bias(NPC).pdf

Chpt 5: Appendix 7_ Report PT Std sol bias(NPC).pdf

CHAPTER 6

Chpt 6: Appendix 1_ANOVA UK Quaternary sequence.xlsx

Chpt 6: Appendix 2_amino acid correlations

Chpt 6: Appendix 3_Thames sequence summary.xlsx

Chpt 6: Appendix 4_DL uncertainty spreadsheet tool.xlsx

CHAPTER 7

Chpt 7: Appendix 1_DL calibration curve v5_using RFOK.xlsx

Chpt 7: Appendix 2_DL calibration curve v6_DL aligned with RFOK.xlsx

Acknowledgements

This thesis is completed with the support of my supervisors Matthew Collins, Kirsty Penkman, James Cussens and Norman MacCleod, who have lead me along paths of uncharted territory. Particular thanks go to Matthew and Kirsty for their time, availability and constant willingness to help. Also for their tolerance of my evasion of meetings or frequent late arrivals, especially to Kirsty, to whom I owe a mountain of cakes.

I would also like to thank all the other AARers who assisted with the proficiency test including; John Wehmiller, Darrell Kaufman, Katherine Sides, Jordon Bright, Rick Oches, José Ortiz, Colin Murray-Wallace and Terry Lachlan. Also to Richard Bintanja at Utrecht University, The Netherlands, for providing reconstructed palaeoclimate temperature data.

Thanks to Matthew Collins and Ken Mathieson (FERA) for spreadsheet assistance and much gratitude is owed to Bea Demarchi and Richard Allen for analytical technical support and training. But particularly for their patience, friendship and companionship, together with other weary travellers who have accompanied me on this journey, including Cynthianne, Bella, Nienke and Molly.

A special acknowledgement goes to my two sisters, Lyn and Claire who have constantly reminded me of my conviction to undertake this research, owing to their utter bemusement and frequent challenging of my motivation. Also with thanks and in memory of my Mum and Dad, who I have no doubt will have been watching and cheering me on from the other side.

Last but by no means least, I would like to acknowledge, the love and support of my two long-suffering boys, Matty and Nicki. For putting up with their Mum when she was tired and grumpy, stressed and confused and always having a hug and a smile for me at the end of the day. For being my purpose, my strength, my hope and my light at the end of the tunnel, this was for you.

Declaration

I hereby certify that the work described in this thesis is my own, except where otherwise acknowledged, and has not been submitted previously for a degree at this or any other university.

Joanne Powell

"*The Goddess Uncertainty was born, like Athene, from the brow of her parthenogenetic parent, the supreme god Iso*[1]*. The pregnancy was not an easy one. There were conflicting pressures from the muses of physical metrology that affected the development of the embryo and gave Iso a headache. However, in 1993, after several years gestation, the new deity Uncertainty was finally born, fully armed, and intent on helping the mortals in a largely unsuspecting analytical community. News of the event was carried to the Britons by a local deity called Namas*[2]*, and the Headache was passed onto analytical chemists*."

(Thompson, 1995 p 117N)

[1]ISO; is the International Standards Organisation, author of Evaluation of Measurement data – Guide to the expression of uncertainty in measurement, known as the GUM (JCGM 100, 2008).

[2]Namas; National Measurement Accreditation Service, later renamed UKAS; United Kingdom Accreditation Service.

Chapter 1. Introduction

This thesis concerns the evaluation of measurement uncertainty in amino acid racemisation (AAR), and its potential use as a quantitative geochronological dating technique. The following chapters explore the determination of uncertainty estimation from three different perspectives; intra-laboratory, inter-laboratory and site based. However, before a more detailed look at uncertainty estimation is undertaken, it is important to first set the research within its appropriate context. The current chapter therefore is aimed at providing an over-view of the Quaternary, the time period most relevant to the research due to the frequent climate oscillations and their impact on the temperature record. Definitions and subdivisions are first considered, followed by a look at the importance of the marine cores in providing a global reference chronology. The problems in correlating the fragmented terrestrial record are discussed with emphasis on the need for independent dating methods. An overview of AAR is given with a look at its current use as a relative dating technique and considers the potential for quantitative AAR. The chapter ends with a summary of the aims and objectives of the research, the thesis structure and some useful terminology.

1.1 Quaternary Geochronology

It is currently believed that the earliest hominin genus *Australopithecus* emerged out of Africa 4.5 million years ago, with fossil evidence of our own genus *Homo*, appearing 2.3 million years ago from sites in Kenya and Tanzania (Renfrew and Bahn, 2012). In Northern Europe, the earliest evidence of human occupation can be traced back to the British Pakefield site in Suffolk, and dated to about 700 kyr based on event stratigraphy, lithostratigraphy, palaeomagnetism, amino acid geochronology and biostratigraphy (Parfitt *et al.*, 2005). Thus the last two and a half million years of geological time, that spans the Quaternary, has been a critical period in which *Homo* developed and migrated out of Africa. Today, the Quaternary is known for its oscillating glacial/interglacial cycles, extinction of the megafaunal species and human evolution and migration. Detailed knowledge of these

climate changes are therefore crucial to our interpretation of the archaeological record and early man's response to environmental change. For the more recent archaeological sites, material evidence such as the remains of built structures, landscape and site features (e.g. post holes and hearths) and excavated archaeological deposits can be sequenced. Common styles can be identified in recovered artefacts (e.g. metal jewellery and weaponry, pottery and stone tools) and cultural sequence chronologies, based on typologies, derived. However, the further back in time we go, less and less material evidence is recoverable and the archaeological archive merges with the geological one. Therefore, in order to understand the archaeological record, we need to understand the geological record too.

1.1.1 Defining the Quaternary

The Quaternary was first used to describe sediments and evidence of exotic boulders and extinct animals that lay on top of Tertiary rocks by the Italian geologist Arduino in 1759. Later it was formally used by Desnoyers in 1829 when describing sediments in the Seine Valley, and predates the use of the term Pleistocene by Lyell in 1839 (Gibbard and Kolfschoten, 2006; Gibbard and Head, 2010). The use of the term Quaternary in more recent times, has however been contentious. The need to standardise a formal stratigraphical boundary stratotype for the Pleistocene and Quaternary was recognised in 1948, but it was not until 1982 that a Global Stratotype Section and Point (GSSP) was proposed and finally ratified in 1985 by the International Union of Geological Sciences (IUGS) (Bassett, 1985). This was set at the Vrica section in Calabria in southern Italy and dated to 1.64 Ma (MIS[1.1] 63) (Aguirre and Pasini, 1985; Gibbard *et al.*, 2009) but subsequently revised to 1.806 Ma by astronomical tuning (Lourens *et al.*, 2005). However there was a strong feeling that the boundary should be placed earlier to reflect mounting evidence of significant cooling occurring between 2.8 – 2.4 Ma depending on region (Versteegh, 1997; Monegatti and Raffi, 2001; Roveri and Taviani, 2003), and that some of the cold climate faunal indicators such as the ostracod *Cytheropteron testudo* and the bivalve *Artica Islandica,* had appeared before 1.8 Ma (Arias *et al.*, 1980; Aiello *et al.*, 1996; Gibbard *et al.*, 2009; Gibbard and Head, 2010). Consequently, in 2009, a revised scheme was presented and ratified in June 2009 which redefines the base of both the Quaternary System/Period and Pleistocene Series/Epoch to bring them in line with the Gelasian Stage GSSP at Monte San Nicola, Sicily in Italy (Rio *et al.*,

[1.1] MIS = Marine Isotope Stage; a numbering system derived from deep sea sediment cores and based on changing oxygen isotope ratios in marine microfossils. Fluctuations in the isotopic signal is believed to reflect changes in the land ice volume and correlates with warm / cold climate oscillations observed through the Quaternary. (See section 1.2).

1998), and dated to 2.588 Ma (MIS 103) (Gibbard *et al.*, 2009; Gibbard and Head, 2010), which lies just 1 m above the Gauss-Matuyama palaeomagnetic reversal (Rio *et al.*, 1998; Lourens, 2008).

> "*A base-Quaternary boundary at 2.6 Ma will strengthen recognition within terrestrial as well as marine sections owing to major global changes in the terrestrial biota, including humans, and in sedimentation particularly with respect to loess deposition across northern Eurasia. Such major global changes are lacking around 1.8 Ma."* (Gibbard and Head, 2010, p155).

The top of the Neogene is now defined by the base of the Quaternary and the top of the Pliocene, by the base of the Pleistocene (Gibbard and Head, 2009b, 2009a; Finney, 2010; Gibbard and Head, 2010), The Quaternary now comprises both the Pleistocene and Holocene (defined by reference to the Greenland ice core NGRIP GSSP and dated to 11.7 ka), (Walker *et al*., 2009).

1.1.2 Subdivisions of the Quaternary

In 1829, Ignaz Venetz-Sitten first recognised the signs of ice erosion in regions beyond the Swiss Alps, but it was Cuvier's former student, Louis Agassiz, who in 1840, first attributed the diluvium sediments and sculpted Scottish U shaped valleys to glacial activity. Later, Agassiz' single glacial episode was replaced in 1854 by Merlot's two glacial stages, separated by a warmer diluvial stage (Stringer, 2006; Elias, 2007; Renfrew and Bahn, 2012). However, gradually it became recognised that there may in fact have been more than two cold phases. In 1874 James Geike suggested there had been a series of alternating glacial and interglacial episodes in his book *The Great Ice Age and its Relation to the Antiquity of* Man. Later in 1909, four glacial stages were identified in *Die Alpen im Eiszeitater (The Alps in the Ice Age)* by Albrecht Penck and Eduard Brukner. The original divisions of the Quaternary were based on lithological glaciofluvial accumulations that could be traced back to terminal moraines. These layers were immediately underlain by fossil bearing sediments attributed to warmer conditions and characterised the alternating pattern of the Quaternary. These were named as Würm, Riss, Mindel and Günz with intervening warm phases and became widely accepted as a basis of global stratigraphy (Stringer, 2006; Elias, 2007; Renfrew and Bahn, 2012) for more than 50 years, with comparable schemes appearing in Europe, Russia, USA, Africa, Patagonia and New Zealand (Gibbard, 2007) and attempts to correlate it with pluvial lakes of more arid regions in North America and Africa (Lowe and Walker, 1997; Renfrew and Bahn, 2012)

Thus, with the identification of glacial and interglacial sediments and gravels, a new climatic based stratigraphic framework emerged. Categorisation by inferred climatic conditions is known as climatostratigraphy (Lowe and Walker, 1997). However, application of the Alpine sequence to non-Alpine environments resulted in some difficulties due to a varied and incomplete terrestrial record. An inferred climatostratigraphy could be deduced by looking at climate induced environmental changes. Terrestrial proxy indicators such as pollen sequences, glacial varves and loess profiles enabled linkage with Palaeolithic chronologies (Aitken and Stokes, 1997). It also soon became evident that a localised terrestrial record provided a much higher resolution and additional intervening warm and cold stages could be identified.

Whilst the Quaternary is now defined biochronologically (Gibbard and Head, 2010), the working subdivisions of the geological timescale however, are considered to be stages. A stage should enable intra-regional classification, with a succession of time-parallel boundaries (Hedberg, 1976; Gibbard and Kolfschoten, 2006).

Early efforts to formalise the climate based stratigraphical terminology resulted in geological-climate units being proposed. Units of the geologic-climate classification were defined by The American Commission on Stratigraphic Nomenclature in 1961 (Gibbard and Kolfschoten, 2006, p447), thus;

> "*A **Glaciation** is a climatic episode during which extensive glaciers developed, attained a maximum extent, and receded. A **Stadial** ('Stade') is a climatic episode, representing a sub-division of a glaciation, during which a secondary advance of glaciers took place. An **Interstadial** ('Interstade') is a climatic episode within a glaciation during which a secondary recession or standstill of glaciers took place. An **Interglacial** ('Interglaciation') is an episode during which the climate was incompatible with the wide extent of glaciers that characterise a glaciation*."

Glacials or cold stages tend to exist for a prolonged period perhaps tens of thousands of years, where temperatures in the mid to high latitude regions promoted ice formation. Stadials tend to be shorter in duration, perhaps 10,000 years or less. In comparison, interglacials or warm / temperate stages, may have been comparable to temperatures of today, or higher, with a duration of 10,000 years or more, whilst interstadials, are short lived warm periods within a glacial of 5,000 years or less (Walker, 2005). The distinction between a glacial and a stadial or an interglacial and an interstadial is not always clear. Evidence for the different episodes were originally derived from the terrestrial proxy indicators. Cold

episodes indicated by glacial deposits and periglacial sediments, whilst evidence for warm phases was generally indicated by the fossil record such as pollen, insect or mammalian assemblages and biogenic lake sediments (Walker, 2005).

Whilst the Late Quaternary 0-125,000 years is readily correlated with the terrestrial record in the Northern Hemisphere (MIS 1-5), the Middle Quaternary 125,000 – 780,000 years, is less straightforward. There are a number of interglacial and glacial stages but some have no formal designation, with some warm or cold stages containing both warm and cold episodes, thus designation becomes increasingly uncertain (Walker *et al.*, 2012) For example, MIS 3 although a warm stage is only analogous with an interstadial, whilst MIS 5 has several oscillations (5a, 5c and 5e are warmer) with 5b and 5e being cooler although it is 5e that is recognised as the last interglacial (Lowe and Walker, 1997). Stage 7 is also similarly divided with 7a and 7c being warmer sub-stages and 7b again being cooler. For the Early Quaternary, from 780,000 yrs and earlier, correlation of the MIS record with the terrestrial record becomes increasingly speculative. The most dramatic changes are referred to as Terminations . Termination 1 is between 2/1 and Termination 2 between 6/5 (since 3 isn't fully recognised as an interglacial), and can be useful for providing correlations between stratotypes, as can the palaeomagnetic record using boundaries described by major magnetic reversals (Lowe and Walker, 1997).

However, the terrestrial stratigraphic record is highly fragmented, glacial conditions in one region may not be glacial in another but simply just a cold stage, and similarly a warm interglacial in one region may only be an interstadial in another. Sections of the Quaternary record may be represented differently in different regions, perhaps due to differences in deposition rates, or completely missing due to erosion from glacial melt waters or removal by later advancing ice. Temporal resolution between regions and between different proxy climate indicators may vary, be time-transgressive, or respond to climate change differently.

Piecing together the terrestrial record is one of the biggest challenges to Quaternary scientists, which is why regional stratigraphies became fundamental for the Quaternary.

1.2 A Continuous record

1.2.1 Deep Sea cores

Facing difficulties with terrestrial chronologies, researchers turned to the marine environment; a depositional setting that should accumulate more continuous records. A

turning point came with the ability to drill deep ocean cores back into the Tertiary ocean floor sediments. Individual cores provide evidence of uninterrupted sedimentation for hundreds of thousands of years, which, when pieced together, provide a continuous sequence stretching back beyond the Quaternary. Marine microfossils in the cores gave an oxygen isotope signal reflecting the ratio between the lighter ^{16}O and the heavier ^{18}O isotopes. Changing isotopic signals indicate changes between glacial and interglacial. During glacial episodes, the lighter ^{16}O would have evaporated from the ocean surface and been incorporated into the expanding ice sheets, leaving behind the heavier ^{18}O. During these phases, the oceans would have been enriched with ^{18}O, which would then have then been taken up by the developing foraminifera and coccoliths (Shackleton and Opdyke, 1973), thus raising the $\delta^{18}O$ signal in the sediment cores. Similarly, during the interglacials the ^{16}O returns, diluting the ^{18}O and so the $\delta^{18}O$ signal drops. Although originally thought to reflect changing temperature (Emiliani, 1955, 1966a, 1966b) the balance between the two oxygen isotopes is now considered to be influenced by changes in the land ice volume (Shackleton and Opdyke, 1973). Reflecting Emiliani's original numbering system, a series of marine isotopic stages (MIS) can be identified starting from the top of the ocean bed. Glacials are evenly numbered while the intergacials are assigned odd numbers, the current warm stage being MIS 1. The first oxygen isotope sequence was derived from Caribbean and Atlantic sediment cores giving a sequence of 16 isotope stages (Emiliani, 1955, 1966a, 1966b). These were later extended to 22 following the analysis of V28-238 Pacific Ocean core (Shackleton and Opdyke, 1973). Remarkably, the isotopic signal appears geographically consistent, providing a continuous sequence of oscillating warm and cold stages making it a unique proxy for climate change across the globe. Over the course of the Quaternary, 2.6 Ma, more than 100 stages can now be identified, with cores from The Deep Sea Drilling Programme (ODP677 and ODP 846) defining stages from the Middle Pleistocene and earlier (Shackleton and Hall, 1989; Shackleton *et al.*, 1990; Shackleton *et al.*, 1995).

Time-series analysis of these oscillations indicate periodicities associated with the Astronomical Theory of climate change, originally developed by Croll in the Nineteenth Century and expanded upon by Milutin Milankovitch in the 1920's (Lowe and Walker, 1997). Previously, the hypothesis was largely rejected during the 1940s and 50's with the advent of radiometric dating, however the cyclic nature of the oxygen isotopes from marine cores awakened new interest (Imbrie and Imbrie, 1979, cited in Lowe and Walker, 1997) Milankovitch hypothesised that global surface temperatures, affected by radiant solar

energy, would be influenced by regular and predictable changes in the Earth's orbit and axis (Lowe and Walker, 1997). These influences include;

1. the precession of the equinoxes (movement of seasons around the sun, two cycles; periodicities ≈ 23 ka and 19 ka),
2. obliquity of the ecliptic (tilt of earth's axis; periodicity ≈ 41, ka),
3. eccentricity of the orbit (changes in shape of earths obit; periodicity ≈ 100 ka)

Spectral analysis revealed evidence of 100 ka, 43 ka 24 ka and 19 ka cycles, with the eccentricity of the orbit exerting the largest effect whilst the shorter ones modulated the longer term changes (Hayes, 1976, cited in Lowe and Walker, 1997). Subsequently, similar profiles were also seen in other terrestrial proxy records. Taken together, these astronomical influences are known as *orbital forcing* and this is now seen as a primary driving force behind global climate change.

A time-scale for the oxygen isotope record was first derived using an established date for the Brunhes-Matayama geomagnetic reversal at MIS 19, with radio carbon dating used in MIS 2 and Uranium-series dating of Termination II (MIS 6-5 transition). Dates were subsequently interpolated assuming a constant but predicted sediment accumulation rate. Having used dating to establish the validity of oribital forcing, the timescale was then *tuned* using a constant lag for each dominant cycle (Aitken, 1990). Thus, inferred ages can be ascribed to individual stages and their boundaries by extrapolating back from the present. Imbrie *et al*. (1984) used several stacked isotope records to derive a time scale for the last 800,000 years, called the SPECMAP timescale. A comparison between different tuned timescales indicates ages for Termination II of 128 ka (Shackleton and Opdyke, 1973; Imbrie et al., 1984) or 130 ka (Martinson et al., 1987, cited in Lowe and Walker, 1997) compared to 127 ka ±6 ka by U-series dating. For the Brunhes-Matuyama boundary, tuning gave 734 ka (Imbrie et al., 1984), compared to radiometric date of 730 ka ± 11 ka. All orbital tuning errors are reported as ± 5 ka (Aitken, 1990).

Stage boundaries are set mid-way between maximum and minimum $\delta^{18}O$ signals. Due to fairly rapid ocean mixing, slow sedimentation rates and bioturbation smoothing out short term effects, these boundaries are considered time equivalent and can be used as chronostratigraphic markers, enabling correlation between the marine and terrestrial Quaternary records.

1.2.2 Ice Cores

In addition to deep sea coring, improved technology has enabled deep coring of the polar ice sheets. Ice cores were first recovered from Greenland in the 1960s, however it was the European Greenland Ice Core Projects 'GRIP' in 1992 (GRIP Project Members, 1993) and 'NGRIP' in 2003 (North Greenland Ice Core Project Members, 2004), and US Greenland Ice Sheet Project 'GISP2' in 1993 (Hammer, *et al*., 1997) that have been some of the most scientifically important. Hitting bedrock at 3029 m, 3085 m and 3053 m respectively, they spanned over 100,000 years, encompassing MIS 5e. In the 1980s Antartic Vostock Ice cores went back to 420 ka (Petit *et al*., 1999) and the more recent EPICA (Concordia Station, Dome C, Antartica) provides a 740 ka record at a depth of 3270 m (EPICA Community Members, 2004).

Ice cores are unique in that they provide a high-resolution land based atmospheric record, extending beyond the last glacial cycle, and provide a multi-proxy record. Air bubbles trapped in accumulating snow provide information on climate forcing atmospheric greenhouse gases, aerosolic dust and volcanic ash plus several isotopic profiles are also present. $\delta^{18}O$ records work counter to those of the ocean sediments, where higher ratios now represent higher temperatures at the time of formation. The heavy hydrogen isotope deuterium (D or ^{2}H) relative to ^{1}H is also used as an indicator of surface air temperature, (denoted as δD).

Prior to the Greenland ice cores, there was a generally accepted view that for the 100 ka before the Holocene, there was a single, extended glacial stage, interspersed with warmer, short-lived interstadials. However, it has been shown that there were up to 25 significant climate oscillations, with as much as a $15^{o}C$ amplitude, known as Dansgaard-Oescher events (Lowe and Walker, 1997; Johnsen et al., 2001). $\delta^{18}O$ records from GISP2, suggest that these were characterised by rapid warming and slower cooling, each lasting between 500-2000 yrs (Stuiver and Grootes, 2000).

Although the Vostock core is longer than the Greenland ones, climatic fluctuations appear less pronounced. However both the Greenland and Antarctic cores suggest that the last interglacial lasted longer than indicated by the marine record, up to 20 ka compared to 10 ka, beginning at 133 ka rather than 125 ka, a delay in glacier melting accounting for the delay in the marine record (Dansgaard et al., 1993, cited in Aitken and Stokes, 1997)

Comparison of the EPICA core with the previous ones, together with a 340 ka record from Mt Fuji (Watanabe *et al* 2003) showed close agreement in the measured properties for

the last four glacial cycles. However the longer record from the EPICA core indicates that between 740,000-430,000 yrs, interglacials in Antarctica were cooler but lasted longer. Similarities are observed between Termination V (transition between MIS 12 and 11) with the most recent, Termination 1 (transition between MIS 2 and 1), and tentatively suggest that MIS 11 may be a close analogue for the present and future climate (without human interference). If so then this may suggest that a similarly stable climate might be expected for another 17,000 yrs, (EPICA Community Members, 2004).

Ice cores can be dated by counting annual layers, however the further back in time one goes the less distinct the records become, with problems from low accumulation, compression and diffused seasonality markers. Ice-flow models or orbital tuning to other climate proxies can be used for correlation of isotopically defined events in ice cores with independently dated events in marine records or speleothems have also been used (Shackleton *et al*., 2004; Svensson *et al.*, 2006)

Counting errors reported by Walker (2005) for the GISP2 core are; ± 1-2% up to 12 ka; ± 2% to 40 ka; ± 5% to 45 ka; ± 10% at 50 ka; ± 20% up to 110 ka (Meese *et al*., 1997). For the EPICA core, errors of ± 10 yrs back to 700 years; ± 200 years to 10 ka; ± 2 ka back to 41 ka (Schwander *et al.*, 2001), ± 10 ka at 807 ka; and estimated to be ± 20 ka at 960 ka. Vostok cores report all errors of less than 15 ka; better than 10 ka for most of the record and < 5 ka for the last 110 ka.

1.2.3 Marine-Terrestrial Correlation

The advent of orbitally tuned oxygen isotope records from deep sea sediment cores provided the first continuous, global, geochronological sequence for the Quaternary. The realisation that there were considerably more climatic events indicated by the marine sediments and higher resolution ice cores, has resulted in considerable efforts to correlate localised fragmented terrestrial stratigraphies with the global isotopic stages and time-scale (Kukla, 1977). This is hardly surprising but as Gibbard and Kolfschoten (2006) remark, poses some practical difficulties. The only way this can be achieved is either by curve fitting of a terrestrial record to the marine chronology or by applying quantitative dating methods. The process of curve fitting relies on there being a long, continuous stratigraphic record, perhaps from pollen, loess, palaeosols or glacial lake varves. However, for land based sequences these are rare and probably unreliable in the absence of litho or biostratigraphic markers. For shorter duration stratigraphies, event markers such as volcanic tephra layers or magnetic

reversals, might be used (Lowe and Walker, 1997; Gibbard and Kolfschoten, 2006; Walker, 2008)

The extent to which a given geologic-climate unit is represented in a proxy stratigraphic record will depend on the amplitude and duration of the climate change and on the sensitivity of the proxy records. For example, the response shown in a pollen zone may be different compared to that of an insect or mammalian assemblage zone. Response rates to climate change by different proxies will vary and may not be immediately evident in the terrestrial record, for example. This is demonstrated by the boundary between MIS 1 and 2, now known to pre-date the Holocene-Pleistocene boundary by 2000-4000 years (Gibbard and Kolfschoten, 2006)

However, Blaauw (2012) points out the dangers of aligning proxy records and challenges the notion of continuous pollen zones, peat layers and tephras, even across relatively small geographic regions. Aligning terrestrial sequences to previously tuned records is in effect 'double tuning' (Blaauw *et al*., 2010; Blaauw and Christen, 2010; Blaauw, 2012). Therefore any errors incorporated into the initial tuning (e.g. SPECMAP (Shackleton and Opdyke, 1973)) will also need to be taken into account in the second tuning event, thus accumulating uncertainty. It is further argued that

> "*Although independent radiometric dating of sea level changes has largely confirmed SPECMAP's timing (Thompson and Goldstein, 2006), its chronological uncertainties are on the order of several thousand years (Martinson et al., 1987; Lisiecki and Raymo, 2005)*." (Blaauw, 2012, p41).

Tuning of a marine core with the Greenland GRIP one has lead to problems of synchroneity between the two due to the poor resolution of the marine sediments (Cayre *et al*., 1999). It is observed that many published uncertainties that may have originally accompany tuned sequences start to be dropped once in use, and what was once uncertain now becomes fact (Blaauw, 2012). The same effect is also observed in reviewing the British chronologies. Original independent dating becomes largely overlooked and is replaced by the presumption of confirmed fact and totally ignores any uncertainty associated with the original dates. Even tie points that might assume zero uncertainty between proxies, will retain the original tuning uncertainty associated with them. Objective statistical methods for probability based peak comparison are sadly lacking and there is a need for perhaps a grey-scale to reflect uncertainty regions (Blaauw and Christen, 2010). Whilst these approaches would tend to lose the finer detail and short-term event correlations, they would provide a

realistic basis for comparison that could be refined as dating accuracy improves. Blaauw suggests that perhaps the safest route is to adopt a 'null-hypothesis' approach, where results are assumed uncorrelated until proven different by independent dating (Parnell *et al*., 2008; Charman *et al*., 2009; Blaauw *et al*., 2010; Blaauw, 2012). The INTIMATE project (Integration of ice core, marine and terrestrial records) relies on correlation of event stratigraphies (Walker *et al.*, 2001; Blockley *et al*., 2012). However they also stress the importance of independent dating after first identifying local events from independent evidence and correlation;

> "*The third step (which is perhaps the most difficult but perhaps also the most important) is to use independent dating evidence to establish the degree of synchroneity between local and GRIP events."* (Walker *et al*., 1999)

Dating methods fall into one of two categories generally referred to as ***absolute*** or ***relative*****.** *Relative* methods rely more on stratigraphic correlation and independent calibration whilst absolute methods require a time-dependent, quantifiable physical or chemical change to occur (Wagner, 1998). To assume that a date is *absolute* is probably overly optimistic, since it relies entirely on the application of a technology at a given moment in time and allows no room for improvement and change. For this reason, in this thesis, so called *absolute* methods will be referred to as ***quantitative.*** Sadly, even quantitative methods bring with them their own set of complications.

The terrestrial record may only reflect those precise moments in time when material was deposited, such as volcanic tephra, and may not reflect the entire duration of a warm or cold stage, unlike perhaps lake varves that may represent continual accumulation. Thus the location and timing of marine isotope stage boundaries in proxy records is a particular problem. For older geological sequences, due to poorer stratigraphic resolution, climate changes will appear to be reflected in the geologic record almost instantaneously. Such boundaries within the more recent Quaternary are difficult to determine and correlate. The transition between a warm to a cold stage may take several thousand years, therefore at which point does the boundary occur? Lowe and Walker (1997) suggest this may be at any of three points, i) the start of warming after a temperature minimum, ii) perhaps when it reaches a temperature equivalent to today, or iii) perhaps when it crosses a thermal threshold level, reflected by the occurrence of an indicator species. Thus it is likely that there will be a disagreement between chronostratigraphic and geochronologic boundary markers.

Uncertainties associated with quantitative dates, generally refer to the analytical uncertainty occurring in the laboratory (with perhaps the exception of layer counting). Further uncertainty due to sampling is of equal and potentially greater influence on the measurement result and must be reflected in the final uncertainty estimate. The problem is, realistic uncertainty estimates often make uncomfortable reading.

1.2.4 The British Chronological Framework

One of the first suggestions to subdivide the British Pleistocene based on climate change was proposed by The Geological Society of London with the publication of a British chonostratigraphical scheme in 1973. This scheme recognised four interglacials based upon palaeobotantical and sedimentary evidence (Mitchell *et al.*, 1973; Morigi *et al.*, 2011). However, it wasn't long before the marine cores (Shackleton and Opdyke, 1973; Hayes *et al.*, 1976) provided a detailed chronological framework with which to correlate the terrestrial record. Since then there have been numerous revisions and stratigraphical sequences reported, which still place heavy dependence on biostratigraphical and lithostratigraphical methods (for example, Gibbard, 1994; Bowen, 1999; Westaway *et al.*, 2002; Bridgland *et al.*, 2004a; McMillan *et al.*, 2005; Bridgland, 2006; Stringer, 2006; Cohen and Gibbard, 2011). For the most part there is a general consensus regarding the allocation of stages, MIS 1 representing the Holocene (Flandrian); MIS 2-5d the Devensian (cold stage); MIS 5e the Ipswichian Interglacial; MIS 12 the Anglian cold stage and MIS 13 down to the Brunhes-Matuyama reversal at MIS 19 representing the Cromerian complex. However there was a long running debate between Gibbard (1994) who advocated fewer subdivisions during the 'Wolstonian' (MIS 10-6, maybe 11) than Bridgeland's Thames Terrace sequence (1994, cited in Bowen, 1999). This lack of resolution between MIS 10-6 was also reflected in The Geological Society's 2nd edition of The Geology of England and Wales (Catt *et al.*, 2006), the British Geological Survey's report of Britain's Quaternary and Neogene deposits (2005) and the International Commission on Stratigraphy's (ICS) Subcommission on Quaternary Stratigraphy's (SQS) most recent global chronostratigraphical correlation v2011 (Cohen and Gibbard, 2011). However recent work by the AHOB group (Ancient Human Occupation of Britain), appears to have favoured Bridgland's original suggestion and has referred to MIS 7 as the 'Aveley Interglacial', MIS 9 as the 'Purfleet Interglacial' and MIS 11 as the 'Hoxnian Interglacial' (Stringer, 2006, 2011). It would seem the debate continues.

Whilst some reference is made to the use of independent dating methods in the construction of these chronologies/stratigraphies (Bowen, 1999; Bridgland *et al.*, 2004a), the

emphasis remains heavily in favour of stratigraphic correlation, often without reference to independent dating. Bridgland *et al*., (2004a, p206) remark that,

> "*Accurate age estimation is not only important in establishing an absolute chronology of stratigraphical units and an aid to regional-global correlations, but it also provides the necessary constraints to reconstruct the rates of change in fluvial activity and in some cases provides a framework for detailed assessment of lead-lag effects in landscape evolution triggered by major environmental change evidenced by associated faunal/floral changes. However, better resolution of dating techniques is required before this potential can be fully realised*."

1.3 Dating methods

The measurement of time requires a time-dependent, quantifiable change to occur, whether that be physical or chemical (Wagner, 1998), prior to radiometric methods, this meant layer counting. Carbon 14 is the most common radiometric method. Its range of suitable materials includes almost anything containing organic carbon, for example shells, seeds and bones, but the method is limited to relatively young material, around 50 ka. The technology has been greatly improved over recent years with the introduction of Accelerator Mass Spectrometry (AMS) (Bronk Ramsey *et al.*, 2004), an extended calibration curve (Reimer *et al.*, 2009) and the application of a Bayesian statistical approach (Heaton *et al.*, 2009).

Other radioactive isotopes that can be used have longer half-lives and enable older material to be dated. These include Uranium-series isotopes, most applicable to carbonate materials such as speleothems and corals from about 100 -500 ka with an uncertainty perhaps as low as 1% (Walker, 2005). Argon isotopes ($^{40}K/^{40}Ar$ or $^{40}Ar/^{39}Ar$) can be used for dating volcanic material (igneous rock and tephra), useful as a chronologic control across regions and strata and potentially applied across the whole Quaternary when present as a continuous record. The difficulty is, they occur rarely in Britain. $^{40}K/^{40}Ar$ dating is only applicable to older samples, >100 ka due to very high uncertainties associated with younger material (approx 100%). $^{40}Ar/^{39}Ar$ dating by comparison has far better precision and can provide age estimates of 10 ka or less (Walker 2005). Cosmogenic nucleide (CN) dating is based on the accumulation of cosmic ray induced radionuclides on exposed rock surfaces and includes, ^{10}Be, ^{26}Al, ^{36}Cl ^{3}H and ^{21}Ne. Applicable dates range anywhere from a few thousand years to a few million. Shorter-lived isotopes applicable to periods of a few hundred years

include ^{210}Pb, ^{137}Cs and ^{32}Si (Walker, 2005). Relevant minerals include quartz, olivine and garnet. However, there are inherent problems associated with CN dating due to the need to *zero the clock* and exposure history for glacial and fluvial deposits is difficult to predict. A final group of quantitative methods are the luminescence or radiation exposure dating methods. These include OSL (optically stimulated luminescence), TL (thermoluminescence) and ESR (electron spin resonance) and work on the principle of measuring freed electrons that have been trapped in the crystalline rock matrix after exposure to radiation. They differ only in the method used to excite or free the electrons, either using light energy, heat or a magnetic field. A related method is Fission track dating which counts the number of damage trails left by the ^{238}U isotope. All methods are appropriate for dating sediments, rocks, speleothems, flint, tooth enamel and even pottery. Applicable age ranges for these methods is anything from 100 yrs to several hundred ka (Walker, 2005).

Thus, opportunities for independent dating are dependent on the availability of appropriate materials and there being in the appropriate age range. However, whilst amino acid racemisation (AAR) still has similar issues regarding appropriate matrices, it is unique in that it possess the potential to cover the entire Quaternary and beyond (Miller *et al.*, 1979). AAR is generally recognised as a relative dating method, based on the relative ordering of D/L values within a limited geographic area (Wehmiller and Miller, 2000; Miller and Clarke, 2007). Aminostratigraphy (Miller and Hare, 1980) has been an important influence in the development of the current British geological succession (Miller *et al.*, 1979; Bowen *et al.*, 1989; Gibbard, 1994; Bowen, 1999; Bowen, 2000; Bridgland *et al.*, 2004b; Bridgland, 2006; Stringer, 2006; Penkman *et al.*, 2011). Bowen was able to correlate D/L values for isoleucine in non-marine molluscs, with marine isotope stages using independently dated deposits (Bowen *et al.*, 1989; Bowen, 2000). The resolution provided by the D/L value gave convincing evidence for the applicability of AAR to geochronology, most recently evidenced by the publication of a revised AAR chronological framework based on *Bithynia* opercula, correlated against additional archaeological and biostratigraphical sequences (Penkman *et al.*, 2011).

1.3.1 Amino Acid Racemisation

1.3.1.1 Background and application

Amino acid racemisation, or epimerization for molecules with two carbon centres, is a diagenetic process that occurs naturally following protein synthesis (the more general term ‘racemization’ will be used hereafter to refer to both racemization and epimerization). The process involves the slow inter-conversion between the two chiral forms of amino acids, the

building blocks of proteins, from the original *laevo* (L-form) in life to the *dextro* (D-form). Conversion of the L to D form continues until equilibrium is reached, which for most amino acids is usually equal to 1, although values of 1.3 are reported for isoleucine ratios (Miller and Clarke, 2007). This process can take many hundreds of thousands of years, thus the D to L ratio or D/L value can be used as an indicator of time (Miller and Hare, 1980).

Philip Abelson (1954) was the first to recognise the persistence of amino acids in fossil shell, later supported by kinetic experiments that derived a half-life for the decarboxylation of alanine at room temperature of "*about 10 billion years*" (Conway and Libby, 1958). Abelson proposed that over time, the hydrolysis of proteins might release free amino acids, which themselves might be retained within a biomineral matrix (Abelson, 1955). In 1962, Ed Hare identified AAR in fossil samples whilst working on his doctoral research but it was his work with Abelson (Hare and Abelson, 1968) and Mitterer (Hare and Mitterer, 1967, 1969) that AAR was first proposed as a dating technique. Some of the first applications of AAR as a dating method include marine core sediments (Bada, 1970; Wehmiller and Hare, 1971) using the epimerisation of isoleucine. Initial results were encouraging, with agreement of sedimentation rates in the marine cores finding close agreement with palaeomagnetic data (Bada, 1970) although racemisation rates were found to be non-linear beyond a value of 0.25 (Wehmiller and Hare, 1971) or 0.3 in foraminifera. (Bada and Schroeder, 1972). In 1972, the work of King and Hare (1972) recognised that rates of racemisation varied between different species of forams and Kvenvolden *et al*. (1973) determined that the rates varied for different amino acids, with aspartic acid, alanine, and phenylalanine being amongst the fastest and isoleucine and valine being the slowest. The idea that amino acids existed as different fractions (free, peptide bound and protein bound) was then proposed as a possible explanation of the non-linearity over time (Bada and Man, 1973).

However, it was its application to archaeological bone that was potentially the most exciting. Initial efforts gave mixed results when compared against radiocarbon dates. Investigation of animal bones by Turekian and Bada (1972) indicated discrepancies although results determined using rates derived from kinetic experiments (Bada, 1972) showed better correlation. It was during this early phase of AAR dating that the dating of Palaeolithic remains from La Jolla in California, caused some unwelcome press. The La Jolla bones were assessed using aspartic acid racemisation, and compared to two calibration samples; a bone less than 200 yrs old and one radiocarbon dated to 17,000 yrs. Consequently an extrapolated age for La Jolla man was put at between 30-50ka (Bada *et al.*, 1974). In 1984, both the 17 ka bone and that of La Jolla, were reanalysed by an improved radiocarbon

technique using Accelerator Mass Spectroscopy (AMS). This time far more recent dates were derived; 5,000 years for the calibration sample and 5,540 ± 400 yrs for La Jolla. Using the revised date, this would have given an AAR date of approximately 7000 years (Bada *et al*., 1984).

Nonetheless, because of the *openness* of bone and the folding and complex nature of collagen (Collins *et al* 1999), bone is not considered a suitable material for dating by AAR. AAR can potentially be applied to any material where amino acid residues persist over geological time, but is most favourable with materials where the organic component is protected by a carbonate biomineral, providing an intra-crystalline closed system for the protein to break down predictably (Brooks *et al.*, 1990; Penkman *et al.*, 2008). Recent applications to fossil biominerals include terrestrial mollusc shell (Hearty and Kaufman, 2009; Marković *et al.*, 2011), opercula (Penkman *et al.*, 2011; Briant *et al.*, 2012), mollusc shells (Demarchi *et al.*, 2011; Wehmiller *et al.*, 2012) ratite egg shells (Clarke *et al.*, 2007; Magee *et al.*, 2009), corals (Hendy *et al.*, 2012), ostracods (Bright and Kaufman, 2011) foraminifera (Kaufman, 2006) and teeth (Dobberstein *et al.*, 2008; Griffin *et al.*, 2010).

The rates of racemization for the 20 or so naturally occurring amino acids are highly temperature dependent, matrix and species specific (Wehmiller and Miller, 2000; Miller and Clarke, 2007). As the thermal history of a site is rarely known, it can be difficult to use AAR kinetic modelling to determine absolute age estimates (Clarke and Murray-Wallace, 2006; Kosnik *et al.*, 2008). For this reason, much research tends to apply the technique as a relative stratigraphic tool (Miller *et al.*, 1979; Bowen *et al.*, 1989; Bowen, 2000; Wehmiller *et al.*, 2010; Penkman *et al.*, 2011), (with numerical ages only being assigned to samples within a defined locality using independently calibrated material (e.g.Hearty and Kaufman, 2009; Murray-Wallace *et al.*, 2010; Demarchi *et al.*, 2011), or by adopting a dual approach using both calibration and kinetic modelling (e.g. Wehmiller *et al.*, 2010; Wehmiller *et al.*, 2012). The assumption is that if sites share the same temperature history, any observed D/L differences can be interpreted as relative age differences. Similarly, it becomes possible to use D/L values for palaeothermometry, (as indicators of relative temperature variation between same age sites), once independently dated using appropriate techniques (e.g. Kaufman, 2003; Owen *et al.*, 2007; Bright *et al.*, 2010; Reichert *et al.*, 2011).

The last 30 years have seen significant changes in AAR analysis. Early research based on ion-exchange liquid chromatography (IEx) was able to separate L-isoleucine from its diastereomer D-alloisoleucine, yielding a D-Aile/L-Ile value, or often termed A/I value. As

methods developed, it became possible to detect and measure increasing numbers of chiral pairs of amino acids, from six or seven using gas chromatography (GC) to ten or more routinely determined today using reverse-phase HPLC (RP or rpHPLC) (Kaufman and Manley, 1998) Further improvements in preparative methods and materials (Sykes *et al.*, 1995; Penkman *et al.*, 2008) have greatly improved the resolving capabilities of the technique and proven potential for developing widespread chronologies (e.g. Africa: (Brooks *et al.*, 1990); Australia: (Murray-Wallace, 1995); USA: (Wehmiller *et al.*, 2010); eastern Europe: (Oches and McCoy, 2001); western Europe: (Ortiz *et al.*, 2004; Penkman *et al.*, 2011). AAR now requires mg sample sizes, is relatively fast and with inexpensive preparation and analytical costs further supports its application in routine analysis.

1.3.1.2 Precision

Associated with methodological advances are improvements in reported intra-laboratory analytical precision estimates, often reported as less than ± 1% (Penkman *et al.*, 2011). However significant inter-laboratory and method differences have long been known (Kvenvolden, 1980; Wehmiller, 1984). Whilst the precision and internal consistency of an individual laboratory's data may be excellent, the lack of comparability limits the full exploitation of the technique and its wider applicability.

Clearly, the accuracy of age estimates relies heavily on the accuracy of the analytical data. Accuracy is comprised of both precision and trueness (measured as bias) elements. Precision can be determined through repeated measurements of the same or similar substance under repeatability or reproducibility conditions. However bias requires evaluation against a true or reference value, a material that does not currently exist for AAR. For this reason, most AAR uncertainty estimation focuses on precision evaluation in the absence of defined reference materials.

Published **intra**-laboratory precision estimates are often excellent. Wehmiller and Miller (2000) have reported intra-laboratory precision estimates of 2% for repeated instrumental determinations by gas chromatography (GC) of the same hydrolysate, between 3-5% for multiple analyses of different fragments of the same material, and between 5-10% for multiple samples from the same sample location. More recently, in an evaluation of marine molluscs from the North Carolina coastal plain (Wehmiller *et al.*, 2010), analytical precision for most amino acids was reported as being better than 2% (based on D/L values from multiple chromatograms of the same derivative using GC). CV% values based on

multiple shells approximated to about 6%, but the range varied for different amino acids, on a few occasions exceeding 30%.

Uncertainty estimates for repeated analyses by RP of intra-laboratory reference solutions (approximate D/L of 0.5) carried out over several years have been reported as 1.5% for aspartic acid D/L values and 1.4% for glutamic acid D/L values (Kosnik and Kaufman, 2008). For a reference solution with a lower D/L ratio, (approximately 0.09), higher uncertainty estimates were obtained; 3.7% and 3.8% respectively, although an average of 1.4% was suggested as being representative of the analytical uncertainty for both aspartic acid and glutamic acid based on the mid-range D/L values.

By comparison, studies between laboratories and different methods (i.e. GC vs RP) report greater imprecision. In an early inter-laboratory comparison study (Wehmiller, 1984) the precision estimates achieved by individual laboratories were reported, but precision estimates **between** participating laboratories were not provided. However, significant differences **between** laboratories' results were commented on, in some cases resulting in greater than 25% differences in estimated age, and called for the need for reference standards in routine analysis to ensure comparability and reproducibility of results. Bakeman (2006) recorded a 6.8% higher systematic offset for A/I ratios by GC compared to RP, and 1.9% compared to IEx. A further 4.6% difference for glutamic acid D/L values and as large as 25% for valine D/L values between GC and the higher RP values in both cases is also observed.

Important unaccounted for differences between AAR age estimates and other dating methods have also been observed (Wehmiller, 1992). 30% imprecision is reported for age estimates from *tidied* aspartic acid and glutamic acid data by Kosnik and Kaufman (2008), of which the analytical uncertainty is reported to account for only 5%. Even wider age precision estimates up to ±40-50% have also been reported, determined using A/I ratios where the age equation was not calibrated locally, (McCoy, 1987). Whilst these effects may be due to a number of reasons, clearly in the presence of such large discrepancies, the control of bias and the accurate reporting of analytical data become paramount.

1.3.1.3 Measurement Uncertainty (MU) in AAR

As already mentioned, the absence of defined reference materials has been a serious draw back to the control of systematic errors and the proper reporting of uncertainty estimates in AAR analysis. Uncertainty estimates that are reported in the literature are given only as precision estimates, at times representing only the instrumental precision between

repeated injections of a given sample. Such precision estimates do not provide a realistic estimate of the precision for the method or sampling, let alone a full evaluation of uncertainty.

Research into the area of analytical uncertainty in AAR is grossly lacking. Efforts have been made to derive uncertainty estimates (Kosnik and Kaufman, 2008; Westaway, 2009), but demonstrate a lack of general understanding about uncertainty sources and evaluation. Kosnik and Kaufman (2008) evaluate long-term standard solution data but assume that all analytical uncertainty should be the same for all amino acids. On finding disagreement, a process for culling the data is presented (Y-criteria) in order to make it fit. This in effect is imposing a confidence interval in order to force agreement. What Kosnik and Kaufman fail to appreciate is that whilst bias effects acting on each of the amino acids will be the same, the precision of each of the amino acids very likely won't be. This will be due to inherent differences in the physical and chemical properties between the amino acids, resulting in different instrumental effects and detector sensitivities acting on each with additional long term stability issues potentially influencing individual amino acids too. Similarly, Westaway (2009) presents an algorithm for evaluating standard uncertainties (standard errors) using the number of replicate values to improve precision estimates. However, Westaway fails to appreciate that the standard deviations used for this analysis, represent only the injection precision and do not represent the uncertainty of the method, sample or site. Consequently any conclusions regarding sub-stage resolution are likely to be far too tight and unrealistic.

Documentation providing guidance on measurement uncertainty evaluation has been in circulation within the analytical community for many years within industrial and service sectors. Perhaps because the research community have not been constrained by the same commercial pressures, (for example requirements for accreditation), it would seem that at least in respect of the understanding and expression of uncertainty, they may have been left behind.

Quantitative AAR age estimates can be achieved by calibrating against samples of known age and interpolating between tie points (Wehmiller and Miller, 2000; Miller and Clarke, 2007). However this approach carries with it potentially large uncertainties arising from inaccurate curve fitting combined with additional uncertainty associated with the method used to derive the reference dates, which are not themselves, included in any uncertainty estimate.

A further complication is added by the nature of the temperature dependency of racemisation. This is potentially particularly troublesome due to the history of climate oscillations and extreme temperature variability. Thus during very cold stages, the rate of racemisation is so slow, there may be no significant change in D/L value from the end of one warm stage to the beginning of the next. Clearly this presents a problem for numerical dating, as each glacial may last thousands of years. This therefore requires interpretation using secondary evidence and correlation with other stratigraphic markers; in addition, a very large uncertainty needs to be incorporated into final quantitative dates. For this reason, AAR is generally not used for high-resolution work (Wehmiller and Miller, 2000).

However, one of the purposes for this research is to develop a quantitative integrated dating method, including uncertainty estimates, utilising the differential rates of protein decomposition of the individual amino acids. Whilst it is fully appreciated that any models developed will be entirely dependent on the accuracy of kinetic models and palaeoclimate reconstructions, a single dating technique covering the whole Quaternary, could potentially have a very significant impact.

1.4 Aims & Objectives

Marine and ice cores have provided a valuable stratigraphic and chronologic framework with which the fragmented terrestrial record might be correlated. However, independent dating and uncertainty determination is essential to avoid mis-interpretation. From the previous discussions, it can be seen that AAR has played an important role in the British Quaternary stratigraphy. However, D/L values are currently used without accompanying uncertainty estimates. The inability to correct for bias also prevents wider correlations and potentially hemispheric chronostratigraphies from being achieved. The purpose of this thesis is therefore to address these two important issues and investigate the potential for quantitative AAR dating.

The original aim of the project was to retrospectively evaluate the measurement uncertainty (MU) in AAR D/L values using an extensive analytical RP data archive held by BIOARCH at the University of York, integrating the covariant relationships between the different amino acids based on protein degradation patterns. However, MU means different things to different people and determination is often multi-layered, multi-faceted and often dependant on requirements and perspective.

Ultimately it is the uncertainty associated with the expression of the D/L value that is required, since it is the D to L ratio (D/L value) that is used in AAR geochronology. However, error influences are introduced during the preparative stages and analysis of the individual L and D isomers. MU estimates are only valid providing measurement results have been derived using a measurement procedure that is under statistical control. In the absence of reference materials to correct for bias and known performance parameters, normally determined as part of the initial method validation, statistical control could not be assumed. Thus, from a quality management perspective, it is at this level that uncertainty first needs to be controlled and evaluated.

With this in mind, a three tiered approach to the research was adopted and is reflected in this thesis.

1. Evaluate Intra-laboratory analytical precision estimates.
    i. Compare 'Bottom-up' and 'Top-down' approaches to uncertainty determination
    ii. Evaluate uncertainty estimates using data from the AAR archive
    iii. Consider implications for routine analysis and internal quality control
2. Coordinate an Inter-laboratory Proficiency Test as an indicator of analytical bias.
    i. Determine individual laboratories' relative bias estimates for different amino acids in different test materials.
    ii. Compare RP bias estimates with GC and IEx methods
    iii. Compare bias estimates between different amino acids
3. Determine Site D/L uncertainty estimates from *Bithynia* opercula data.
    i. Derive D/L uncertainty estimates using ANOVA for individual locations.
    ii. Model the covariant relationships for amino acid decomposition.
    iii. Derive an integrated uncertainty model based on the joint probability density.
    iv. Develop a model that could determine quantitative ages with uncertainty estimates using racemisation kinetics and palaeoclimate models.

Whilst reference to marine isotope stages have been made accompanying specific examples, it is important to stress that **it is not the aim of this thesis to assess the validity or assignment of any named site**. Rather the emphasis is in the development of a model, based on existing information, to give uncertainty estimates, which, with further refinement, could potentially be used for chronological purposes without the need for independent calibration.

1.4.1 Thesis Structure

Chapter 1 has considered the context of the research presented in this thesis, including the definition and subdivisions of the Quaternary, correlation of the terrestrial record with marine and ice cores, the need for independent quantitative dating, together with the suitability of AAR to provide correlation going back through the Quaternary. Following on from the introduction, **Chapter 2** provides an overview of measurement uncertainty based on international guidelines, considering potential sources of error and an outline to approaches used in its determination. **Chapters 3 and 4** then go on to consider MU from the Intra-laboratory perspective. Chapter 3 focuses on the theoretical evaluation of MU in the context of AAR analysis and Chapter 4 then presents results from the evaluation of standard solutions and other available solid matrix materials. Results of these analyses are considered with regard to quality control activities, including repeatability estimates, control charts, instrument response factors and calibration. The focus then changes to Inter-laboratory uncertainty assessment in **Chapter 5**. For part of this study, an inter-laboratory proficiency test was coordinated between eight AAR geochronology laboratories in the USA, Australia, Spain, Germany and the UK. Results of this work have been compiled into a set of 6 individual reports that were circulated to participants. However, due to the enormous amount of data generated, a summary paper has been prepared and submitted for publication (Chapter 5). In this paper a summary of precision estimates derived from individual participants' results is presented to enable a direct comparison with previous inter-laboratory comparison studies (that have focused solely on precision estimates). A summary of the relative bias is also provided, but for detailed coverage of the results for each of the six test materials used, readers are directed to the anonymous copies of these reports, which have been included as Chapter 5 Appendices. A subsequent paper, combining precision and bias data into overall estimates of uncertainty has also been drafted in anticipation of being submitted for publication. However, due to word restrictions in this thesis, this has been placed as Appendix 1 to Chapter 5. Having considered analytical MU from both an intra- and inter-laboratory perspective, site based MU is considered in **Chapter 6**, including influences from sampling. Correlations between archived amino acid D/L values, based on protein decomposition rates, are evaluated and predictive curves used to derive uncertainty estimates for known valine D/L values . These are then combined using a Bayesian approach for known variances, to give combined uncertainty estimates for valine D/L values for samples of opercula of the freshwater gastropod *Bithynia* from previously sampled sites within the Thames Terrace sequence. Using racemisation kinetics and a palaeoclimate

reconstruction, predicted rates of racemisation have then been used to develop a model that could potentially be used to derive quantitative AAR age estimates with an uncertainty range in **Chapter 7**. However, ages currently derived are purely illustrative as they depend entirely on using appropriate kinetic parameters and climate reconstructions, which themselves carry uncertainty influences that also need to be incorporated. Nonetheless the model demonstrates the potential for quantitative AAR dating. Future Work given in the final **Chapter 8**, concludes the thesis.

1.4.2 Terminology

A Glossary of frequently used, accuracy and uncertainty related terms and abbreviations is provided at the end of this thesis. However, before discussions concerning amino acids are presented, it is helpful to provide a brief summary of some frequently used terms at the start. Thus for the following amino acids the following three letter abbreviations may be used interchangeably in the text;

aspartic acid (ASP); asparagine (ASN); alanine (ALA); arginine (ARG); glutamic acid (GLU); glutamine (GLN); isoleucine (ILE); alloisoleucine (AILE); leucine (LEU); methionine (MET); phenylalanine (PHE); serine (SER); tyrosine (TYR); valine (VAL).

Asparagine (ASN) and glutamine (GLN) both naturally rapidly and irreversibly deaminate to aspartic acid (ASP) and glutamic acid (GLU). Their occurrence is therefore rare and usually undetectable by RP. However, so as not to ignore the existence of asparagine and glutamine, abbreviated references ASX and GLX will be used to indicate the combined ASP+ASN and GLU+GLN respectively, although any full references in the text will be to aspartic acid and glutamic acid only, unless otherwise shown.

All AAR results are determined as the concentration of the D isomer divided by the L isomer, referred to as the DL ratio or D/L value, with one exception. For isoleucine, the D form is referred to as alloisoleucine, thus the D/L value has historically been referred to as the D-AILE/L-ILE or A/I value, and either form may be seen in the text.

Other abbreviations that may be seen throughout the text include measurement uncertainty (MU), proficiency test (PT), collaborative trial (CT), standard deviation (std dev), standard solution (std sol) and intervals of time expressed per thousand years as either ka or kyr, or per million years as Ma or Myr.

Chapter 2. Measurement Uncertainty

2.1 Introduction

Measurements are a fundamental requirement of modern living. However, whilst for the majority, the information provided by a measurement value is assumed to be the real value or true value, a single measurement or even a group of measurements simply represents one (or several) of many possible values for the given measurand. The result is thus only a representation or our best estimate given the limitations of the equipment, conditions, expertise etc, the true value remains unknown. For this reason it is necessary to assess the dispersion of other possible values of our estimate for the same measurands, and report it alongside our measurement result. This parameter is known as the **measurement uncertainty** and provides a quantitative expression of the level of doubt associated with a reported result.

> *"Unfortunately there is no unique way to express quantitatively the 'doubt' that the uncertainty represents. As a consequence, different and in some cases conflicting uncertainty evaluation procedures were developed over the years*" (Lira, 2002 p xi).

For over a century, international metrology laboratories have developed and maintained a global measurement system ensuring accuracy and uniformity in international measurement standards. The need for an international convention for units of exchange was first recognised in the mid Nineteenth century to enable the growth of international trade and finally agreed upon with the signing of the Convention du Metre in Paris in 1875 (Lira, 2002). The international metric system of units created by the Convention and subsequently maintained by the Bureau International des Poids et Mesures (BIPM), is an intergovernmental treaty and established a common structure by which governments could act harmoniously regarding metrology. In 1978, the Comite International des Poids et Mesures (CIPM), recognising the lack of uniformity in the handling of uncertainty measurement, requested BIPM to establish common fundamental principles. In 1980, the

BIPM together with eleven National Metrology Institutes published recommendation INC-1 and a specialist ISO (International Organisation for Standardisation) Technical Advisory Group, TAG 4, was set up to expand on the basic principles therein and produce a practical guidance document. The result was the authoritative document the Guide to the Expression of Uncertainty in Measurement in 1993, published by ISO in the name of, BIPM; the International Electro technical Commission (IEC); the International Organisation of Legal Metrology (OIML); the International Federation of Clinic Chemistry (IFCC); the International Union of Pure and Applied Chemistry (IUPAC) and the International Union of Pure and Applied Physics (IUPAP) (Lira, 2002).

The Guide or GUM (JCGM 100, 2008) as it has come to be known is still commonly accepted as the international definitive guidance document for uncertainty measurement, although since then various supporting documents have been written to assist in its interpretation and implementation at bench level (EURACHEM / CITAC, 2000; Magnusson *et al.*, 2004; EUROLAB, 2006, 2007) and several other alternative methodological approaches have been proposed (RSC Analytical Methods Committee, 1995; Barwick *et al.*, 2000; ISO 21748, 2010). However it was the publication of the GUM that has resulted in the global consensus on reporting uncertainty associated with measurements and has enabled comparison and standardisation of those results in calibration, accreditation, and analytical service around the world.

It is concerning the frequency that the terms accuracy, error, precision and uncertainty are used synonymously in the literature, resulting in confused interpretations by the reader, not helped by the changing emphasis and use in guidance documents. In many respects, archaeology is no longer a discrete discipline, drawing more and more on a multi-faceted approach and an interdisciplinary perspective in the analysis and interpretation of our ancestral remains. Today's archaeologists have to become experts not only in their own field but also draw on expertise in botany, ecology, zoology, osteology, medicine, disease and diet, geography, geology, climatology, chemistry, biochemistry, physics, sociology, statistics, to name but a few. Clearly this is an impossible task for a single individual. It therefore seems hardly surprising that much valuable information gets innocently overlooked and a multi-disciplinary approach becomes essential. The application of natural and physical science to answer archaeological questions is broadly termed archaeometry and reflects better the concept of the quantitative measurement of things archaic. As such, it is appropriate that such laboratory analyses are carried out to the same specifications and quality standards to which the rest of the analytical community routinely

subscribe. The evaluation of measurement uncertainty is one such requirement which has become an inseparable part of chemical analysis, fundamental to method evaluation, development and comparison and enabling correct interpretation of data thus derived.

This chapter presents the principal approaches recommended for the evaluation of measurement uncertainty for chemical analysis and considers their applicability to evaluating an extensive archive of amino acid racemisation data collected over several years by the BIOARCH team.

The chapter begins with consideration of measurement uncertainty as a fundamental component of laboratory Quality Assurance, traceability and as a measure of fitness for purpose. Differences between essential concepts such as the error and uncertainty, accuracy and precision will be considered before taking an overview on the processes of uncertainty evaluation. These include the "bottom-up", uncertainty budget approach described by ISO's GUM (JCGM 100, 2008) the "top-down" inter-laboratory method validation approach as described by the Royal Society of Chemistry's (RSC) Analytical Methods Committee (AMC) (RSC Analytical Methods Committee, 1995) and the intra-laboratory method validation approach (Barwick *et al.*, 2000).

2.2 Accuracy and Fitness for Purpose

Before looking at how measurement uncertainty is evaluated, it is appropriate to first clarify some fundamental concepts and define terms that will be referred to later in the chapter. The text from which the definitions below are based is taken from the latest edition of the International Vocabulary of Metrology – *Basic and general concepts and associated terms* or VIM (Vocabulaire international de metrologie, (JCGM 200, 2008).

Measurements are never made without first having a purpose, perhaps to answer a question, solve a problem, ensure compliance or investigation. In practice, the result from a single measurement is unlikely to be the actual or 'true' value for that measurand, it is merely an estimate of it. Measurements are subject to errors and there will always be some doubt associated with a result. If this analysis was to be repeated, a slightly different value would most likely be obtained. If this analysis was to be repeated over and over again, the dispersal of the data representing the range of possible values for our measurand would represent the amount of doubt associated with our mean value. In order to interpret the data correctly any reported result needs to be accompanied by an indication of the level of doubt or uncertainty concerning that value in order to ensure the value is fit for its intended

purpose. Technical fitness for purpose is generally expressed as a statement of accuracy, (Ellison and Williams, 1998). Accuracy is defined in the VIM (JCGM 200, 2008), as;

> *(VIM 2.13)* ***measurement accuracy****; (accuracy of measurement; accuracy):- closeness of agreement between a measured quantity value and a true quantity value of a measurand.*
>
> *NOTE 1 The concept 'measurement accuracy' is not a quantity and is not given a numerical quantity value. A measurement is said to be more accurate when it offers a smaller measurement error.*
>
> *NOTE 2 The term "measurement accuracy" should not be used for measurement trueness and the term measurement precision should not be used for 'measurement accuracy', which, however, is related to both these concepts.*
>
> *NOTE 3 'Measurement accuracy' is sometimes understood as closeness of agreement between measured quantity values that are being attributed to the measurand.*

Accuracy is a qualitative concept made up of both the precision and trueness (bias) elements of the analytical method applied, reflecting both the random and systematic error effects respectively.

> *(VIM 2.14)* ***measurement trueness****; trueness of measurement; trueness:- closeness of agreement between the average of an infinite number of replicate measured quantity values and a reference quantity value.*
>
> *NOTE 1 Measurement trueness is not a quantity and thus cannot be expressed numerically, but measures for closeness of agreement are given in ISO 5725.*
>
> *NOTE 2 Measurement trueness is inversely related to systematic measurement error, but is not related to random measurement error.*
>
> *NOTE 3 Measurement accuracy should not be used for 'measurement trueness' and vice versa.*
>
> *(VIM 2.15)* ***measurement precision****; precision:- closeness of agreement between indications or measured quantity values obtained by replicate measurements on the same or similar objects under specified conditions.*

NOTE 1 Measurement precision is usually expressed numerically by measures of imprecision, such as standard deviation, variance, or coefficient of variation under the specified conditions of measurement.

NOTE 2 The 'specified conditions' can be, for example, repeatability conditions of measurement, intermediate precision conditions of measurement, or reproducibility conditions of measurement (see ISO 5725-3, 1994).

NOTE 3 Measurement precision is used to define measurement repeatability, intermediate measurement precision, and measurement reproducibility.

NOTE 4 Sometimes "measurement precision" is erroneously used to mean measurement accuracy.

Precision does not relate to a true or reference value, it depends only on the distribution of random error effects (ISO 3534: 3.14) (RSC Analytical Methods Committee, 2003a). The "specified conditions" in Note 2 relates specifically to the repeatability and reproducibility conditions of analysis (see Glossary) and imply different meanings to the interpretation of precision. In addition the reference to intermediate conditions of precision has been included in the VIM document to better reflect the conditions of analysis encountered by a single laboratory in routine internal quality control.

The contribution of random effects to the overall error cannot be anticipated and gives rise to increased variability in repeated analyses of a measurand, broadening the dispersion of results. Random error can be minimised by increasing the number of measurements taken. Note; The standard deviation of the arithmetic mean of a set of data is NOT a measure of the random error of the mean, rather, it is a measure of the uncertainty on the mean due to random effects (EURACHEM / CITAC, 2000).

The systematic contribution to the overall error is the component which remains constant or varies predictably over the course of a series of measurements and is often referred to as the bias. It affects all results in the same way and is independent of the number of analyses carried out. Constant systematic errors can be determined and results should be corrected accordingly using reference materials and standard solutions, to correct for recovery or recalibration to bring the system back into analytical control. Note; the uncertainties of these standards and the uncertainty of the correction must be taken into account (EURACHEM / CITAC, 2000). "*The uncertainty of a correction applied to a result to compensate for bias, is NOT systematic error. It is the uncertainty of the result due to*

incomplete knowledge of the required value of the correction" (JCGM 100, 2008 p5, note to 3.2.3)

Note that in ISO 3534-1; Statistics – Vocabulary and Symbols (cited in (RSC Analytical Methods Committee, 2003a), Trueness is described as "*closeness of agreement between the average value obtained from a large series of test results and an accepted reference value*" (3.12), whilst bias is given as "*the expectation of the test results and an accepted reference value*" (3.13). Thus bias is the opposite equivalence of trueness, trueness is the absence of bias (RSC Analytical Methods Committee, 2003a), see Figure 2.1.

In considering the definition of accuracy, De Bievre (2006, p654) observes that the reference to the '*True value*' may be an impractical or even misleading concept.

> "*Moreover, the insight is growing in an increasing part of the measurement community that one cannot determine the 'true value' as a matter of principle. [Thus if we] cannot know what 'reality' is……a model of reality is a less ambitious but better concept*, "

He suggests that to define accuracy by its '*true value*' is to suggest that the inaccuracy of the true value can be determined. "*If we could determine (in)accuracy quantitatively, then we could calculate the 'true value' from our measurement results!"* (De Bièvre, 2006), (which we cannot!). In attempting to resolve this conundrum he proposes two alternative possibilities, either we create a mental construct and define a conventional truth (in order to perpetuate the need for exactness and the truth) or we adopt the concept of measurement uncertainty. If accuracy (precision plus trueness or bias) is a characteristic of the measurement result, "*measurement uncertainty is a characteristic of the process leading to its statement*", i.e., the measurement's statement of accuracy, and that requires critical evaluation of the process using the skill and expertise of the analyst. De Bievre continues;

> "*And evaluation is a process. A process of thinking, not a characteristic. Measurement uncertainty conveys more correctly the slight doubt which is attached to any measurement result. Thus a doubtful meaning of 'accuracy' (doubtful because tied to 'true value') is replaced by a practical one: 'measurement uncertainty'*". (2006, p 645)

Thus in summary, the purpose of measurement uncertainty is to evaluate a measurement process, and combine the effect of all error contributions into a single value as an indication of accuracy, within a specified level of confidence (NMS, accessed 2009a). This process is summarised in Figure 2.2.

Figure 2.1: The influence of precision and Trueness on Accuracy and Uncertainty

(after RSC 2003, AMC Technical brief No 13)

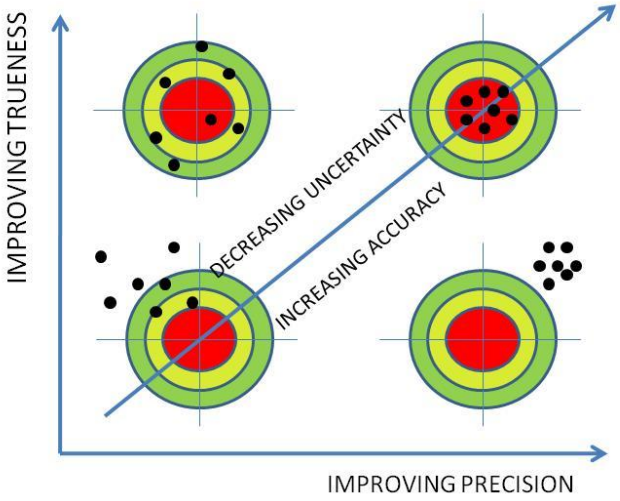
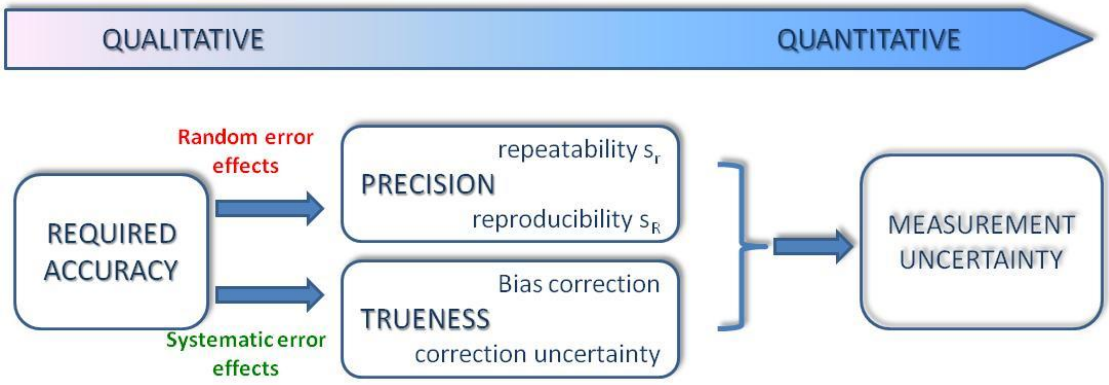


Figure 2.2: Measurement Uncertainty as a function of Accuracy.



2.3 Defining Measurement Uncertainty

Measurement uncertainty is a range in which the true value is most likely to lie and is represented as a standard deviation. It is the measure of the effect of analytical error on the measurement result. Measurement uncertainty cannot correct for analytical errors, it merely provides a means for quantifying their effect.

(VIM 2.26) ***measurement uncertainty****; (uncertainty of measurement; uncertainty):-non-negative parameter characterizing the dispersion of the quantity values being attributed to a measurand, based on the information used*

NOTE 1 Measurement uncertainty includes components arising from systematic effects, such as components associated with corrections and the assigned quantity values of measurement standards, as well as the definitional uncertainty. Sometimes estimated systematic effects are not corrected for but, instead, associated measurement uncertainty components are incorporated.

NOTE 2 The parameter may be, for example, a standard deviation called standard measurement uncertainty (or a specified multiple of it), or the half-width of an interval, having a stated coverage probability.

NOTE 3 Measurement uncertainty comprises, in general, many components. Some of these may be evaluated by Type A evaluation of measurement uncertainty from the statistical distribution of the quantity values from series of measurements and can be characterized by standard deviations. The other components, which may be evaluated by Type B evaluation of measurement uncertainty, can also be characterized by standard deviations, evaluated from probability density functions based on experience or other information.

NOTE 4 In general, for a given set of information, it is understood that the measurement uncertainty is associated with a stated quantity value attributed to the measurand. A modification of this value results in a modification of the associated uncertainty.

Measurement uncertainty can arise from a number of sources, sampling inhomogeneity, inaccurate weighing or volume measurement, uncertainty of reference materials, matrix interference, instrument sensitivity, analyst bias, temperature effects, etc, etc. It is not always possible to measure or correct for such influences but without knowledge of measurement uncertainty and the reliability of data, it is difficult to draw meaningful interpretations, make appropriate comparisons or ensure compliance with legislative limits. In short, knowledge of the measurement uncertainty does not create doubt about the validity of the measurement result, rather it provides confidence that the data is fit for its intended purpose.

2.3.1 Measurement Uncertainty and Measurement Error

Uncertainty and error are often mistakenly used synonymously in the literature. Uncertainty should never be considered to represent the error. "*Error is an idealised concept and cannot be known exactly*" (EURACHEM / CITAC, 2000). It is also perfectly possible for the result after applying a correction, to be close to the (theoretical) *'true value*' and have a negligible error but for the uncertainty to remain very large due to the associated doubt of the analyst.

In a more recent editorial, De Bievre (2008 p429) observes that whilst the concept of determining analytical measurement error was replaced by the formal introduction of measurement uncertainty with the arrival of the GUM (1993), over fifteen years ago, there remains a reluctance to convert to more current thinking and asks whether "*.... 15 years is still too short for such a change of paradigm?*". He concludes,

> " *The (r)evolution from thinking in terms of **error** (deviation from a presumed 'true value') to **doubt** about the degree of knowledge of a measurement result, occurred around the millennium change: one could point to it as having occurred in the year 2000 ± 10.*" (Ibid, p 430)

2.3.2 Fitness for Purpose and Quality Assurance

We have seen in the above section how measurement uncertainty can be used as a quantitative expression of accuracy and how this can provide confidence in the fitness for purpose of the analytical result for its intended use. Clearly if results are unreliable there are also financial implications for the laboratory to take into account either through the risk of non-payment or the expense of repeating the analysis (Thompson and Fearne, 1996; Marschal, 2004; RSC Analytical Methods Committee, 2008b).

However, in order for the analyst to arrive at this point it is essential that the method used is capable of producing a result of suitable accuracy, and is indeed fit for its intended use. Fitness for purpose of the analytical method within the laboratory is demonstrated by the control of these influencing factors. It is therefore important that a framework is established that will verify the analysis is being performed under analytical control and provide both to the analyst and any third party, assurance of quality. Such measures are normally implemented within the context of a Quality Management System and encompass Quality Assurance and Quality Control procedures to ensure process stability. Quality assurance is the overarching system which plans and documents the processes involved in

ensuring a quality product. Quality Control refers to the activities carried out that ensures the quality of the routine processes (NMS, accessed 2009b).

> *"Thus fitness for purpose tells us how much uncertainty is acceptable, quality assurance ensures that an acceptably small uncertainty is achievable and internal quality control demonstrates that the sufficiently low uncertainty is actually achieved."* (Thompson, 1995 p117N*)*.

A consensus of general requirements ensuring the competence of analysis, have long been recognised (Mesley *et al.*, 1991; NMS, accessed 2009c). These include;

- use of validated methods
- properly maintained and calibrated methods
- the use of reference material to calibrate methods
- effective internal Quality control (control charts, etc)
- participation in inter-laboratory check sample schemes
- independent audits of quality control procedures
- external assessment by accreditation or other compliance schemes
- properly trained staff

However, whilst the need to account for the effect of errors was covered by the use of validated methods to determine precision, the concept of measurement uncertainty wasn't introduced as a separate requirement until 1993 with the introduction of the ISO guide to uncertainty measurement (GUM) and took a further six years before it became embedded in laboratory protocol with the publication of the accreditation standard ISO/IEC 17025 (2005) – General Requirements for the Competence of Testing and Calibration Laboratories. Laboratory accreditation to ISO 17025 formalises the requirements for analytical assurance, into a Quality Management Framework. These are based on internationally agreed principles covering both management and technical aspects of laboratory competence. Technical requirements include;

- Personnel
- Accommodation and Environmental conditions
- Method selection; - validation and uncertainty
- Equipment
- Measurement Traceability
- Sampling
- Handling of test and calibration items

- Assuring the quality of test and calibration results;
  - quality control and proficiency testing
- Reporting the results

Through inference, if the technical aspects of analysis listed above are necessary to evaluate and control, it follows that these are also the processes that could potentially contribute most influence on the measurement result, introduce the largest error effects and make the most contribution to the uncertainty estimate. It is therefore appropriate to look at these processes in more detail to better understand the potential sources of uncertainty in analysis.

2.3.2.1 Personnel

Competence of staff undertaking the analysis is fundamental to the quality and accuracy of the measurement result and in controlling systematic, random and gross error effects. Different analysts will get different results through different interpretations of the instructions (only from a poorly written SOP), personal bias in reading equipment and instruments such as the meniscus on a graduated pipette, different reaction times, colour judgement etc. It is not always possible to account for the differences between analysts but these effects will increase the variability of the data even when the method is under statistical control and will contribute to the **intermediate precision** of the method.

2.3.2.2 Accommodation and Environmental conditions

The conditions in which the analysis is carried out can have a significant impact on the overall contribution to uncertainty. Not only does this include the physical storage conditions in which the material is kept but also the environmental conditions where the analysis is carried out. This primarily involves issues of temperature, humidity, light, air pressure, ventilation, risk of contamination etc.

Inappropriate storage of material can have a direct affect on the stability of a matrix and or analyte itself but an indirect effect on the measurement of the analyte through the temperature effect on glassware for example, invalidating its calibration.

The incomplete definition of a measurand such as specifying ‘room temperature’ without a specified temperature value and acceptable tolerances, can have a major influence. Insufficient knowledge of the effects, imperfect measurement or uncontrolled environmental conditions can make significant contributions to between-run variability. This can affect long-term **intermediate precision** but also introduce **laboratory bias.**

2.3.2.3 Method selection, validation and uncertainty

Appropriate method selection and definition is essential prior to analysis, for example is it organic or inorganic arsenic required as the measurement result, free or total hydrolysable amino acids etc. Having determined the analyte required it is then necessary to define any specific criteria, precisely, such as total hydrolysable amino acids at 110^{o}C, 140^{o}C or 180^{o}C and for how long, concentrations of reagents etc., can all have serious consequences on the measurement result due to **method and laboratory bias** and associated uncertainty. The effect of imperfect realisation as discussed above should also be evaluated. The effect of sample preparation such as any pre-treatment, grinding, sieving, drying, extraction, digestion and extraction phases, incomplete recovery of the analyte or different recoveries between reference standard and matrix, all need to be evaluated during the development and validation of the method. However, variations in repeated measurements under apparently identical conditions (**repeatability**) will still occur, due to unaccounted for random effects, instrument noise, environmental fluctuations and fluctuations in homogeneity. Sample effects due to the matrix such as assumed stoichiometry, incomplete reactions or interference from side reactions, changes in matrix stability and other uncontrollable (and often indefinable) random error effects, will all contribute uncertainty to the final value, This also includes uncertainty in reference materials or inaccuracies in assumed constants and rounding effects.

Thus, in light of all these potential sources of error and in order for the chosen analytical procedure to be deemed of sufficient quality and fit for purpose, the method will have had to have undergone validation. Validation is the process whereby an analytical method is evaluated to determine the exact limits and range of applicability and define working parameters. It is of preeminent importance that the method has undergone validation prior to its establishment as a routine method, as the determination of measurement uncertainty assumes two fundamental prerequisites;

> "*a) that a validated (characterised) method is used for the determination, and b) an assurance that the material analysed falls within the scope of the method validation. If these criteria cannot be fulfilled, it is unlikely that a meaningful uncertainty can be associated with a measurement."(RSC Analytical Methods Committee, 1995)*

The parameters defined in validation include; traceability (to reference materials and calibration of equipment), sensitivity, selectivity / specificity, limit of detection (LOD), limit of

quantification (LOQ), linearity, precision (defined as repeatability, internal reproducibility (intermediate precision) or reproducibility), trueness (method and laboratory bias), measurement uncertainty, ruggedness/robustness, establishment of QC limits.

2.3.2.4 Equipment

Random instrument fluctuations and systematic limitations in the measuring equipment used such as calibration limits of a balance etc, temperature control with defined specification, auto-analyser carry over effects, finite resolution of discrimination thresholds, peak overlap errors, LOD and LOQ, graduation of scale and effect of heat on volumetric glassware causing changes in equipment characteristics and instrument performance since the previous calibration will all contribute to the variability of **repeatability** and **intermediate precision** evaluation or a systematic **laboratory bias**. Other aspects such as ownership and use should be considered especially if it is being borrowed or used by other analysts together with maintenance, service and calibration requirements.

2.3.2.5 Measurement Traceability

One of the critical attributes of valid analytical measurement is the concept of comparability, with other data produced within the same laboratory, between laboratories, between different methods for method development, to ensure compliance with legal standards etc. Comparability is demonstrated through traceability back to international standards through an unbroken chain of reference. This is usually achieved through calibration of laboratory equipment and instrumentation during the validation stage, ensuring that the values generated by the measurement system and specified conditions are related back to reference materials.

Quality issues that affect this process and consequently impact on measurement uncertainty include reagent purity and uncertainty of reference values (e.g. certified reference material (CRM) specifications). These affect **laboratory bias,** whereas instrument drift between runs or between calibrations, introduce variability to **intermediate precision**. Computational effects such as using a straight line calibration on a curved response, leads to poor fit and higher uncertainty. In addition, non-certified reference materials used to spike samples and internal standards used in the determination of recovery and quality control charts are all add to the uncertainty, including volumetric solutions which will have doubt associated with the assay of the concentration values.

2.3.2.6 Sampling

Sampling can introduce significant uncertainty and depending on the requirements may be excluded from the validation process and protocol if this is beyond the control of the analyst, analysis is therefore carried out on the sample as received. However, often sub-sampling from the material supplied is a requirement and issues surrounding provenance, sample bias and representativeness, poor homogeneity and contamination, need to be carefully considered. Clearly sampling affects precision under **repeatability** and internal **reproducibility conditions** and if unrepresentative can introduce **laboratory bias** into the overall system.

2.3.2.7 Handling of test and calibration items

This primarily involves the sampling, transport, receipt and storage of materials to protect against deterioration, damage, contamination and instability. Critical stages should be included in the validation process. Overlapping with these considerations are operator influences in sub-sampling, reading instruments, reaction times, interpretation of instructions, dilutions and weighing errors, storage conditions and stability etc.

2.3.2.8 Assuring the quality of test and calibration results (QC and proficiency testing schemes)

Having validated a method, it is important to monitor its stability to ensure it remains under statistical control during routine use and that analyses are carried out within the scope of validation to ensure reliable data. Uncertainty estimates should only be made when the method is in statistical control i.e. the performance is consistent with that established during method validation (including the use of calibration and control charts).

Fundamental to this is the establishment of internal quality control (IQC) and external quality control (EQC) procedures. IQC includes the use of blanks, calibrants, reference materials, quality control materials, spiked samples or internal standards, replicate analyses, control charts etc. EQC includes participation in proficiency testing schemes and possibly inter-laboratory studies / collaborative trials. IQC enables the measurement system to be monitored on a routine basis and will flag up anomalies and non-conforming behaviour affecting **repeatability** and **intermediate precision**. Sometimes unrecognised systematic effects exist but can't be accounted for, EQC is a mechanism that enables comparison with other laboratories and permits the monitoring of **laboratory bias** and **method bias**, not otherwise possible

2.3.2.9 Reporting the results

Results need to be reported in such a way so as to be readily understood and interpreted by the end user. Measurement results should always be accompanied by a statement of accuracy, usually given as an uncertainty value with a defined confidence level quoted. It is essential that the uncertainty is a true reflection of all the above mentioned potential contributions of doubt, which, clearly can be a very daunting prospect for the analyst to be able to account for and quantify all the contributions.

Because of these difficulties, ISO (International Standards Organisation) published a set of guidelines to assist in this task, now known as the GUM (JCGM 100, 2008). However, as illustrated by the above simple review of potential sources of error effects, the task has often been criticised as being unmanageable within a chemical laboratory due to the often lengthy and complex procedures involved. Consequently other different approaches have been proposed, in keeping with the principles of the GUM, but more applicable to the procedures commonly encountered in an analytical laboratory, utilising data derived from validation studies and collaborative trials.

The following sections will now look in more detail at some of the shared processes common to all methods together with the individual quantitative approaches recommended by the GUM and other alternative sources.

2.4 Measurement Uncertainty Evaluation

In principle, the process involved in evaluation of measurement uncertainty is straight forward. The GUM identifies the following steps (JCGM 100, 2008);

- Specify Measurand
- Identify Uncertainty Sources
- Quantify Standard Uncertainty components
- Evaluate combined uncertainty
- Evaluate expanded uncertainty
- Report uncertainty

With the exception of step 3, all steps in the process are common to all laboratory based approaches of uncertainty evaluation, and will be looked at in more detail in the following Chapters. However it is the actual process of uncertainty quantification that has caused most conflict within the analytical community. Consequently a number of alternative

approaches have arisen, primarily that of utilising existing data from validation (Barwick *et al.*, 2000), inter-laboratory comparisons (RSC Analytical Methods Committee, 1995) or intra-laboratory ANOVA methods (ISO 21748, 2010).

2.4.1 Quantifying Standard Uncertainty components

Original guidelines for the evaluation of measurement uncertainty were developed by metrologists and physicists. The guidelines were accepted uneasily by analytical chemists as being too technical, inappropriate and too complex to administer to often lengthy and multi layered analytical processes (Lira, 2002). To assist the analytical community EURACHEM interpreted the GUM as it came to be known, for analytical chemistry and published a more practical version with worked examples in their own guide, Quantifying Uncertainty in Analytical Measurement in 1995. The ISO or GUM approach is based on developing a comprehensive mathematical model of the entire measurement procedure and evaluating the uncertainty contributions associated with every input quantity both individually and combined into a single expression. This approach views the propagation of uncertainty from the grass roots, often referred to as the "bottom-up" approach and accounts for each uncertainty contribution at source, hence it is also sometimes referred to as an "uncertainty budget" approach. Whilst the GUM allows for other approaches to be utilized, the modelling approach marked a significant shift in perception regarding error treatment and expression and has become unequivocally associated with the "bottom-up", uncertainty budget, modelling or simply the GUM approach.

Since then a number of alternative or "empirical" methods of evaluation have been described such as a factorial approach (Julicher *et al.*, 1999; Hill and von Holst, 2001b, 2001a). Those that have received greatest attention typically propose "whole method" approaches based on method performance indicators from studies designed to encompass as many effects from uncertainty sources as possible. These data can either be derived from inter-laboratory (between laboratories) or intra-laboratory (single laboratory) method validation studies and takes an overall view of the effect of uncertainty on the analytical data. The characterization of method performance parameters (repeatability and reproducibility) through collaborative trials have long been recognized (Wernimont, 1985; ISO 5725, 1994; Parts 1-6). The same year that Eurachem published their original guide to the GUM (EURACHEM, 1995), The Analytical Methods Committee (AMC) of the Royal Society of Chemistry, published their own "top-down" approach to uncertainty measurement (RSC Analytical Methods Committee, 1995) based around the collaborative trial design and

discussing the applicability of in-house method validation and internal QC and external QC. Single laboratory or in-house method validation measurements of precision and bias are generally routine and familiar activities to a large number of laboratories. Thus, in (1999), Maroto *et al*. proposed an intra-laboratory approach, followed shortly afterwards by a laboratory protocol for measurement uncertainty based on experimental design and method validation (Barwick and Ellison, 2000a; Barwick *et al.*, 2000). It is this later approach that is often favoured by analytical chemists, as it requires little additional effort, time or money. More recently ISO published a guidance document (ISO/TC 21748:2004) for the use of repeatability and reproducibility and trueness estimates, linking both the inter- and intra-laboratory approaches as a unified "top-down" approach. In addition, suggestions utilizing results from external quality control activities such as proficiency tests have also been proposed (Magnusson *et al.*, 2004; EUROLAB, 2006, 2007).

The intra-laboratory approaches are the subject of Chapters 3 and 4. Results from a proficiency test (PT) are summarised in Chapter 5. These have been used to provide a combined estimate of uncertainty after the Nordtest Report TR537 (Magnusson *et al*., 2004) and the EUROLAB reports. However, due to word restrictions in the presentation of this thesis, results have been provided as separate appendices to Chapter 5, which include copies of the PT reports and a draft paper currently in preparation for submission.

Figure 2.3: Routes for measurement uncertainty determination

(after Désenfant and Priel, 2006 and; EUROLAB, 2007)

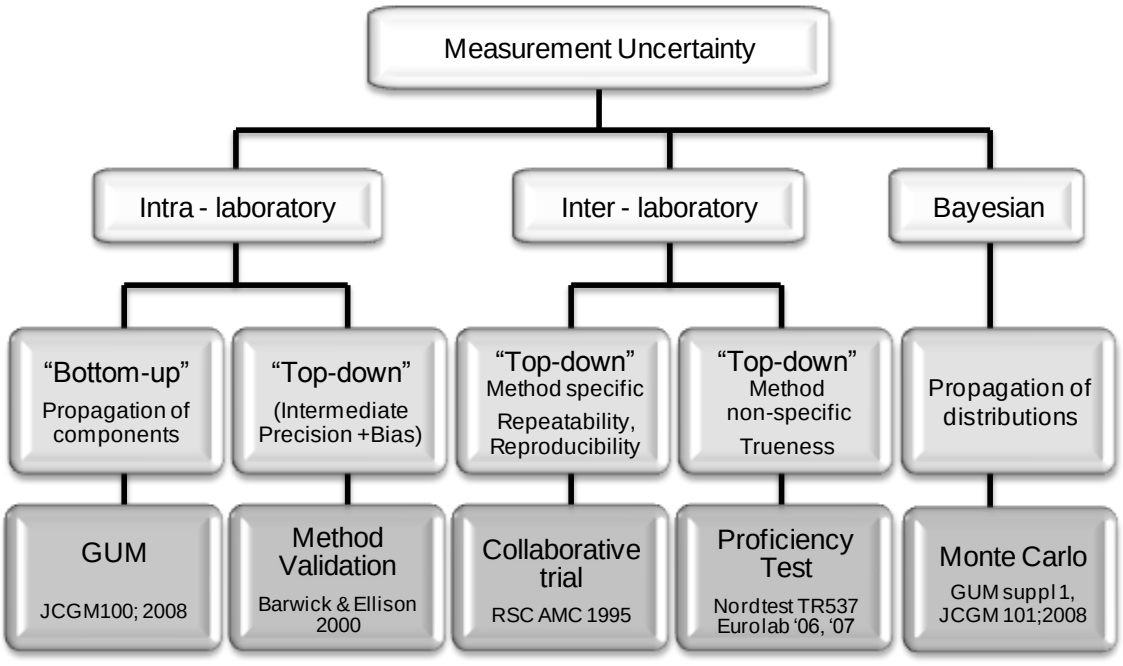


Figure 2.3 illustrates the various possible routes to determining MU. However, one final approach to MU evaluation is shown in Figure 2.3 but has not so far been mentioned. This uses a Bayesian statistical approach to model the propagation of theoretical uncertainty distributions, and include evaluations known as Monte Carlo methods. Although these techniques are not new to archaeology (Buck *et al.*, 1996), they have only been introduced into the realms of measurement uncertainty in analytical chemistry, relatively recently, as a supplement to the original GUM document (JCGM 101, 2008). Bayesian analysis is given further consideration in Chapters 6 and 7, but for now, the remainder of this Chapter will focus on the classical approaches.

2.5 The Modelling Approach, (Uncertainty Budget, "Bottom-up" or GUM Approach)

The GUM recommended approach (JCGM 100, 2008), requires all uncertainty components to be expressed in the same form, as a standard deviation, prior to combination and expansion. Standard deviations of uncertainty components are referred to as standard uncertainties, '*u*'. However, different sources of uncertainty can report their uncertainty component in different ways.

Thus **Type A** uncertainty estimates tend to "...*be evaluated from the statistical distribution of the results of series of measurements and can be characterised by standard deviations*" whilst **Type B** uncertainty estimates tend to be derived by other means such as certificates and "....*are evaluated from assumed probability distributions based on experience or other information*." are which are considered equivalent to the corresponding standard deviation (EURACHEM / CITAC, 2000, p4, 2.1.1)

2.5.1.1 Type A evaluation of standard uncertainty (after JCGM 100, 2008, p10, 4.2)

The best estimate of the true or expected value μ_q, of a quantitiy *q,* which is a random variable, for which *n* independent observations have been taken under repeatability conditions, q_k is the arithmetic mean, $\bar{q}$.

$$\bar{q} = \frac{1}{n}\sum_{k=1}^{n} q_k \qquad \textbf{(2.1)}$$

The influence of random error effects result in variation of individual observations of q_k. This experimental variance $s^2(q_k)$, estimates the variance σ^2 of the probability

distribution of *q,* and together with the positive square root $s(q_k)$, giving the experimental standard deviation, defines the variability or dispersion of observed values q_k about the mean. Thus*;*

$$s^2(q_k) = \frac{1}{n-1}\sum_{j=1}^{n}(q_j - \bar{q})^2 \qquad \textbf{(2.2)}$$

However, the standard deviation required as the standard uncertainty is that of a single value, thus if the estimate of q is taken as the mean, the best estimate of the variance of the mean is given by;

$$s^2(\bar{q}) = \frac{s^2(q_k)}{n} \qquad \textbf{(2.3)}$$

The best estimate of the standard deviation of the mean is given by;

$$s(\bar{q}) = \frac{s(q_k)}{\sqrt{n}} \qquad \textbf{(2.4)}$$

The standard deviation of the mean $s(\bar{q})$ describes how well $\bar{q}$ estimates the true or expected value μ_q and can be used as the measurement uncertainty of $\bar{q}$ such that $s(\bar{q}) = u(\bar{q})$.

To ensure that the deviation of the sample $s(\bar{q})$ provides a reliable estimate of the true or expected population standard deviation (σ_q), *n* must be large. The difference between $s(\bar{q})$ and σ_q needs to be taken into account when calculating confidence limits through use of the t-distribution to accommodate smaller *n* values and ensure the sample data approximate to a normal distribution.

2.5.1.2 Pooled experimental standard deviation

Where data is available from a series of repeated measurements, a pooled experimental variance, s_p^2 , or standard deviation, s_p, may better represent the dispersal of the mean, such that $u = s_p/\sqrt{m}$; where m = total number of independent observations (ie, $n_1 + n_2 ... + n_i$).

$$s_p^2 = \frac{\sum_{i=1}^{N}((n_i - 1)s_i^2)}{\sum_{i=1}^{N}(n_i - 1)} \qquad \textbf{(2.5)}$$

Where s_i^2 is the variance of the ith series of n_i independent repeated observations.

2.5.1.3 Type B evaluation of standard uncertainty (after JCGM 100, 2008, p11, 4.3)

Type B standard uncertainties are evaluated from certificates or specification sheets, literature or assumed, *a priori* probability distributions and experience, or very simply anything else that doesn't constitute a type A, statistical evaluation from a series of observations. Guidance in the GUM emphasises that Type B evaluations of uncertainty should be considered equally reliable as Type A, especially where a Type A evaluation is based on comparatively small number of observations.

Uncertainty can be reported in a number of different ways and will need to be converted to a standard uncertainty format. The following examples are based on those presented in the GUM, p11, section 4.3,

a. If the uncertainty is given as a confidence limit or interval.

Example; the concentration of a standard solution quoted by a supplier as 1000 ±3mg/L at 95% confidence.

Conversion: divide the half range (± value), by the appropriate student t-value if degrees of freedom are known, otherwise assume a value of 1.96 for 95% CI, e.g. 3/1.96 = 1.53 mg/L (1.64 for 90% and 2.58 for 99% (GUM 2008, 4.3.4)).

b. If the uncertainty is given as an expanded uncertainty,

Example; a certified reference material (CRM) quotes a concentration of 1000 ±3mg/L, representing the half width of the expanded uncertainty, calculated from a coverage factor *k*=2, or at the 2 standard deviation level, giving a level of confidence approximating to 95%.

Conversion: divide the half range (± value), by the stated coverage factor, e.g. 3/2 = 1.5mg/L.

c. If a stated range is given whereby the true value is equally likely to occur across the entire range, the probability that the value lies within the interval a^- to a^+ is 1, and describes a rectangular distribution with the probability of the value falling outside the range is zero. In This case, x_i, the expectation (of the expected value for X_i), is the midpoint of the interval, $x_i = (a^- + a^+)/2$ and the variance is;

$$u^2(x_i) = (a^+ - a^-)^2/12$$

If then the difference between a- and a+ is equivalent to 2a, then;

$$u^2(x_i) = (2a)^2/12 \equiv a^2/3$$

And the standard uncertainty expressed as a standard deviation is given by;

$$u(x_i) = a/\sqrt{3}$$

Example; the purity of a substance used to prepare a calibration standard is given as 99.9 ±0.1%.

Conversion; assume a uniform or rectangular distribution and divide the purity uncertainty by $\sqrt{3}$, e.g. 0.1/$\sqrt{3}$ = 0.058%.

d. If the stated range where the values closest to the mean are more likely than those at the extreme, the distribution is described as triangular;
Example; manufacturer's tolerance for a volumetric flask is given as 100 ±0.8 mL.

Conversion; assume a triangular distribution and divide the tolerance by $\sqrt{6}$, e.g. 0.8/$\sqrt{6}$ = 0.33 mL.

e. Given as a probability;
Example; there is a 50:50 chance that the value lies between the interval defined by – a to +a, that is a 0.5 or 50% probability that a result measures 10.11 ±0.04mm in length.
Conversion; assume a normal distribution such that the best estimate of X_i is the mid-point and the half width interval is denoted by $a = (a^+ - a^-)/2$, with expectation μ and standard deviation σ, 50% of the interval is denoted by $\mu \pm \sigma/1.48$, $u(x_i) = 1.48a$, therefore $u(l) = 1.48\text{x}0.04 = 0.06mm$.

When considering Type B evaluations it is important not to double count uncertainty components, i.e. where any Type B effect does not already contribute to the variability of observations already accounted for in the statistical evaluation of Type A uncertainty.

Having evaluated and expressed as standard deviations each of the uncertainty contributions included in the model, the next stage in the ISO GUM approach to uncertainty estimation is the combination of these contributions into a single value. This process is considered further in section 2.8. However, this approach to uncertainty estimation has received much criticism from the analytical community as being overly complex and unwieldy and not representative of what actually happens in routine analysis. For this reason the "top-

down" approach was published in 1995 (RSC Analytical Methods Committee) making use of collaborative trail method performance characteristics.

2.6 Inter-laboratory Collaborative Trial or "top down" method

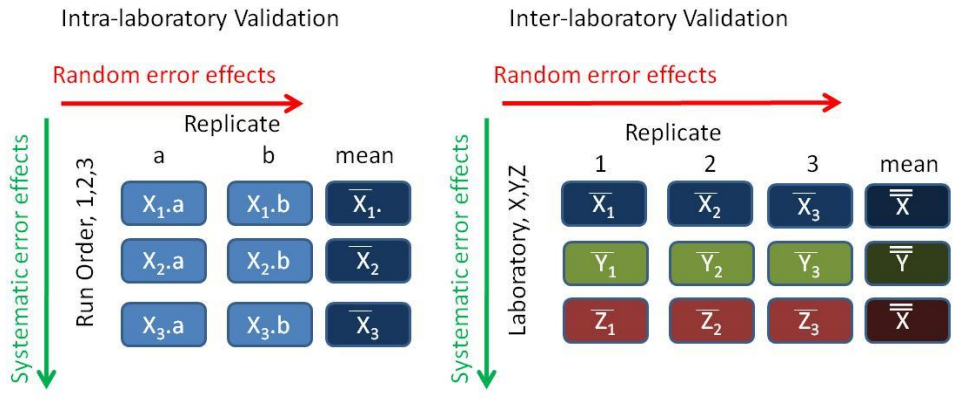
An alternative approach to the ISO "bottom-up" approach is the method proposed by the Analytical Methods Committee (AMC, 1995) of the Royal Society of Chemistry, the so called "top-down" approach. Based on the principles of inter-laboratory studies or collaborative trails to formally validate analytical methods, (the use of validated methods is a fundamental pre-requisite for valid uncertainty measurement). The theory behind this is to view the laboratory from a "*higher level", i.e.., as a member of a population of laboratories*" (RSC Analytical Methods Committee, 1995, p2304), so that random and systematic error effects in a single laboratory become random error effects between laboratories when seen from a "*higher perspective*" which can be more simply evaluated (see Figure 2.4**Error! Reference source not found.**).

More often than not the emphasis of analysis has tended to focus on precision elements of accuracy, trying to minimize the variability between observations by reducing the between-run influences and attempting to control random error effects, often at the expense of trueness, influenced by systematic bias. The accuracy of any analytical measurement x, can be shown as;

$$x = X_{true} + \delta_{method} + \delta_{lab} + \delta_{run} + \varepsilon \qquad \textbf{(2.6)}$$

Where; X_{true} is the (theoretical) true value, δ_{method} is the method bias, δ_{lab} is the laboratory bias, δ_{run} is the between-run bias and ε is the random measurement component (RSC Analytical Methods Committee, 1995)

Figure 2.4: Relationship between Intra- and Inter-laboratory Random and Systematic Error Effects



The systematic uncertainty within a single run is a fixed level, but when viewed as one of a number of successive runs, it becomes a random variable with variance σ_{run}^2. Similarly for a particular laboratory, the bias is fixed but when seen as one of a number of laboratories, again it becomes a random variable with variance σ_{lab}^2. However, method bias is not quite so easily handled since the purpose of most inter-laboratory trials is to determine method specific parameters, not compare methods.

However the uncertainty of the method bias can be measured from the use of reference materials that have a defined value and associated uncertainty, $X_a \pm u_a$. The method bias is an estimate of the difference between the consensus or assigned value for the reference material analysed by laboratories and the certificated value, $\bar{x} - X_a$, with a standard deviation $\sigma(\bar{x} - X_a)$, thus the standard uncertainty of the method bias, u_b, is given by;

$$u_b = \sigma(\bar{x} - X_a) = \sqrt{\sigma_{\bar{x}}{}^2 + u_a^2} \qquad \textbf{(2.7)}$$

Thus if the variance of the reference material is small compared to the variance of the assigned value (less than one tenth (RSC Analytical Methods Committee, 1995), and the assigned value is close to the certified value, then u_a can be omitted from the uncertainty calculation for method bias which simply becomes $\sigma_{\bar{x}}$.

u_b already contains the uncertainty of the assigned value, so the overall uncertainty for a single measurement x becomes;

$$u_x = \sqrt{\sigma_\varepsilon{}^2 + \sigma_{run}{}^2 + \sigma_{lab}{}^2 + u_b^2} \qquad \textbf{(2.8)}$$

For empirical methods, i.e. method defines the analyte, the (theoretical) true value becomes the consensus or assigned value, thus the method bias and its associated uncertainty become zero, giving;

$$u_x = \sqrt{\sigma_\varepsilon{}^2 + \sigma_{run}{}^2 + \sigma_{lab}{}^2} \qquad \textbf{(2.9)}$$

In circumstances where the uncertainty only needs to be determined within any single laboratory for its own purposes, laboratory bias can also be discounted, giving;

$$u_x = \sqrt{\sigma_\varepsilon{}^2 + \sigma_{run}{}^2} \qquad \textbf{(2.10)}$$

Re-interpreting the above in terms of the parameters defined by a collaborative trail; the method bias is often discounted as by definition the method is empirical and being

evaluated, repeatability standard deviation is a measure of the random error effects so $\sigma_r = \sigma_\varepsilon$ and reproducibility standard deviation is a measure of the overall accuracy of the trial, $\sigma_R^2 = \sigma_L^2 + \sigma_r^2$ where $\sigma_L^2 = \sigma_{run}^2 + \sigma_{lab}^2$ and is a measure of the between laboratory variability, (σ_{run}^2 is rarely evaluated in collaborative trails). Thus, not including method bias, σ_R is a single measure of the variability or uncertainty of the measurement procedure at all levels, including the often neglected laboratory bias (ISO 21748, 2010)

$$\sigma_R = \sqrt{{\sigma_L}^2 + {\sigma_r}^2} \quad \textbf{(2.11)}$$

Values for S_R, S_L, and S_r, are obtained by a one way analysis of variance (ANOVA).

$$s_r = \sqrt{within\ group\ mean\ square} \quad \textbf{(2.12a)}$$

$$s_L = \sqrt{\frac{between\ group\ mean\ square - within\ group\ mean\ square}{n}} \quad \textbf{(2.12b)}$$

$$s_R = \sqrt{s_r^2 + s_L^2} \quad \textbf{(2.12c)}$$

Further details on the calculations of s_R, s_L and s_r are given in ISO 5725 (ISO 5725, 1994) and their use in measurement uncertainty in ISO 21748.

The reproducibility standard deviation (s_R) is often used relative to the concentration of the analyte in question, i.e. relative standard deviation of reproducibility, RSD_R , When such data is available from collaborative trials this value can be used directly as the combined standard uncertainty (EURACHEM / CITAC, 2000). When uncertainty estimates are taken from previous inter-laboratory studies, it is necessary to demonstrate that the method as carried out in the laboratory is capable of achieving comparable precision and that the bias data remains justified, determined by measurement of bias through the analysis of appropriate reference materials, recovery analysis or proficiency testing. It is also necessary to demonstrate that the measurement procedure remains in statistical control using regular QC samples. Where these conditions are met and the method is being operated within its scope of validation and field of application, it is acceptable to apply reproducibility data from previous studies directly to uncertainty estimates in the laboratory.

ANOVA methods have now been described for single laboratory applications (ISO 21748, 2010).

2.7 Single Laboratory Method Validation Approach

So far we have considered the ISO "bottom – up" approach which requires an exhaustive account of individual uncertainty components of the measurement process. Alternatively the AMC "top-down" approach still requires all the contributory factors to be taken into account, but takes an overview of the process and evaluates the output of analysis rather than individual inputs. The former approach has been criticized as being difficult to apply on a routine basis whilst the latter is costly to organize and coordinate in regards to time and money, is often inflexible regards to concentration and matrix specificity and unresponsive to the needs of method development and improvement. As a compromise, a third approach to uncertainty evaluation has been proposed (Maroto *et al.*, 1999; Barwick and Ellison, 2000a; Barwick *et al.*, 2000). This approach takes advantage of utilising the simpler "top-down" perspective but applied to the evaluation of accuracy parameters at individual laboratory level and allows for in-house method validation data, often routinely carried out by competent staff, to be used, with little additional effort, time or cost.

Intra-laboratory (in-house) method validation and Quality Control (QC) activities are principal requirements which ensure that the data that are released are fit-for–purpose. Validation is usually a one off activity or carried out at infrequent intervals and provides information about the expected performance of the method. QC provides a way of observing that performance over a period of time. Evidence of validation is a requirement prior to accreditation of the specific method which defines the scope (matrices, analytes, concentration range), of the method's applicability (ISO / IEC 17025, 2005). Method stability and statistical control are prerequisites for uncertainty measurement, without which the evaluation of uncertainty is a pointless exercise and has no meaning, thus validation and QC are fundamental to a laboratory's routine activities.

2.7.1 Method Validation

Typically, method validation characterizes the performance of a specific method with regard to "...*applicability, selectivity, calibration, trueness, precision, recovery, operating range, limit of quantification, limit of detection, sensitivity and ruggedness*" (Thompson *et al.*, 2002, p839). The relationship between uncertainty and random (within-run measurement precision, i.e. repeatability) and systematic (run, laboratory and method bias) error effects have already been discussed in the previous sections. At the single laboratory level, within-run variability reflects random error and is usually unaccountable variability in the measurement process. This might include gravimetric and volumetric errors, relative

inhomogeneity of samples and slight variations in carrying out repeat analyses by the same person etc. Between-run effects reflect day to day variability in the measurement system, including change of analyst and batches of reagents, calibration drift and recalibration of instruments, environmental effects such a temperature, air pressure and humidity. Run to run variability can be determined directly by carrying out repeated analyses on different days. Taken together, the within-run and between-run variability reflects the typical variability in the operation of the measurement system and is often referred to as the intermediate precision or internal reproducibility. Laboratory variation is due to factors such as the variations in calibration standards, instrument differences and reference material supplier and environmental conditions. Laboratory bias is highlighted through collaborative trails, but where such studies have not been carried out, may be determined from calibration evaluation and comparison against certified reference materials. Where collaborative trials have been carried out, it is often helpful to compare single laboratory validation with reproducibility estimates as this can help reveal whether significant effects have been unaccounted for by the laboratory or require justification for better performance. Method bias is usually only identified through comparison of different methods such as through proficiency tests or other method specific collaborative trials. However, as previously discussed method bias can be discounted where the method is considered empirical. The contributions of the remaining three influences (random measurement, run and laboratory bias) are often of a similar magnitude and need to be taken into account when determining the uncertainty evaluation of a method, (Thompson, 1995; Thompson *et al.*, 2002). In addition, an important factor is the way variability of data is often inversely related to the concentration of the analyte, i.e. dispersion decreasing as concentration levels increase. This particular effect was observed and reported by Horwitz (1985). For chemical analysis it is therefore possible, in the absence of collaborative trial data, to predict the reproducibility value using the Horwitz equation. However, caution should be exercised for new measurands with uncharacterised performance, as it has been found that this is not always true for every analyte such as found in the analysis GMO material (Powell and Owen, 2002).

However, it is not within the scope of the current thesis to consider all aspects of method validation. Those affecting the uncertainty estimation such as precision and trueness, are covered in greater depth in the following chapters.

For a fuller description and examples of the treatment of uncertainty in method validation, the reader should refer to Barwick *et al.* (2000), Thompson *et al*. (2002) and ISO 21748 (2010).

2.7.2 Quality Control Activities

2.7.2.1 Internal Quality Control

From the sections above, it can be seen that method validation provides a means of determining method performance capabilities and its limitations that may be expected in routine analysis. However, these characteristics are only consistent as long as the measurement process remains in statistical control. Generally speaking, validation of the method is often carried out using known material, whereas routinely the method will be applied to samples of unknown material. Thus in order to ensure the process remains in control, it is important to be able to run stable materials with known performance characteristics, alongside the unknown samples in order to provide confidence in the results of the unknown materials, i.e.., to control the quality, and are thus referred to as quality control materials.

Consideration towards the scope of QC materials and their application in control charts, and calibration is given in detail in Chapters 3 and 4.

2.7.3 Proficiency Testing (External Quality Control)

Participation in proficiency tests (PT) provides an external control of analytical procedures and enables comparability on a much wider scale with other laboratories. Results of proficiency tests can be a good indicator of laboratory bias and a check on laboratory uncertainty. The spread of results from a laboratory over a period of time should be compatible with that laboratory's evaluation of uncertainty. The differences between the laboratory values and the assigned values provide a means of evaluating the uncertainty for those elements of the method, ie "*the standard deviation of the differences would give the standard uncertainty*", (EURACHEM / CITAC, 2000). The participant's result is compared to the assigned value for the round based on the consensus value of participating laboratories, and using the target value for standard deviation obtained usually from the reproducibility standard deviation given in collaborative trails or by using the Horwitz function to predict expected laboratory behaviour. Test materials left over after the end of a proficiency test act as a suitable matrix specific reference material in the absence of a CRM, as the value of the analyte has been determined by a consensus, it has minimal bias associated with it. X-charts can be used to observe performance in individual rounds, long term trends or unexpected error influences needing investigation. In recent years the use of PT in evaluation of bias and measurement uncertainty has been developed (Magnusson *et al.*, 2004; EUROLAB, 2007).

The evaluation of PT data and its use in deriving uncertainty estimates is considered in detail, in Chapters 5 and the accompanying Appendices.

2.7.4 Method Comparisons

Uncertainty measurement is a complex subject area. Clearly there is a need for different approaches to uncertainty evaluation as important information and significant influences arise from different sources. "*However, there is a risk with this accumulation of theory and terms: it can overwhelm comprehension,*" (Alvarez-Prieto *et al.*, 2009, p624).Whatever the method used for its evaluation, it is not difficult to appreciate the ease with which contributions could be omitted from the modelling approach or the effect of variation of significant factors not built into the validation design. For example, if the measurand is the average concentration of an analyte in a large batch of material, sampling uncertainty needs to be included, if it is the concentration of the laboratory sample, random effects influencing repeatability and run to run variability need to be determined. In addition to within-laboratory influences there are between-laboratory differences which will only be highlighted or accounted for through inter-laboratory studies such as participation in collaborative trials and proficiency tests which may additionally identify method bias. Horwitz (1998, 2003) has commented on the ease of overlooking important variables whilst double counting others and the presence of unknown interactions and interferences. Visser observes that the ISO uncertainty budget approach does not produce comparable uncertainty estimates with those derived from validation or inter-laboratory studies (2002; 2004). Hund *et al.* (2001) comment that the ISO GUM uncertainty budget approach might be well suited for physical measurements but poses significant difficulty for laboratories attempting to construct a model that adequately reflects complex analytical methods and strongly recommend the use of validation and QC. Hund *et al*. (2003) later observe smaller uncertainty estimates using the GUM approach compared to others when evaluating the analysis of tylosin by reverse phase HPLC.

Several studies have shown that measurement uncertainty is often significantly underestimated. "….*Given the present lack of comparability and reliability in uncertainty evaluation in testing, the way forward is to compare uncertainty estimates obtained using different approaches*", (EUROLAB, 2007, p8). Indeed perhaps a mixed design becomes crucial in order to identify the omission of significant contributions by comparing one method against the other, then at least there will be some control to ensure all influencing factors have been accounted for. The issue of unaccounted uncertainty is raised in one of De

Bievre's editorials (2008) when he observes that often Type A contributions are focused on by analysts having determined accuracy profiles for the measurement process, to the exclusion of other Type B effects, which may account for the non-equivalence of comparative uncertainty results. Often confusion relating to the understanding of standard deviation of the sample and the uncertainty of the sample mean is a frequent cause of confusion (De Bievre, 2008)

Specification of measurement conditions becomes fundamental to the correct interpretation of measurement uncertainty information and perhaps lack of clarification on this matter is another source of variability between methods or of the same method carried out by different laboratories. Inter-laboratory values will be affected by systematic and laboratory effects which can give "well performing" laboratories a pessimistic estimation of uncertainty (de Silva *et al.*, 2006), whilst estimations not including reproducibility contributions represent an unrealistic evaluation (RSC Analytical Methods Committee, 2003a) and in-house validation of non-standardised methods that have not characterized all the potential influencing factors, could be criticized as being overly optimistic (Magnusson *et al.*, 2004).

2.8 Combining Standard Uncertainties

Whatever method is adopted, in order to ensure that all uncertainty contributions are accounted for, it can be helpful to refer to a relevant uncertainty model (EUROLAB, 2007), i.e. such as used for the cause and effect diagram, where contributions from sampling, test items, instrument effects, operator, method, etc are listed. Perhaps a hierarchical scheme such as the classification of uncertainty according to repeatability, run bias, laboratory bias, method bias, referred to as the "ladder of errors" (Thompson, 2000) might be applied.

When all the individual components of uncertainty have been determined, standard uncertainties have to be combined. For simple models involving only a sum or the difference of values, i.e.., $y = (p + q + r + \cdots)$, the combined standard uncertainty $u_c(y)$, is given by (EURACHEM / CITAC, 2000, p26, 8.2.6);

$$u_c\big(y(p, q \ldots)\big) = \sqrt{u(p)^2 + u(q)^2 + \cdots..} \qquad \textbf{2.13}$$

Or for models involving only multiplication or division i.e.., $y = p/(q \times r \times \ldots)$,

$$u_c(y) = y.\sqrt{\left(\frac{u(p)}{p}\right)^2 + \left(\frac{u(q)}{q}\right)^2 + \cdots..} \qquad \textbf{2.14}$$

However, more generally, in all but the simplest models as above, the expression for combined standard uncertainty of a value y that is dependent on a number of independent variables; $y = f(x_1, x_2 \ldots..x_n)$, and based on a first-order Taylor series approximation (Taylor and Kuyatt, 1994) is given by the equation;

$$u_c\big(y(x_1, x_2 \ldots.)\big) = \sqrt{\sum_{i=1,n} c_i^2 u(x_i)^2} = \sqrt{\sum_{i=1,n} u(y, x_i)^2} \qquad \textbf{2.15}$$

Where c_i is the sensitivity coefficient given as the partial differential of y with respect to x; i.e. $c_i = \partial y/(\partial x_i)$ and known as the law of propagation of uncertainty. Sensitivity coefficients may be derived through experiment as described by Thompson *et al*. (2002), and involves changing x_i and observing the effect on y. It recommends taking an additional two values of x_i and determining the gradient which provides an approximation for c_i.

Where variables are not independent, the covariance between variables needs also to be taken into consideration. A more detailed discussion with examples is given in section 4.4.3 but for full details the reader is referred to the EURACHEM / CITAC Guide CG4 (2000), the GUM (JCGM 100, 2008) and Chapter 6 of this thesis.

2.9 Expressing MU as an Expanded uncertainty (95% CL)

The combined uncertainty calculated above provides a value representing the expected dispersion for the measurement value, equivalent to one standard deviation either side of that value. For a normal distribution where x represents our best estimate of the true mean value, μ, the interval $\mu - \sigma < x < \mu + \sigma$ equates to 68% of possible outcomes, i.e. about a third of the time you might expect to get a result outside of this range but still be an acceptable value, within the range of the normal distribution. For this reason, a coverage probability equal to approximately 95% of the population is a preferred interval to use when quoting associated uncertainty. For a large, representative sample, 95% is given by a coverage factor *k*=2, representing 2 standard deviations, (although in fact this is actually 95.45% coverage probability, 95% is given by *k*=1.96). For samples where the degree of freedom is small, typically below about 50, the normal distribution broadens and flattens and is better represented by a t-distribution. Equivalent *k* values can be found from t-distribution

tables, specifying the appropriate coverage interval required and the relevant degrees of freedom.

Thus for reporting purposes, the measurement value should be stated $\pm$ the expanded uncertainty (U), with the coverage factor used and level of confidence, i.e. $x \pm U_x$ (expanded uncertainty using *k*=2 at 95% confidence).

2.10 Conclusions

Clearly, the summation of all the significant contributions to uncertainty can place a heavy demand on the uninitiated and relies inextricably on the knowledge and skill of the analyst as to where the contributions originate. Having considered all the possible contributions of uncertainty, the final result, can to the dismay of many analysts, be much larger than originally anticipated when compared to the simple standard deviation of the values, traditionally used as a measure of error, and bring into question the validity of such a result. This is simply demonstrated by the consideration of a set of data. For example; consider a set of eleven samples, each analysed in duplicate;

34.43	33.53	35.98	35.37	34.82	35.47	34.46	34.97	34.81	35.25	35.04
33.41	32.87	35.33	33.18	34.28	34.16	34.94	34.28	34.59	34.86	34.43

The mean $\bar{x}$ and standard deviation, s_x of all the data are; 34.566 and 0.782 respectively, (n=22).

The standard deviation of the mean, $\bar{s}_x$ (also referred to as standard error of the mean, or standard uncertainty) is $\bar{s}_x = {}^{s_x}/_{\sqrt{n}}$ and equals 0.167 (n=22, 1 std dev), and a relative value of (0.167/34.566)x100 = 0.48%. Thus an analyst seeing that their relative uncertainty is less than 0.5% might be very happy. However, this is not the complete picture and does not take into account both the within and between run variability (or indeed lab or method bias). Thus we turn to ANOVA to derive a value for the reproducibility standard deviation s_R of 0.787 or as a relative value, 2.3%, considerably larger than 0.48% naively derived originally. Thus as observed by the RSC Analytical Methods Committee (RSC Analytical Methods Committee, 1995);

> *"…uncertainty would not be greatly reduced by averaging measurements collected under repeatability conditions. The n repeatability measurements would not have a standard uncertainty of* $u_{\bar{x}} = {}^{u_x}/_{\sqrt{n}}$, *but*

$u_{\bar{x}} = \sqrt{\sigma_{\varepsilon}^2/n + \sigma_{run}^2 + \sigma_{lab}^2 + u_b^2}$, *which may not be much smaller than* u_x*."* (p2305).

> *"When all the separate contributions are combined the resulting uncertainty will sometimes be an unexpectedly large proportion of the measurement. This is often worrying for those not accustomed to a realistic appraisal of errors, and sometimes for those who are. However, analytical chemists must be prepared to apply realistic criteria for fitness for purpose in all circumstances. All too often analytical chemists seek to achieve a quality of data that is unnecessarily high for the application. This stems from early training, when we are encouraged to produce the most accurate result possible. Such a strategy is appropriate for training students in skilful manipulation, but in real life is rarely germane to the demands of fitness for purpose."* (RSC Analytical Methods Committee, 1995, p2303)

Perhaps because of this or for other reasons, a Bayesian approach based on probability densities is becoming increasingly popular. Bayesian methods have been servicing the archaeological community for more than twenty years (Naylor and Smith, 1988 cited in; Buck, 2004), primarily to aid the interpretation of radiocarbon data. However they have only relatively recently started to filter down to the commercial sectors, as demonstrated by the first GUM supplement to Monte Carlo simulation only a few years ago (JCGM 101, 2008). However, the (chemical) analytical community still await the arrival of user-friendly guidance documents for mere mortals to be able to apply the methods described routinely. It is also noticed how, almost without exception, current guidance documents make minimal, if any mention of performing weighted calculations in the determination of uncertainty estimates. I am certain that if they had, the chemical analytical community would have been very quick to pick this up. A weighted uncertainty that favours the smallest uncertainty values, would be every chemists dream, compared to the current guidance which seems in favour of reporting the largest!

However, for the purpose of this thesis, the focus for the most part will be towards the more traditional approaches, with an emphasis towards the evaluation and control of uncertainty influences at the intra-and inter-laboratory levels (Chapter 3, 4 and 5). A Bayesian approach is applied in developing an integrated expression for protein decomposition in Chapter 6 and compares these uncertainty estimates with those derived solely by ANOVA. The Bayesian derived values are then used in Chapter 7 for the development of sequence chronology.

Chapter 3. Analytical Uncertainty in AAR; an Intra-Laboratory Perspective

3.1 Introduction

In the last chapter, the subject of measurement uncertainty was introduced in the context of chemical analysis and a variety of evaluative approaches for its practical determination were presented. The approach adopted must be specific to meet the needs of the measurement system on a case by case basis. The "bottom-up" approach is often criticized as being too unwieldy and impractical for many chemical analyses with complex and lengthy extraction and pre-treatment stages, leading to an under representation of the true level of uncertainty associated with results. In such a situation the "top-down" approach is often favoured as determination of precision estimates, encompass the entire measurement process. Precision estimates of reproducibility may be determined either through an organized inter-laboratory collaboration, or at a more local intra-laboratory level, as it requires little further work than that usually already undertaken by the laboratory in the validation of the test method prior to its adoption in routine analysis.

This chapter will now consider the intra-laboratory evaluation of measurement uncertainty as specifically applied to amino acid racemisation determination as carried out by the University of York. The emphasis of the current research was always to evaluate retrospectively, the mass of information previously generated by the AAR laboratory, and not to undertake chemical analysis directly. As the method had been in regular use for several years, it had been assumed that the method had undergone thorough validation prior to the start of this project. For these reasons, no additional analytical measurements were scheduled into the original workplan or carried out. Evaluations presented in this and the next chapters are therefore derived using existing data determined by researchers at the University of York and do not, unless stated, use the author's own analytical measurements.

The chapter starts by considering the sources of uncertainty in the AAR measurement system and which factors contribute to the final uncertainty of D/L values.

Having identified potential sources of error using the GUM approach (see Chapter 2), the chapter then considers the use of reference materials in monitoring and controlling these influences, and how precision and bias evaluation as part of method validation can be used to control measurement quality.

3.2 Evaluating Sources of Uncertainty

Prior to the determination of individual uncertainty contributions, the GUM requires that there is initially a clear statement about what is being measured, a description of the measurement procedure and measurement steps, with a quantitative statement for the expression of the measurement result that reflect the parameters on which it depends (JCGM 100, 2008). Based on this mathematical model for the measurement result, a cause and effect diagram can then be constructed. Using each of the key components as the main branches on an ishikawa or fishbone diagram, additional factors are added to each stage of the method, working outwards until error influences seem sufficiently remote (Ibid). The diagram can then be simplified by grouping together similar contributions (such as the effect of temperature on volume and the use of the same weighing instrument to prevent over-counting), or combining influences into a single branch such as a single precision branch. Having identified all the important sources of potential error, the mathematical model can be updated to incorporate additional terms as required.

3.2.1 Specification of the measurand

Eurachem (EURACHEM / CITAC, 2000) point out the importance of identifying measurement systems where results are independent (where the result does not depend on the method) or dependent (where the result does depend on the method, i.e. empirical methods) of the method. Distinguishing between these two effects could be significant and Eurachem stress that only those effects relevant to the result should be included. For example, where there is known method bias or matrix effects, then the results will normally be reported with reference to the method or matrix. It is therefore unnecessary to consider bias contributions intrinsic to the method and results are reported *uncorrected* (EURACHEM / CITAC, 2000)

For AAR, the dependence or independence of results on method, (i.e. RP, gas chromatography (GC) or Ion Exchange chromatography (IEx)) have not been fully established. Within the AAR community, there is currently no correction for laboratory or method bias.

For the purposes of this thesis, the method is considered empirical, with results being specific to the method **and laboratory**, in the absence of external reference materials.

Thus for AAR the specification of the measurand might be something like; "Pleistocene opercula from the terrestrial gastropod *Bithynia tentaculata*, sampled from [site / horizon details] on [date] by [person], for the determination of amino acid L and D isomers by reverse-phase HPLC, expressed as a ratio and reported as a D/L value".

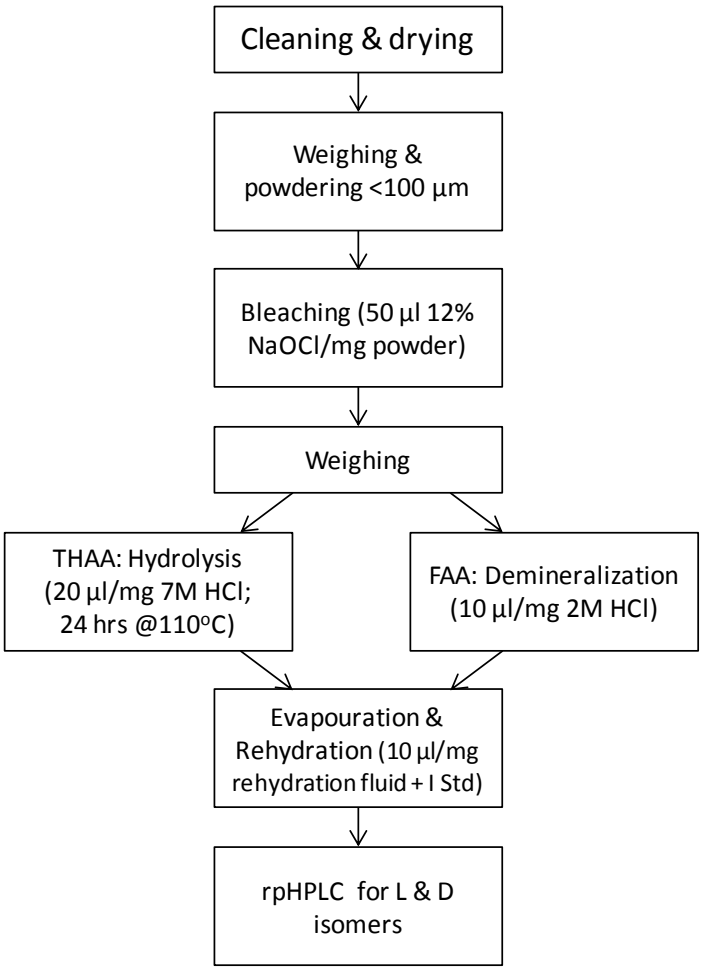
3.2.2 Description of the measurement procedure

Details of each step of the measurement procedure are given in the standard operating procedures in the laboratory (SOP) and summarised below, together with an overview of possible uncertainty sources.

The measurement procedure can be simply represented diagrammatically, Figure 3.1, and is briefly discussed below. Initially samples to be tested are cleaned and washed by sonication using ultrapure water, until the water remains clear. Samples are then dried at room temperature and powdered, before bleaching with sodium hypochlorite for a total of 48 hours. The weighing and particle size of the finely ground material will affect the volume of bleach added (50 µL/mg) and the surface area exposed to the bleach. This could affect the removal of inter-crystalline protein, which may add errors in the quantification of the intra-crystalline fraction later, contributing uncertainty to the final measurements. After removal of the bleach, the dried material is again weighed (approx 1-10 mg) into sterile glass vials prior to hydrolysis for total hydrolysable amino acids (THAA) or demineralisation for the free amino acid fraction (FAA). Once again weighing errors and balance calibration uncertainty accumulate here. The powder then has a measured volume of acid added to the vial, or for the FAA, sufficient acid to ensure the powder fully dissolves. For some biomineral matrices (such as ostrich eggshell), this can take a relatively large amount of acid and it is essential that the total volume required is recorded. Uncertainties arising from the dilution and making up of the acid to the correct concentration, together with inaccurate recording, measurement and volumetric errors will all add further uncertainty to this stage of the process.

For THAA, the acidified sample is then heated under an enriched nitrogen atmosphere in an oven at 110^{o}C for 24hours. Here, oven calibration and temperature fluctuation, including removing the samples too early or too late could all have an effect and add further uncertainty contributions. The samples are then evaporated to dryness.

Figure 3.1: AAR measurement process schematic



The time taken to achieve the desired "pellet" will vary depending on the volume of acid used and the efficiencies of the centrifugal evaporators, but generally samples are left to evaporate overnight. Whilst these instruments are operated without heating, temperatures within the centrifugal evaporator are frequently elevated above room temperature ~ 30-40 ^{o}C, during the process. The effect of gentle but prolonged warming on samples is a factor that probably requires further evaluation. Racemisation is a temperature dependent process and whilst the effects of time and temperature during the preparative steps are probably negligible compared to racemisation on geological timescales, nonetheless, they should be considered in the design of a validation programme together with effects from heating during hydrolysis. The final stage before analysis is rehydration of the dried sample. A stock supply of rehydration fluid is made up intermittently when supplies run low (perhaps once or twice a year) and includes a measured quantity of 0.01 mM L-homoarginine, used as an internal standard for the quantification of individual amino acid L and D isomers (see section 3.2.3 below). Thus there are uncertainties associated with the preparation and

concentration of the internal standard required for quantification later. Samples are rehydrated using an appropriate volume of rehydration fluid per mg of original sample. There are several influences that will lead to errors in determining the appropriate volume of rehydration fluid required, and consequently uncertainty of the final result. These include; inaccuracies in the weighing of the original sample, potential losses during the extraction stages, and undissolved sample remaining in the bottom or attached to sides of the vial. In addition, the actual volume occupied by the dried residue "pellet" at the bottom of vial, will add a further influence affecting the interpretation of chromatogram peak areas after analysis and the correct determination of the true concentration.

Whilst random error is exactly that...random, and cannot be predicted, many of the effects raised above will be present, systematically in every sample prepared at the same time, and will contribute towards method bias. Similarly, instrumental settings, reagents and solvent batches, temperatures, pressures, columns, volumes etc. will systematically affect all samples run during that batch. For this reason, individual runs are said to possess run bias, in addition to random error contributions. However, the extent of the effect of this variability on individual samples, can be determined by measuring the variability between multiple samples (of the same thing), in the same run. This is the repeatability precision. If replicate injections are measured from the same sample vial, all that will be measured is the instrumental variability. If separate samples of the same material are prepared, side by side, then the within-run precision will also reflect the variability in carrying out the method, which is what is required. At a higher level, between-run precision will also reflect changes in individual run bias that will occur from day to day, or operator to operator etc. This enables bias uncertainty contributions which might not be easily determined on their own, to be more simply quantified as a precision estimate when taken together.

3.2.3 Quantitative Expression

Determination of the L and D isomer concentrations use the internal standard present in the rehydration fluid (L-homoarginine), as a reference value. The quantitative expression used for the result, links the key parameters, in this case the mass of the sample taken for hydrolysis or demineralisation (M_s), the concentration of the internal standard (C_{is}), the volume of the rehydration fluid (containing the internal standard) used (V_{is}) and the chromatogram peak areas for the particular amino acid L or D isomer, for example L-valine (A_{LVal}) and the internal standard (A_{is}). Whilst the final value required for geochronology is

the D/L value, **the result from the analytical measurement of the matrix, (i.e. the measurands) is the peak area, reinterpreted as the concentration of an amino acid L or D isomer (C_{LVal}).**

The formula used to derive the concentration of the unknown isomer is obtained with a little rearrangement, thus;

$$\text{From;} \quad \frac{A_{L\,Val}}{\mathrm{C_{L\,Val}}} = F \times \frac{A_{is}}{C_{is}} \qquad \textbf{(3.1a)}$$

$$\text{we get;} \quad \mathrm{C_{L\,Val}} = \frac{A_{L\,Val}}{F \times (A_{is}/C_{is})} \qquad \textbf{(3.1b)}$$

where, F is a correction factor called the Response Factor.

In equations (3.1a & b), the concentrations C_{is} and $C_{L\,Val}$ are both expressed in terms of mM, (since C_{is} = 0.01 mM). This is equivalent to mmoles/L. However, what is required are the number of moles of the unknown isomer, present in the volume of rehydration fluid used, expressed as µL. This will also be the same as the number of moles present in the powdered sample originally hydrolysed. Therefore the mM (or 0.01 mmoles/L) is divided by 1000,000 to give 0.01 mM/µL, and multiplied by the appropriate volume (µL) of rehydration fluid used (V_{is}). This value is then divided by the weight of sample to give the number of moles present per mg of sample.

The concentration of the isomer (in this example L-Valine), is more appropriately expressed in nmoles or pmoles per mg, thus;

$$Result = C_{L\,Val} = \frac{F \times A_{L\,Val}((C_{is} \times V_{is})/A_{is})}{M_s} \qquad \textbf{(3.2)}$$

where; $C_{L\,Val}$ (pmoles/mg), C_{is} (mmoles/µL), M_s (mg) and V_{is} (µL) and peak areas are in arbitrary units.

The factor currently used for fluorescence detection correction, (the Response Factor (RF)), was originally determined from previous studies on amino acids in collagen (Collins pers. comm.).

Whilst this section has considered potential sources of uncertainty arising in the measurement procedure, other sources may also need to be considered when determining an uncertainty statement for the end result. Once all uncertainty contributions have been

identified, the overall uncertainty budget (GUM) needs to be evaluated, ensuring that over-counting common contributions is avoided and with due regard to cancelling effects.

3.2.4 Weighing up the uncertainty budget

In considering sources of uncertainty it is important to consider whether sampling is an important factor that needs to be built into the model. Often in a commercial laboratory, sampling is not the responsibility of the analyst. However, very often the homogeneity of the raw material presented by a client for analysis may be an issue and an uncertainty contribution estimated from representative sub-samples taken for evaluation. In the case of individual opercula, sub-sampling from a bulk isn't an issue as it is single opercula that are analysed. However, uncertainty related to the distribution of individual opercula within a single horizon and the homogeneity of the sediment sample will be more of a problem. When considering the uncertainty of material recovered from an archaeological / palaeontological / geological site, independent repeated measurements on different opercula will be necessary to reflect additional sampling uncertainty. However, for the purpose of this chapter, uncertainty contributions will be restricted to the analytical process and site sampling uncertainty will be considered in more detail in the Chapter 6

Many of the primary sources of uncertainty have already been mentioned in section 3.2.2. Figure 3.2 is a cause and effect diagram, suggested by the GUM, illustrating the main sources of uncertainty in the analysis of amino acid isomers by RP. Note the inclusion of both the hydrolysis and demineralisation branches, although in practice only one would be relevant to a specific analysis (THAA or FAA). Note also the inclusion of a homogeneity branch, which may or may not be relevant depending on the matrix under investigation. Eurachem suggest that an additional recovery branch is always added to represent *"...a nominal correction for overall bias, usually as recovery,..."* (EURACHEM / CITAC, 2000).

Having carried out an exhaustive analysis and identified all possible sources of uncertainty in the method, the analyst is then required to gather all individual uncertainty contributions together, to end up with a final, single combined uncertainty estimate for the method.

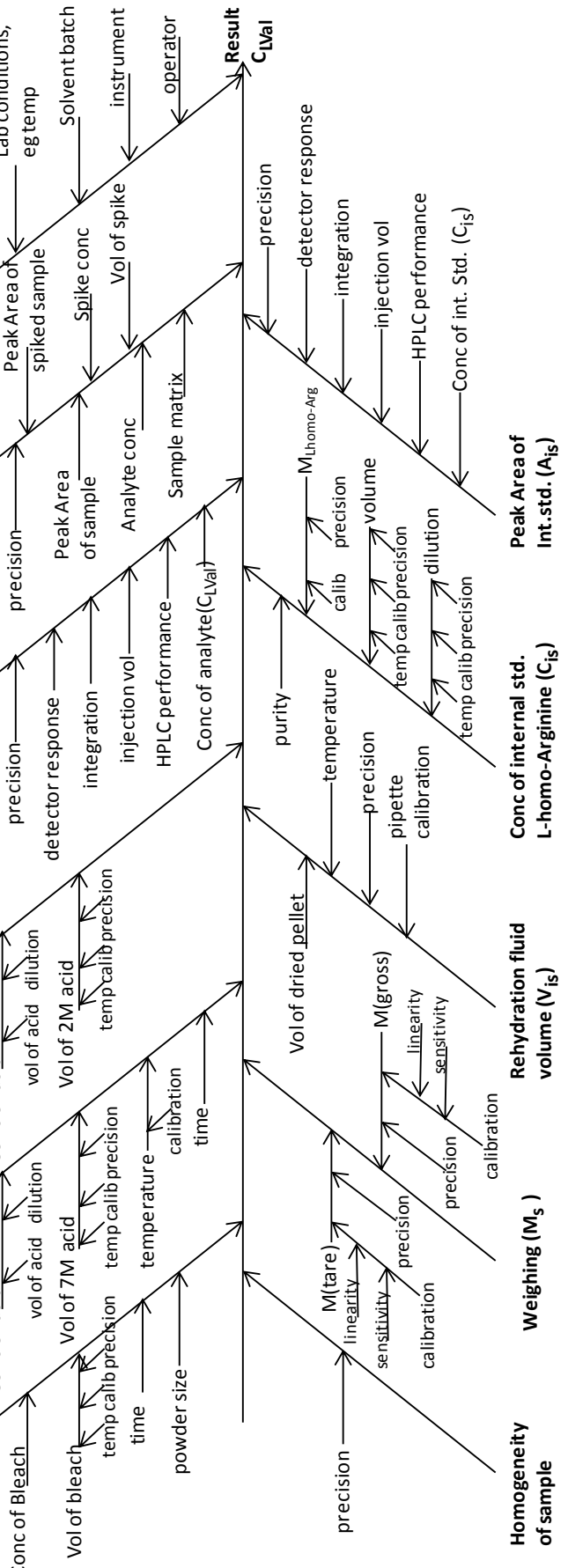


Figure 3.2: Main uncertainty sources for AAR analysis by RP

3.2.5 Combining individual uncertainty contributions

For AAR geochronology, the final measurement result required, is not the concentration of the amino acid but the D/L value derived from the ratio of the D isomer concentration to that of the L isomer concentration. This now presents a dilemma; how best to combine the individual uncertainty contributions? There are three possible choices;

1. Accept that the final calculation is a quotient (conc. of D / conc. of L) and combine all contributions (i.e.; twice) for both D and L isomers, according to the principles of uncertainty propagation for models with x or ÷ in them;

$$u_c(y) = \text{y}.\sqrt{\left(\frac{u(p)}{p}\right)^2 + \left(\frac{u(q)}{q}\right)^2 + \cdots}$$

2. Avoid double counting common uncertainty contributions twice and only count them once in the overall combined uncertainty calculation, or,
3. Cancel both. Measurement results for both L and D isomers are obtained from the same sample extract, therefore, it could be argued that both random and systematic effects are acting equally on both the numerator and denominator and common uncertainty components cancel.

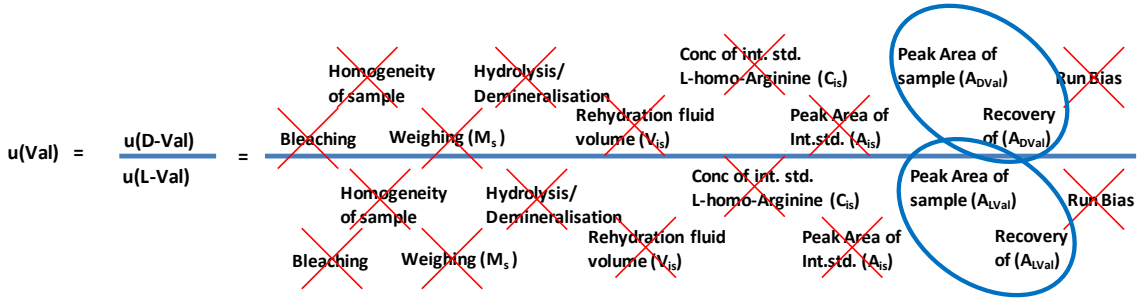
To help resolve this issue, advice was sought from LGC, one of the UK's National Measurement Institutes who share the responsibility of delivery of the UK's Chemical and Biological Metrology programme, an initiative funded by the Government's National Measurement Office[3.1]. Sadly after several emails and attempted phone calls all that was received was the promise of a response.

Figure 3.3 helps to illustrate the effect of adopting the third of the three options above, i.e. cancellation. Many of the sources of uncertainty affect both L and D isomers equally. Examples of common influences will include those associated with the physical preparation of the test sample taken for analysis, (i.e.; homogeneity, bleaching and weighing), those that originate from the preparation of the extract (i.e.; hydrolysis or demineralization), or those that affect the quantification of the isomer concentrations (i.e. volume of rehydration fluid, concentration of internal standard (L-homo-arginine), peak area of internal standard and run bias). So although components such as the mass of the bleached sample used, volume of rehydration fluid, peak area and concentration of the internal standard, all contribute towards the calculation of the unknown isomer concentration, because their values and respective standard uncertainty contributions are

[3.1] http://www.nmschembio.org.uk/GenericArticle.aspx?m=92&amid=3409

fixed for both the L and D isomers, then the effect of these uncertainties cancel each other out and theoretically, can be ignored in the calculation of the ratio.

Figure 3.3: Suggested cancellation of shared uncertainty sources for D/L values



If it is correct that these uncertainty contributions cancel out, then there only remains uncertainty associated with the determination of each of the individual isomer peak areas and their respective recoveries, used to correct for analyte losses during extraction and analysis (shown by the circled factors in Figure 3.3).

Using the cancellation approach, Figure 3.4 simplifies the cause and effect diagram and demonstrates how uncertainty contributions, resulting from only peak areas and recoveries, could be combined for each amino acid's D/L value.

3.2.6 A "Top-Down" perspective

Schematics such as Figure 3.2, that model the theoretical uncertainty budget, become intricate and unwieldy for all but the simplest measurement processes. The uncertainty budget approach is in principle very procedural, focussing on the propagation of uncertainty from the method and analytical steps. However, in practice, an analyst will need to be able to report an uncertainty estimate to a customer that would encompass the expected variation intrinsic to the method over time. From a customer's perspective, they need to know that the laboratory would be able to produce the same result whether they presented them with a sample in January, July or November, irrespective of the instruments used or an individual staff absence.

For this reason, opinions regarding the application of the GUM approach are divided. On the one hand it provides a visual representation of all possible sources of uncertainty that can be readily appreciated, whilst on the other it is often criticised for underestimating combined uncertainties, as it is very easy to omit important contributions.

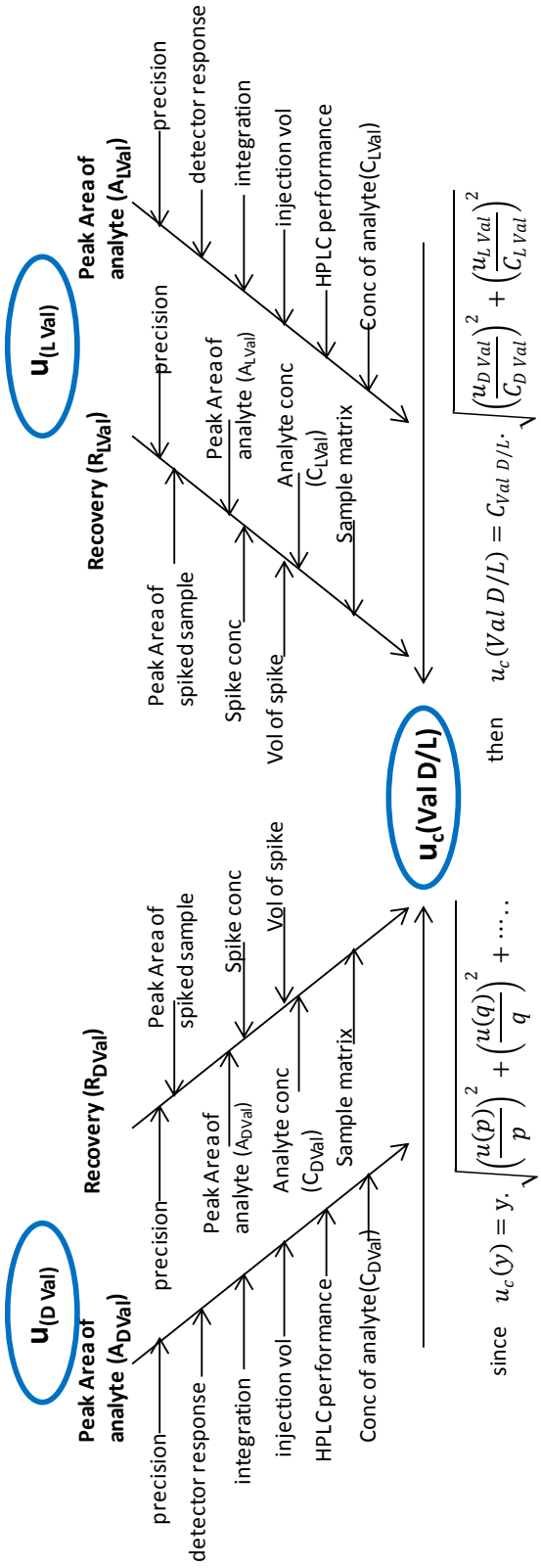
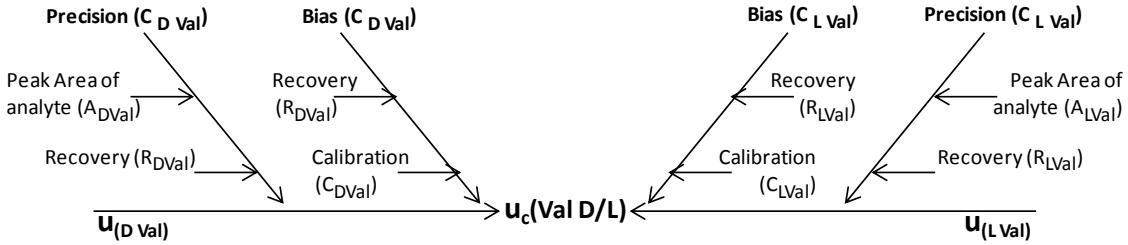


Figure 3.4: Simplified model for sources of uncertainty for Amino Acid D/L Values

For AAR analysis, the GUM approach seems overly burdensome, and additional attention would need to be given to the effect of matrix and amino acid concentration. Consequently, laboratories have favoured "top-down" approaches, incorporating inter-laboratory reproducibility precision to account for laboratory bias where possible, or simply focusing on data generated as part of single laboratory method validation. This utilises both precision and bias data, and avoids additional work, time and expense.

Figure 3.5, provides an alternative model that allows us to circumvent the theoretical construct in favour of an all inclusive evaluation of standard uncertainties, and avoid underestimating contributions from inaccurate models.

Figure 3.5: Simplified model based on accuracy parameters for D/L values

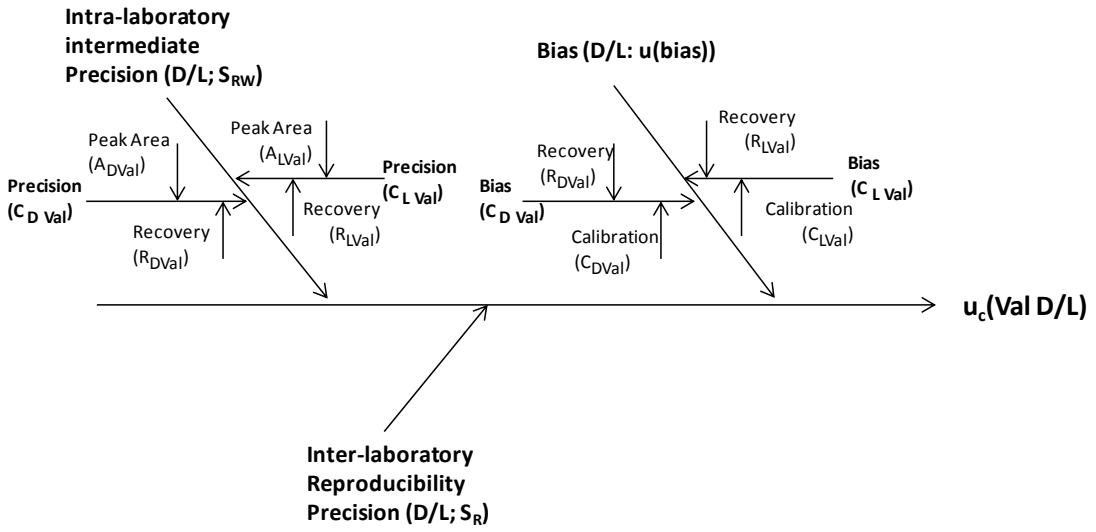


In the above diagram, **precision and bias are seen as properties of the individual isomer concentrations,** since these are the end products of the measurement process. As such, these should be evaluated individually and the combined L and D uncertainties (($u_{c(D\ Val)}$ and $u_{c(L\ Val)}$) further combined for the expression of uncertainty relating to the D/L value, if that is what is required for geochronological purposes.

However, this is not common practice in AAR geochronology, which adopts an even higher perspective and evaluates the uncertainty (limited to precision analysis) of the D/L value itself. So, although this moves the estimation of the final result further away from the measurement process, it could be argued that in principle, the uncertainties associated with the concentrations of the L and D isomers are simply branches lower down the analytical tree. On this basis, this approach does not appear to contradict the principles of the GUM. Therefore the final cause and effect diagram might look something like Figure 3.6.

Here estimates for the combined uncertainty $u_{c(Val\ D/L)}$, are determined either as the combined intra-laboratory intermediate precision (s_{RW}) plus uncertainty due to bias u(bias), or as a single measure derived as the inter-laboratory reproducibility (s_R) (Magnusson *et al.*, 2004).

Figure 3.6: Final uncertainty model for D/L values



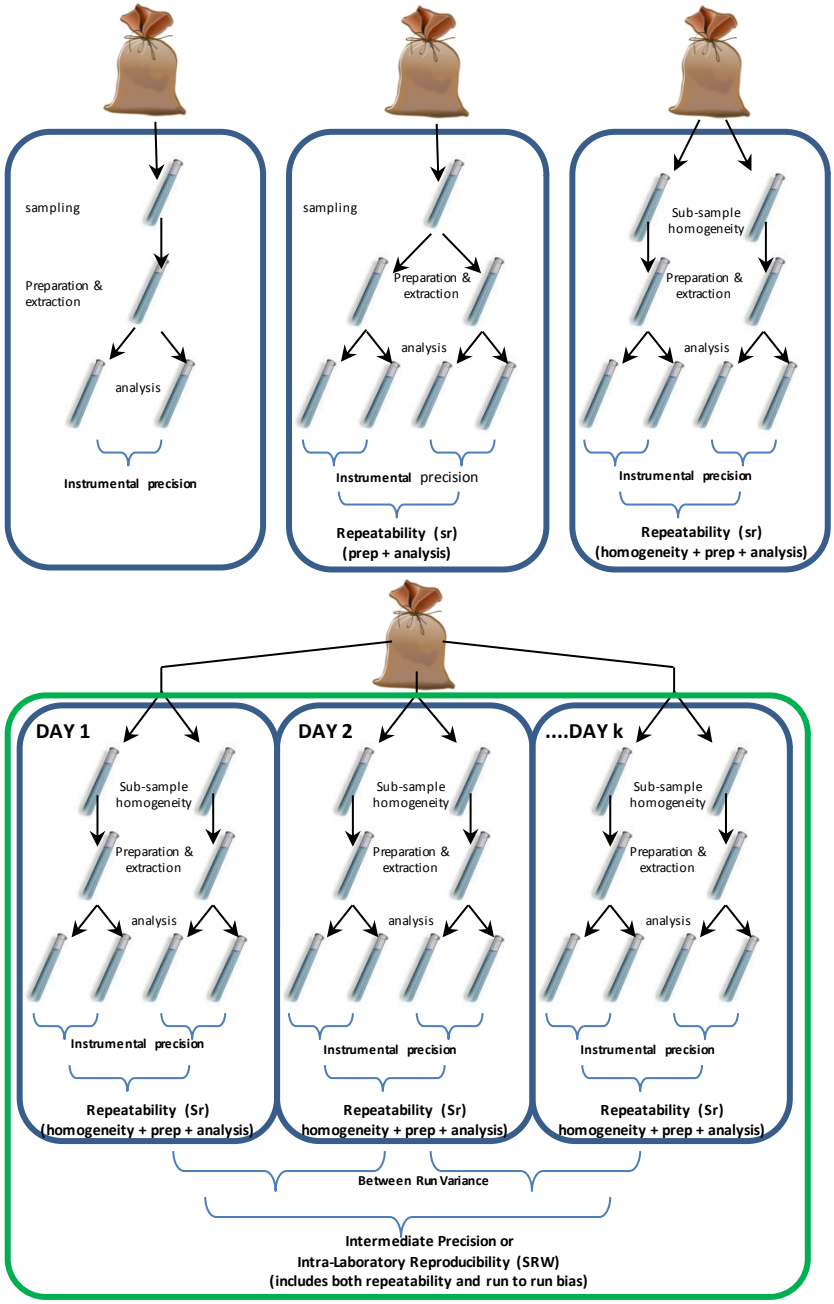
For a single laboratory, intra-laboratory estimates of **repeatability precision (s_r)** are derived from repeated measurements of the same or similar material, determined in a single analytical run and reflect **random measurement errors (ε).** Uncertainty contributions may be from any part of the measurement procedure that could potentially vary within the run when applied to different samples. This might include gravimetric and volumetric inaccuracies, sub-sample heterogeneity, mixing effects, hydrolysis and drying effects etc., and will be reflected in the distribution of individual measurement results (Thompson *et al.*, 2002). In addition to factors that affect individual samples, there are factors that affect the whole analytical run, and every sample in that run equally. Included in this group are particular batches of reagents and solutions used in the extraction stages (such as the acid concentration, age and effectiveness of the bleach, oven temperatures, laboratory temperatures, instrumental conditions, calibration, operators etc.). Simply, anything that affects the whole run, systematically, contributes to the **run bias (δ_{run}).** Within a single analytical run, the run bias is fixed. However, after having carried out a number of separate runs on different days, each with their own slight differences (perhaps a new batch of acid or buffer, different analyst, a warmer day etc.) when viewed over time, the run bias becomes a random variable. As such the effects can be more simply determined as a precision estimate across the runs, the between-run precision. This, combined with an estimate of the repeatability precision (which, if determined from a single run should be sufficiently representative of most of the anticipated variability due to random effects), provides the **intra-laboratory reproducibility**, an **intermediate precision** estimate **(s_{RW}).** s_{RW} therefore reflects all the uncertainty likely to be experienced by the laboratory due to the application

of the measurement process. This is especially important if batches of samples are prepared together but split and run on different days or instruments. If however, replicate instrumental measurements were made on the same extract from a single sample (rather than working up separate portions of the material, taking each through all the preparation and extraction stages separately), then repeatability precision will only reflect the instrumental precision and not the precision of the whole measurement process. Replicate analyses of sub-samples taken from the bulk can also be used to include uncertainty due to any inhomogeneity of the raw material too.

Figure 3.7 illustrates the effect of replicate measurements on precision estimates. Whilst replicate instrumental measurements are included for completeness, they are not completely necessary where instrumental stability is not a problem. A laboratory with limited time and money would do better by adding additional replicates at a higher level, as instrumental precision is likely to be negligible on automated HPLC instruments compared to the uncertainty introduced from the preparation and extraction or sub-sampling stages. It is therefore preferable to have better control of these influences by increasing the number of replicate sub-samples being worked up through the measurement process. In Figure 3.7, the three scenarios at the top demonstrate the precision estimates achievable from a single run and the uncertainty contributions included in each. The lower part of the illustration demonstrates the effect of expanding the precision estimate across several runs over several days.

There are two further levels of bias that need to be considered in an analytical system. Run bias has the lowest level of influence, the next highest level would be **laboratory bias** (δ_{lab}) followed by **method bias** (δ_{method})**.** Laboratory bias reflects any systematic offset that might be present whilst applying the measurement process by a particular laboratory. It includes influences due to the use of specific instruments, separation columns and settings, along with particular lab-specific details found beneficial in the preparation and extraction stages that might not be applied by a different laboratory. Simply, it includes anything that might affect all the measurement results systematically and is not dependent on operator, reagent batch etc. Method bias is the systematic effect that using a given method might have on all the measurement results, regardless of the laboratory carrying it out. For example, preparative method A might require hydrolysis at 110^{o}C whilst method B might use 120^{o}C, or perhaps differences in the analytical method used, (i.e. RP vs GC).

Figure 3.7: Effects of replicate measurements on precision estimates



The IUPAC Harmonized Guidelines for Single-Laboratory Method Validation (Thompson *et al.*, 2002), stress the importance of including these effects in the evaluation of the fitness-for-purpose of a method, and ensuring that the resulting uncertainty is included in the overall uncertainty for that method. It should perhaps be noted that bias is assumed to be negligible within any measurement system. The purpose of evaluating bias is to assess its significance in relation to the measurement result. Where bias is found to be significant by using a t-test, results need to be appropriately corrected. However, any uncertainty that arises as the result of assessing the bias (such as purity or reference material uncertainty),

should be included in the combined uncertainty estimate (EURACHEM / CITAC, 2000; JCGM 100, 2008). In cases where there has been no correction, then the bias itself also needs to be included in the overall combined uncertainty estimate.

Generally speaking, laboratory and method bias are difficult to separate for a single laboratory. The evaluation of any bias requires the evaluation against a reference, whether that is a reference material, another laboratory or another method. Use of reference materials to assess intra-laboratory bias during validation, will usually evaluate run, laboratory and method bias combined. To evaluate laboratory bias a laboratory needs to compare its results against those of another. This is most commonly done by participation in an inter-laboratory collaborative trial (CT) which is method prescriptive (Horwitz, 1995). By reducing down as much variability from the method as possible, the only other source of variability in the measurement results will be from individual laboratory bias. When seen from a higher perspective still, the individual laboratory bias (fixed for a given laboratory), becomes another random variable when seen across several laboratories, and can be evaluated as a precision estimate, the **reproducibility precision (s_R)** (in the same way as run to run bias is viewed at the single laboratory level). Further, because method influences are in effect removed, s_R, provides a value that represents the uncertainty for any laboratory carrying out the method and covers both random and systematic effects. For this reason, reproducibility precision is often favoured by analysts as a single measure of uncertainty providing their own intra-laboratory repeatability doesn't exceed that given by the collaborative trial. Comparison against the consensus value from a collaborative trial also enables an individual laboratory to assess its own individual laboratory bias.

Method bias could potentially be determined from a collaborative trial if a certified reference material (CRM) was used as a test material, with a known reference value. A comparison of the consensus value with this would then provide an estimate of the overall method bias as carried out by any laboratory.

Other external comparisons such as proficiency testing, can also provide valuable information about laboratory + method bias combined. This can be particularly valuable as it can highlight trends over time if carried out frequently enough. Where significant method differences are reported between participants, comparisons of the consensus values between different groups could provide method bias information, whilst comparison of an individual laboratory's data with the consensus from others using the same method, can give laboratory bias information.

Use of a CRM or other reference method by a single laboratory would only provide an estimate of the combined laboratory + method bias. However, often suitable, matrix-matched CRMs are not available for all types of analysis or matrices and the organisation of a collaborative trial can be costly. The absence of a method and matrix specific collaborative trail prevents comparability of a laboratory's own s_r values and adoption of the CT's s_R values. In this case regular participation in a coordinated proficiency test becomes incontrovertible.

Other higher levels of bias exist, such as the effect of matrix and concentration, which are often neglected by laboratories. It is noted that the use of recovery analysis to determine bias can be especially affected by concentration and should be reflected in the over-arching uncertainty budget (Thompson *et al.*, 2002). Understandably, a laboratory might be reluctant to increase its uncertainty estimation by expanding it to include other matrix effects. For this reason, matrices are often grouped or specifically identified under the remit of validation which can be seen from the schedules of accreditation held for testing laboratories on the UKAS United Kingdom Accreditation Service) website; http://www.ukas.com/about-accreditation/accredited-bodies/Testing-laboratories-schedules.asp

In Figure 3.6, the effect of calibration has also been included along with recovery on the bias branch. Recovery is used to correct for bias resulting in analyte loss during extraction and analysis and is one method that can be used to assess trueness (bias) in validation. Calibration is one of several additional criteria also required for evaluation as part of method validation and allows for instrument effects, detector sensitivity etc. to be corrected enabling arbitrary peak area values to be converted into useable concentration units. On its own, Thompson *et al*. (2002, p846) note that calibration errors are often small and are usually included under the umbrella of other "top-down" methods (with perhaps the exception of the calibrant uncertainty): "....*random errors resulting from calibration are part of the run bias, which is assessed as a whole, while systematic errors from that source may appear as laboratory bias, likewise assessed as a whole*." However, where gross errors occur in calibration, these can have a significant systematic impact on measurement results.

It is suggested that the repeatability, run effect and laboratory effect are all of an equivalent level, therefore none should be omitted during validation (Thompson *et al.*, 2002). However, higher level laboratory and method bias contributions have not been included in Figure 3.6, since for the purpose of this thesis, AAR analysis is considered empirical. Thus all measurement results reported relate specifically to the method as carried out by the

University of York and as such, may not be directly comparable with another laboratory's measurement result of the same item.

3.3 Uncertainty estimates at the University of York

The absence of commercially available, independently certified reference materials suitable for AAR analysis has been a considerable set back to the assessment of measurement uncertainty in AAR. Evaluation of the stability and suitability of the method has therefore not been possible, and has prevented an impartial demonstration that the method is fit for its intended use. As a consequence, there has been no formal method validation carried out, there are no complete uncertainty estimates and procedures such as recovery and calibration have not been rigorously evaluated. Although standards and blanks are analysed routinely in all the AAR laboratories, in essence, the methods are beyond the scope of statistical control and this requires urgent attention.

Measurement uncertainty determined through single laboratory method validation (or even by a collaborative trial) is usually a one-off event and establishes the performance characteristics for the method. The stability of the measurement system is then monitored once the method has been brought into routine use by a process of internal and external quality control (IQC and EQC). IQC encompasses those processes carried out by a laboratory, to monitor precision and bias of the measurement results, after analysis, recovery correction and calibration. It ensures that the measurement system remains in a state of statistical control. Clearly it isn't possible to know whether the measurement process is under statistical control from the analysis of an unknown test sample, since the appropriate performance parameters for the unknown test sample are unknown! Therefore reference materials (RMs) with known characteristics are run alongside test samples. As such, uncertainty estimates that accompany measurement results can be trusted. EQC are those processes coordinated outside the normal laboratory environment, such as participation in proficiency testing or a collaborative trial.

EQC requires the use of matrix-appropriate, homogeneous test materials. As part of this research, a proficiency test was coordinated with other international AAR laboratories and this will be the subject of Chapter 5.

IQC requires the use of appropriate RMs, with known characteristics or reference values. Whilst efforts have been made at York to use an internal standard for calibration and incorporate in-house standard solutions into routine analysis, further guidance is required to

ensure they are used to gain their maximum benefit, together with perhaps the use of matrix-similar (in the absence of matrix-matched) reference materials.

Taken together, validation and quality control form the backbone of quality practices.

Because the method had already been in routine use at York for several years, it had been assumed that the method had already been fully validated. The emphasis of the current research was therefore to evaluate, retrospectively, accumulated analytical data and derive uncertainty estimates. However, research has shown that whilst certain aspects of validation have been evaluated (i.e. limits of detection and quantification, sensitivity, linearity etc), others are lacking and require attention (i.e. precision analysis, ruggedness testing and bias analysis). Nevertheless, a full and formal method validation was beyond the scope of the current study. Consequently in the absence of this, data derived from current RMs used at York have been evaluated as far as possible. The following chapter is therefore given to these evaluations and ends with suggestions for their future use.

However, before the data are presented, it is worth first considering the role of RMs and how they are used in validation (calibration, precision analysis and bias control) and IQC.

3.4 Use of reference materials (RMs)

Determination of measurement uncertainty assumes two fundamental prerequisites; the use of a validated method and that the measurement process remains in statistical control. In order to accomplish this, both processes require the use of reference materials.

Emons (2006, p690) points out that the ISO Guide 35 (relating specifically to reference materials) defines a RM as,

> "....a *material sufficiently homogeneous and stable with respect to one or more specified properties, which has been established to be fit for its intended use in a measurement process. NOTE 1: RM is a generic term. Note 2: Properties can be qualitative or quantitative, e.g.; identity of substances or species. Note 3: Uses may include the calibration of a measurement system, assessment of a measurement procedure, assigning values to other materials and quality control. Note 4: An RM can only be used for a single purpose in a given measurement."*

The use of RMs can cause much confusion as reference to them often implies intended use rather than a definition of their intrinsic properties (Emons, 2006). Thus a

reference material, certified or not, may also be referred to as an in-house standard or laboratory control material. Emons refers to there being a family of RMs and suggests that the term Quality Control Material (QCM) should be distinguished from a certified reference material (CRM). A QCM would then refer to a material of suitable homogeneity and stability so as to be fit for its intended purpose but not of sufficient characterization to be used for calibration or provide metrological traceability. CRMs in comparison, are accompanied by a certificate indicating their accepted reference value and associated uncertainty and provide metrological traceability back to the international system of units (SI). Measurements on CRMs effectively calibrate the whole procedure to a traceable reference, determining the combined effect of many sources of uncertainty (EURACHEM / CITAC, 2000). Finally a third group, referred to as calibration standards or calibrants (CAL), may or may not be CRMs, but in addition to certification and traceability, they have sufficient characteristics suitable for calibration (Emons, 2006).

The International Harmonized Protocol for IQC of analytical chemical laboratories (Thompson and Wood, 1995) suggests that where a CRM is not appropriate (or available), then it is up to individual laboratories or groups of laboratories to produce their own in-house RM. This situation is also recognised by the European Commission who provide guidance on producing CRMs via a collaborative trial approach (Quevauviller, 1998) and guidance on the production of in-house RMs can be found on the LGC website.

3.4.1 RMs in Validation

3.4.1.1 Precision

One of the primary roles of validation is the determination of uncertainty on results obtained from the application of a measurement method under prescribed conditions; those being repeatability, reproducibility and intermediate conditions. There are different suggestions as to how this should be carried out. Estimates for the standard deviation for a single result from results of duplicate analyses (single run), are achieved by taking the standard deviation of the differences between the measured pairs and dividing by the square root of n (i.e.; 2) (Barwick *et al.*, 2000). However, of most use is an estimate of the total, combined precision that takes into account both the repeatability precision (s_r) and run to run precision (s_{run}), since both of these sources are operating on an individual sample. Estimates of the combined total precision are then given by $s = \sqrt{(s_r^2/n + s_{run}^2)}$ (Thompson *et al.*, 2002). Alternatively it is suggested that s_{tot} could be measured directly by the analysis

of a single sample in successive runs and simply taking the standard deviation of results in the usual way (Thompson *et al.*, 2002). Presumably S_{tot} is then equivalent to S_{RW}, more usually derived by the use of an analysis of variance, ANOVA, although a comparison of differences in values thus derived has not been carried out in this study. Precision representing the within-run and between-run variation can be obtained by using ANOVA to separate the different contributions to uncertainty (Barwick *et al.*, 2000; Thompson *et al.*, 2002).

However, all the above approaches agree that data used should be sufficiently representative of all likely variability that might be reasonably expected to occur during routine application of the method, including changes of instrument and operator where relevant. It is essential that for each sample analysed, separate portions of the material have been taken and worked up through the whole method, otherwise precision estimates will not mirror the full extent of variability on a test sample. The number of replicate injections taken is dependent on the method protocol and should mimic the method exactly as applied to routine samples.

Reference materials used for precision analysis should be equivalent to those tested in routine analysis; i.e. be of the appropriate matrix and concentration and suitable to undergo all stages of the measurement procedure, including early preparation and extraction, in addition to instrumental analysis, recovery correction (if applied) and calibration. Suitable materials that could be used include CRMs (if available) or other quality control materials such as those left over after a proficiency test which therefore posses a consensus value and an uncertainty estimate. Further, the cost of using CRMs on a regular basis may be prohibitively high and both CRMs and prepared test materials may show tighter homogeneity than material typically presented for analysis and as such may not be truly representative. Alternatively, a test sample of sufficient quantity, homogeneity and stability, could be used for the development of an in-house RM. Guidance on the production of in-house RMs is freely available from LGC (Brookman, 1998) and makes the recommendation that an in-house RM should be calibrated initially against a CRM. This would enable precision estimates to be determined and ensure absolute accuracy of the material under analysis which otherwise may not be known and may be subject to bias influences. In the absence of CRMs, chacterisation of a candidate material can be achieved through collaborative trail. Precision may be expressed as relative precision estimates but the effect of concentration should be checked, since precision very often varies with concentration.

Regardless of the material used, standard solutions are not suitable for precision evaluation of routine materials for a number of reasons; the analytes may not be in the same form, will not be subject to matrix effects and do not represent the total variation from the application of the whole measurement procedure on a test sample. However, if they are sufficiently well characterised, they can potentially be used as CALs in calibration, (see section 3.4.1.3) or be used as QCMs to monitor analytical stability (see section 3.4.2.5).

3.4.1.2 Bias

Trueness is defined as the closeness of agreement between an expected result and the true value, which in practice is replaced by a reference value. Trueness is usually expressed in terms of the bias, or the difference, which represents a measure of the systematic error effects in a measurement system. Bias is determined through a bias study and again should be representative of the range of concentrations and matrices to which the method will be applicable. Therefore a bias study should include at least a representative random sample or routine materials, and perhaps those matrices or concentrations at the extreme of the analysis, i.e. which present the greatest challenges.

In determining bias the usual approach is to compare results of the analysis with a RM, alternatively comparison against a referenced method is also possible. When considering which type of reference material should be used, it is important to consider the application of the method. For example for compliance and regulatory purposes a certified reference material has the highest level of traceability and a stated concentration with a known level of uncertainty. If a suitably matrix-matched CRM is available then this should be the preferred option. In which case, the bias is simply the difference between the mean of the measurement results and the certified reference value; i.e. $bias = \bar{x} - x_o$. However, often the CRM is not of a suitable matrix that will reflect the behaviour of the analyte in the sample and may not respond to the measurement procedure in the same way. A commonly used alternative approach is to prepare a stable in-house reference material that can be used for long term work and trend analysis. For short-term or non-critical work a spike may be sufficient either added to a previously analysed sample or a second sample with the analyte of interest. Where a referenced method is being used (i.e. one previously validated through a collaborative trial), then the results of a test method can be compared to those of the reference method. It should be noted that in most instances bias measurements constitute both the laboratory and method bias components, although for an empirical method, method bias contribution is zero. In situations where a suitable commercial RM is not

available and an in-house RM has not been prepared, then it is necessary to compare measurement results with the consensus derived from other laboratories in a proficiency test. Regular participation can be invaluable in providing long term monitoring of measurement stability and providing impartial evidence of fitness-for purpose to potential clients.

Measurement procedures do not always extract all the analyte under investigation, but it is not always possible to determine how much is actually present. This is where spiking into a test sample is helpful. Although the analyte may not mimic the exact behaviour of the analyte in the matrix it is the most common approach used (EURACHEM, 1998). The significance of the bias effect needs to be evaluated.

When using a reference material with an assigned concentration value c_R; and the mean of repeated measurement observations of the sample c_o, bias can be expressed simply as the difference between the two or as the relative value;

$$Bias = c_o - c_R \qquad or \qquad \%Bias = \frac{c_o - c_R}{c_R} \times 100 \tag{3.3}$$

Measurement bias is often referred to as the recovery (R), and expressed as a ratio or a percentage;

$$Recovery = \frac{c_o}{c_R} \quad or \quad \%R = \frac{c_o}{c_R} \times 100 \tag{3.4}$$

The bias, or recovery is used to correct measurement results, whilst it is usually only the uncertainty component that gets included in the uncertainty budget (see below). The measurement uncertainty for %Recovery determined from a reference value is given by;

$$\frac{u_{\%R}}{\%R} = \sqrt{\left(\frac{u_{c_o}}{c_o}\right)^2 + \left(\frac{u_{c_R}}{c_R}\right)^2} \tag{3.5}$$

Where u_{c_o} is the uncertainty in the mean value from repeated measurements of the reference material, i.e. $u_{c_o} = c_o/\sqrt{n}$ and u_{c_R} is the uncertainty of the reference material obtained from the supplier or certificate (for a CRM).

Where recovery has been determined through the spiking of a test portion, recovery is calculated as;

$$Recovery = \frac{c_{sp} - c_b}{c_s} \quad and \quad \%R = Recovery \times 100 \tag{3.6}$$

where; c_{sp} is the mean value of repeated measurements of a test portion after the addition of the spike,

c_b is the mean value of repeated measurements of a test portion before the addition of the spike,

c_s is the increase in concentration of the test portion after addition of the spike.

The associated %Recovery is determined by;

$$\frac{u_{\%R}}{R} = \sqrt{\frac{\left(u_{c_{sp}}^2 + u_{c_b}^2\right)}{\left(c_{sp} - c_b\right)^2} + \left(\frac{u_{c_s}}{c_s}\right)^2} \quad \textbf{(3.7)}$$

where; $u_{c_{sp}}$ is the uncertainty in the mean value from repeated measurements of the sample after addition of the spike, i.e. $u_{c_{sp}} = c_{sp}/\sqrt{n}$

u_{c_b} is the uncertainty in the mean value from repeated measurements of the sample after before of the spike, i.e. $u_{c_b} = c_b/\sqrt{n}$

u_{c_s} is the uncertainty in the mean value from repeated measurements of the increase in concentration of the sample after addition of the spike; calculated from the reference material uncertainty value and weighing and volumetric activities involved in the preparation of the spike.

Having determined the recovery and its associated uncertainty, there are three possible scenarios;

1. Recovery is not significant and results are not corrected
2. Recovery is significant and results are corrected
3. Recovery is significant and the results are not corrected.

To determine the contribution of recovery to the combined uncertainty for the whole method, the estimate is compared to 1 using the test statistic t (Barwick & Ellison, 1999, 2000), where u_R is the recovery uncertainty, not expressed as a percentage, thus,

$$t = \frac{|1 - R|}{u_R} \quad \textbf{(3.8)}$$

If the number of degrees of freedom are known for u_R (a GUM Type A uncertainty derived from repeated measurements) then the t value can be compared to the 2-tailed

critical value, t_{crit} for the relevant degrees of freedom at 95% confidence. If $t < t_{crit}$, then recovery is not significantly different from 1 and need not be counted.

If, however, the number of the degrees of freedom are not known due to the contribution of a reference value (a GUM Type B uncertainty derived from a certificate or probability model) then t is compared with k, the coverage factor used in the calculation of the expanded uncertainty for the measurement result. If $t < k$, then recovery is not significantly different from 1 and can again be ignored.

If there is a significant difference, then the GUM states that values must be corrected for bias in the determination of the measurement result, this is achieved through applying a recovery correction factor. If the bias is significant but the method does not require a correction to be applied, then the uncertainty contribution to the overall combined uncertainty must reflect this additional associated doubt and uses the equation below to account for it (Barwick & Ellison 2000);

$$u'_R = \sqrt{\left(\frac{1-R}{k}\right)^2 + u_R^2} \qquad \textbf{(3.9)}$$

The exception to this is where the method is empirical and the bias is assumed to be equal to 1 and no correction is necessary

3.4.1.3 Calibration

Calibration as part of the method validation process is a far more extensive process that is usually employed as part of routine analysis (Horwitz, 1995) However, calibration evaluation as part of validation, can provide some important information that may subsequently affect the quantitative reporting of measurement results (such as linearity, whether the correlation passes through the origin, and matrix effects). RMs used for calibration (calibrants or CALs) may be pure substances, standard solutions (if sufficiently well characterised) or matrix-matched materials such as CRMs, depending on requirements and availability. CALs, may or may not be CRMs. Whether they are pure substances, standard solutions or matrix specific, in all cases they need to be sufficiently defined to provide a reliable reference value with stated uncertainty which can be used to accurately translate instrument response (such as peak area values), into useable concentration units. The benefit of using a matrix CRM is that it requires treatment in the same way as a test sample and as such, will mimic matrix effects. Measurement results from unknown samples will therefore be automatically corrected for recovery of analyte losses encountered during the

measurement stages. Use of a pure substance (i.e. a standard solution) will only correct for losses during separation and derivatisation on the HPLC.

Calibration, as described by the IUPAC Harmonized Guidelines relate solely to what is termed **external calibration**. This requires the evaluation of several, (perhaps five or six) suitable reference materials and the instrument response (y-axis) plotted against the concentration (x-axis) and a calibration curve fitted. If the line is straight and goes through the origin, then it may be appropriate to only use a single point check during routine use, otherwise three or more may be required (Jones, n.d). The use of known value QCMs can be run straight after the calibrant(s) to check the calibration and make sure it is giving the expected result, within the known acceptable range (see section 3.4.2.5). External calibration is therefore determined during validation but monitored and adjusted when necessary (perhaps with fresh calibrants or after instrument servicing and repair). However Jones notes that this form of calibration is acceptable provided there is no fluctuation in injection volume and there are no changes in HPLC conditions during the run (i.e. after the calibrants have been run).

To eliminate the effect of injection volume fluctuation, an **internal standard (IS) calibration** can be used. An IS is a compound, similar to the analyte(s) in question but sufficiently different so as not to interfere with the native species i.e. a non-naturally occurring compound. This will then be added to all standards and samples at a fixed level. At York, a known concentration of L-homoarginine is used and incorporated into the rehydration fluid, thus; $\frac{Area(IS)}{Conc\ (IS)} \propto \frac{Area(S)}{Conc\ (S)}$ where (S) relates to the unknown sample. However, this assumes that the ratio of Area(IS)/conc(IS) is the same as that for the naturally occurring sample, and that the relationship is constant for all concentrations and all amino acids. However, this is not calibration as it does not involve a calibrant. Cuadros-Rodríguez *et al*. state "...*Furthermore, calibration using just the so-called internal standard cannot be made*.", (2001 p627).

Internal standard calibration requires the spiking of the IS into external standards, (the standard solution or pure substance). Jones (n.d.) describes a method similar to that of external calibration, where several RMs are spiked and a calibration curve derived. However, this time using CAL values normalised by the IS equivalents; peak area (std sol)/peak area (IS) on the y-axis, against conc (std sol)/conc (IS) on the x-axis. Thus the ratio values will remain constant even if injection volume fluctuates. The calibration can then be checked in each run

using either a single-point (spiked standard solution) or multi-point if the calibration curve is non-linear, or doesn't go through the origin.

An alternative approach to single-point IS calibration is described by Cuadros-Rodríguez *et al*., (2001), and is perhaps a more practical solution for AAR analysis. A calibrant is spiked in the same way as previously described (or already used as a constituent of the standard solution as in the case of AAR) and a calibration response factor (F) is obtained;

$$F = \frac{Conc_{(std\ sol)}/Area_{(std\ sol)}}{Conc_{IS}/Area_{IS}} \quad \textbf{(3.10)}$$

The unknown concentration of the sample is then given by (see also equation (3.2));

$$Conc_{sample} = F \cdot \frac{Conc_{IS}}{Area_{IS}} \cdot Area_{sample} \quad \textbf{(3.11)}$$

This way, individual response factors can be determined for each isomer separately in each analytical run using known concentrations of the amino acids in the standard solution. Although the use of standard solutions does not take into account analyte losses during the preparative stages, it does at least permit detector response and instrumental losses to be accounted for (assuming that amino acid isomers in solution behave in the same way as they do in a matrix extract). This subject is expanded on with examples of AAR data in section 4.5.3.

However, it should be noted that random errors in the determination of the response factor should be minimised by taking replicate measurements, and single-point calibration is only acceptable if the observed scatter is small enough compared to the expected precision for the method Cuadros-Rodríguez *et al*. (2001). De Bievre contests (1999) that single point calibration only partially corrects a measurement result as it is derived from only a single calibrant and not from interpolation of a series, thus there is no check on the accuracy of the calibrant's data point(s).

Nonetheless, IS calibration is still an accepted practice in the analytical community, especially where the expected analyte concentration range is limited (Cuadros-Rodríguez *et al*). Another approach to calibration that may be especially relevant to AAR is that of, standard addition. When the response of the detector to the matrix is not known, (compared to that of the calibrant), the method of **standard addition (SA) calibration** may be useful. SA calibration is linked to the evaluation of bias (section 3.4.1.2), but Emons (2006) warns that the same RM should not be used for the evaluation both bias and calibration. In SA

calibration, two aliquots of the same sample extract are taken, one has a known volume of calibrant added (containing analytes at a known concentration), whilst the second sample is diluted with the same volume of an appropriate analyte-free solution or water. The instrument responses to the spiked and unspiked samples are measured. The concentration of the analyte in the unknown sample is then given as;

$$Conc_{sample} = \left(Conc_{(std\ sol)} \cdot \frac{Vol_{sample+spike}}{Vol_{sample}} \right) \times \left(\frac{Area_{sample}}{Area_{sample+spike} - Area_{sample}} \right) \quad \textbf{(3.12)}$$

3.4.2 RMs in Internal Quality Control (IQC)

According to the RSC Analytical Methods Committee, the role of IQC is "*to check that the uncertainty at validation does not deteriorate after validation*..." (RSC Analytical Methods Committee, 2010, p1). Method validation provides a means of determining method performance parameters that may be typically expected in routine analysis. However, these characteristics are only consistent as long as the measurement process remains in statistical control. Generally speaking, validation of the method is often carried out using known material, whereas routinely the method will be applied to samples of unknown material. Thus in order to ensure the process remains in control, it is important to be able to run stable materials with known performance characteristics, alongside the unknown samples in order to control the quality and provide confidence in the results of the unknown materials, hence the use of the name quality control materials (QCMs). The purpose of IQC therefore, is to monitor the stability of the entire method, not just the instrumental analysis. IQC includes the use of blanks, calibrants, spiked samples, replicate analyses, QC samples and control charts, and enables the performance of a method to be monitored over time.

3.4.2.1 Blanks

The analysis of blank samples is designed to identify issues with contamination. Reagent blanks (procedural blanks) contain all chemicals added during analysis (except the test sample) and go through the procedure as if samples themselves. These blanks help to identify problems of contamination from reagents, glassware and vials etc. Sample blanks are made up from material very similar to the test samples (if available) but do not contain the analytes of interest. These can be used to monitor interferences from the sample matrix, resulting in false positives. In either case, detection of analytes, significantly above zero or

the limit of quantification (LOQ), imply that test samples may require correction and the source of contamination requires identifying and eliminating.

3.4.2.2 Calibrants

Calibration is covered in greater detail in the earlier section 3.4.1.3. Full calibration is generally only evaluated and determined during validation, with the introduction of new reference materials, or after instrument servicing or repair. However, use of spot checks using a single calibrant ensures that the existing external calibration remains in control and there has not been any instrumental drift requiring recalibration. By comparison, an internal standard calibration will generally be carried out with each run, to correct for fluctuations in injection volume and other instrumental based losses and detector response. Internal standards added to pure substances or standard solutions will only correct for the analytical (instrumental) effects. Internal standards added to matrix-matched calibrants at the start of the measurement procedure (and therefore undergo the same treatment along with the sample), will additionally account for method losses as long as the measurement process will not degrade, denature or otherwise affect the properties of the internal standard.

3.4.2.3 Spiked samples

Spiking (or fortification) involves the addition of the analyte(s) in question to either a previously analysed test material or a duplicate sample. The sample to which the spike is added may or may not contain the analytes in question. If the sample material is free from analyte, then in effect a matrix-matched RM is being made with a known analyte level. Spiked samples are especially useful for recovery checking when analytes or sample matrices are considered unstable. When no suitable QCM is available it can be used to identify bias and is particularly helpful for one-off analyses which may not fall under the scope of existing validation. The recovery of the added analyte (the marginal recovery) is then the difference between the spiked and unspiked samples, divided by the amount of analyte added. However, there is an underlying assumption that the recovery of the spiked analyte is equivalent to the recovery of the matrix bound analyte but it is "....*difficult to ensure speciation, binding and physical form of the added analyte is the same as the native analyte*...." (Horwitz, 1995). Further it is essential that the RM used for the spike and calibration are not traceable back to the same stock solution, as separate sources of error will not be detected.

3.4.2.4 Replicate analyses

The use of the occasional replicate analysis provides a check on repeatability. Assuming the measurement system has already undergone validation, replicate measurements of all samples are not necessary. However it should be stressed that in this situation, both replicates are independent, that is, they have been worked up from separate sub-samples of the material being used. This is not the same as replicate injections for the same extract from the same vial. This only provides an estimate of instrumental repeatability and likely to be extremely small compared to the other uncertainty influences encountered whilst undergoing preparation and extraction.

The purpose of running duplicate samples is to "*...ensure that the differences between paired results are consistent with or better than the level implied by the value of* σ_0 *used by the laboratory for IQC purposes*." (Horwitz, 1995) In this context σ_0 is the repeatability standard deviation s_r. Such information can be helpful interpreting control charts, and especially helpful when running non-standard samples such as matrices for which precision parameters may not have been previously evaluated.

In this instance; $S_r = \sqrt{\sum d^2/2n}$, where d is the difference and is given as; $|d|=|x_1-x_2|$. However, care needs to be taken regarding the appropriate concentration range and the control limits used for comparability. If the concentration range of the samples being duplicated are the same as those used for QCMs and control charts then the 95% probability (approx 2 std dev) control limit is set as the repeatability, r, where r is the value less than or equal to the absolute difference between two measurement results obtained under repeatability conditions.

$r = t \times \sqrt{n} \times s_r$ where t is the t-value for a normal distribution at 95% probability, i.e. 1.96 , rounded to 2, and n is the number or replicates , i.e., 2, (Horwitz, 1995).

This subject is expanded on further with detailed examples in section 4.5.1.

3.4.2.5 QCMs and Control Charts

Having carried out a method validation, checks need to be made that ensure the measurement system doesn't significantly deviate from the predicted range. Control is provided by the use of quality control materials (QCMs) used during an analytical run. QCMs have known precision parameters and may be CRMs, the same material used during validation or standard solutions. Often cost prohibits the use of CRMs as QC materials and

in-house, matrix matched materials that can be produced in bulk with sufficient homogeneity and stability may be preferable. The only requirement is that the material is available in a sufficiently large enough quantity to provide continuity over time, that the material is homogeneous and stable for the duration over which the bulk is intended to be used. If matrix-matched QCMs are used then they are worked up through the whole method along-side the test samples. This permits the stability of the whole measurement procedure to be monitored. If standard solutions are used, control is limited to the stability of the instrumental analysis.

The most effective way of monitoring QC materials is through the use of control charts, typically a Shewhart chart. Shewhart charts plotting results of individual results are known as X-Charts, the mean of replicate analyses are called X-bar charts, ranges (R-charts) and standard deviations as (s-charts) (RSC Analytical Methods Committee, 2010) Values are plotted in a time ordered manor with warning and action level indicators represented by ± 3 and $\pm 2\ s_{RW}$ lines respectively. These can be derived either from external fitness for purpose requirements (such as legal limits) or validation precision (Ibid). X-charts can help monitor both random and systematic effects and an R-chart can also provide repeatability control. Data from control charts can also be combined to calculate measurement uncertainty (Nordtest 2007). The use of control charts is considered in further detail with AAR examples in section 4.5.2.

For an X-chart plotting individual values, $s_{RW} = \sqrt{(s_r^2 + s_{run}^2)}$

For an X-Bar chart plotting the mean of *n* replicates, $\bar{s}_{RW} = \sqrt{(s_r^2/n + s_{run}^2)}$. However, the number of replicates must always stay the same, otherwise the $\bar{s}_{RW}$ value will vary with *n.* For this reason the X-chart for individual values is probably preferable.

It is possible that for routine use where the range of concentration may be limited that only a single QC material would be needed. For legal or threshold testing, a QC sample close to the limits is suggested and for an analyte whose concentration range varies, possible two different control materials representing the typically expected range could be used. For short runs with few samples, at least one QC sample should be used, for longer runs with more samples perhaps 1 every 10 test samples might be preferred (AMC 2010).

Guidance on the interpretation of Shewhart charts is provided in Appendix 3 of the IUPAC protocol (Horwitz, 1995), which presents the Westgard Rules which have been detailed below;

For a single chart, an "out of control system" is indicated if;

- the value falls outside the action limit (± 3),
- the previous and current values fall outside the warning limit (± 2) but inside the action limit,
- 9 successive values fall on the same side of the mean line.

For 2 control charts;

- at least one value falls outside the action limit (± 3),
- both values are outside the warning limits,
- the previous and current values of the same chart fall outside the warning limit (± 2) but inside the action limit.

3.5 Conclusions

This chapter has looked at the principles of measurement uncertainty as applied to AAR analysis, both from the "bottom-up" and "top-down" perspectives. It is clear that for all but the simplest of measurement procedures, the GUM's "bottom-up" approach of identifying and accounting for all uncertainty contributions is hugely time consuming, at times bewildering and potentially grossly inaccurate in all but the most experienced hands. By comparison, "top-down" approaches take an overview of uncertainty. These approaches adopt the policy that all uncertainty influences accumulated by a measurement process will be reflected by the uncertainty of the final result, regardless of whether these components can be individually accounted for, and avoids the risk of under or over counting contributions. The use of reference materials in method validation and quality control activities employed to minimise error influences has also been briefly discussed but will be further expanded on in the next chapter.

Chapter 4 will now take a retrospective look at intra-laboratory uncertainty determination from the AAR archive of RM data. A "top-down" approach has been adopted, with consideration firstly given towards the normality of the distribution, in order to confirm underlying statistical assumptions prior to analysis.

Chapter 4. A Retrospective Analysis of MU in AAR

4.1 Introduction

The original RP method used for AAR analysis at York, was presented by Kaufman & Manley in 1998. It has been subsequently refined and developed by Penkman *et al*. (2008) to include an initial bleaching pre-treatment stage after Sykes *et al*. (1995) and the routine analysis of both the free (FAA) and total hydrolysable (THAA) amino acids. As the modified method has now been in use in the laboratory for several years with an accumulation of several thousand sets of results, it had been assumed that the method had already underdone vigorous testing as a result of application in routine use. Thus it was not within the scope of the current research to undertake a full method validation, but to review data retrospectively.

Currently, the reference materials (RMs) used routinely in AAR analysis include;

- reagent blanks,
- three stock standard solutions classified by D/L value; 0.167d, 0.5d and 0.91d
- an internal standard (L-homo-arginine or LhArg) present in the rehydration fluid at a level of 0.01 mM.
- three sets of in-house biomineral matrix RMs; ILC-A, ILC-B and ILC-C

The biomineral matrix RMs, were produced by Wehmiller for an AAR inter-laboratory comparison in 1984. Each of the three bulk stocks were produced from ground mollusc shells, and are similar in composition (i.e. a calcium carbonate biomineral matrix), but not identical to, the opercula matrix, which forms the basis of this study. Nonetheless, analysis of these materials requires the application of the entire method, and as such they could be considered as in-house RMs. However, the properties of these materials have not been characterised by precision experiments and are not currently used routinely as part of the analytical IQC. The original inter-laboratory study (Wehmiller, 1984) was conducted predominantly using Gas Chromatography (GC) and Ion Exchange (IE) analysis. However,

differences between GC and RP measurement results have been recognised (see Chapter 5) and consequently, any previous consensus values cannot be relied upon for RP applications.

Due to the absence of commercially available or otherwise defined reference materials, assessment of bias has proven impossible for the laboratory. For the last few years, the laboratory has carried out a large number of analyses on individual *Bithynia* opercula. Because of their minute size, this has meant that developing a sufficient quantity of a matrix-matched reference material impossible too. As a result, there has been no formal evaluation of precision parameters carried out either. Uncertainties are expressed as precision estimates, either from repeated injections for individual opercula or the average injection precision from multiple opercula. However, Injection precision represents only the instrumental component of uncertainty and is not representative of the whole method.

Due to the difficulties in assessing bias, for the purposes of this chapter, AAR analysis is assumed empirical, that is, the method defines the output. Therefore results are specific to the individual laboratory carrying out the analysis, in order that bias contributions can be assumed negligible. AAR uncertainty estimates are therefore limited to estimates of precision. In the absence of validation data, the derivation of precision estimates by other means is considered below.

4.1.1 Evaluating precision in AAR

Site-specific AAR analysis usually involves repeated measurements being made for a given location. Therefore measurements are often made using multiple individual opercula (i.e. multiple samples) taken from the original sediment (primary sample). If precision estimates are derived from multiple measurement results (i.e. multiple samples), this would describe the distribution and represent the precision for that sediment. However, a review of past data indicates that the measurements made on multiple different samples (taken from the same primary sample) are not always measured during the same analytical run or on the same instrument, although they might be. Resulting measurement values (from multiple individual samples) therefore represent a mixture of repeatability and reproducibility conditions, i.e. repeatability conditions with the odd reproducibility value thrown in, or vice versa. This mixing of precision conditions results in an inherent inconsistency in the nature of the precision estimates derived for each site and will affect uncertainty comparability between sites of similar ages and temperature histories. For example, assuming all repeated injections, (n=2) represent the instrumental (repeatability) precision, the sample to sample variance may be derived either under repeatability or

intermediate reproducibility conditions, depending on whether additional samples were analysed during the same run or different runs perhaps on a different instrument. This makes separating out sources of uncertainty and determining precision estimates from multiple individual unknown samples, problematic.

The ILC samples are not run routinely and data are limited. However, potentially they could be used as indicators of precision, even though the ILC matrices are not exactly the same, being a calcite/aragonite mix, whist the opercula are predominantly calcitic. Further complications arise due to the ILC materials having been made from a number of individual shells, therefore the homogeneity may not be comparable to that found routinely in individual opercula.

This brings us to the internal standard solution and the D/L standard solutions. The concentration of the internal standard, L-homoarginine, is set as 0.01 mM in the rehydration fluid. Rehydration of dried samples following hydrolysis or demineralisation uses a stock solution that lasts many months. For the calculation of the unknown concentrations, it is assumed that there is no significant variability in the concentration of the LhArg between the infrequent batches, although there will be slight variability due to weighing and volumetric error influences as suggested in the last chapter (section 3.2.2). However, even though concentration is assumed to be fixed for the purpose of the method, fluctuations in the peak area of a "fixed" concentration will provide an indication of the level of stability in the instrumental determinations (see section 4.2.1). If introduction dates are known, this could also provide valuable information on batch-to-batch variability of the internal standard and identify systematic offsets.

Records of the D/L standard solutions, provides the most data. At least one sample of standard solution is analysed every 24 hours as a visual check on measurement stability, the sample used depending on the expected D/L range of the samples under analysis. Each vial of standard solution can be used for up to 5 HPLC injections, being refrozen in between runs if necessary.

4.1.1.1 Analysis of Variance, ANOVA

ANOVA is a statistical technique frequently used in hypothesis testing. It evaluates the significance of variation due to one or more experimental factors, compared to the effect of purely random influences on the variability of observed data. ANOVA is a powerful tool that can separate and determine the contribution from different sources of variation (Miller

and Miller, 2005) and it is this function that is exploited in the calculation of analytical precision estimates (ISO 5725, 1994; ISO 21748, 2010).

An evaluation of the standard solutions by ANOVA can reveal information regarding the stability of analytes over time. The repeatability precision represents the instrumental precision of multiple determinations from a single vial i.e. $n \leq 5$. The between-run precision therefore represents the level of agreement between individual vials (although taken from the same original bulk stock solution). Whilst it can be appreciated that this data is not representative of solid matrix materials or the uncertainty arising from the application of the whole measurement procedure, it provides a baseline and characterises the AAR precision estimates without the interference of matrix effects.

The use of ANOVA in hypothesis testing, shares a prerequisite of parametric statistics, that the data being evaluated obey certain assumptions, these being,

i) Independence
ii) Homoscedasticity or equality of variances
iii) Approximate to normality

However, for the purposes of this research, the interest is not so much in the determination of significant differences between the groups (analytical runs) but rather in the numerical determination of the variation components, i.e. the within and the between-run variability.

Nonetheless, independence of observations is provided by the use of different standard solution sample vials used on different instruments over time. This factor variation represents the between-run variability. The within-run variability or repeatability represents the variation due to random effects only acting on repeated measurements of the same vial. Thus although these data are not strictly independent (i.e. different vials analysed in a single run), data are determined from separate injection chromatograms and is sufficient for a retrospective evaluation of instrumental random error.

Homoscedasticity or equality of variances implies that the variance of the random component of variability is independent from the factor variability. This is important in order to be able to pool the within sample variances when calculating the overall random variability (Miller and Miller, 2005). With regard to standard solutions, it is a reasonable assumption that the instrumental variance would not change significantly from day to day. For the purposes of these evaluations, even if there is a slight difference in repeatability

estimates between instruments, then the pooled data will reflect the extent of this variability in the precision estimate, and need not be an issue.

The final assumption of normality is important in the application of an *F*-test when determining the significance of variances between two samples assumed to have been drawn from a normal population (Miller and Miller, 2005). In the absence of this assumption being true, the risk of obtaining a false positive increases, i.e. the null hypothesis, H_0, is rejected, that is a significant difference is observed when in fact there isn't one (Type I error) or H_0 is retained, that is there is no significant difference observed when in fact there is one (Type II error) (Miller and Miller, 2005).

However, Miller and Miller (2005, p61) continue *"...the F-test as applied in ANOVA Is not too sensitive to departures from normality of distribution*".

McDonald (2009, p151) explains,

> *"Fortunately, an anova is not very sensitive to moderate deviations from normality; simulation studies, using a variety of non-normal distributions, have shown that the false positive rate is not affected very much by this violation of this assumption (Glass et al., 1972; Harwell et al., 1992; Lix et al., 1996). This is because when you take a large number of random samples from a population, the means of those samples are approximately normally distributed even when the population is not normal."*

It is further suggested that although the *F*-test has "*assumptions and practical limitations*", there are no assumptions required for the general use of ANOVA (Montgomery, 2001).

Therefore, whilst it may appear that close adherence to these assumptions may not be implicit for the application of ANOVA in the determination of precision estimates, it would be good practise to consider the level of agreement observed between the two instruments, or indeed, the normality of the distributions prior to the calculation of precision estimates from pooled data by ANOVA. Consequently, the majority of section 4.2 is given to the comparison of standard solution D/L values using t-test evaluations of significant differences between instrumental means, the identification of outliers and determinations of central tendency and normality.

Section 4.3 presents an evaluation of precision estimates by ANOVA for D/L values in standard solutions, with further consideration given towards outliers, repeatability limits,

sample size and expanded confidence intervals. Section 4.4 is similar but with an emphasis on biomineral data, using both D/L values and concentrations from ILC data and also proficiency test data (see Chapter 5). Finally section 4.5 looks at the role of AAR RMs in routine quality control, (including; repeatability, control charts, bias control, response factors and calibration) and an explanation offered as to why observed A/I values of 1.3 or higher, may really only be equal to 1.0.

However, section 4.2 first begins with a brief evaluation of the internal standard (IS) LhArg data, plotted in run order on both Hew and Gilly HPLC instruments. Observations of the internal standard peak areas, provides a visual presentation of the stability of the analytical system. In principle, peak areas should be approximately equivalent between machines for a given moment in time, since the same batch of IS present in the rehydration fluid is being run on both instruments.

4.2 Reference solutions

4.2.1 Instrumental stability in uncalibrated data: LhomoArginine

Figure 4.1 and Figure 4.2 show plots of the internal standard (IS) L-homoarginine chromatogram peak areas determined on “Gilly” and “Hew“, the two HPLC instruments. Data have been taken as the mean of repeated peak area values from single vials of the rehydration fluid (LhArg blank), obtained during individual runs, between 2003-2010. Charts show the mean as a solid horizontal line, with ± 2 standard deviation confidence intervals as dashed lines either side. The linear trendline is also shown as a dotted line as an indication of the general trend.

It would seem from this data that Gilly has been the most stable of the instruments over time, whilst Hew has experienced some significant fluctuations. In 2003 both instruments appeared to be giving approximately similar peak area readings of about 700. However in Hew, there appear to have been two significant changes in the instrumental settings, around 2005 and 2007, so that by 2010 peak area values of LhArg on Hew had doubled to 1500. The confidence limits reflect this variability, with the RSD for Gilly being approximately 15% whilst that for Hew is 23%. For Hew the confidence range is exaggerated due to the change in instrument response over time and would no doubt, be tighter for shortened periods of time. It nonetheless highlights the differences in long-term peak area precision due to changes in instrumental calibration.

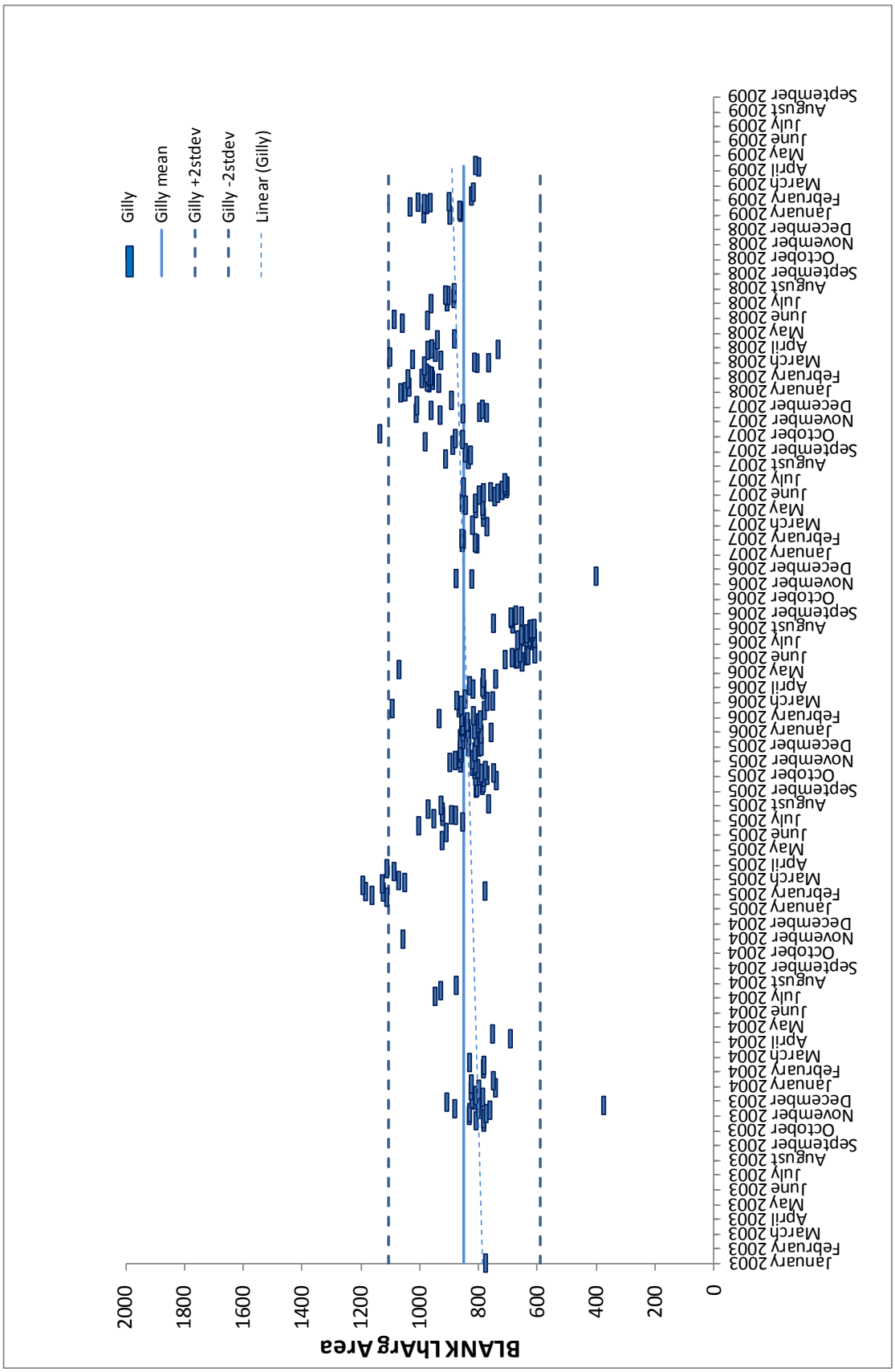


Figure 4.1: Peak Areas of LhArg in rehydration fluid (0.01mM) run on "Gilly"

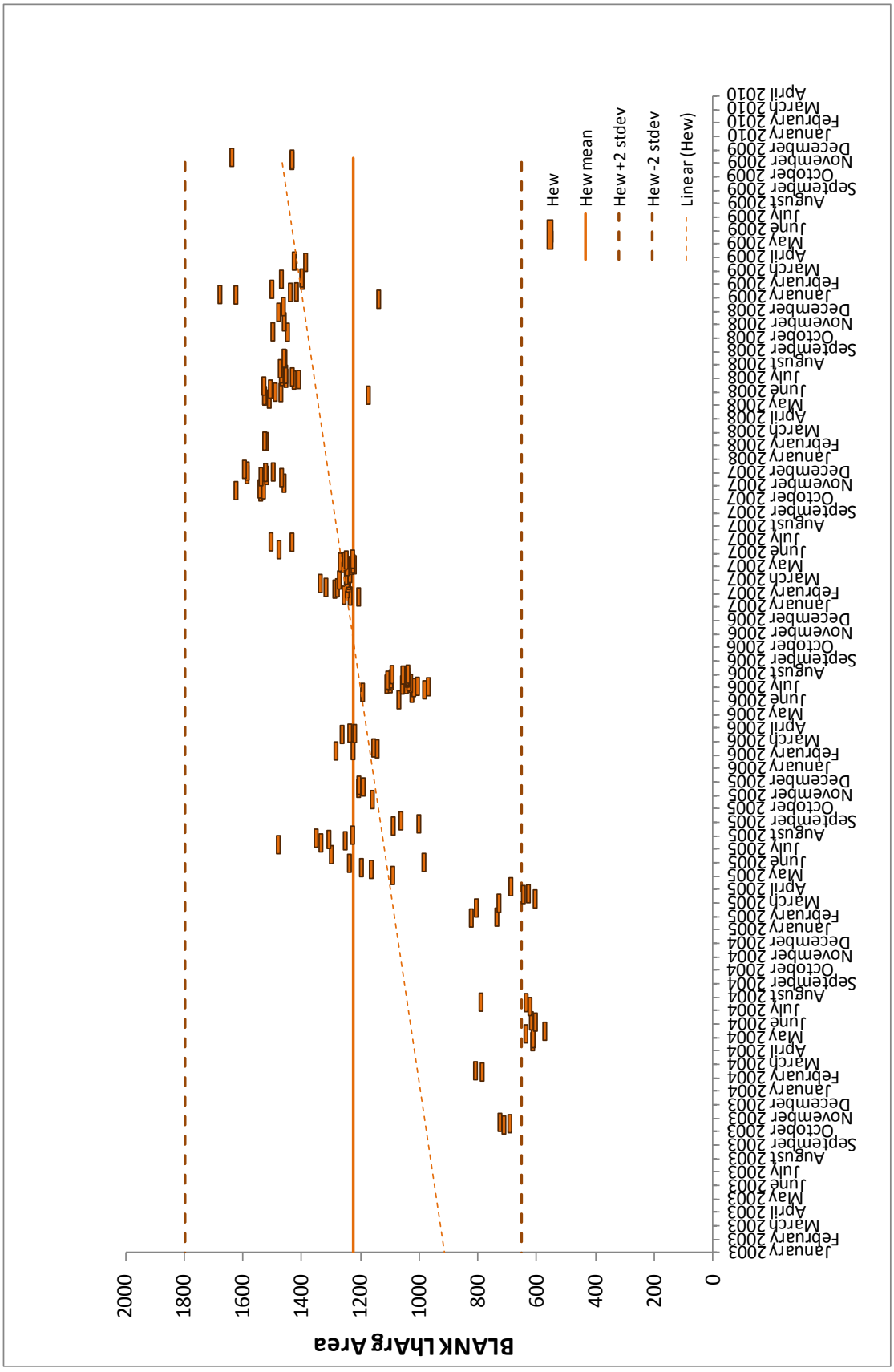


Figure 4.2: Peak Areas of LhArg in rehydration fluid (0.01mM) run on "Hew"

The cause of instrument fluctuation is thought to be due predominantly to bulb fluctuation over time, and they are replaced when levels get too low. These effects will have also acted on the standard solutions and test samples and influenced the peak areas in a similar way. Therefore, whilst these findings appear concerning, it should be borne in mind that standard solutions and sample extracts are all made up in the same rehydration fluid used for the LhArg blanks. These can then be normalised using the internal standard peak area values, thus enabling D/L values to be derived by the ratio of normalised D to L area ratios;

$$sample\frac{D}{L} = \frac{sample\ Area(D)/IS\ Area(LhArg)}{sample\ Area(L)/IS\ Area(LhArg)} \quad \textbf{(4.1)}$$

According to the chemical suppliers Sigma-Aldrich, L-homoarginine is described as an *unnatural arginine analog.* It produces a unique chromatogram peak that doesn't compete or interfere with the other naturally occurring amino acids. However its use as an internal standard in this context is dependent on three critical assumptions;

1. Detector sensitivity is the same between the amino acids and the internal standard.
2. Detector sensitivity is the same for both the D and the L isomers of a given amino acid,
3. Detector sensitivity is the same for both the D and the L isomers of different amino acids.

Issues related to instrument response factors will be considered in further detail in section 4.5.3. However, for now attention moves to the evaluation of the distributions of measurement results in standard solution.

4.2.2 Evaluating Normality and Identifying Outliers in Standard Solutions

In routine AAR analysis, both Hew and Gilly are used synonymously. Therefore intermediate reproducibility uncertainty estimates should be derived using pooled data. Before uncertainty estimates are derived from this data, it is important to establish that the pooled data approximate to normality and that there is no significant differences observed between the means obtained from one instrument compared to the means obtained by the other in order that precision estimates might be determined by ANOVA.

These evaluations of normality are based on the analysis of D/L values derived from the measurement results of standard solutions used in routine analysis. Each standard

solution is comprised of amino acid isomers at given concentrations, sufficient to produce D/L values defined by the individual solution, i.e. 0.167 D/L, 0.5 D/L and 0.91 D/L. Results for ten amino acids, for which both the L and D isomers can be reliably determined and are routinely quantified in test samples, have been assessed. If data is inherently non-normal, this allows the opportunity to apply an appropriate transformation to normalize the distribution, prior to evaluating precision estimates. Evaluation of standard solution precision provides an ideal baseline representing laboratory variability without matrix or age interactions.

Standard solutions are run routinely, one measured approximately every five test samples. Three standard solutions were available and each was made up of D/L mixtures of Asx, Glx, Ala, Arg, Ser, Val, Met, Phe, Leu, and Ile. Original evaluation of concentrations suitable for analysis lead to a range of trial solutions (Penkman, 2005) and it was solution 'd' that was found to be most appropriate. Original solutions were also made up in L-homoarginine but subsequently have been made up in water, resulting in two separate sets of data at the lower D/L level. Thus data have been evaluated for both 0.167d (LhArg) and 0.167dH20 standard solutions where indicated, together with the 0.5d and 0.91d solutions.

Original analytical data was accessed from the BioArCh Excel data archive. Each analytical RP run is given a unique reference, i.e. g002-6103.xls, where "g" designates the specific instrument (i.e. G or g for Gilly and H or h for Hew), 002 is a unique sequence or run reference, autosampler well position no. 61, and injection sequence order no. 03. This therefore provides a means or sorting the data to give replicate analyses for each sample using Excel's inbuilt pivot table facility.

4.2.2.1 Student's t-Test for Significant Differences

Before precision estimates could be calculated, it was first necessary to determine that data generated between the two Agilent 1100 HPLC instruments (Hew and Gilly), were comparable. If it could be assumed that values were statistically equivalent, i.e.; from the same population, data from the two instruments could be pooled and evaluated as a single data set. For this, mean D/L values for each amino acid in each standard solution, analysed on both instruments, were used and evaluated using t-tests to determine the significance of the difference between group means.

Significance tests enable us to determine whether an observed difference between two sets of values such as the experimental mean and the true value (if it could be known) or between two group means is significant or can be simply attributed to random error. Using

statistical theory, a significance test calculates the probability of getting the observed data if the null hypothesis (H_o) is true, i.e. that there is no significant difference between two sets of data. The lower the probability, the less likely it is that the observed differences occurred by chance and the less likely it is that H_o is true, thus the more likely the alternative hypothesis (H_1) is valid, i.e. that there is a significant difference between the observed group means, and the null hypothesis is rejected. Simply, the higher the probability value, the more likely any observed differences occurred by chance and there is no significant difference between the group means. In general practice, it is usually assumed that if the probability of the difference occurring by chance is less than 5% (i.e. $\alpha = 0.05$ or 1 in 20), then H_o is rejected and H_1 accepted, that is, there is a significant difference and it " *is said to be significant at the 0.05 or 5% level*" (Miller & Miller 2005, p39). However, it can also be said that there is a 1 in 20 chance of the null hypothesis being rejected when in fact it is true. Thus if an even greater level of confidence is required, a higher level of significance can be used such as 1% (0.01) or 0.1% (0.001). Note; "*that if the null hypothesis is retained it has not been proved that it is true, only that it has not been demonstrated to be false*". (Miller & Miller 2005, p40).

Traditionally, whether the H_o is retained or rejected has been determined by the calculation of the t-statistic which is then compared to a critical value at the relevant probability level. Thus if t(stat) is greater than t(critical), H_o is rejected. However, in addition to this it is now possible to calculate the actual probability value using most software packages, allowing the actual level of significance to be accurately determined.

For each amino acid, t-test evaluations were carried out on individual vial means from both Hew and Gilly, and the t-statistic compared to the t-critical value at the 0.05 probability level for a 2-tail distribution. Although random error variances are assumed to be equal on the two instruments (same instrument, same material), because the Hew and Gilly data were not generated as paired values, and that the number of samples analysed are different on the two instruments, for these evaluations, unequal variances have been assumed. Results of these evaluations are given in Table 4.1 and presented as a histogram in Figure 4.3. **Red** data (Table 4.1) indicate probability levels falling below the $\alpha = 0.05$ probability level and where the t-statistic exceeds the critical value (ignoring the direction of the sign), thus rejecting the null hypothesis that there is no significant difference. **Orange** values are close to the limit where H_o is retained and should be viewed with caution.

Figure 4.3: t-Test (two tail, unequal variances). Probability of there being no significant difference between instruments in standard solutions.

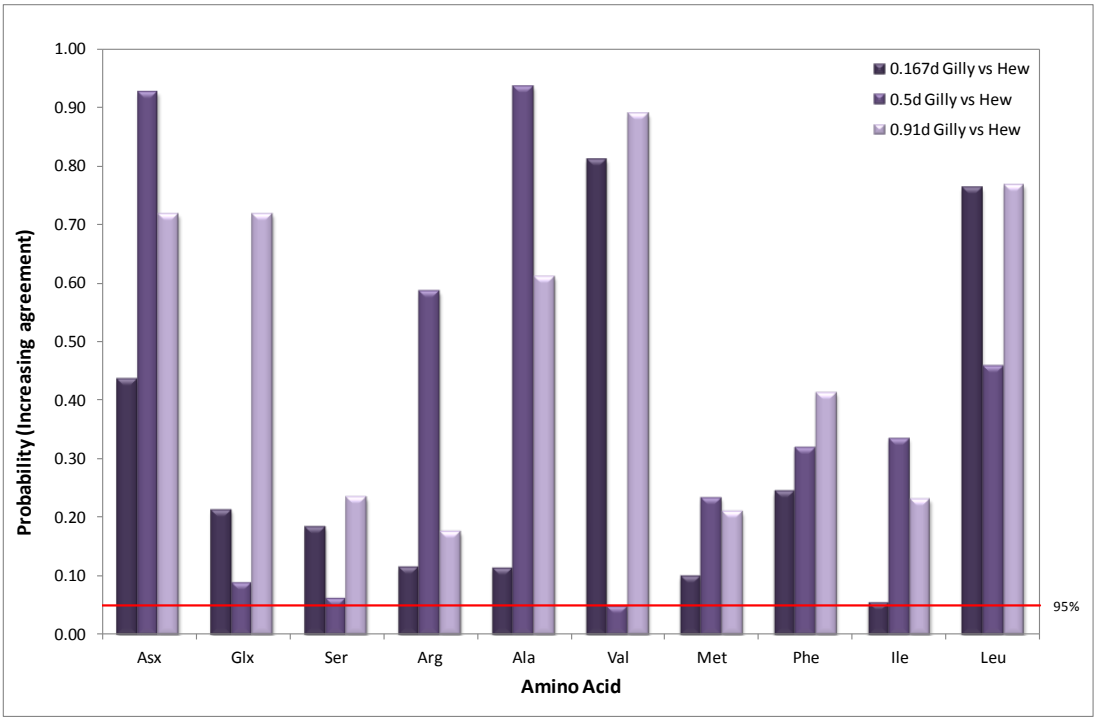


Table 4.1: Hew vs Gilly t-Test analysis (p=0.05)

Hew vs Gilly t-Test comparisons								
Std sol	**amino acid**	**P**	**t-stat**	**t-crit**	**amino acid**	**P**	**t-stat**	**t-crit**
0.167d	Asx	0.435	0.786	2.00	Glx	0.213	1.26	2.01
0.5d	Asx	0.926	0.094	1.97	Glx	0.089	-1.72	1.98
0.91d	Asx	0.718	0.362	1.97	Glx	0.718	0.362	1.97
0.167d	Ser	0.184	1.35	2.01	Arg	0.115	1.6	2.00
0.5d	Ser	0.062	-1.89	1.98	Arg	0.586	-0.546	1.98
0.91d	Ser	0.235	-1.19	1.97	Arg	0.176	1.36	1.97
0.167d	Ala	0.113	1.62	2.01	Val	0.810	-0.242	1.99
0.5d	Ala	0.935	0.081	1.98	Val	**0.048**	**-2.00**	1.98
0.91d	Ala	0.611	0.51	1.98	Val	0.889	0.140	1.98
0.167d	Phe	0.244	-1.18	2.03	Met	0.1	1.68	2.02
0.5d	Phe	0.319	1.00	1.98	Met	0.233	-1.20	1.98
0.91d	Phe	0.413	-0.823	1.99	Met	0.210	-1.26	1.99
0.167d	Ile	0.054	-1.96	2.00	Leu	0.761	0.306	1.67
0.5d	Ile	0.335	-0.969	1.99	Leu	0.458	0.745	1.98
0.91d	Ile	0.232	1.20	1.98	Leu	0.768	0.296	1.98

note; the polarity of the t-stat is ignored, only the absolute value should be compared to the t-critical value

red data indicate t-stat>t-crit

orange data indicates t-stat is close to t-crit

Figure 4.3 plots the probability level for each amino acid. The horizontal line at the 5% level indicates the threshold at which H_O is retained. p-Values falling above this line imply that observed differences between the two sets of results occur through random error.

Results falling below this line suggest that there is a less than 5% chance that observed differences are due to random error and greater than a 95% chance that it is due to a genuine systematic bias.

The t-test evaluations are based on large data sets and as such should be reliable. However, initial results gave some unexpected results. Although just below the permitted threshold, valine in the 0.5d std sol indicates differences between the instruments may be significant. Other results that were acceptable but close to the threshold were Glx and Ser again in the 0.5d std sol, and Ile in 0.167d std sol.

Where differences between Hew and Gilly are suggested, this may be due to a genuine difference between the two data sets generated from different instruments or the presence of extreme values which are influencing the calculated mean and variance of the data.

4.2.2.2 Scatter Plots

To answer this question, a series of scatter charts were plotted for each of the amino acids, using every individual replicate result as a separate value, as typically practiced by the laboratory. To illustrate these charts, data for valine have been used and shown in Figure 4.4 - Figure 4.6. The first of these charts (Figure 4.4) shows the Val D/L plotted against Glx D/L. As these data show variability in both the x and y directions, Glx D/L values were then replaced by instrument to make the spread of valine D/L values clearer (Figure 4.5). Further charts like these can be found in Chpt 4: Appendix 1 for all the amino acids evaluated.

These scatter charts simply plot all data points as separate values. They provide a clear visual comparison of D/L values for a given amino acid in each of the std sol concentrations, run on both instruments. They also show the presence of extreme values, often indicating the incorrect reporting of results, i.e. recording 0.91 D/L value for a 0.167 D/L standard (Figure 4.5). Charts suggest that, generally speaking data from the two instruments are comparable and suggest that the discrepancies observed in the t-tests are likely to be due to the influence of mis-reported extreme values.

Figure 4.4: Scatter Plot of Val D/L vs Glx D/L

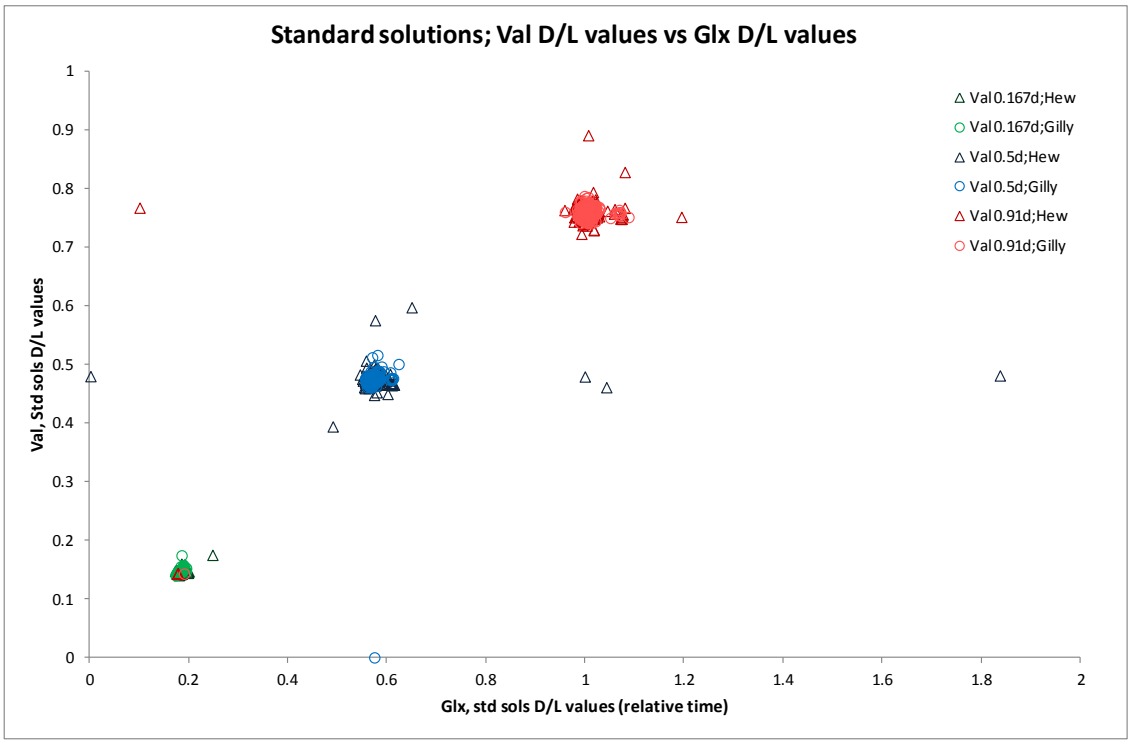
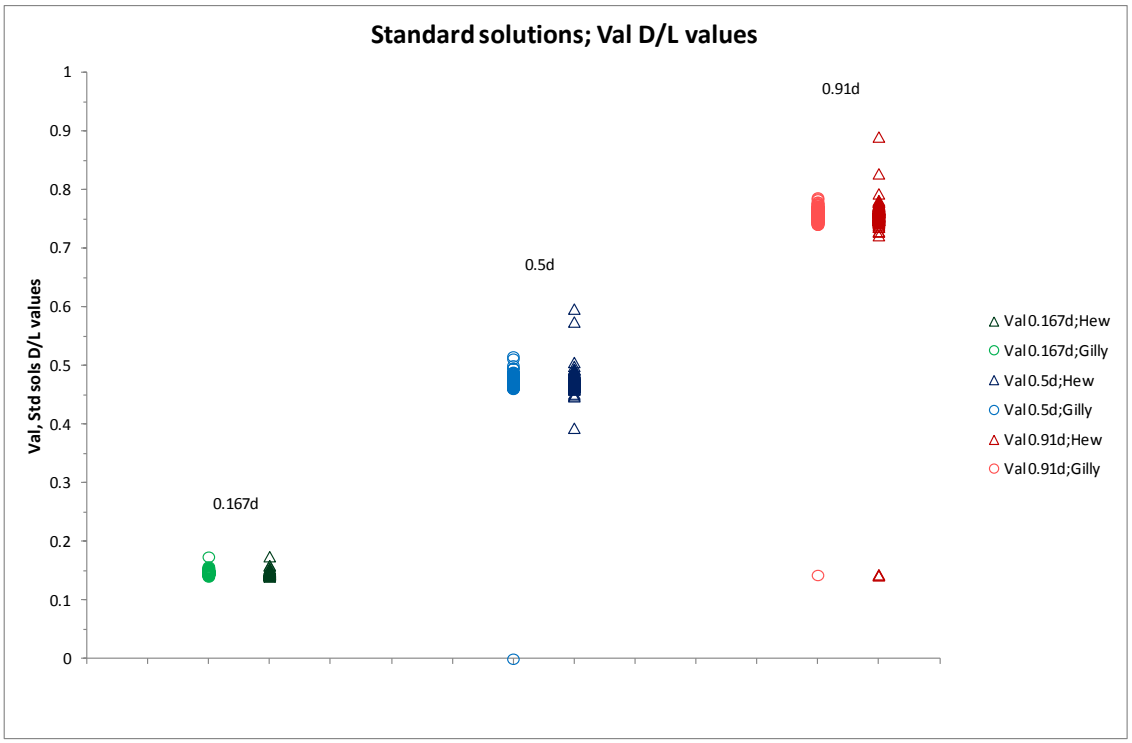


Figure 4.5: Scatter Plot of Val D/L vs Expected D/L, by instrument.



4.2.2.3 Repeatability

In considering the data, it is perhaps misleading to include every individual analytical result as an independent value. This will include separate results of repeat injections from the same sample vial. As such any precision estimates derived from these measurement results will include the instrument variability, and will therefore be much tighter than that expected between separate samples alone.

Replicate data for valine D/L values from individual sample vials of 0.5d std sol are shown, (Figure 4.6a and b). Similar plots of the other amino acids in all three standard solutions are given in Chpt 4: Appendix 2. These charts provide a visual presentation of the within-sample variability or repeatability observed in routine analysis of the standard solutions. In order to show sufficient resolution of individual values, the D/L y-axes on some charts have had to be truncated.

Consequently the larger value outliers are not present on many of the charts. It is also important to point out that in order to fit the charts on the page, the y-axis scales are not all exactly equivalent, although efforts have been made to ensure that major divisions are comparable.

4.2.2.4 Boxplots

An alternative representation of the variability of D/L values would be to consider the mean of repeat injections as the representative value for any given sample (as already done for the t-test evaluations). To interpret the overall distribution of the replicate means, the use of Minitab's Boxplot chart function provides an easy comparison; indicating central tendency, highlighting variability and asymmetry in the distribution, and identifying potential outliers. Figure 4.7 provides an example for 0.5d std sol, with valine D/L data circled. Similar diagrams are provided in Chpt 4: Appendix 3 for amino acids in the other standard solutions.

In each chart (Chpt 4: Appendix 3) comparisons of all the amino acids quantified within a specific std sol are presented and data from two instruments, Hew and Gilly are compared to illustrate the differences in mean and median values together with variability between amino acids within the same standard solution. Each standard solution has two charts; the first provides an overall picture of the distribution and the extent of outlier values, the second shows close up detail of the central region for each amino acid, allowing for a better comparison of the means, medians and inter-quartile ranges.

Figure 4.6: Replicate injection D/L values are shown for valine in 0.5d std solutuion

Figure 4.6a: Run on Gilly

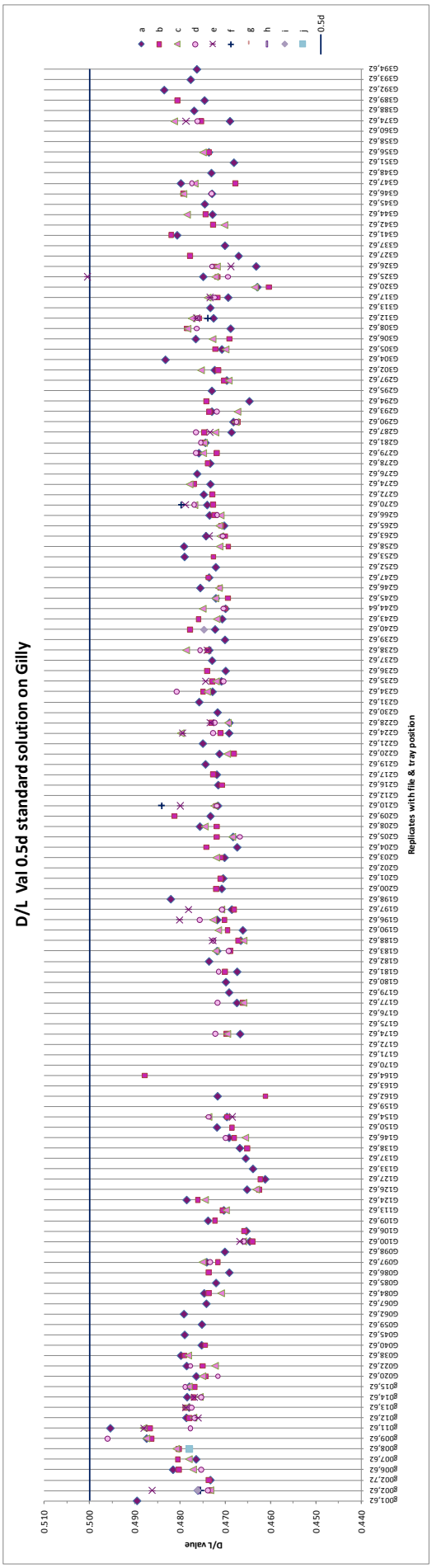
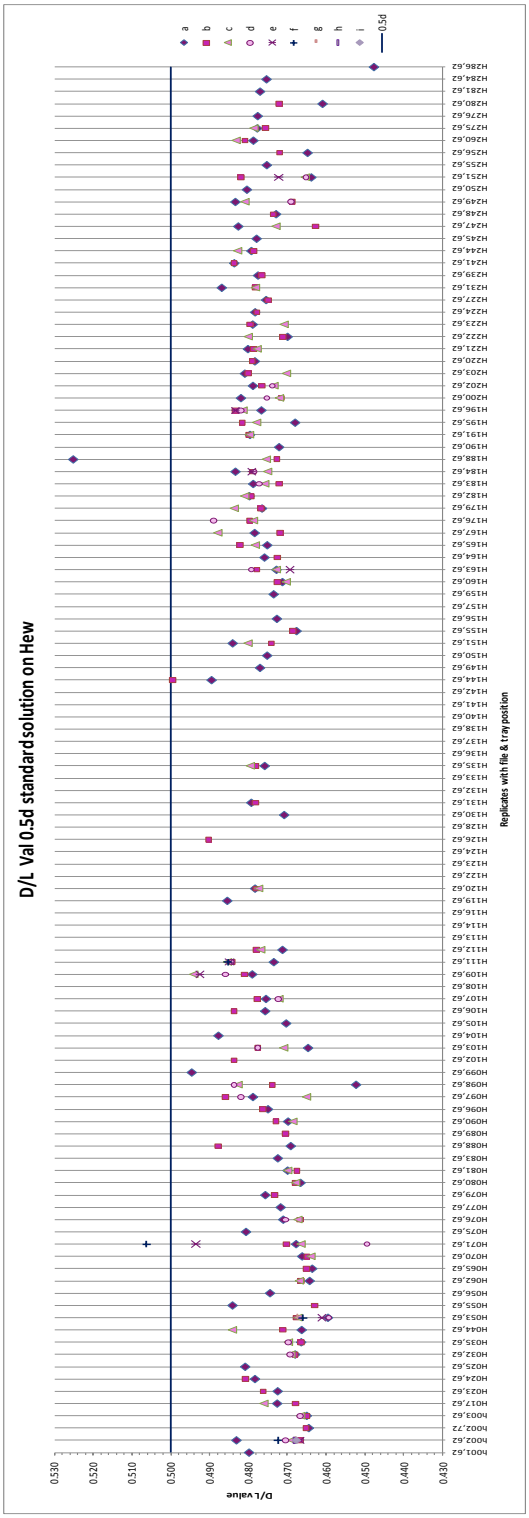


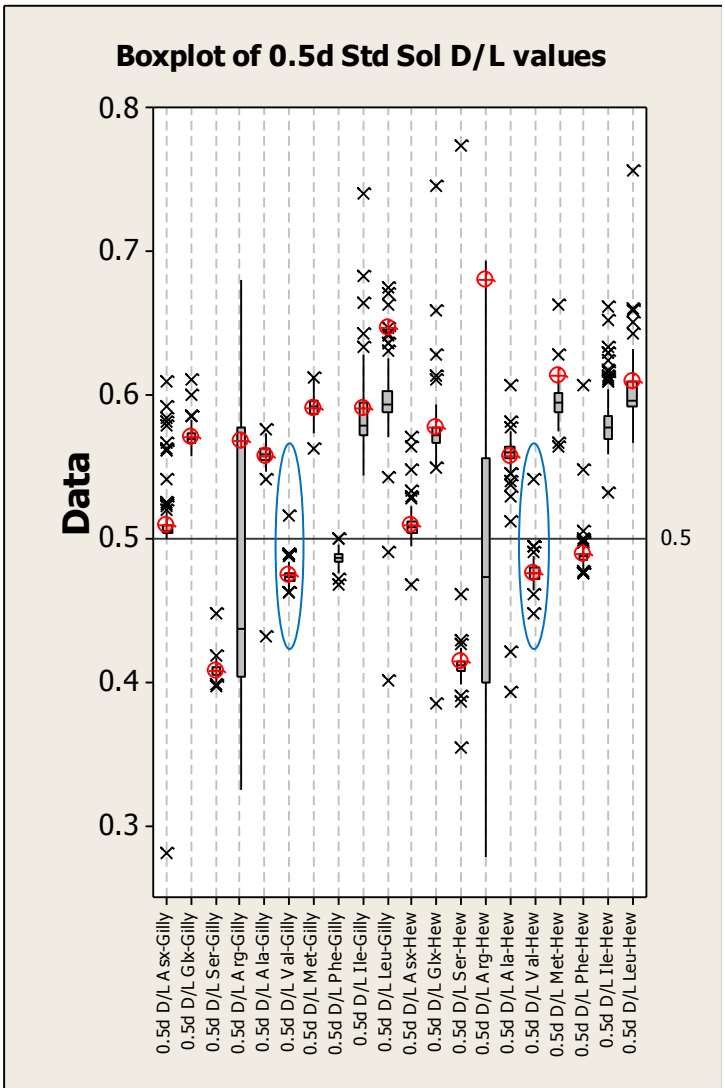
Figure 4.6b: Run on Hew

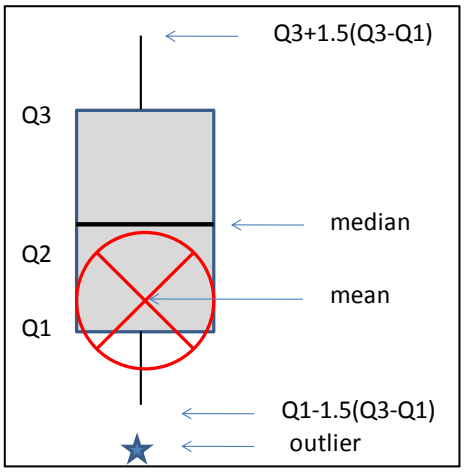


For valine, the locations of the median and mean appear almost identical with little observable difference between Gilly and Hew data. For other amino acids it becomes apparent that the large numbers of outliers are significantly influencing the calculation of the means. In the majority of cases, these means all lie above the median, with some amino acids also having noticeable high-tail skews to their distributions based on the position of the median.

It is important to determine whether there is a genuine bias or skew in the data and seek to identify possible causes. If for example, data are found to be log-normal, it then becomes possible to transform values by log transforming them prior to carrying out further parametric evaluations. It thus becomes helpful to observe individual density distributions and to carry out tests for normality.

Figure 4.7: Boxplot (with key) for valine D/L values comparing Gilly and Hew data.





The shaded grey box presents the middle 50% of data, i.e. the inter-quartile range, that is 25% of all data lying below and above the median or middle value.

The two vertical whiskers extending either end represent the remaining data with the exclusion of outliers. Whisker limits are determined as the inter-quartile range limits, determined as ± 1.5 x inter-quartile range.

Individual data points shown beyond theses limits are considered outliers.

4.2.2.5 Frequency histograms & Kolmogorov-Smirnov Normality test

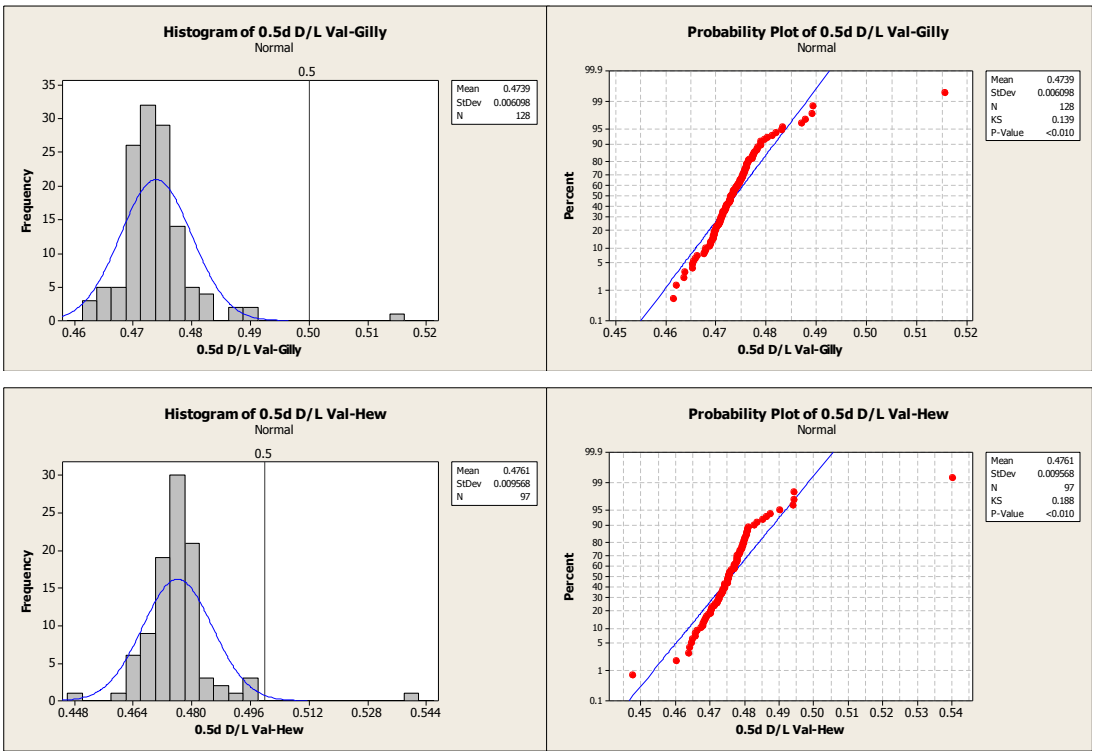
Histograms were plotted for every group of amino acid D/L values at each of the std sol concentrations with normal curves superimposed. These, together with results for Kolmogorov-Smirnov (K-S) tests for normality (Miller and Miller, 2005) are given in Chpt 4: Appendix 4, with example charts shown for Gilly and Hew valine data in Figure 4.8.

The histograms allow the distribution densities to be more clearly visualized. Whilst in a few examples (Chpt4: Appendix 4), data appear to show a normal distribution, in a great many more cases, the full extent of the influence from high-end tails and extreme values on central tendency and their approximation to normality, can be better appreciated. For data to be considered normally distributed, the plotted red data should lie along the diagonal blue line on the K-S probability plot (Figure 4.8) and have a p-value greater than the chosen level of significance for the test, or α level (usually $\alpha = 0.05$ for a 95% confidence range). Minitab does not provide the functionality that permits α to be adjusted and simply assesses the data against fixed criteria. Data are considered to follow a normal distribution when the p-value >0.15. No further information on the closeness of fit or otherwise is given by the software. From the descriptive evaluations of the raw data, it is clear that the influence of outliers needs to be minimised. However, when handling potential outlier data caution must be exercised. There is often a fine line between genuine sample values sitting at the extreme of a non-normal distribution and those which should be considered aberrant and dropped from the sample data set.

4.2.2.6 Outlier removal

In circumstances where the homogeneity of material can be assured, such as in a solution, aberrant values (such as data out by a factor or 10 or more) can be easily considered as misreported results. Where long tails are evident it becomes less easy. On this occasion, in order to observe the effect of removing outliers on the distribution, rather than a need to accurately determine alternative exclusion criteria, Minitab's approach to determining outliers used on the boxplots has been used as an exclusion guide. Thus data were re-evaluated, with values greater than Q3 + 1.5 (Q3-Q1) and less than Q1 - 1.5 (Q3-Q1) being excluded from the data set.

Figure 4.8: Distributions and K-S plots for valine D/L values



4.2.3 Re-evaluating Normality

Using this revised dataset, t-tests were re-applied to Gilly and Hew data. These new t-test results are given in Table 4.2. In addition, all boxplots, histograms and K-S normality tests were repeated. These revised charts are given in Chpt 4: Appendix 5 and Appendix 6 respectively. However revised charts for valine data are given below for comparison.

Disappointingly, results of the repeated t-tests were not generally improved but frequently made worse. Table 4.2 shows that after the removal of outliers, both Asx and Ser D/L data in all standard solutions, gave significant differences between instruments, together with several of the other amino acids shown in red with others being close to the limits of acceptability shown in orange.

This data is illustrated by the revised histogram in Figure 4.9, which shows the increase in the number of bars falling below the 95% probability level. The revised boxplots demonstrate that removal of outliers does not fundamentally alter the position of the median or inter-quartile range although it does affect the positioning of the mean, bringing it closer towards the median (Figure 4.10).

Table 4.2: Re-evaluated t-Test analysis (p=0.05), after outlier removal

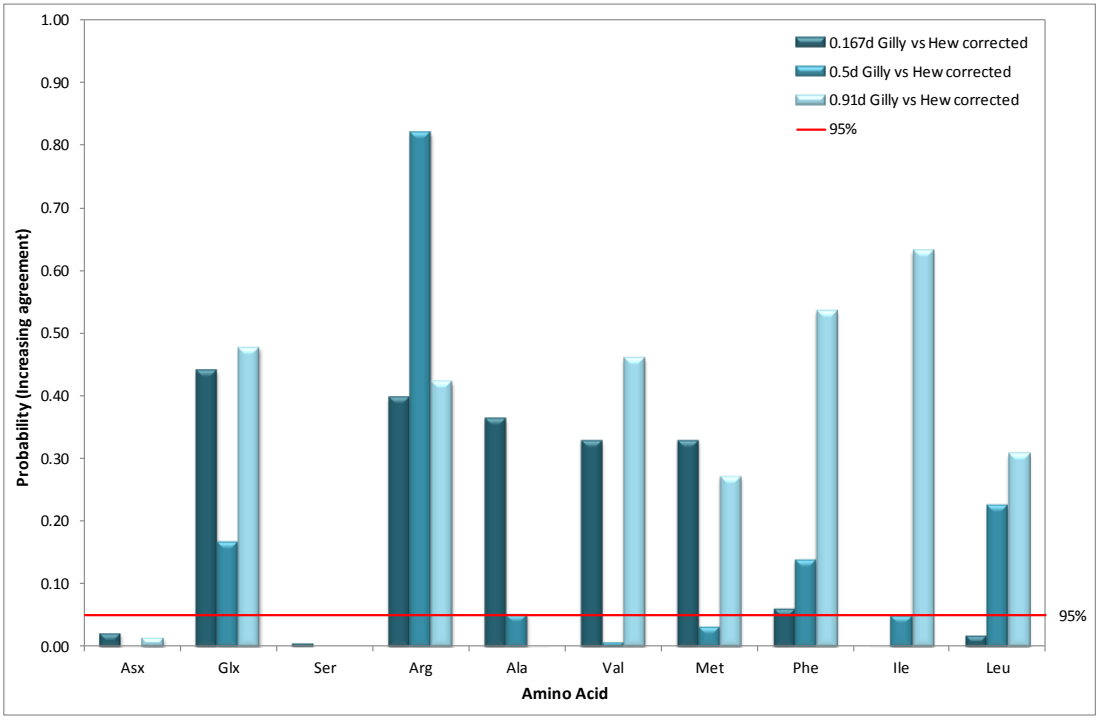
Revised Hew vs Gilly t-Test comparisons								
Std sol	**amino acid**	**P**	**t-stat**	**t-crit**	**amino acid**	**P**	**t-stat**	**t-crit**
0.167d	Asx	**0.022**	**-2.40**	2.03	Glx	0.440	-0.78	1.99
0.5d	Asx	**3E-04**	**-3.73**	1.97	Glx	0.168	-1.39	1.97
0.91d	Asx	**0.014**	**2.47**	1.97	Glx	0.477	-0.71	1.98
0.167d	Ser	**0.004**	**-2.95**	2.00	Arg	0.397	0.85	1.99
0.5d	Ser	**2E-09**	**-6.32**	1.97	Arg	0.820	0.23	1.97
0.91d	Ser	**2E-08**	**-5.90**	1.97	Arg	0.422	0.80	1.97
0.167d	Ala	0.364	-0.91	2.00	Val	0.328	-0.98	2.00
0.5d	Ala	**0.052**	**-1.95**	1.97	Val	**0.008**	**-2.69**	1.97
0.91d	Ala	**0.001**	**-3.48**	1.98	Val	0.459	-0.74	1.98
0.167d	Phe	**0.060**	**-1.95**	2.04	Met	**2E-04**	**3.96**	1.99
0.5d	Phe	0.138	-1.49	1.97	Met	**0.032**	**-2.17**	1.97
0.91d	Phe	0.535	-0.62	1.98	Met	0.271	-1.10	1.97
0.167d	Ile	**4E-04**	**-3.92**	2.03	Leu	**0.018**	**-2.46**	2.02
0.5d	Ile	**0.049**	**1.98**	1.97	Leu	0.226	-1.21	1.97
0.91d	Ile	0.632	0.48	1.98	Leu	0.308	1.02	1.97

note; the polarity of the t-stat is ignored, only the absolute value should be compared to the t-critical value

red data indicate t-stat>t-crit

orange data indicates t-stat is close to t-crit

Figure 4.9: Revised t-Test (two tail, unequal variances). Probability of there being no significant difference between instruments in standard solutions, after outlier removal



Revised frequency distributions for valine now focus on the central region of the distribution, and K-S plots show data better fitted to the expected trendline (Figure 4.11). Chpt 4: Appendix 6 shows similarly revised plots for other amino acids in the standard solutions.

In the Chpt 4: Appendix 6 K-S charts, it can be seen that distributions often remain affected by high-end tails, for example Asx, Ile and Leu, despite removing extreme values, and that in a few instances data suggest bimodality, particularly in the case of Arg.

Table 4.3a-c, provide a comparison of mean D/L values between the raw data and the trimmed data after removal of outliers, together with measures of the dispersion of the data sets and K-S p-values. Comparative median values for 0.167d std sol data are given later in Table 4.4.

Figure 4.10: Revised Boxplot for amino acid D/L values comparing Gilly and Hew data, after outlier removal

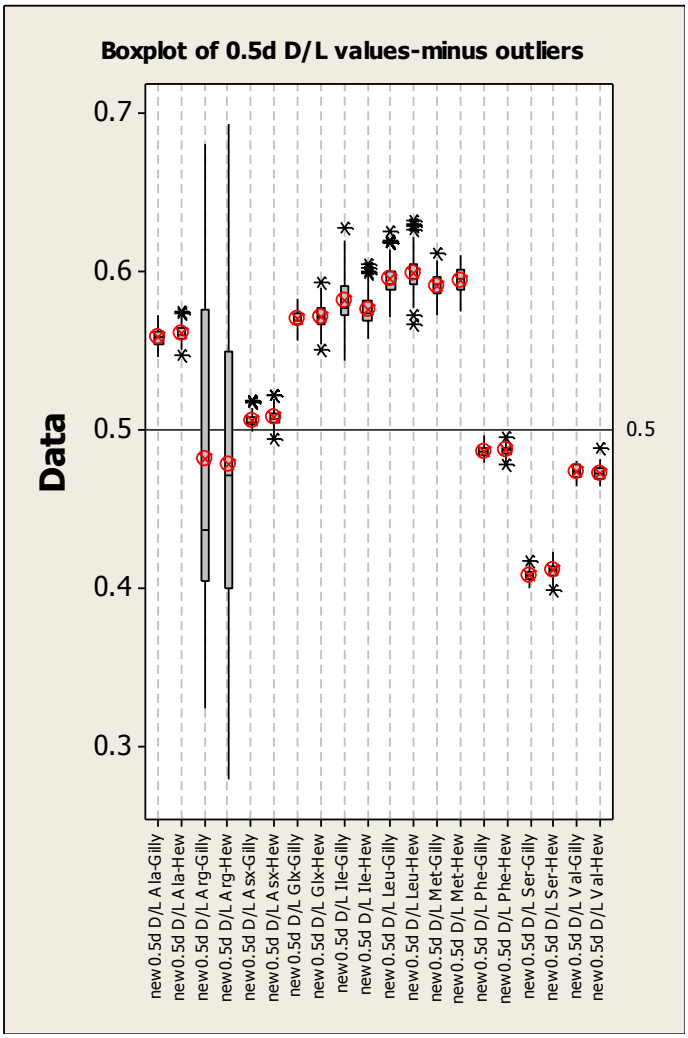
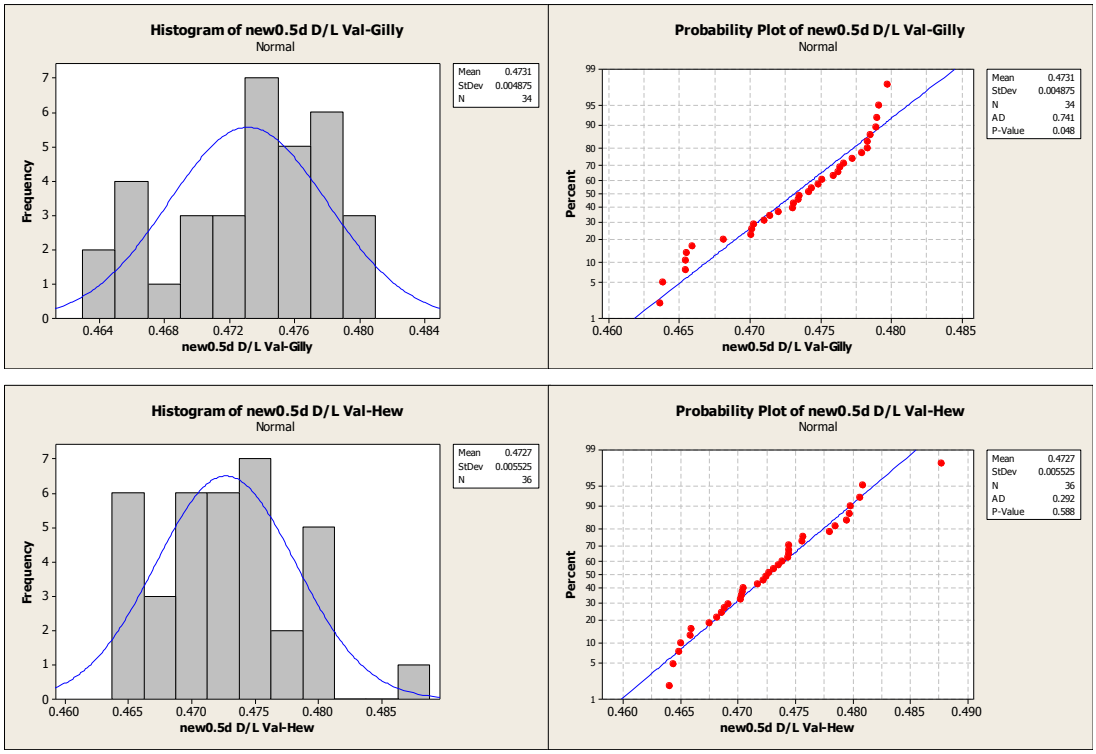


Figure 4.11: Revised Distributions and K-S plots for valine D/L values, after the removal of outliers.



4.2.3.1 Considerations of Outlier removal

After exclusion of the outliers, although high-end tails can sometimes be seen, the central region of the Hew and Gilly boxplots, the inter-quartile range, tend to become better aligned for individual amino acids. Table 4.3a-c demonstrate that in many cases the mean D/L values are in closer agreement with tighter distributions given by the standard deviation and CV% and that there is a general tendency for the K-S p-values to increase indicating that distributions are tending to become normalised after outlier removal.

Having now removed outliers, any remaining disagreement between instruments may be due to either;

i. genuine differences between sample means, or,
ii. a function of the t-test, where trimming the data has tightened the distributions such that the two sets of data (Gilly and Hew) appear to represent independent populations for a given amino acid, when in fact they are not.

Table 4.3: Comparison of Amino Acid Mean D/L Values, standard deviations and standard uncertainties

Table 4.3a: Raw data vs data with outliers removed., in 0.167d Std Sol

Std sol	0.167d	Mean of all replicates						Mean of replicates after outliers removed					
amino acid	instrument	n	mean	std dev	%CV	std u	KS p value	n	mean	std dev	%CV	std u	KS p value
Asx	Gilly	49	0.203	0.1436	70.70	0.0205	<0.010	41	0.172	0.0019	1.10	0.0003	0.024
Asx	Hew	38	0.186	0.0358	19.20	0.0058	<0.010	33	0.173	0.0050	2.90	0.0009	0.023
Glx	Gilly	49	0.217	0.1659	76.50	0.0237	<0.010	47	0.183	0.0044	2.40	0.0006	<0.010
Glx	Hew	38	0.187	0.0119	6.30	0.0019	<0.010	32	0.184	0.0027	1.50	0.0005	>0.150
Ser	Gilly	49	0.156	0.1113	71.60	0.0159	<0.010	47	0.133	0.0017	1.30	0.0002	<0.010
Ser	Hew	38	0.134	0.0032	2.40	0.0005	0.042	36	0.134	0.0021	1.60	0.0004	>0.150
Arg	Gilly	48	0.228	0.1586	69.60	0.0229	<0.010	47	0.193	0.0551	28.60	0.0080	0.045
Arg	Hew	38	0.190	0.0442	23.30	0.0072	0.048	38	0.190	0.0442	23.30	0.0072	0.048
Ala	Gilly	48	0.215	0.1544	71.90	0.0223	<0.010	43	0.180	0.0051	2.80	0.0008	<0.010
Ala	Hew	38	0.179	0.0086	4.80	0.0014	<0.010	33	0.181	0.0041	2.30	0.0007	>0.150
Val	Gilly	42	0.148	0.0052	3.50	0.0008	<0.010	41	0.147	0.0030	2.10	0.0005	0.367
Val	Hew	36	0.148	0.0054	3.70	0.0009	<0.010	33	0.147	0.0022	1.50	0.0004	0.923
Met	Gilly	42	0.214	0.0517	24.20	0.008	<0.010	41	0.206	0.0073	3.50	0.0011	0.021
Met	Hew	36	0.200	0.0067	3.30	0.0011	<0.010	35	0.199	0.0056	2.80	0.0010	0.031
Phe	Gilly	42	0.160	0.0049	3.10	0.0008	<0.010	40	0.159	0.0029	1.80	0.0005	>0.150
Phe	Hew	36	0.166	0.0276	16.70	0.0046	<0.010	33	0.160	0.0024	1.50	0.0004	>0.150
Ile	Gilly	41	0.197	0.0128	6.50	0.002	<0.010	36	0.193	0.0030	1.50	0.0005	>0.150
Ile	Hew	35	0.203	0.0153	7.50	0.0026	<0.010	29	0.197	0.0073	3.70	0.0014	>0.150
Leu	Gilly	41	0.216	0.0832	38.50	0.013	<0.010	39	0.202	0.0070	3.40	0.0011	0.066
Leu	Hew	35	0.212	0.0198	9.30	0.0033	<0.010	35	0.212	0.0198	9.30	0.0033	<0.010

Values given in red indicate distributions do not approximate to normality

Table 4.3b: Raw data vs data with outliers removed., in 0.5d Std Sol

Std sol	0.5d	Mean of all replicates						Mean of replicates after outliers removed					
amino acid	instrument	n	mean	std dev	%CV	std u	KS p value	n	mean	std dev	%CV	std u	KS p value
Asx	Gilly	139	0.509	0.0270	5.30	0.0023	<0.010	124	0.506	0.0037	0.70	0.0003	<0.010
Asx	Hew	114	0.509	0.0114	2.20	0.0011	<0.010	107	0.507	0.0054	1.10	0.0005	0.032
Glx	Gilly	139	0.570	0.0071	1.20	0.0006	<0.010	135	0.57	0.0055	1.00	0.0005	>0.150
Glx	Hew	114	0.578	0.0453	7.80	0.0042	<0.010	105	0.571	0.0077	1.30	0.0007	>0.150
Ser	Gilly	139	0.408	0.0050	1.20	0.0004	<0.010	135	0.408	0.0034	0.80	0.0003	>0.150
Ser	Hew	114	0.414	0.0352	8.50	0.0033	<0.010	107	0.411	0.0049	1.20	0.0005	>0.150
Arg	Gilly	138	0.568	1.0190	179	0.0867	<0.010	137	0.482	0.0883	18.30	0.0075	<0.010
Arg	Hew	107	0.680	1.9090	281	0.1845	<0.010	105	0.478	0.0875	18.30	0.0085	<0.010
Ala	Gilly	139	0.557	0.0123	2.20	0.0010	<0.010	136	0.558	0.0056	1.00	0.0005	>0.150
Ala	Hew	107	0.557	0.0231	4.20	0.0022	<0.010	96	0.560	0.0053	0.90	0.0005	>0.150
Val	Gilly	128	0.474	0.0061	1.30	0.0005	<0.010	34	0.473	0.0049	1.00	0.0008	0.048
Val	Hew	97	0.476	0.0096	2.00	0.0010	<0.010	36	0.473	0.0055	1.20	0.0009	0.588
Met	Gilly	129	0.591	0.0086	1.50	0.0008	0.131	127	0.591	0.0080	1.40	0.0007	>0.150
Met	Hew	97	0.613	0.1859	30.30	0.0189	<0.010	92	0.594	0.0091	1.50	0.0009	0.093
Phe	Gilly	129	0.832	3.8900	468	0.3425	<0.010	121	0.486	0.0034	0.70	0.0003	>0.150
Phe	Hew	97	0.489	0.0144	2.90	0.0015	<0.010	84	0.487	0.0032	0.70	0.0004	<0.010
Ile	Gilly	128	0.590	0.0476	8.10	0.0042	<0.010	121	0.582	0.0149	2.60	0.0014	<0.010
Ile	Hew	96	0.864	2.7650	320	0.2822	<0.010	81	0.576	0.0097	1.70	0.0011	0.033
Leu	Gilly	129	0.646	0.5673	87.80	0.0499	<0.010	116	0.595	0.0100	1.70	0.0009	<0.010
Leu	Hew	96	0.609	0.0478	7.90	0.0049	<0.010	89	0.599	0.0124	2.10	0.0013	<0.010

Values given in red indicate distributions do not approximate to normality

Table 4.3c: Raw data vs data with outliers removed., in 0.91d Std Sol

Std sol	0.91d	Mean of all replicates						Mean of replicates after outliers removed					
amino acid	instrument	n	mean	std dev	%CV	std u	KS p value	n	mean	std dev	%CV	std u	KS p value
Asx	Gilly	141	0.897	0.0691	7.70	0.0058	<0.010	124	0.894	0.0051	0.60	0.0005	<0.010
Asx	Hew	100	0.893	0.0806	9.00	0.0081	<0.010	85	0.895	0.0079	0.90	0.0009	<0.010
Glx	Gilly	141	1.000	0.0694	6.90	0.0058	<0.010	136	1.004	0.0072	0.70	0.0006	>0.150
Glx	Hew	100	0.996	0.0870	8.70	0.0087	<0.010	93	1.005	0.012	1.20	0.0012	>0.150
Ser	Gilly	141	0.688	0.0738	10.70	0.0062	<0.010	136	0.699	0.0057	0.80	0.0005	>0.150
Ser	Hew	99	0.699	0.0588	8.40	0.0059	<0.010	95	0.705	0.0075	1.10	0.0008	0.036
Arg	Gilly	139	0.847	0.5028	59.40	0.0426	<0.010	137	0.812	0.1516	18.70	0.013	<0.010
Arg	Hew	89	0.785	0.1546	19.70	0.0164	<0.010	88	0.792	0.1403	17.70	0.015	<0.010
Ala	Gilly	138	0.928	0.0670	7.20	0.0057	<0.010	135	0.934	0.0076	0.80	0.0007	>0.150
Ala	Hew	89	0.922	0.0875	9.50	0.0093	<0.010	81	0.938	0.0078	0.80	0.0009	>0.150
Val	Gilly	132	0.755	0.0542	7.20	0.0047	<0.010	128	0.759	0.0065	0.90	0.0006	0.891
Val	Hew	80	0.754	0.0699	9.30	0.0078	<0.010	77	0.761	0.0073	1.00	0.0008	0.059
Met	Gilly	132	1.014	0.0734	7.20	0.0064	<0.010	127	1.021	0.013	1.30	0.0012	0.011
Met	Hew	80	2.121	7.8270	369	0.8751	<0.010	76	1.023	0.0133	1.30	0.0015	>0.150
Phe	Gilly	132	0.938	1.6130	172	0.1404	<0.010	123	0.806	0.0049	0.60	0.0004	>0.150
Phe	Hew	81	1.656	7.7520	468	0.8613	<0.010	73	0.806	0.0057	0.70	0.0007	>0.150
Ile	Gilly	132	1.175	1.8750	160	0.1632	<0.010	121	0.988	0.0127	1.30	0.0012	0.045
Ile	Hew	80	0.979	0.0923	9.40	0.0103	<0.010	77	0.988	0.0202	2.00	0.0023	<0.010
Leu	Gilly	132	1.054	0.0895	8.50	0.0078	<0.010	111	1.059	0.0127	1.20	0.0012	<0.010
Leu	Hew	80	1.050	0.1020	9.70	0.0114	<0.010	72	1.059	0.0131	1.20	0.0015	>0.150

Values given in red indicate distributions do not approximate to normality

Without the ability to accommodate expectation in the t-test, based on previous analytical results or a value that is a reasonable estimate taking into consideration prior knowledge and experience, applying a purely objective comparison such as this without judgment, may be inappropriate.

Traditional outlier evaluations such as the use of "trimmed" data or median and IQR, originally developed when data were analysed by hand, are frequently insensitive to outlier removal and inappropriate compared to more sophisticated computerised Robust models (RSC Analytical Methods Committee, 1989). Sixty-eight of these approaches are described by Andrews *et al.* (1972, cited in RSC Analytical Methods Committee, 1989). Frequency histograms (Figure 4.8 and Figure 4.11) can provide a clear visual interpretation of the distribution of data, inferring a population density from the sample of data points used. However, as can be seen from the revised valine data (Figure 4.11), interpretation is often limited by the divisions or "bin" sizes applied. Figure 4.11 demonstrates the loss of detail and absence of the two tails from using bins that are two wide for the data. However, with the aid of a freely available Excel add-in or Minitab macro developed by the RSC Analytical Methods Committee (Ellison, 2002a), kernel density estimates can be determined. A kernel density replaces individual data points with probability densities, resulting in a distribution fitted to the data (RSC Analytical Methods Committee, 2006). This will be expanded on further in the next section.

4.2.3.2 Robust Mean Evaluations

In a paper published by the Royal Society of Chemistry's (RSC) Analytical Methods Committee (1989, p1693) they comment; "*The almost universal practice amongst analytical chemists has been to regard outliers as errors, and to delete them from the set of data*". They go on to observe a change in perspective over the years from outlier removal to outlier accommodation. For data that are clearly mis-reported, these are easily identified and removed, however data does not always behave so conveniently particularly for small data sets and decisions have to be made about how to justify exclusion of specific values. The use of robust statistics provides an alternative approach. In its simplest form, the median could be considered a robust estimator ($\hat{\mu}$), as it is the position it occupies that is emphasised rather than the influence from and size of an individual outlier, unlike the mean. Similarly, the robust standard deviation ($\hat{\sigma}$), is derived from the median absolute difference (MAD) by taking the differences between the values and the median, ordering them and finding the

median of those values. The robust standard deviation is then MAD x 1.5 (RSC Analytical Methods Committee, 2001)

Frequently, it is *Huber's h15* method that is applied when calculating robust statistics (RSC Analytical Methods Committee, 1989). It is an iterative calculation that down weights the effect of outliers using a process known as winsorisation (RSC Analytical Methods Committee, 2001), progressively transforming the data until data converge, giving the robust mean and robust standard deviation. Taking initial estimates of the central tendency $\hat{\mu}_o$ and deviation $\hat{\sigma}_o$, if a value, x_i, falls above $\hat{\mu}_o + 1.5\,\hat{\sigma}_o$ then its value is changed to $\tilde{x}_i = \hat{\mu}_o + 1.5\,\hat{\sigma}_o$. Similarly, if a value falls below $\hat{\mu}_o - 1.5\,\hat{\sigma}_o$, its value changes to $\hat{\mu}_o - 1.5\,\hat{\sigma}_o$, otherwise $\tilde{x}_i = x_i$ (RSC Analytical Methods Committee, 2001).

Robust estimations are ideal when the underlying distribution approximates to normal, i.e. unimodal and symmetrical, but carries a few large extreme values or heavy tails (RSC Analytical Methods Committee, 2001). However, they are not infallible. Indeed, where noticeable differences are observed between the median and mean, especially the robust variety, then this clearly calls for further analysis of the data (RSC Analytical Methods Committee, 1989, 2001)

In such circumstances, perhaps because data are suspected of being multimodal, a different approach is required. The kernel density estimate is a slightly more elaborate version of the histogram. However, unlike the distribution of a histogram which can be fundamentally altered and misinterpreted depending on the interval criteria applied to "binning" the data, a kernel density relies on probabilities.

> *"The simple idea underlying the kernel estimate is that each data point is replaced by a specified distribution (typically normal), centred on the point and with a standard deviation designated by h. The normal distributions are added together and the resulting distribution, scaled to have a unit area, is a smooth curve, the kernel density estimate,"* (RSC Analytical Methods Committee, 2006)

The value of *h* can be determined automatically by the software or specified. The kernel density estimate is then the highest point of the curve at value *x*;

$$f(x,h) = \frac{1}{nh}\sum_{i-1}^{n} \varphi\left(\frac{x - x_i}{h}\right) \qquad \textbf{(4.2)}$$

where, $f(x,h)$ is the height of the curve at point x, *h* is the standard deviation and $\varphi(\,)$ is the standard normal density.

Free downloadable macros for both the Robust mean and kernel density estimation, for both Excel and Minitab, are available from the AMC website, http://www.rsc.org/Membership/Networking/InterestGroups/Analytical/AMC/Software/index.asp. (Ellison, 2002b, 2002a).

4.2.3.3 Robust Mean and Kernel Density Evaluations

Using the AMC software described above, standard solution amino acid D/L values were evaluated. Initially, untrimmed data were assessed, based reasonably on the assumption that the robust mean would accommodate and minimize the effect of outliers. However, it was found that when kernel density charts were plotted, these outliers were identified as separate modes, affecting the distribution and presentation of data. Even when the x-axis scale was adjusted to exclude the extreme region, the distribution of the central region was affected by a loss of detail due to the total area being scaled to have a unit density (RSC Analytical Methods Committee, 2006). This effect is demonstrated in Figure 4.12a and b using alanine in 0.167d std sol analysed on Gilly. By comparison, when the previously trimmed data is used, a much smoother distribution of the central region can be seen, Figure 4.12c.

The kernel density add-in used in Excel automatically calculates the target value for standard deviation, shown in Figure 4.12a and b as *h0pt* = 3.28E-03. This value is the one applied to the calculation of the kernel density plot and determines the degree of smoothing. Other expected or observed standard deviations could however, also be specified. The lower the h-value, the closer the fit to the observed data, (that is, the less the smoothing). Typically, when evaluating data for normality in inter-laboratory proficiency tests, a target value for standard deviation is set, derived externally to submitted data, and frequently derived from collaborative trail results. If a target value isn't entered into the Excel macro dialogue box, then a default target value is calculated. Unfortunately, further information regarding the calculation of the *h0pt* value was not accessible with current Excel versions. However, in the absence of collaborative trail data, it does act as a convenient target value and tighter than standard deviation values generally derived from observations.

Figure 4.12: Kernel density using default h0pt: Ala D/L in 0.167d, on Gilly

Figure 4.12a: All data

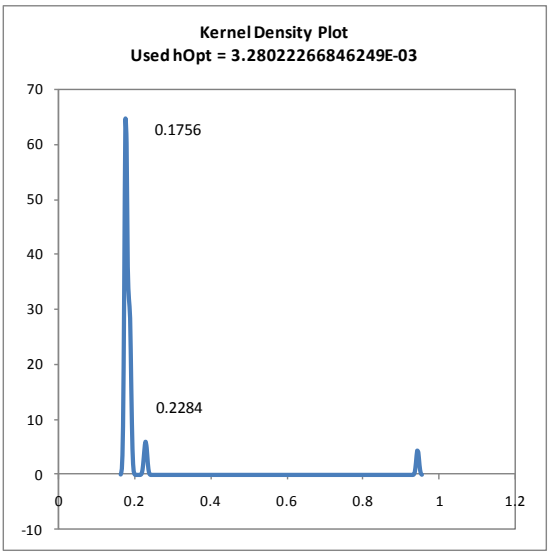
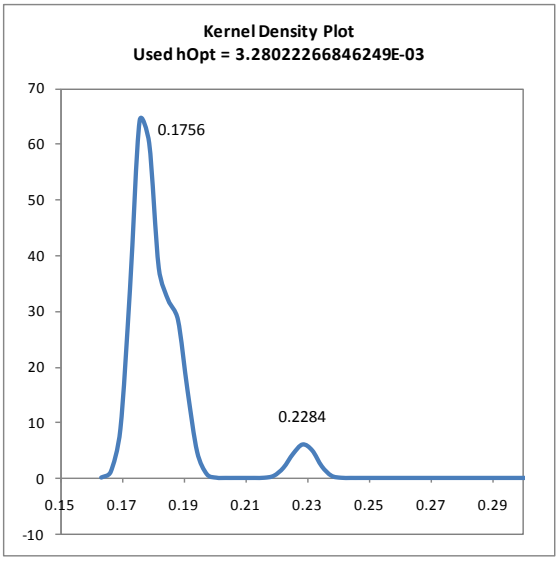


Figure 4.12b: All data, central region



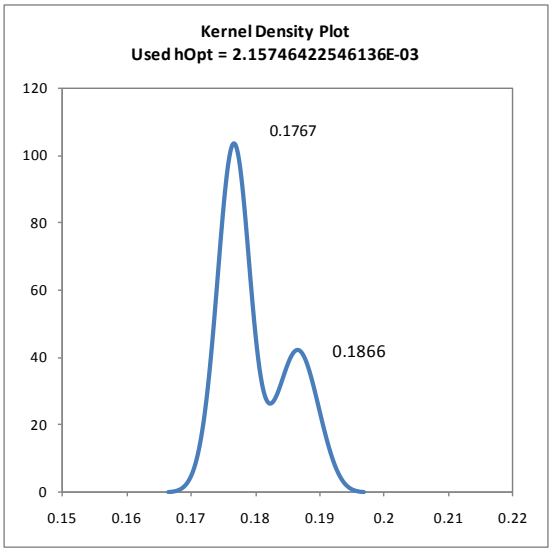


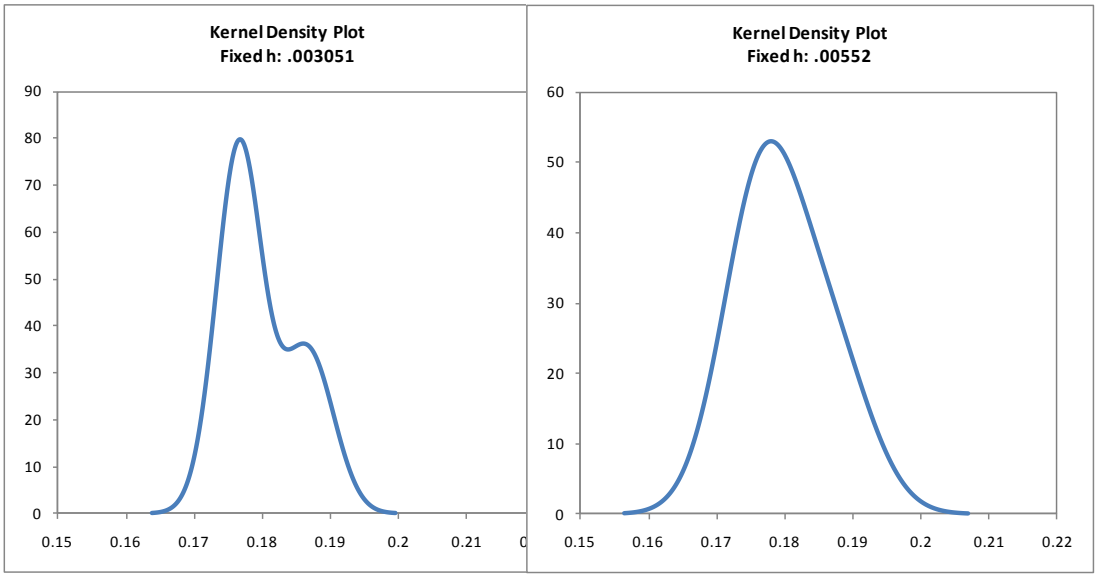
Figure 4.12c: Kernel density: Ala D/L values in 0.167d, analysed by Gilly. Central region after outlier removal.

For example, for replicates of Ala, 0.167d D/L (Table 4.3a), standard deviation values are 8.6E-03 (Hew) and 1.54E-01 (Gilly) compared to 3.28E-03 (*h0pt*, Figure 4.12a and b) and for trimmed data with outliers removed, standard deviations are 4.1E-03 (Hew) and 5.1E-03 (Gilly) compared to 2.16E-03 *(h0pt*, Figure 4.12c).

Applying the robust macro (section 4.2.3.2) (Ellison, 2002b) to determine alternative values for *h*, given as the robust standard deviation (*h15*) and the median absolute deviation (sMAD), enables smoother kernel densities to be achieved (Figure 4.13a and b), since in both cases, *h>h0pt.*

Figure 4.13: Kernel density comparing fixed h: Trimmed Ala D/L in 0.167d, on Gilly

Figure 4.13a: h0pt = sMAD **Figure 4.13b: h0pt = h15 std dev**

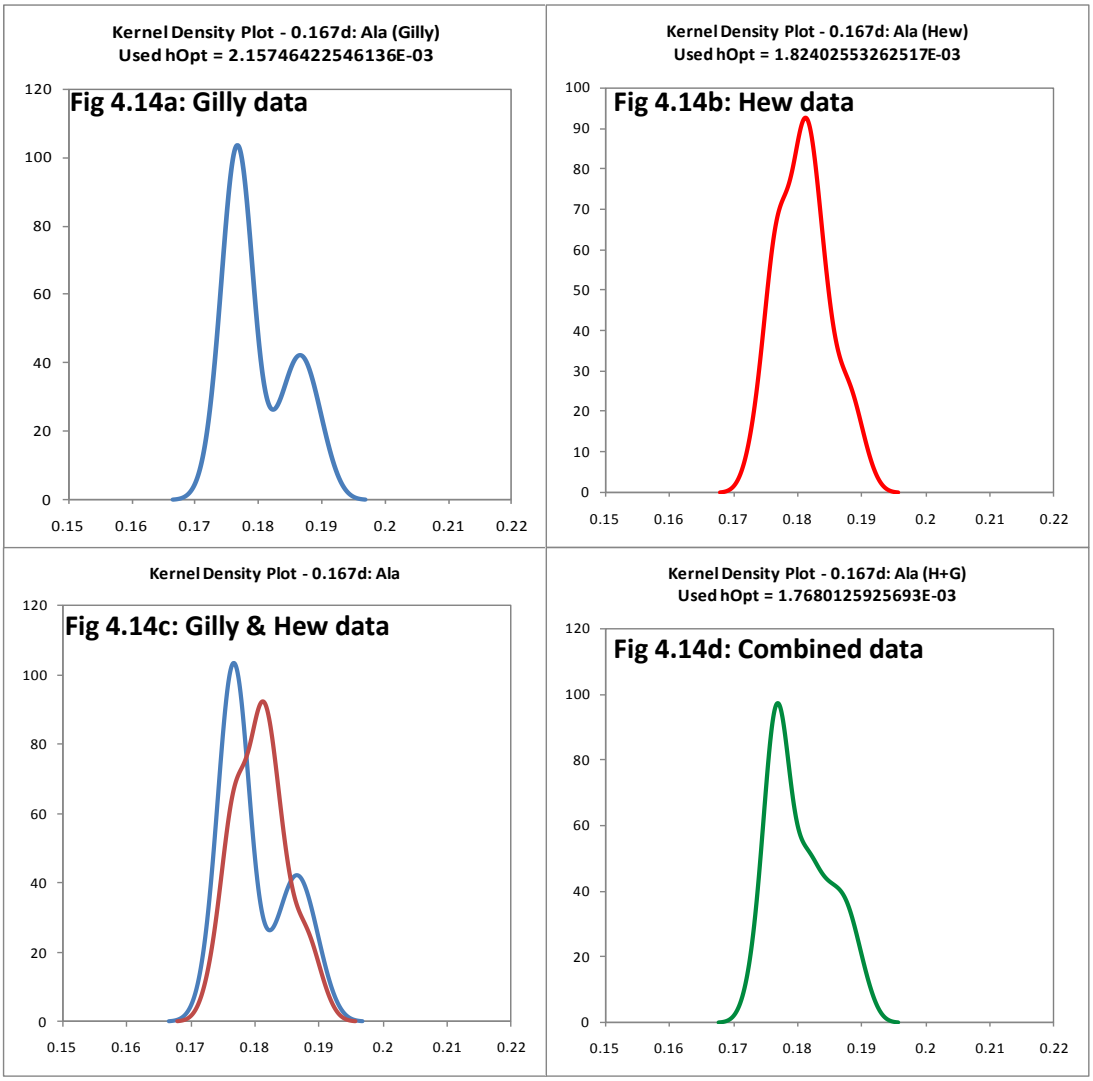


However, as the default *h0pt* value fitted the data well and for trimmed data provided an informative kernel density (Figure 4.12c), the default value for standard deviation has been used for subsequent evaluations. Kernel densities of the other amino acids have been determined and are given in Chpt 4: Appendix 7.

Using the alanine data from previous examples (Figure 4.12 and Figure 4.13), charts show the kernel densities for both Gilly (Figure 4.14a) and Hew (Figure 4.14b) D/L values separately. Distributions are then superimposed for a direct comparison (Figure 4.14c) and finally a combined kernel density for data from both instruments, assessed as a single data set is determined (Figure 4.14d).

For comparison, kernel densities for trimmed valine D/L data in 0.5d standard solution are shown in Figure 4.15a-d. Previously (sections 4.2.2.1 and 4.2.3), valine has demonstrated significant differences between means, although results from the boxplot analyses, might suggest otherwise. Superimposing kernel densities from Hew and Gilly data, indicate that the two sets of D/L values are in close agreement. Given the distribution of both sets of data, it is highly likely that they represent data from the same population, despite earlier t-test results (Table 4.2) and any discrepancy between means may simply be accounted for by the uncertainty of the combined mean.

Figure 4.14: Kernel density summary: Trimmed Ala D/L values in 0.167d std sol.



Results for the robust mean, robust standard deviation, median, sMAD and kernel density modes are given in Table 4.4 for 0.167d standard solution, to demonstrate the differences between measures of central tendency. The standard uncertainty (std u) given in Table 4.4, has been derived from *s/√n.*

Several amino acids have multiple modes of more or less equivalent height and spread, e.g. serine, making it difficult to determine which of these, if any, is dominant (Figure 4.16). It is noticed that in many cases, the default *h0pt* value is substantially smaller than the observed Robust standard deviation or the sMAD value given in Table 4.4. In these instances, perhaps a truer representation of the distributions may have been achieved had the *h0pt* value been relaxed a little, giving a single distribution, rather than a split one.

Figure 4.15: Kernel density summary: Trimmed Val D/L values in 0.5d std sol.

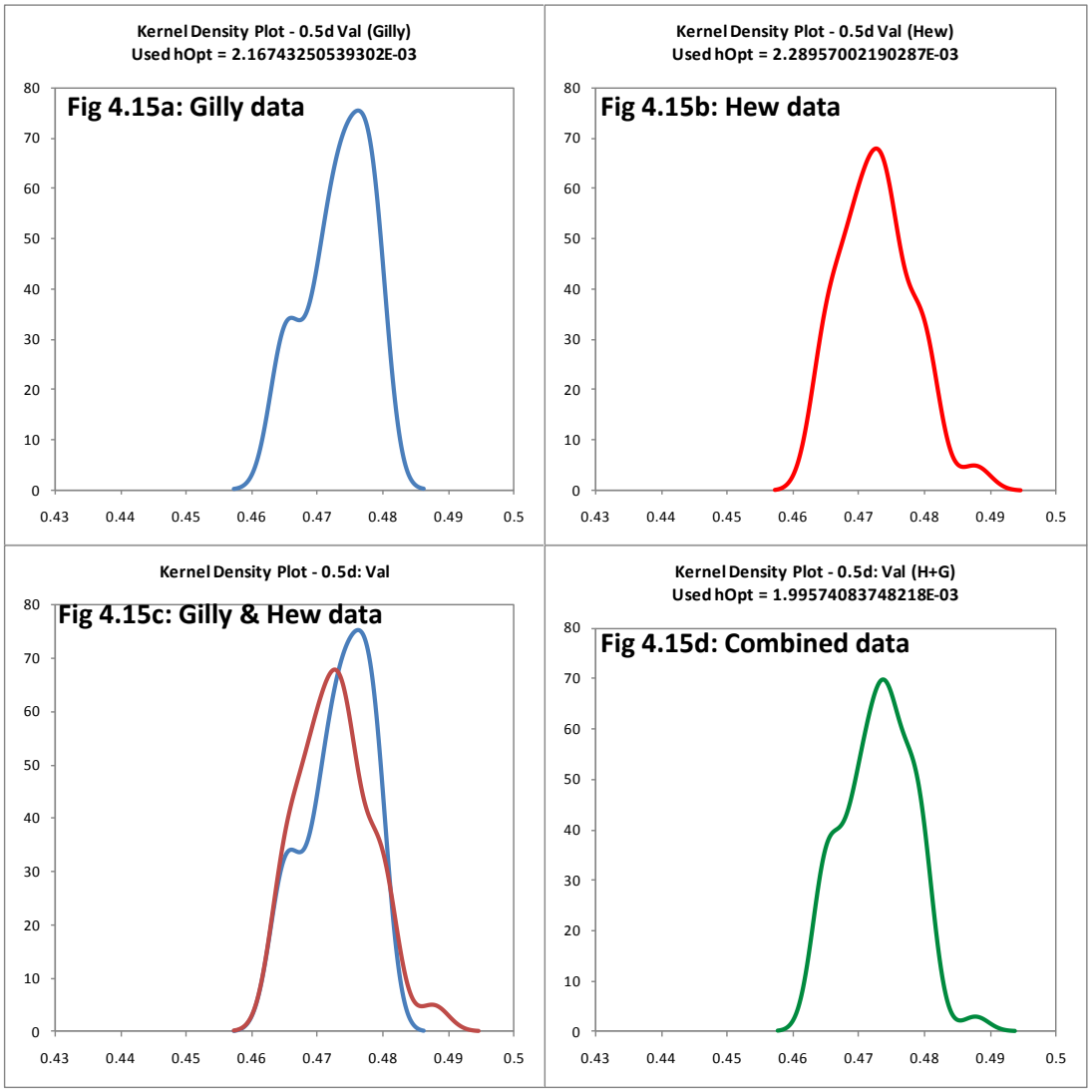
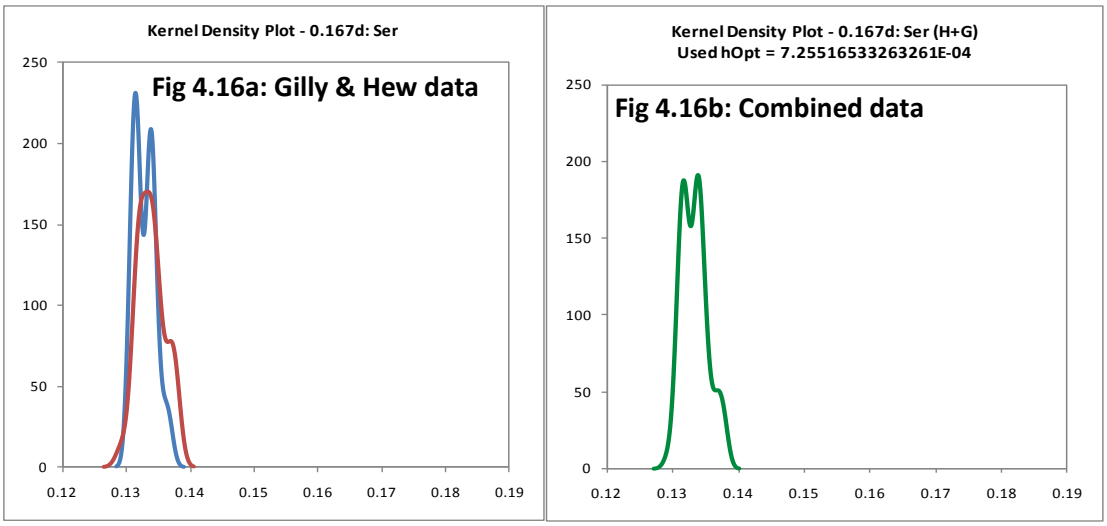


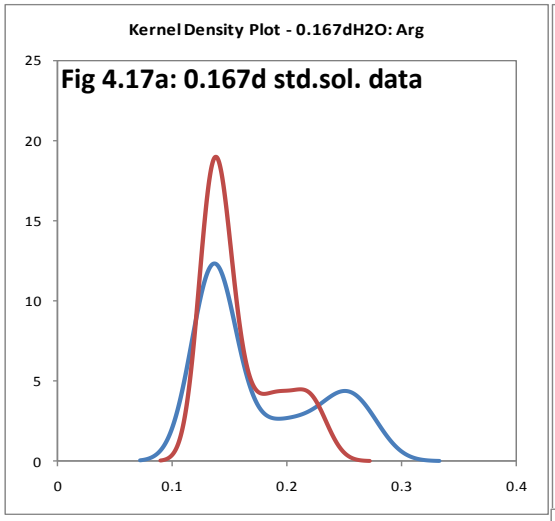
Figure 4.16: Kernel density summary: Trimmed Ser D/L values in 0.167d std sol run on Gilly & Hew

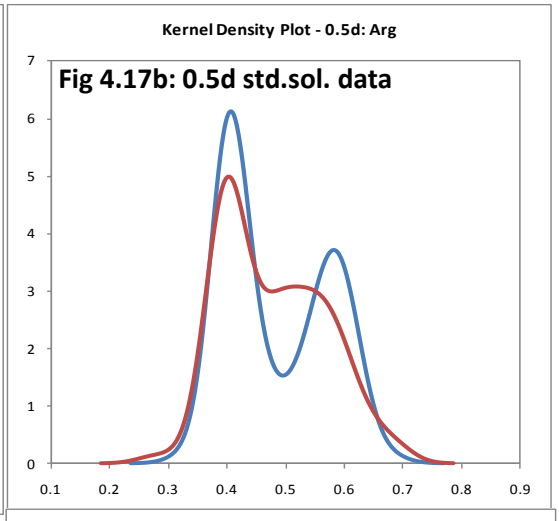


In every instance except for arginine (Arg), the means, medians and primary modes are all in fairly close agreement. Arginine data by contrast is far more varied and this is reflected in the %CV values given (between 20-35% and for H+G mean data, 262%!). This wide disagreement is reflected by the kernel densities, which clearly indicate data become increasingly bimodal as the D/L value increases, Figure 4.17a-c .

For serine, the bimodal twin peaks of the kernel density may be due to a degradation product, but because of the closeness of the peaks to each other, may equally be an artefact of the kernel density and the default *h0pt* being set too tight compared to the observed standard deviations. By comparison, the two modes of the arginine data are widely spread, and likely to represent two genuinely distinct populations. It has been found that arginine closely co-elutes with ammonia (Penkman, pers. comms.) which may be responsible for the apparent bimodality. Whether it is this or another protein degradation product that causes interference hasn't been established, but for this reason, arginine tends not to be included in the suite of amino acids used for quantitative purposes.

Figure 4.17: Kernel density summary: Trimmed Arg D/L values run on Gilly & Hew





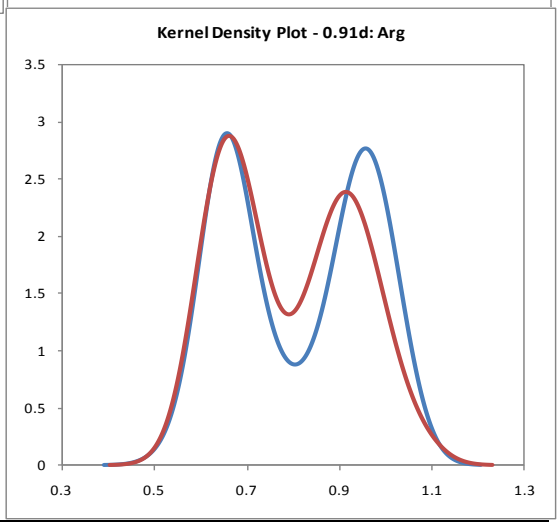


Fig 4.17c: 0.91d std sol

Table 4.4: Comparison between robust mean, median and mode for 0.167d std sol.

Std Sol.		Robust Mean					Median				Kernel Density				
0.167d		n	h15	std dev σ	%CV	std u	Median	sMAD	%CV	std u	mode 1	mode 2	mode 3	mode 4	mode 5
Gilly	Asx	49	0.171	0.002	0.9	2.0E-04	0.171	0.002	0.9	2.0E-04	0.171	0.177			
Gilly	Glx	49	0.183	0.005	2.6	7.0E-04	0.184	0.005	2.9	8.0E-04	0.180	0.186			
Gilly	Ser	49	0.133	0.002	1.3	2.0E-04	0.133	0.002	1.4	3.0E-04	0.131	0.134			
Gilly	Arg	48	0.197	0.054	27.2	7.7E-03	0.186	0.065	34.9	9.4E-03	0.156	0.245			
Gilly	Ala	48	0.180	0.006	3.1	8.0E-04	0.178	0.003	1.7	4.0E-04	0.177	0.187			
Gilly	Val	42	0.147	0.003	2.2	5.0E-04	0.147	0.003	2.3	5.0E-04	0.145	0.148			
Gilly	Met	42	0.206	0.008	3.9	1.2E-03	0.204	0.007	3.4	1.1E-03	0.202	0.214			
Gilly	Phe	42	0.159	0.003	1.7	4.0E-04	0.159	0.003	1.8	4.0E-04	0.154	0.159	0.161		
Gilly	Ile	41	0.193	0.003	1.5	5.0E-04	0.192	0.003	1.6	5.0E-04	0.192				
Gilly	Leu	41	0.202	0.007	3.6	1.1E-03	0.204	0.006	26.1	1.0E-03	0.196	0.205			
Hew	Asx	38	0.173	0.004	2.5	7.0E-04	0.172	0.004	2.1	6.0E-04	0.172	0.183			
Hew	Glx	38	0.184	0.003	1.5	5.0E-04	0.184	0.002	1.2	4.0E-04	0.184	0.188			
Hew	Ser	38	0.134	0.002	1.6	4.0E-04	0.134	0.002	1.6	3.0E-04	0.133	0.137			
Hew	Arg	38	0.187	0.044	23.4	7.1E-03	0.178	0.036	20	5.8E-03	0.163				
Hew	Ala	38	0.181	0.004	2.3	7.0E-04	0.181	0.005	2.5	7.0E-04	0.181				
Hew	Val	36	0.147	0.002	1.6	4.0E-04	0.147	0.002	1.3	3.0E-04	0.147				
Hew	Met	36	0.199	0.006	2.8	9.0E-04	0.198	0.004	2.0	7.0E-04	0.198	0.209			
Hew	Phe	36	0.160	0.003	1.6	4.0E-04	0.160	0.002	1.5	4.0E-04	0.161				
Hew	Ile	35	0.197	0.006	2.8	9.0E-04	0.197	0.006	2.8	9.0E-04	0.194	0.198			
Hew	Leu	35	0.211	0.021	10.1	3.6E-03	0.203	0.010	5.0	1.7E-03	0.200	0.238			
H+G	Asx	87	0.172	0.002	1.4	3.0E-04	0.171	0.002	1.3	2.0E-04	0.171	0.177	0.183	0.184	0.188
H+G	Glx	87	0.183	0.004	2.2	4.0E-04	0.184	0.005	2.5	5.0E-04	0.181	0.185			
H+G	Ser	87	0.133	0.002	1.4	2.0E-04	0.133	0.002	1.8	3.0E-04	0.132	0.134	0.137		
H+G	Arg	86	0.193	0.505	262	5.5E-02	0.179	0.053	29.6	5.7E-03	0.158	0.242			
H+G	Ala	86	0.180	0.005	2.8	5.0E-04	0.178	0.004	2.4	5.0E-04	0.177				
H+G	Val	78	0.147	0.003	1.9	3.0E-04	0.147	0.003	2.1	4.0E-04	0.147	0.155			
H+G	Met	78	0.203	0.008	3.7	9.0E-04	0.201	0.007	3.3	8.0E-04	0.200				
H+G	Phe	78	0.160	0.003	1.6	3.0E-04	0.160	0.003	1.6	3.0E-04	0.161				
H+G	Ile	76	0.194	0.004	2.3	5.0E-04	0.193	0.004	2.1	5.0E-04	0.193	0.214	0.221		
H+G	Leu	76	0.204	0.010	5	1.2E-03	0.204	0.009	4.6	1.1E-03	0.203	0.240			

Underlined values represent primary modes

4.2.3.4 Difference between instruments

A final check was carried out on the differences between D/L means for the two instruments. If the means were from the same population, then in theory, the observed differences should **not be greater** than **twice** the **expanded** combined standard uncertainties (for means) for Hew and Gilly (see Figure 4.18).

That is;

$$Absolute\ difference\ <\ 2\ \times\ (k\ x\ Std\ u),$$

where k= 2 (coverage factor at 95% CI)

Therefore, $Absolute\ difference\ <\ 4\ \times\ std\ u\ (H+G)$

Table 4.5 shows the values for 4 x std u (H+G) for the Robust mean and the median. These are compared to the absolute differences between individual Hew and Gilly data, expressed as an absolute value and as a relative percentage;

$$Absolute\ difference = \sqrt{\left(D/L_{Hew} - D/L_{Gilly}\right)^2} \qquad \textbf{(4.4)}$$

$$Relative\ \%\ difference = \frac{\sqrt{\left(D/L_{Hew} - D/L_{Gilly}\right)^2}}{\left(\left(D/L_{Hew} + D/L_{Gilly}\right)/2\right)} \qquad \textbf{(4.3)}$$

The absolute differences take no account of direction. Values shown in bold (Table 4.5) are the larger of the two comparative values. It can be seen that the two sets of values, (4(H+G std u) vs ab. diff) are generally well matched, with very little difference between them suggesting that the combined (H+G) uncertainty is probably sufficient to account for differences between the individual means. Any differences that do exist, are probably small in comparison to the uncertainty introduced by the application of the full method on test samples, and need not be of concern.

In this case, it can be reasonably assumed that data from the two instruments are derived from the same population and data can be pooled for subsequent evaluations.

Figure 4.18: Significant Difference between individual distribution means compared to the combined standard uncertainty

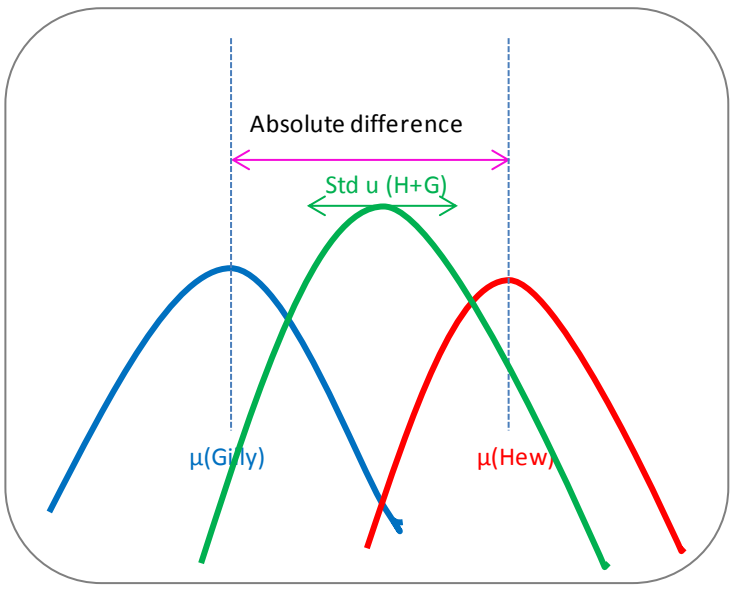


Table 4.5: Differences between Hew and Gilly D/L values

Std Sol.	Difference between Hew vs Gilly D/Ls					
0.167d	Robust mean			Median		
	4(H+G std u)	ab. diff	% diff	4(H+G std u)	ab. diff	% diff
Asx	0.0012	**0.0020**	1.16%	0.0008	**0.0010**	0.58%
Glx	**0.0016**	0.0010	0.54%	**0.0020**	0.0000	0.00%
Ser	0.0008	**0.0010**	0.75%	**0.0012**	0.0010	0.75%
Arg	**0.2180**	0.0100	5.21%	**0.0228**	0.0080	4.40%
Ala	**0.0020**	0.0010	0.55%	0.0020	**0.0030**	1.67%
Val	**0.0012**	0.0000	0.00%	**0.0016**	0.0000	0.00%
Met	0.0036	**0.0070**	3.46%	0.0032	**0.0060**	2.99%
Phe	**0.0012**	0.0010	0.63%	**0.0012**	0.0010	0.63%
Ile	0.0020	**0.0040**	2.05%	0.0020	**0.0050**	2.57%
Leu	0.0048	**0.0090**	4.36%	**0.0044**	0.0010	0.49%

4(G+H std u) vs ab. Diff: bold text represents the larger value

4.2.4 Summing up

This section started by looking at the stability of uncalibrated chromatogram peak areas in the internal standard L-homoarginine. It was found that whilst values fluctuated in both instruments (Gilly and Hew), there were some substantial changes in peak areas derived by Hew over a seven year interval (2003-2010), in fact doubling the peak area values derived for LhArg at a constant 0.01 mM concentration. It is thought that this variability may be due to the fluorescence bulb emissions dropping off over time. If so, then these fluctuations would also affect results derived from standard solutions and biomineral test samples in the same way. For these reasons LhArg is used as an internal standard to normalise and correct peak area values and concentrations prior to the subsequent calculation of D/L values.

Historically, for the purposes of geochronology, it has been assumed that D/L values generated by both instruments are equivalent. For analysis of AAR data and uncertainty determination using ANOVA, it is helpful if data from both instruments can be combined and assessed as a single data set. In order to assess the normality of data, t-tests were applied to D/L values for amino acids in standard solution. Whilst for the majority of cases, significant differences between the two instrument means were not significant in the untrimmed datasets, for valine (in 0.5d std sol), results suggested significant differences were detectable at the 5% confidence level.

Data were plotted to observe the presence of rogue values and potential outliers and evaluated using boxplots (to indicate the position of the means and medians) and frequency distribution histograms and the Kolmogorov-Smirnov test was used to assess normality. Outliers were removed using the same criteria applied by Minitab for the boxplots

(Q3 + 1.5 (Q3 - Q1), where Q = quartile and Q3 - Q1 is the inter-quartile range, (mid 50% of data points) and reassessed. Results of the re-evaluated t-tests were surprising and indicated that rather than improving the agreement between means, in many instances, it made it worse. However, results of the revised boxplots suggested closer agreement between Hew and Gilly data and in the majority of instances the repeated Kolmogorov-Smirnov test gave increased p-values indicating that data were becoming more normalised. Robust means and standard deviations were derived (which minimise the effect of outliers), and kernel density distributions obtained for each instrument separately and combined. The overlap between the superimposed distributions from each instrument and an evaluation of the absolute differences between instrumental Robust means and medians, suggest that data are in fact equivalent.

Any difference between instrumental means can probably be accounted for by the uncertainty of the combined mean, and likely to be small in comparison to uncertainty contributions resulting from the extraction stages of the method on test samples.

Baring in mind that ANOVA's prior assumption for normality may relate only to the application of the F-test (Miller and Miller, 2005; McDonald, 2009), results of these evaluations would indicate that the distribution of data appear to show little deviation from normality and that ANOVA is a sufficiently robust statistical approach to apply to the calculation of precision estimates.

Therefore, for the remainder of this thesis, assessments of uncertainty (unless otherwise indicated) have all been carried out on combined instrument data.

4.3 Precision Evaluation by ANOVA; Standard Solutions

A one-way analysis of variance, **ANOVA,** allows us to separate the uncertainty contributions arising from the **within-sample repeatability (standard deviation, SDr or sr)** and the **between-sample /between-run variability over time.** Taken together, they represent the overall expected uncertainty for carrying out the analytical procedure. When applied to the evaluation of inter-laboratory data, this combined precision estimate is called the **reproducibility standard deviation (s_R)** often expressed as a percentage, as the **relative standard deviation of reproducibility (RSD_R%).** When applied at the single laboratory level, this combined effect is called the **within or intra-laboratory reproducibility or intermediate precision (s_{RW}, or RSD_{RW}%),** and represents the maximum expected variation in results.

The following analysis of standard solution data is based on an archive of data, collected over several years by a single laboratory. Therefore, it can be thought of as an analysis of intermediate precision or intra-laboratory reproducibility, i.e. sufficient for the needs of the laboratory and reflects the level of expected variation for that matrix at a specific concentration, analysed on a routine basis. It includes contributions from the variation due to random errors, method and instrument factors. However, a full measure of precision can only be obtained from an organized inter-laboratory study which, in addition to the above also reflects the between-laboratory variability.

The analysis of data from collaborative trials is based upon an analysis of variance and is described in detail elsewhere (Youden and Steiner 1975, Wernimont 1985, ISO 1994, IUPAC 1995). However if the same underlying principles were to be applied to existing data then this is likely to provide the most informative evaluation of precision estimates so far. This approach is known as the "top-down" method of uncertainty determination and has been incorporated into the guidance document ISO 21748.

ANOVA is simply an analysis of variance, more often used to test hypotheses regarding differences between variances. The F-statistic derived from dividing the between sample variance (between Mean Square or between MS) by the within sample variance (within MS), is then compared to tabulated critical values dependent on the degrees of freedom and required probability level.

However, the calculations employed for arriving at the MS values, provide us with a convenient way of deriving the between-sample and within-sample variances.

Within laboratory (or sample) variance is;

$$s_r^2 = within\ MS, \quad thus;\ s_r = \sqrt{within\ MS}$$

Between laboratory (or sample) variance is;

$$s_L^2 = \frac{between\ MS - within\ MS}{n} \qquad thus;$$

$$s_L = \sqrt{\frac{between\ MS - within\ MS}{n}}$$

And the Reproducibility variance is;

$$s_R^2 = s_r^2 + s_L^2 \qquad \textbf{(4.5)}$$

Where the standard deviation of reproducibility is considered equivalent to the standard uncertainty, (u). Further details of formulae for the calculation of s_r and s_L are given in the ISO Standard 5725, Part 2, 1994, and their application to uncertainty estimation in ISO 21748.

Thus, data for amino acids in each of the standard solutions underwent testing by ANOVA to provide details of within and between sample variability together with intermediate reproducibility estimaters. An Excel spreadsheet previously developed for use in collaborative trial evaluation (Mathieson 2000), was enhanced to accommodate unequal replicates due to variable numbers of repeat injections having been carried out during each run. The calculation for the within MS was therefore adjusted to calculate a pooled variance, rather than the original ISO 5725 design that assumed a uniform replicate analysis. In addition, it was necessary to calculate a representative value for n for use in the calculation of s_L, for the same reason. Both enhancements were in accordance with recommendations given in ISO 5725-2:1994.

For these evaluations, it was considered important to represent the fullest extent of potential variation acting on measurement results carried out by the York laboratory and give realistic intermediate reproducibility estimates. Therefore in this section, evaluations have been carried out including a more recently introduced standard solution that will be identified as **0.167dH2O** and a third newer UHPLC instrument identified as **Chem (C)**. The 0.167dH2O std sol has exactly the same composition as the previously described 0.167d std sol (section 4.2), and has been taken from the same original stock solution. The difference is in the subsequent dilution carried out on all standard solutions, in order to achieve appropriate isomer concentrations for peak area integration and plotting. Standard solutions described in the previous section (i.e. 0.167d, 0.5d and 0.91d), all received a final 10% dilution in rehydration fluid, whereas 0.167dH2O was diluted in HPLC grade water. The diluents used do not affect the resulting D/L values as currently determined by the laboratory. The third instrument Chem, is a UHPLC for which only 10 or so data points were available at the time these evaluations were carried out. Their inclusion, again, do not unduly influence derived precision estimates and their use contributes to a more complete picture of intra-laboratory intermediate reproducibility precision. In the future, a formal method validation with precision analysis would enable comparisons between the performance of HPLC against UHPLC to be made.

4.3.1 Cochran' and Grubb's Outlier Tests

All data values, i.e. all replicates for all samples, were evaluated for a specified amino acid, at a given std sol concentration, on one of three HPLC instruments, together with their combined effect, i.e. H+G and H+G+C. Thus for raw data evaluations, up to a total of 200 (10x4x5) separate analyses were carried.

In addition, in order to avoid the influence of extreme outliers on precision estimates, all data were re-evaluated after the exclusion of major outlier values, identified as being less than or greater than 3 times the initial standard deviation for the all raw data entries prior to ANOVA evaluation.

Finally, a third set of corrected data were evaluated, allowing for the removal of outliers by the recommended methods as detailed in IUPAC Technical Report 1995, and ISO 5725-2:1994; using the Cochran's and Grubbs tests.

Separate evaluations were carried out in order to compare precision estimates between amino acids, between standard solution concentrations and between instruments.

4.3.1.1 Cochran's outlier test

This test is based on the assumption that laboratory repeatability is likely to play a small role compared to other factors and as such, is expected to remain reasonably consistent across replicates for all samples (or laboratories). This process therefore compares the highest replicate standard deviation for the samples (p) by generating a Cochran's statistic, C, and comparing it to a tabulated critical value.

$$C = \frac{s_{max}^2}{\sum_{i=1}^{p} s_i^2} \qquad \textbf{(4.6)}$$

ISO 5725-2:1994, suggests that if the C-value is less than or equal to its 5% critical value, the data is accepted, if greater than 5% critical value but less than or equal to its 1% critical value, the data is identified as a straggler and should be reviewed, if greater than the 1% critical value data is regarded as an outlier and omitted. Data is then subjected to the same evaluation, each time clearing the outliers until the process has exhausted the highest variable values. IUPAC however refer to exclusion of data if values exceed the critical value at 2.5% (one tail) level. For the purposes of this evaluation, data are excluded if they exceed 2.5% level critical value. It should be noted that data with extremely tight deviations are not evaluated in the same way even though they too could overly influence the final precision

estimate. However, it seems a little unfair to penalize a laboratory for demonstrating better performance than anyone else.

4.3.1.2 Grubb's outlier test.

Unlike Cochran's test, Grubbs test looks to compare the largest and/or smallest mean replicate values with the overall mean. Again, a specific value known as the Grubb's statistic (G) is determined and compared in the same way to a tabulated critical value as described for the previous test. Three levels of Grubb's test may be applied to the data, single Grubb's, for a single outlying observation, high or low, double Grubb's to evaluate the two highest or two lowest values at the same time, i.e. hh or ll, and then double Grubb's high and low (hl) value test. Details of the formulae required for these tests are given on ISO 5725-2:1994, p12.

Outlier tests are applied sequentially, initially Cochran's, followed by single Grubb's (SG) when no more Cochran's outliers are found. Following the SG, data are again reassessed for new Cochran's outliers, again SG and if none present, double Grubb's (DG), hh or ll, again if nothing flags, DG hl is applied. Each time an outlier is removed the data undergo a reappraisal by Cochran's test.

Figure 4.19: Mean and Range chart for Ala corrected D/L values, 0.167d std sol, run on Gilly

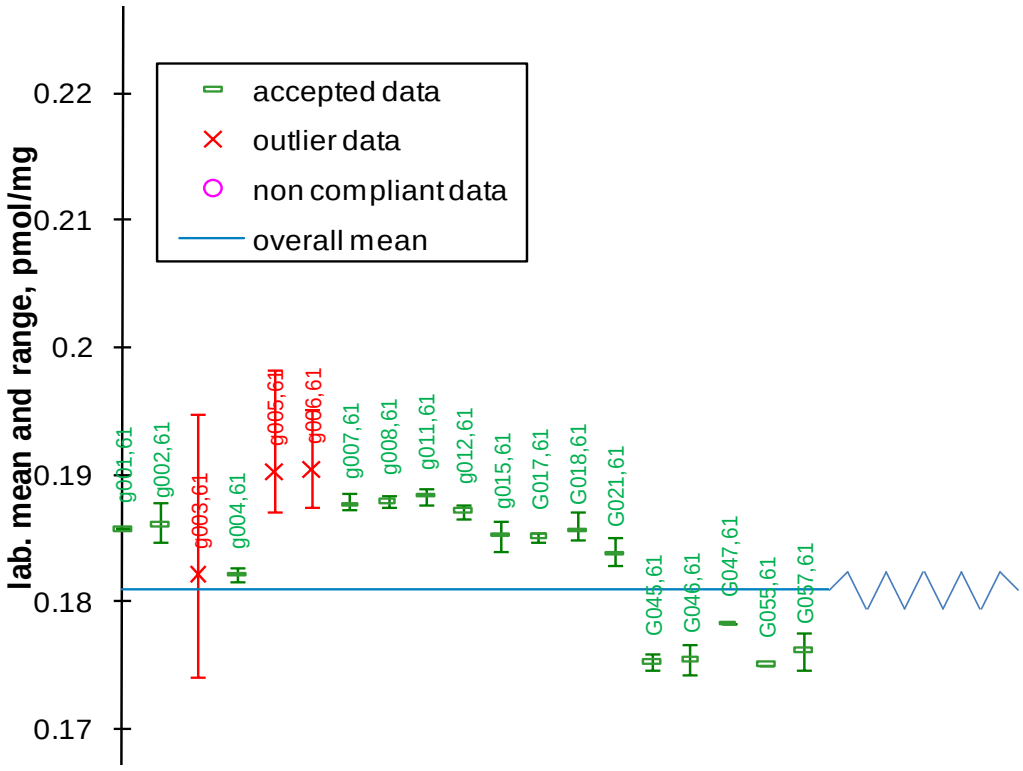
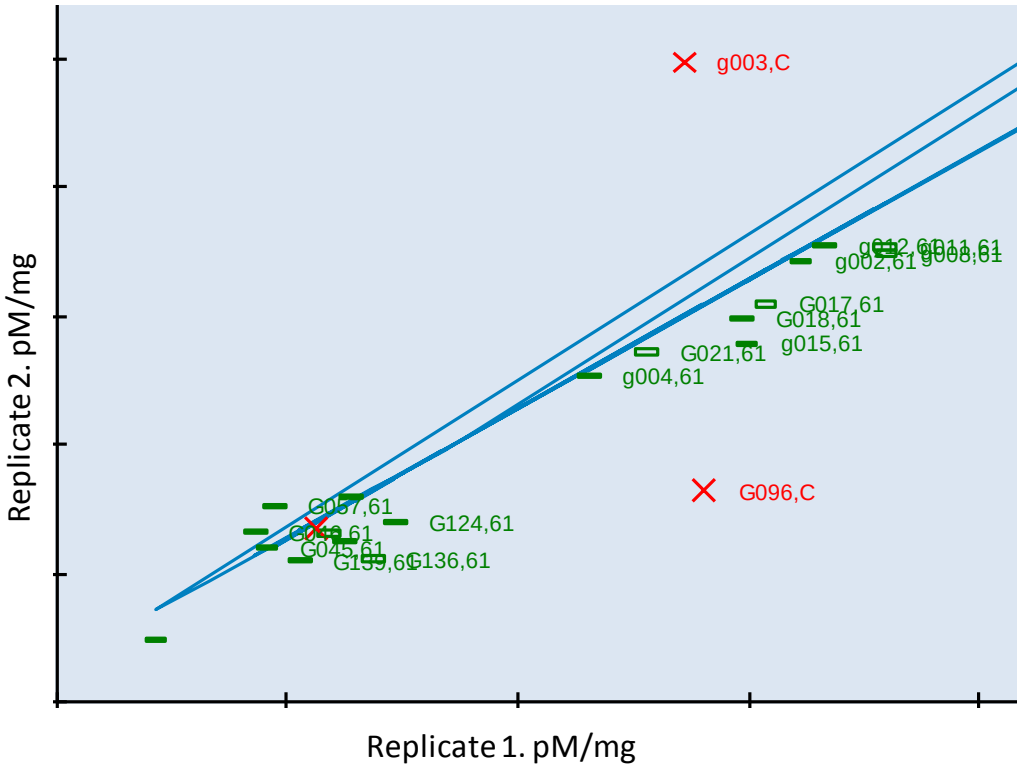


Figure 4.20: Youden Plots of matched replicates (Rep 1 & Rep2) for Ala corrected data, 0.167d std sol, run on Gilly



Caution should also be exercised so as not to remove an excessive number of samples and lose the underlying distribution. This is especially important for small sample (laboratory) numbers and it has been recommended (IUPAC 1995) that outliers should only be removed up to a maximum of a 22.2% reduction in the original number of laboratories / samples, (i.e. 2/9).

Mean and range charts such as that seen in Figure 4.19 can be used to clearly identify replicates whose individual variance is considerably higher than the majority of other results. In addition, Youden plots, Figure 4.20, plot replicate values against each other, and can be used to assist in identifying extreme values when they appear.

Figure 4.21 and Figure 4.22 demonstrate the effect of outlier treatment. Figure 4.21 relates to the effect on the within-lab reproducibility or intermediate standard deviation (i.e., s_{RW}), whilst Figure 4.22 relates to the effect on the within-lab reproducibility or intermediate relative standard deviation (i.e.; RSD_{RW}). Both diagrams show Asx D/L data analysed on Gilly. In both charts, blue lines represent uncorrected raw data, green lines represent data after the removal of gross outliers and red data refer to those data sets subjected to outlier removal using the Cochran's and Grubb's outlier tests.

Figure 4.21: Comparison of the effect of outlier treatment on Reproducibility standard deviations (s_{RW}) of Asx D/L values run on Gilly

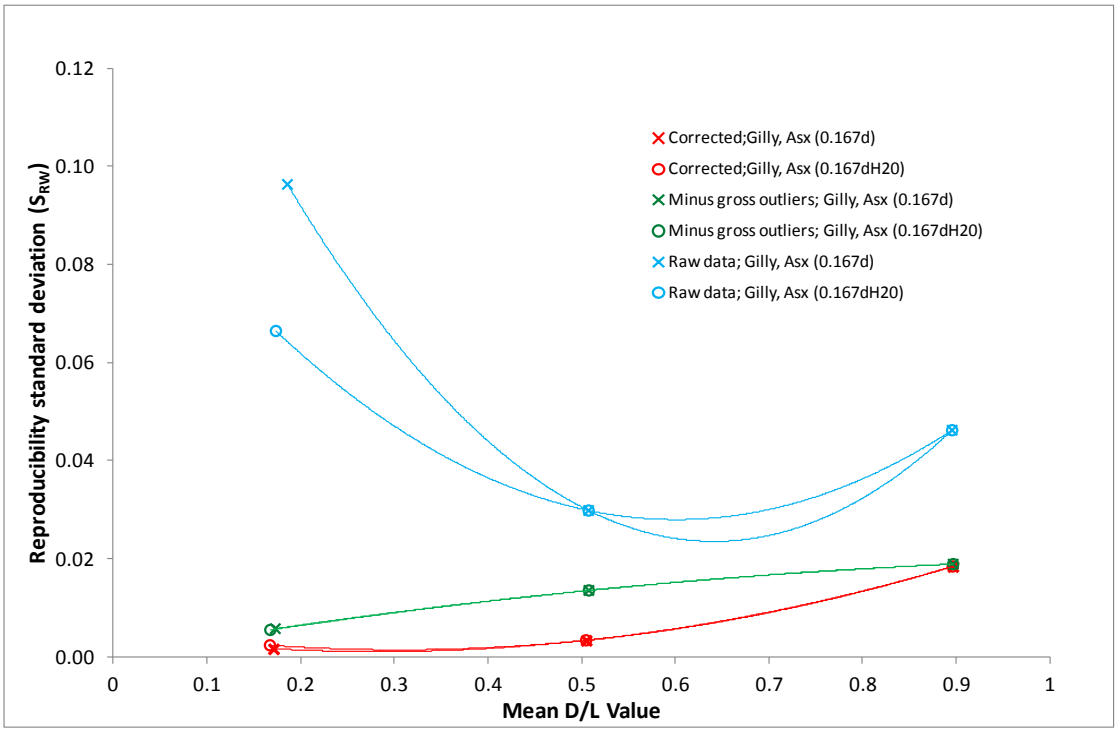
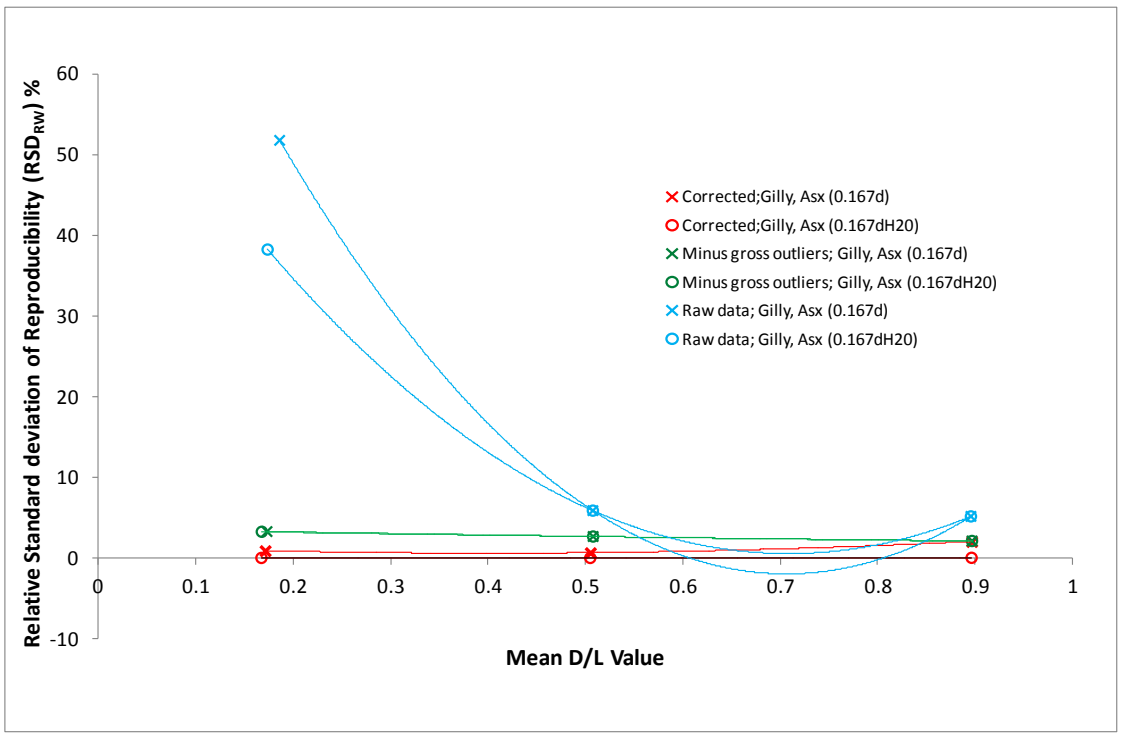


Figure 4.22: Comparison of the effect of outlier treatment on relative Reproducibility standard deviations (RSD_{RW} %) of Asx D/L values run on Gilly



The improvement in precision as outliers are identified and removed can be seen, giving smaller s_{RW} and RSD_{RW} values, as might be expected. Because two separate solutions at the 0.167 level have been evaluated, there are two separate results at the same concentration level. Consequently, two sets of trendlines can be seen, one using 0.167d and the other using 0.167dH20. However, all trendlines use the same 0.5d and 0.91d solution in each chart.

4.3.2 Analytical Precision Estimates

Precision estimates derived using ANOVA on fully corrected values from all three instruments combined, are given in Table 4.6. Data show calculated values for the final or effective replicate number n1, resulting from uneven replicate numbers being reported in the individual runs. p1 are the final number of individual runs used in the ANOVA evaluations after the removal of Cochran's and Grubb's outliers. As might be expected, data demonstrate the smallest standard deviations for repeatability precision, s_r, representing an estimation of random error influences within a single run, with slightly wider estimates for the between-sample, s_L deviations, due to additional variability caused by the changes in day to day run-bias. Together they give the s_{RW}, the overall estimate of expected uncertainty for individual amino acids in standard solution, for any instrument at the York facility.

Data derived from across all three instruments and standard solutions are reasonably consistent, although there are some noticeable differences between the 0.167d and 0.167dH2O std sols, for example; Asx and Glx D/L values. In both instances, the s_L values are slightly wider for the 0.167d std sol by just the two instruments Hew and Gilly compared to the 0.167dH2O data reported for the three instruments together. Differences for other amino acids can also be seen but there is no obvious explanation. Whatever the cause, no further data were identified by the outlier tests and there were no obvious reasons to exclude any additional data from the ANOVA evaluations.

With regard to the repeatability precision, values shown (Table 4.6) represent the level of agreement between repeated instrumental analyses only, i.e. due to replicate injections from the same vial. With the exception of alanine and leucine both in the 0.167dH20 solution, precision estimates for Asx, Glx, Ser, Ala, Val and Phe D/L values are generally less than 1%, for D-Aile/L-Ile, Leu and Met D/L s_r is less than 2% and for arginine s_r increases to 5-6%.

Table 4.6: ANOVA precision estimates for amino acid D/L values in standard solution after removal of outliers using Cochran's and Grubb's tests.ANOVA precision estimates for amino acid D/L values in standard solution after

Amino Acid	Instrument	Std sol	mean	n1	p1	Sr	RSDr%	SL	RSDL%	SR	RSDR%
Asx	G+H	0.167d	0.173	2.3	79	0.0022	0.00	0.0047	2.72	0.0052	3.00
Asx	G+H+C	0.167dH20	0.168	2.6	157	0.0013	0.78	0.0022	1.32	0.0026	1.54
Asx	G+H+C	0.5d	0.506	2.7	236	0.0045	0.90	0.0035	0.69	0.0057	1.14
Asx	G+H+C	0.91d	0.896	2.6	220	0.0040	0.44	0.0172	1.92	0.0176	1.97
Glx	G+H	0.167d	0.184	2.3	84	0.0021	1.12	0.0047	2.55	0.0051	2.79
Glx	G+H+C	0.167dH20	0.193	2.6	163	0.0019	0.96	0.0030	1.58	0.0036	1.85
Glx	G+H+C	0.5d	0.571	2.7	245	0.0045	0.80	0.0067	1.17	0.0081	1.41
Glx	G+H+C	0.91d	1.007	2.7	240	0.0121	1.21	0.0112	1.12	0.0165	1.64
Ser	G+H	0.167d	0.133	2.3	83	0.0012	0.93	0.0017	1.25	0.0021	1.56
Ser	G+H+C	0.167dH20	0.132	2.6	161	0.0012	0.92	0.0017	1.32	0.0021	1.61
Ser	G+H+C	0.5d	0.409	2.7	250	0.0033	0.80	0.0038	0.94	0.0051	1.24
Ser	G+H+C	0.91d	0.700	2.7	238	0.0063	0.90	0.0055	0.79	0.0085	1.22
Arg	G+H	0.167d	0.187	2.3	79	0.0430	6.04	0.0453	24.22	0.0467	24.96
Arg	G+H+C	0.167dH20	0.167	2.6	147	0.0090	5.41	0.0453	27.11	0.0466	27.88
Arg	G+H+C	0.5d	0.483	2.7	237	0.0229	4.73	0.0909	18.81	0.0937	19.40
Arg	G+H+C	0.91d	0.804	2.6	223	0.0403	5.01	0.1542	19.18	0.1594	19.82
Ala	G+H	0.167d	0.179	2.2	74	0.0016	0.90	0.0064	3.57	0.0066	3.68
Ala	G+H+C	0.167dH20	0.159	2.6	153	0.0053	3.31	0.0052	3.28	0.0074	4.67
Ala	G+H+C	0.5d	0.557	2.7	245	0.0045	0.81	0.0077	1.37	0.0089	1.60
Ala	G+H+C	0.91d	0.935	2.6	220	0.0058	0.62	0.0088	0.94	0.0105	1.12

n1 = final number of replicates after removal of Cochran's & Grubbs outliers

p1 = final number of independent samples after removal of Cochran's & Grubbs outliers

Table 4.6: ANOVA precision estimates for amino acid D/L values in standard solution after removal of outliers using Cochran's and Grubb's tests.

Amino Acid	Instrument	Std sol	mean	n1	p1	Sr	RSDr%	SL	RSDL%	SR	RSDR%
Val	G+H	0.167d	0.147	2.4	76	0.0023	1.59	0.0024	1.60	0.0033	2.26
Val	G+H+C	0.167dH20	0.145	2.7	143	0.0013	0.88	0.0023	1.58	0.0026	1.81
Val	G+H+C	0.5d	0.475	2.8	226	0.0041	0.87	0.0053	1.12	0.0067	1.42
Val	G+H+C	0.91d	0.760	2.7	211	0.0060	0.79	0.0070	0.92	0.0092	1.21
Met	G+H	0.167d	0.204	2.3	75	0.0029	1.41	0.0079	3.86	0.0084	4.11
Met	G+H+C	0.167dH20	0.200	2.7	146	0.0027	1.34	0.0052	2.62	0.0059	2.94
Met	G+H+C	0.5d	0.592	2.8	222	0.0051	0.87	0.0088	1.49	0.0102	1.72
Met	G+H+C	0.91d	1.021	2.7	212	0.0091	0.89	0.0122	1.19	0.0152	1.49
Phe	G+H	0.167d	0.160	2.3	66	0.0008	0.53	0.0025	1.57	0.0027	1.66
Phe	G+H+C	0.167dH20	0.157	2.7	138	0.0009	0.57	0.0018	1.14	0.0020	1.28
Phe	G+H+C	0.5d	0.486	2.7	218	0.0028	0.58	0.0038	0.78	0.0047	0.97
Phe	G+H+C	0.91d	0.805	2.7	209	0.0056	0.70	0.0071	0.88	0.0090	1.12
Ile	G+H	0.167d	0.198	2.4	71	0.0026	1.31	0.0115	5.81	0.0118	5.96
Ile	G+H+C	0.167dH20	0.192	2.7	135	0.0022	1.13	0.0097	5.06	0.0099	5.18
Ile	G+H+C	0.5d	0.580	2.8	210	0.0069	1.19	0.0147	2.53	0.0162	2.80
Ile	G+H+C	0.91d	0.989	2.7	206	0.0100	1.01	0.0181	1.83	0.0207	2.09
Leu	G+H	0.167d	0.206	2.4	72	0.0039	1.89	0.0121	5.88	0.0128	6.19
Leu	G+H+C	0.167dH20	0.202	2.7	143	0.0081	4.03	0.0162	8.05	0.0182	9.00
Leu	G+H+C	0.5d	0.601	2.8	218	0.0115	1.91	0.0151	2.51	0.0189	3.15
Leu	G+H+C	0.91d	1.062	2.6	203	0.0108	1.02	0.0163	1.54	0.0196	1.84

n1 = final number of replicates after removal of Cochran's & Grubbs outliers

p1 = final number of independent samples after removal of Cochran's & Grubbs outliers

For Ser, Val, Phe and the majority of Asx and Glx, between-sample precision estimates, s_L, are less than 2%. For Ala and Met, the higher D/L standard s_L is also less than 2% but for 0.167 solutions this increases to less than 4%. Both D-Aile/L-Ile and Leu show wider deviations between samples of between 6-8% and again, Arg gives the least agreement of all the amino acids with between-sample estimates of up to 27%. Overall, the intermediate precision estimates, s_{RW}, follow the same levels of agreement as the s_L values, since these have the largest contribution and will make the biggest impact on the combined value.

As explained in the previous chapter, the **repeatability standard deviations (s_r)** can provide estimates of **repeatability (r),** that is; the absolute difference between replicates, and is helpful in monitoring in-run repeatability, by the analysis of replicate control material samples.

4.3.3 Repeatability limit

Repeatability estimates, are calculated at the 2 standard deviation confidence level and given as $r = t \times \sqrt{n} \times s_r$ where t is the t-value for a normal distribution at 95% probability, i.e.; 1.96 , rounded to 2, and *n* is the number or replicates.

Thus for Asx in 0.167d; $r = 2 \times \sqrt{2.3} \times 0.0022 = 0.0067$

Therefore, in a solution, with an Asx D/L value of approximately 0.17, duplicate measurements (or in this case duplicate injections) should not exceed a 0.007 D/L difference.

Values exceeding a 3 std dev confidence level, ($3 \times \sqrt{2.3} \times 0.0022 = 0.01$ D/L) suggest analytical measurement problems and samples should be re-run or a new QC sample used.

2 standard deviation repeatability limits derived from data given in Table 4.6, are given in Table 4.7. In this example, the repeatability only reflects injection precision between replicate injections. However, ideally, repeatability should reflect the entire measurement process including preparation and extraction stages, for optimal measurement system control, and is why QC materials should ideally be matrix-matched.

Reproducibility limits can also be calculated in just the same way to monitor precision between duplicates in independent runs.

4.3.4 Effect of Sample Size

By convention, the s_R or s_{RW} is regarded as the measurement uncertainty (1 std dev) of the data in question (Magnusson *et al.*, 2004; da Silva *et al.*, 2006) and represents the expected distribution of individual values. However, by convention, where a measurement result is derived from the average of a number of repeated **independent** measurements, then the measurement uncertainty needs also to be a function of the sample size. Therefore,

if; $u = s \approx s_{RW} = \sqrt{s_r^2 + s_L^2}$ for single measurements, when, $u = \frac{s}{\sqrt{n}}$,

then $s_{RW} = \sqrt{s_r^2/n + s_L^2}$ for means

(EURACHEM / CITAC, 2000; EURACHEM/EUROLAB/CITAC/Nordtest/AMC, 2007).

With specific regard to AAR analysis, replicate injection measurements cannot strictly be considered as independent. In order to be independent, separate portions of the original material need to be taken and worked up through the entire method, independently. Further, injection or instrumental repeatability will most likely be much smaller than that derived from independent sample measurements for true repeatability precision determination. Replicate injection measurements might be taken to ensure a more reliable determination of the sample value, but this component of uncertainty would not normally be assessed separately, as instrumental variability will be subsumed into the higher level sample repeatability derived from independent samples (if determined).

Nonetheless, precision estimates derived from existing data must reflect the method as currently practised, so for the purpose of illustration, the replicate injections will be considered as independent measurements in this section.

Using s_r and s_L values from Table 4.6, the effect on s_{RW} can be observed by changing the theoretical sample size, n. Appropriate t-values can be used as correction factors and resultant s_{RW} uncertainties adjusted for small sample sizes, depending on the required confidence level and degrees of freedom (n-1).

Examples of these results are shown below in Figure 4.23a-c for alanine D/L values. s_r and s_L values derived across all instruments for 0.167d, 0.5d and 0.91d standard solutions (Table 4.6) are used to determine revised s_{RW} values with different values of *n*. These revised s_{RW} values are then multiplied by the appropriate t-value, and then added to or subtracted from the mean D/L value for the upper and lower confidence levels.

Table 4.7: Repeatability limits for amino acid D/L values in standard solution

Amino Acid	Instrument	Std sol	mean	n1	Sr	repeatability limit (2r)	
						absolute	as %
Asx	G+H	0.167d	0.173	2.3	0.0022	0.007	3.86
Asx	G+H+C	0.167dH20	0.168	2.6	0.0013	0.004	2.52
Asx	G+H+C	0.5d	0.506	2.7	0.0045	0.015	2.96
Asx	G+H+C	0.91d	0.896	2.6	0.0040	0.013	1.42
Glx	G+H	0.167d	0.184	2.3	0.0021	0.006	3.48
Glx	G+H+C	0.167dH20	0.193	2.6	0.0019	0.006	3.11
Glx	G+H+C	0.5d	0.571	2.7	0.0045	0.015	2.60
Glx	G+H+C	0.91d	1.007	2.7	0.0121	0.040	3.95
Ser	G+H	0.167d	0.133	2.3	0.0012	0.004	2.76
Ser	G+H+C	0.167dH20	0.132	2.6	0.0012	0.004	2.98
Ser	G+H+C	0.5d	0.409	2.7	0.0033	0.011	2.64
Ser	G+H+C	0.91d	0.700	2.7	0.0063	0.021	2.95
Arg	G+H	0.167d	0.187	2.3	0.0430	0.130	69.29
Arg	G+H+C	0.167dH20	0.167	2.6	0.0090	0.029	17.30
Arg	G+H+C	0.5d	0.483	2.7	0.0229	0.075	15.49
Arg	G+H+C	0.91d	0.804	2.6	0.0403	0.131	16.25
Ala	G+H	0.167d	0.179	2.2	0.0016	0.005	2.65
Ala	G+H+C	0.167dH20	0.159	2.6	0.0053	0.017	10.69
Ala	G+H+C	0.5d	0.557	2.7	0.0045	0.015	2.65
Ala	G+H+C	0.91d	0.935	2.6	0.0058	0.019	2.02

Amino Acid	Instrument	Std sol	mean	n1	Sr	repeatability limit (2r)	
						absolute	as %
Val	G+H	0.167d	0.147	2.4	0.0023	0.007	4.84
Val	G+H+C	0.167dH20	0.145	2.7	0.0013	0.004	2.87
Val	G+H+C	0.5d	0.475	2.8	0.0041	0.014	2.89
Val	G+H+C	0.91d	0.760	2.7	0.0060	0.020	2.61
Met	G+H	0.167d	0.204	2.3	0.0029	0.009	4.30
Met	G+H+C	0.167dH20	0.200	2.7	0.0027	0.009	4.38
Met	G+H+C	0.5d	0.592	2.8	0.0051	0.017	2.89
Met	G+H+C	0.91d	1.021	2.7	0.0091	0.030	2.95
Phe	G+H	0.167d	0.160	2.3	0.0008	0.002	1.50
Phe	G+H+C	0.167dH20	0.157	2.7	0.0009	0.003	1.86
Phe	G+H+C	0.5d	0.486	2.7	0.0028	0.009	1.92
Phe	G+H+C	0.91d	0.805	2.7	0.0056	0.019	2.30
Ile	G+H	0.167d	0.198	2.4	0.0026	0.008	4.03
Ile	G+H+C	0.167dH20	0.192	2.7	0.0022	0.007	3.69
Ile	G+H+C	0.5d	0.580	2.8	0.0069	0.023	3.97
Ile	G+H+C	0.91d	0.989	2.7	0.0100	0.033	3.31
Leu	G+H	0.167d	0.206	2.4	0.0039	0.012	5.83
Leu	G+H+C	0.167dH20	0.202	2.7	0.0081	0.027	13.26
Leu	G+H+C	0.5d	0.601	2.8	0.0115	0.038	6.35
Leu	G+H+C	0.91d	1.062	2.6	0.0108	0.035	3.28

Because charts are derived from standard deviations, confidence intervals widen and demonstrate the dependence of uncertainty estimates with increasing D/L values. Charts also show how the uncertainty of the mean diminishes as sample numbers increase, at the different confidence levels. Notice particularly how the effect of sample number becomes critical at values equal to or less than three.

For sample numbers of 5, this gives 4 degrees of freedom, (n=5, df = 4), and a $t_{(\alpha=0.05)}$ value (95% prob) = 2.776. The t-value is used as the coverage factor (k) with which the standard uncertainty estimate (u) is multiplied in order to derive the expanded uncertainty (U) at the required level of confidence. t-values are easily obtained from statistical tables (Neave, 1978). For p=4, df=3, $t_{(\alpha=0.05)}$=3.182, for p=3, df=2, $t_{(\alpha=0.05)}$=4.303, and for p=2, df=1, $t_{(\alpha=0.05)}$=12.71. Thus it can be seen that 3 (5 would be better) is probably the minimal **sample size** that is required for routine use, as this will reduce the uncertainty estimate to an acceptable level, without overstretching a laboratory's resources.

Similar evaluations to those shown in Figure 4.23, for other amino acids using the combined instrument data, (Hew+Gilly+Chem), have been carried out and their confidence interval charts are given in Chpt 4: Appendix 8.

Note: On some of the charts bumps can be observed particularly on the upper and lower 3 standard deviation confidence levels. These are Excel artefacts as it attempts to draw a smooth line between points around a tight bend, and are not a function of the t-values or s_{RW} value used.

4.3.5 Summarising Precision estimates

Two pairs of figures summarizing the overall observed uncertainty for combined data, (i.e. Hew+Gilly+Chem) for each of ten amino acids, are now given. The first pair of charts (Figure 4.24a and b) plot the D/L values obtained for corrected data (outliers removed by Cochran's and Grubb's tests), as a function of the intra-laboratory reproducibility standard deviation (s_{RW}) derived using ANOVA. The first chart (Figure 4.24a) clearly shows the effect of the wide uncertainty associated with arginine observed in the previous sections. The second chart (Figure 4.24b), presents the same data but with a re-adjusted y-axis scale for better resolution. The subsequent pair of charts (Figure 4.25a and b) again display the same data but this time as a function of the relative standard deviation (RSD_{RW}%). Note that the std. sol. 0.167dH20 has been used here due to the absence of 0.167d data from Chem.

Figure 4.23: Effect on Confidence Intervals with changing sample size.

Figure 4.23a: Alanine D/L values, 0.167d std sol

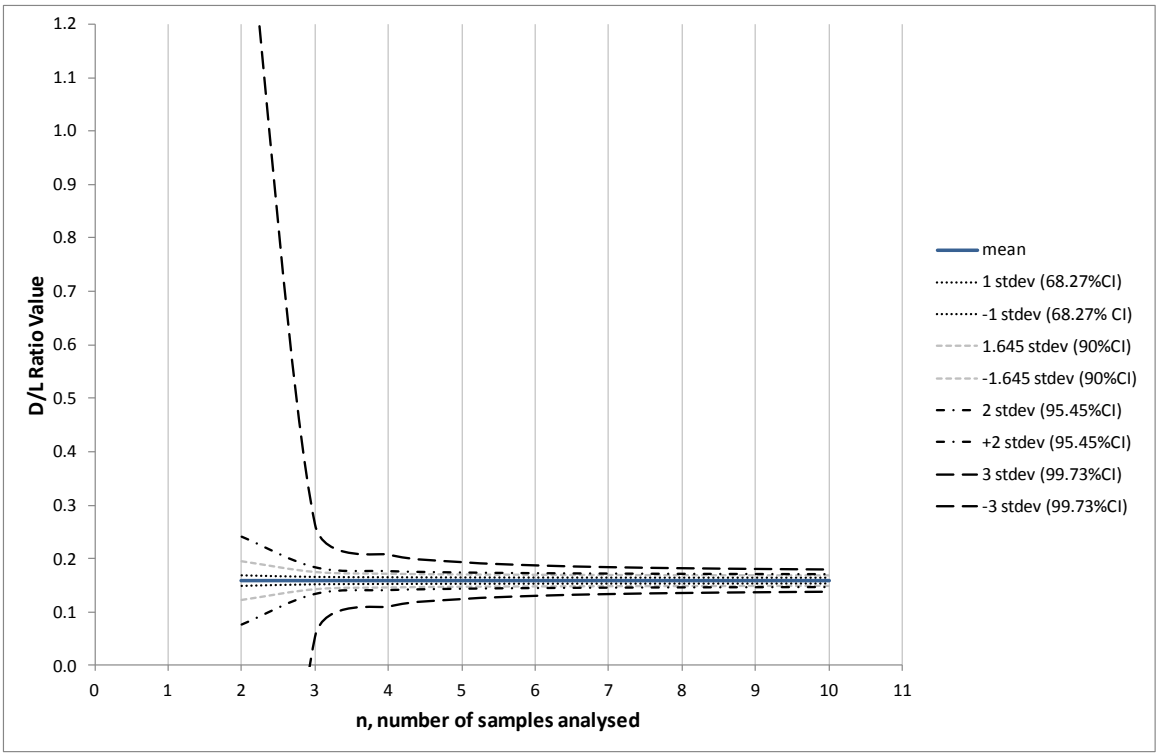


Figure 4.23b: Alanine D/L values, 0.5d std sol.

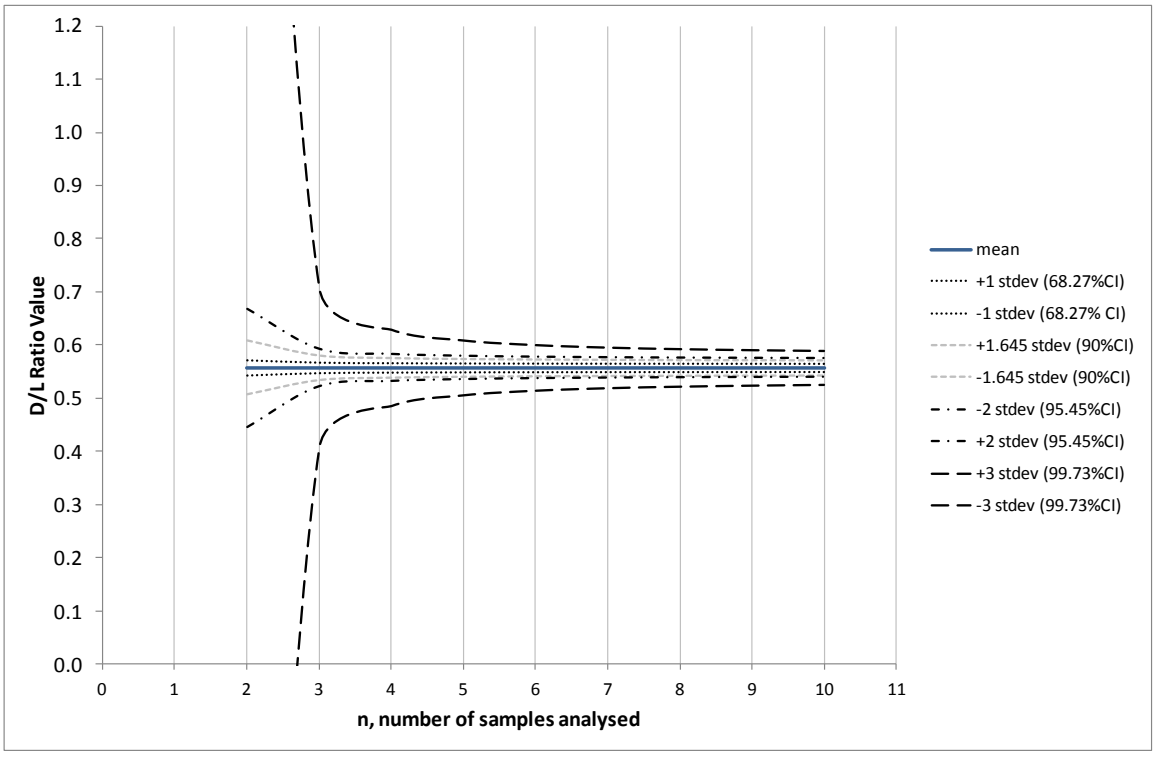
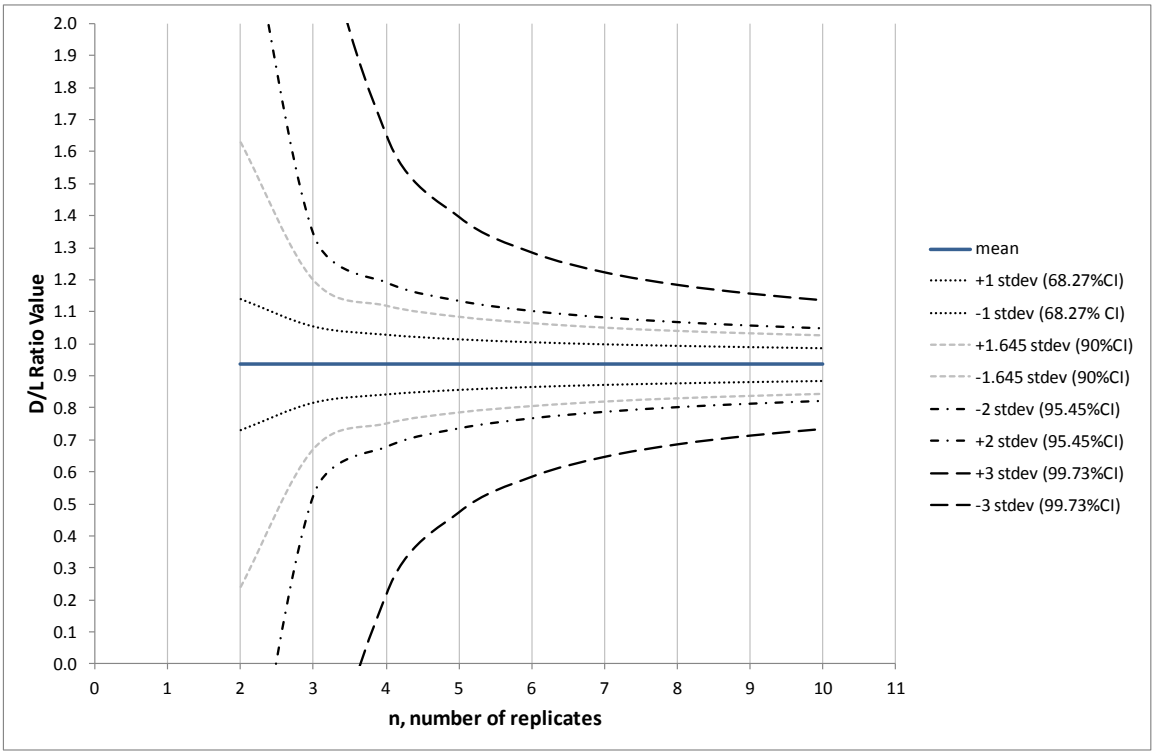


Figure 4.23c: Alanine D/L values, 0.91d std sol.



Establishing relationships between standard deviation relative to the analyte quantity, permits uncertainty estimates for unknown samples to be given with confidence during routine analysis. However this assumes the test samples are of the same or at least similar matrix composition. In this example, trendlines are only appropriate for determining precision estimates for other samples of standard solutions. However, given a sufficient range of D/L values, similar relationships could be determined for biomineral matrices, either through single laboratory validation type precision analyses or an inter-laboratory collaborative trial. Were this not possible, an analyst would otherwise have to rely on successive preparations and measurements, on possibly limited material, and would expect to achieve a much higher expanded uncertainty estimate due to the small sample size.

Figure 4.24: Reproducibility Standard Deviations (s_{RW}) for amino acid D/L values in Standard Solutions (0.167dH2O, 0.5d & 0.91d)

Figure 4.24a: Normal y-axis scale, showing all amino acids

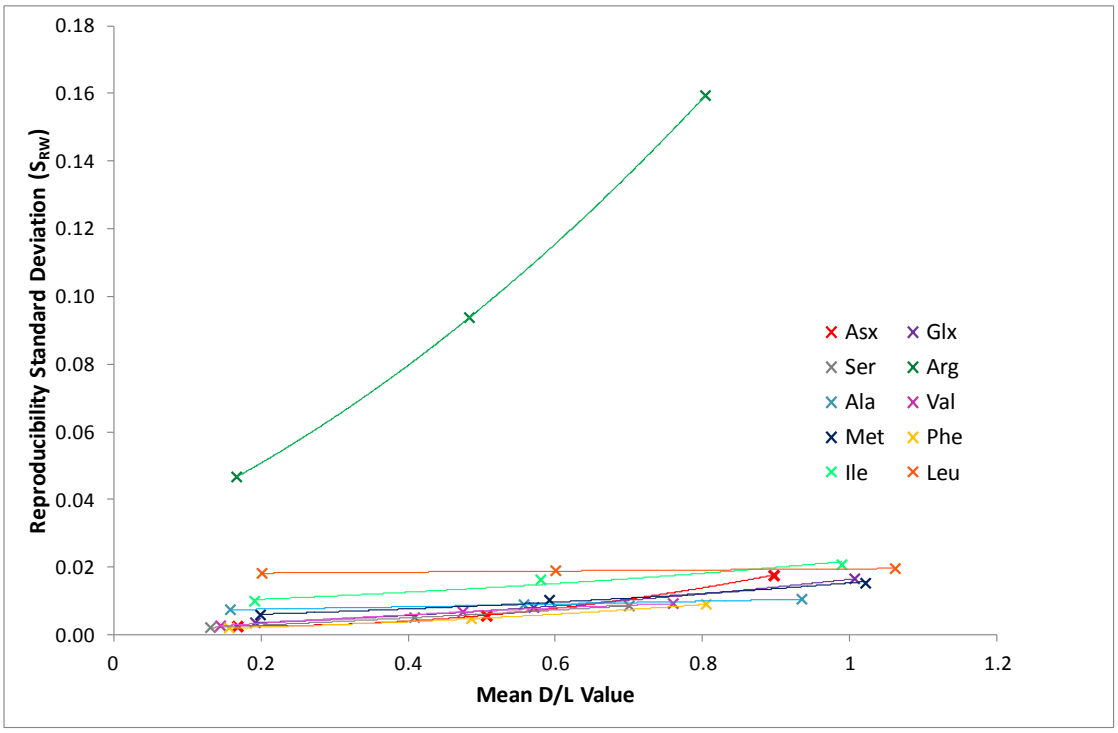


Figure 4.24b: Expanded y-axis scale, arginine data removed

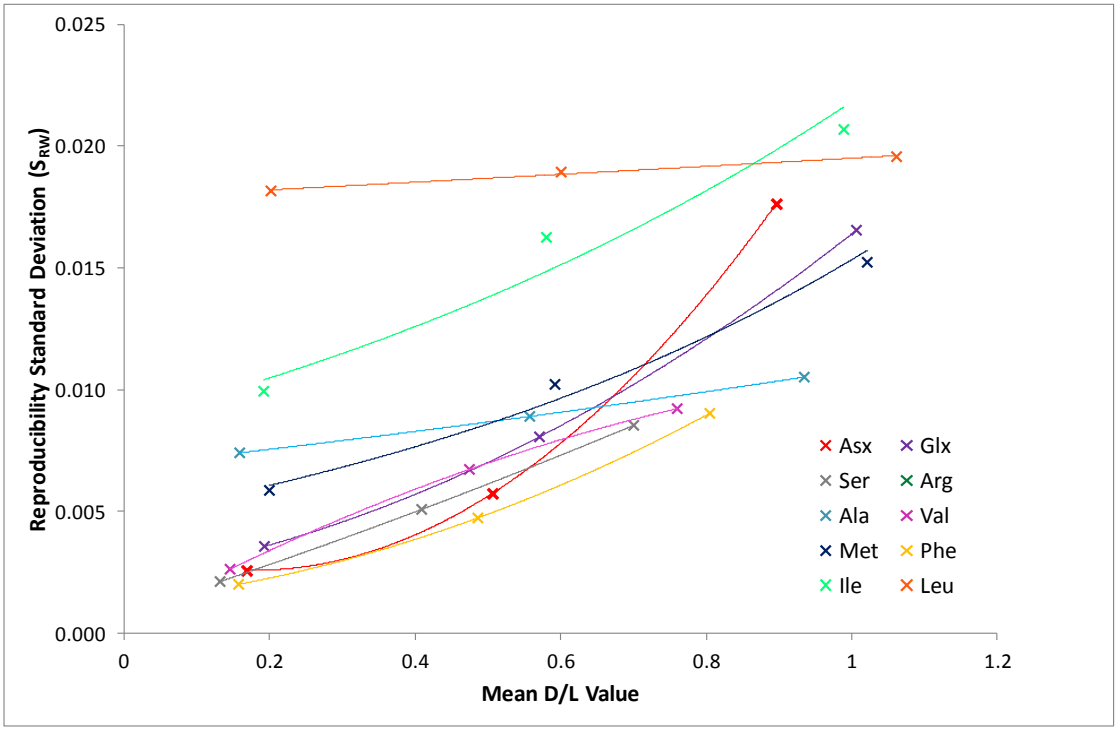


Figure 4.25: Relative Reproducibility Standard Deviations (RSD_{RW} %) for amino acid D/L values in Standard Solutions (0.167dH2O, 0.5d & 0.91d)

Figure 4.25: Normal y-axis scale, showing all amino acids

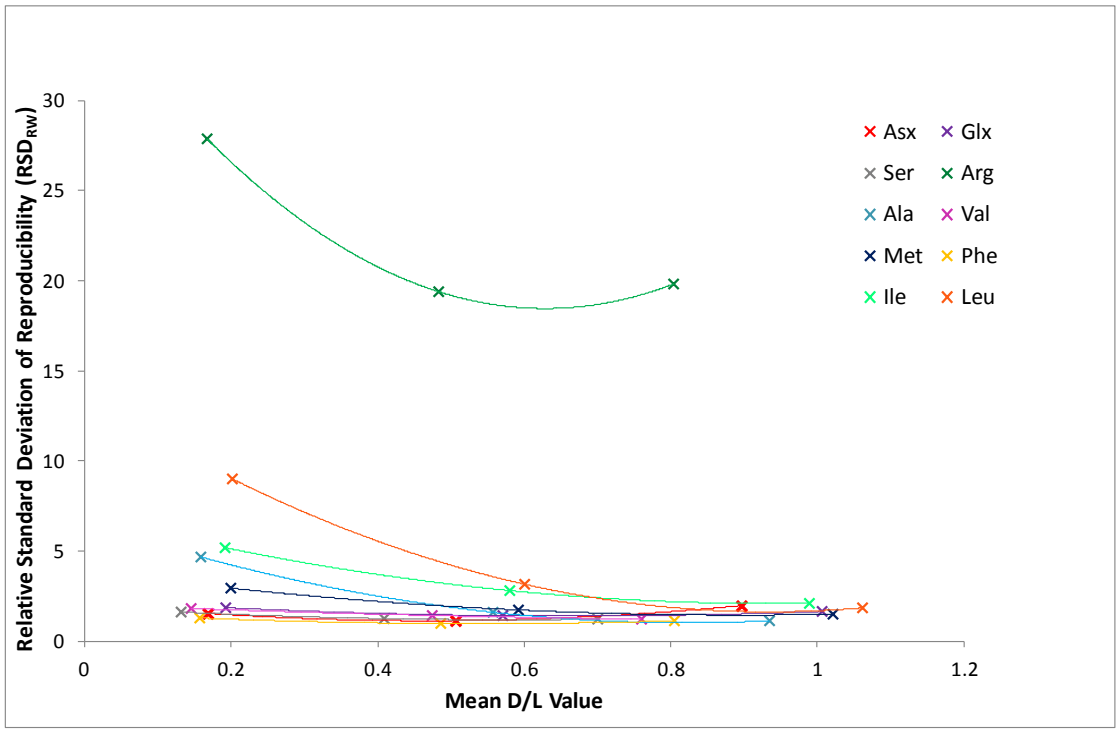
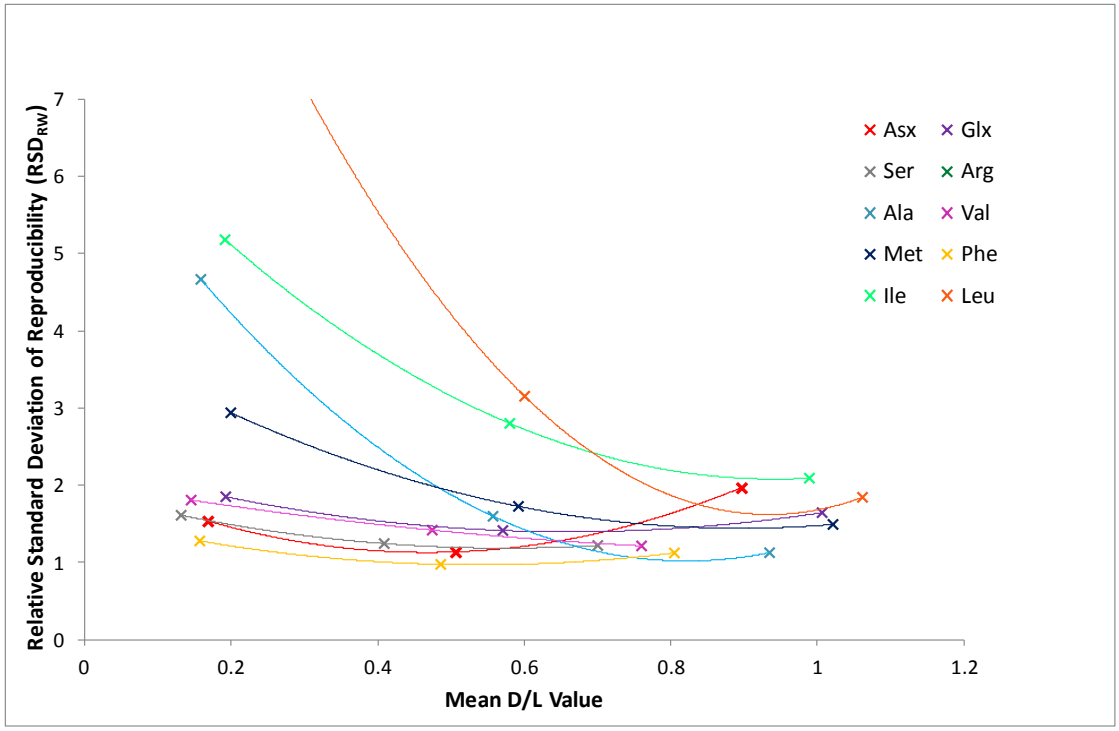


Figure 4.25b: Expanded y-axis scale, arginine data removed



4.3.6 Confidence Intervals

Data used in the evaluation of precision estimates have been derived from a large data set, therefore it can be reasonably assumed that data approximate to normality and there are no sample size effects. Using these relationships, it is possible to calculate 2 and 3 standard deviation confidence intervals (CIs), using the RSD_{RW} as the uncertainty equivalent to 1 standard deviation, and simply multiplying these values by 2 or 3 respectively. These values have been plotted against the expected D/L values present in the standard solutions, Figure 4.26-Figure 4.28. Each amino acid is represented by two charts, (a) and (b), where (a) gives the CIs using standard deviation values, s_{RW}, and (b), presents CIs using the relative standard deviations, expressed as a percentage, RSD_R%. For comparability, y-axis scales are equivalent between amino acids except for arginine, as indicated.

Results demonstrate the variation in precision for different amino acids in standard solutions. They also represent the minimum variability that can be expected from the analysis, since they exclude method preparation and extraction stages and samples have been derived from a homogeneous solution. Thus, uncertainty estimates only reflect the variability generated by a single laboratory due to solution preparation and instrumental fluctuation.

So far, the retrospective evaluation of uncertainty has only covered that of amino acid D/L values in standard solutions. Standard solutions are an ideal starting point since the individual L and D isomers will be as homogeneously distributed throughout the solutions as possible, giving the least possible variability in D/L values, free from matrix interference and extraction influences. However, the measurement procedure determines isomer concentrations in biomineral matrices, therefore control of the measurement process should ideally be at this level too. In the absence of an additional reference material, concentrations in standard solutions are assumed constant. Therefore assessing **concentration** uncertainty in these standard solutions is not possible, nor applicable to the target matrices.

However, limited biomineral data is available, derived from an early inter-laboratory comparison (Wehmiller, 1984) using mollusc shell inter-laboratory comparison materials (ILC) and also from an inter-laboratory proficiency test (PT), carried out as part of this research (see Chapter 5). This data will now be considered in the next section.

Figure 4.26: Confidence intervals derived from standard solutions; Asx D/L values

Figure 4.26a: Using the Standard Deviation of Reproducibility, s_R

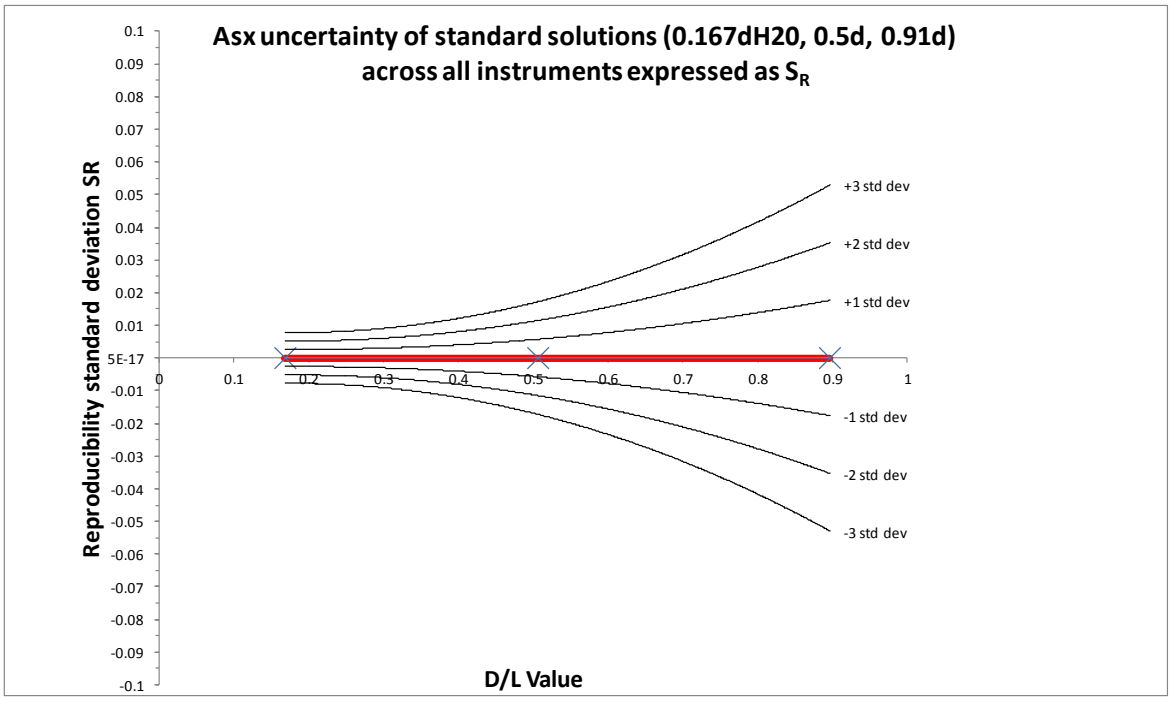


Figure 4.26b: Using the Relative Standard Deviation of Reproducibility, RSD_R

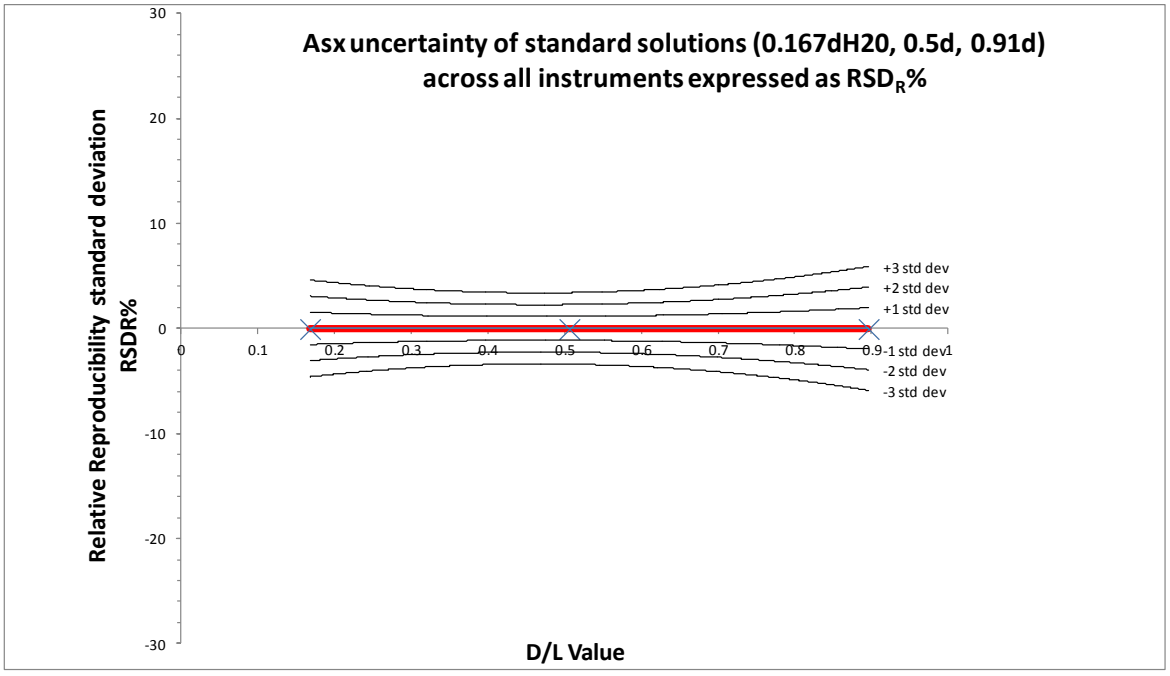


Figure 4.27: Confidence intervals derived from standard solutions; Val D/L values

Figure 4.27a: the Standard Deviation of Reproducibility, s_R

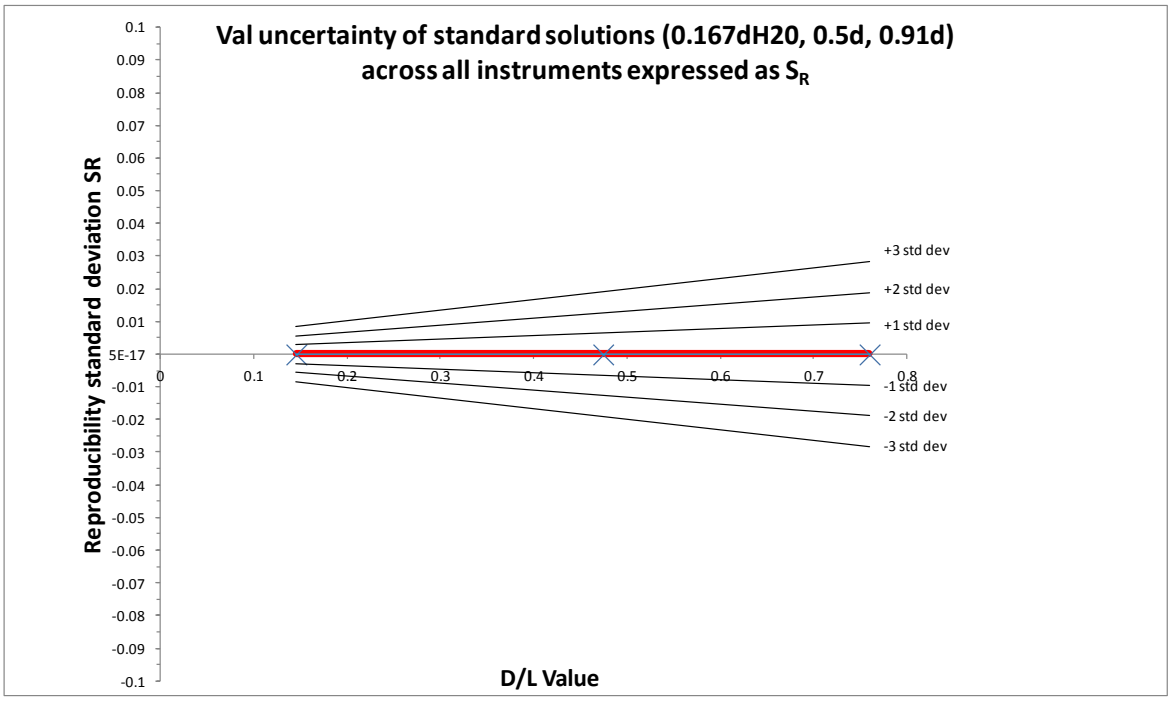


Figure 4.27b: the Relative Standard Deviation of Reproducibility, RSD_R

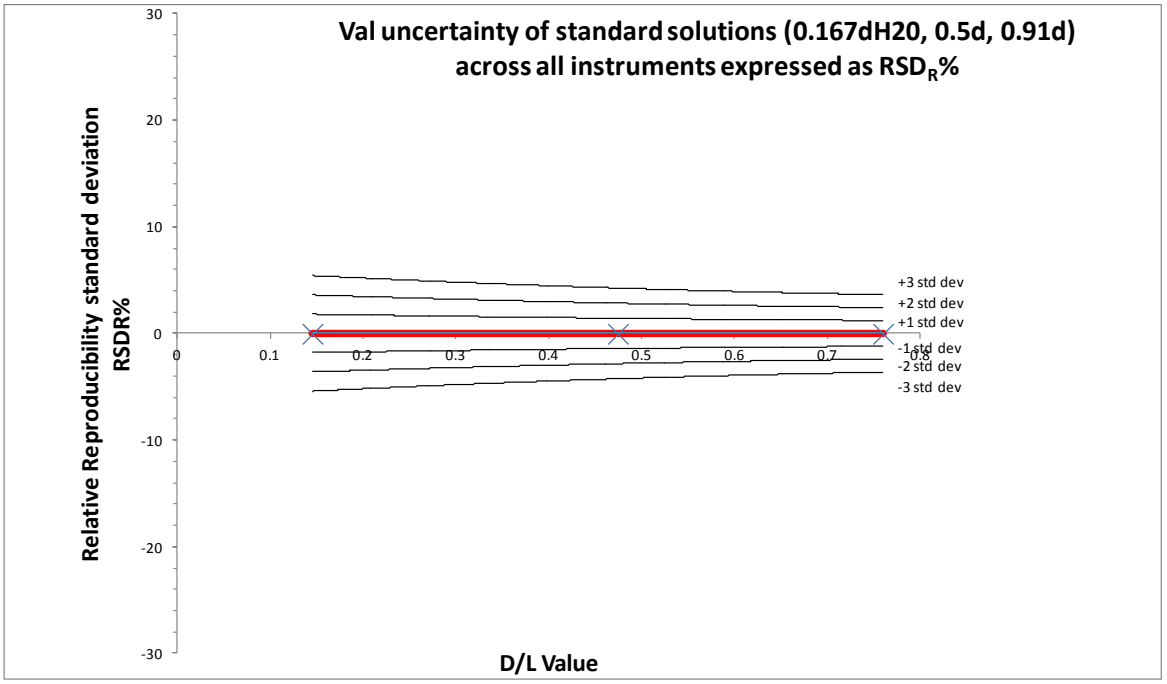


Figure 4.28: Confidence intervals derived from standard solutions; D-Aile/L-Ile values

Figure 4.28a: the Standard Deviation of Reproducibility, s_R

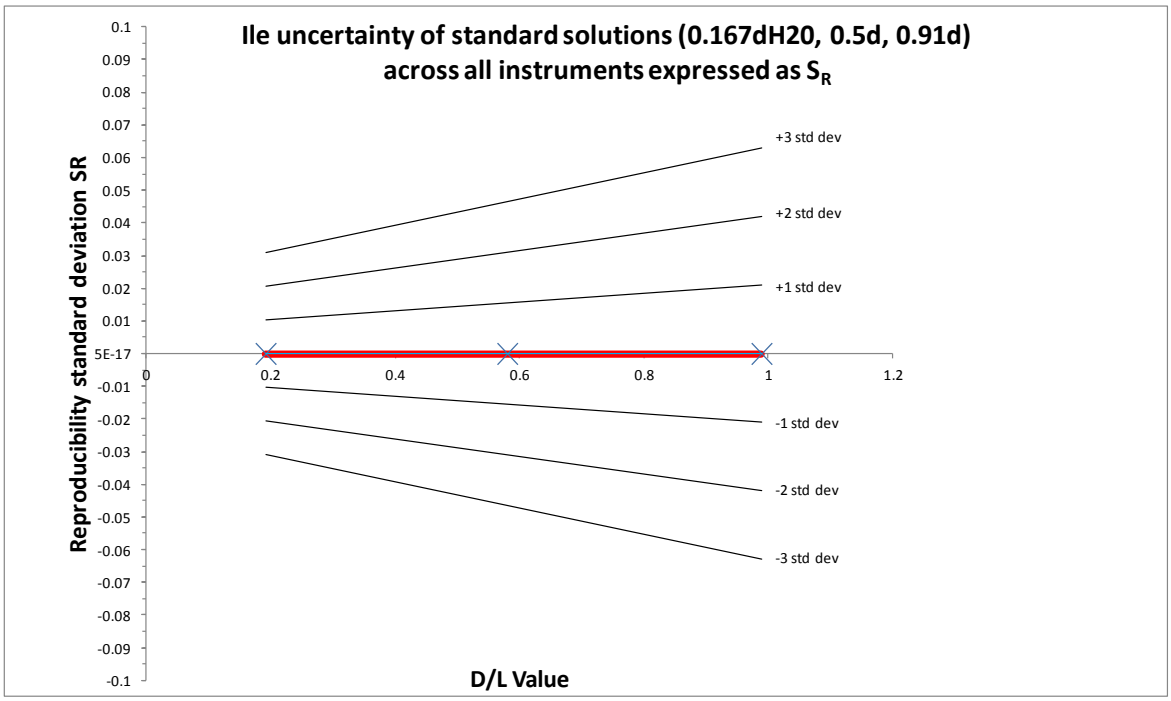
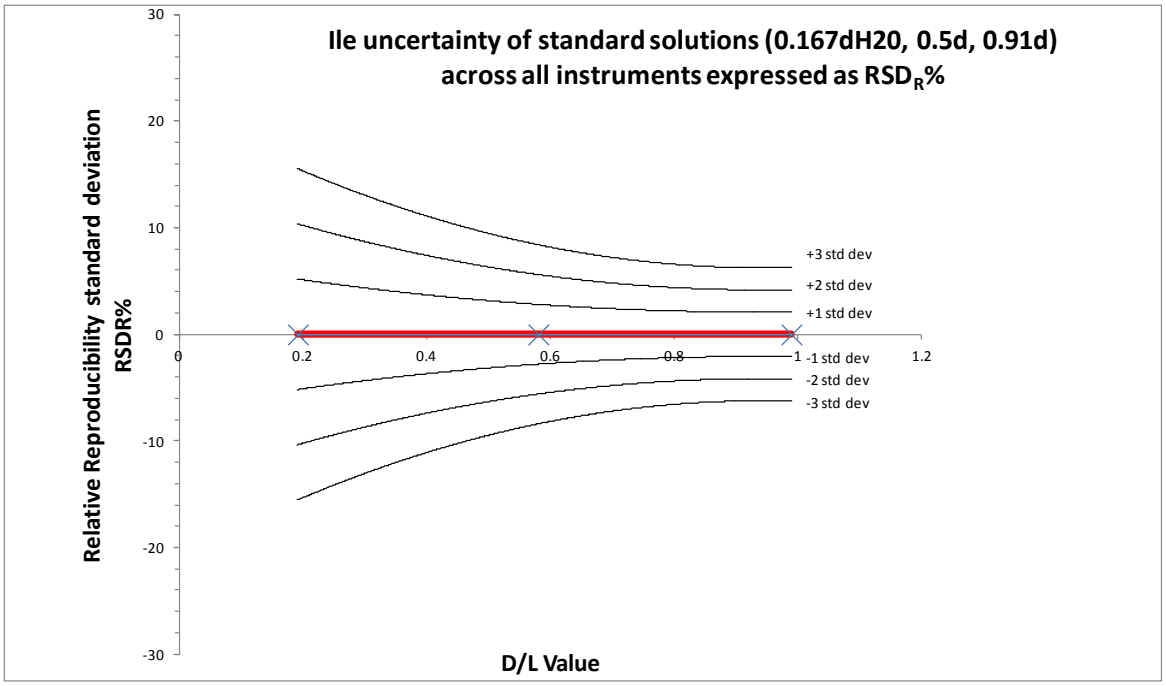


Figure 4.28b: the Relative Standard Deviation of Reproducibility, RSD_R



4.4 Precision Evaluation by ANOVA; Biomineral Matrices

4.4.1 Mollusc shell, ILC-A, B and C materials.

As a result of previous inter-laboratory comparisons (Wehmiller 1984, 2010), data from the analysis of three test materials were available. Test materials were each prepared from a bulk of powdered shell, which for the purpose of the inter-laboratory studies, were considered homogeneous. The inter-laboratory comparison materials A, B and C, (ILC-A, B and C), where originally prepared for Wehmiller's study in 1984. Bulk collections of different aged Pleistocene mollusc shells were each ground to produce quantities of powders from which individual vials of material were measured and given to participating laboratories; ILC-A was prepared from *Saxidomus* shells, and both ILC-B and ILC-C from *Mercenaria* (both Heterodont molluscs of the Family Veneridae).

Table 4.8 shows means and standard deviations for total hydrolysable amino acid (THAA) D/L values in the ILC materials whilst Table 4.9 shows the same data but for individual amino acid L and D isomer concentrations. Large quantities of the bulk material were originally prepared to act as quality control materials for laboratories. However, because the method applied at York incorporated an initial bleaching stage, it was decided not to use the materials routinely as D/L values would not be comparable to those from elsewhere (Penkman, pers.comms). Furthermore, due to the small data sets available for each ILC and, for the most part, an absence of replicate values, it was not possible to perform an evaluation by ANOVA. In this instance, data have been evaluated using a mean and standard deviation of individual sample values (denoted as *p* in Table 4.8, rather than the usual *n* used to denote replicates of an individual sample), or where replicates were reported, only the first replicate value was used.

Each of the *p* samples were analysed on separate occasions, in separate runs. Therefore the standard deviations and RSD% values reflect the between-sample / between-run precision for the whole method on mollusc shell matrix, but do not incorporate a repeatability element. The D/L data in Table 4.8 are then best compared with the s_L and RSD_L values from Table 4.6. The difference in precision for D/L values between mollusc shell and standard solution can clearly be seen. Amino acids in solution are fundamentally free from other matrix constituents and interferences and have not undergone aggressive preparation and extraction stages. Therefore precision estimates in standard solutions only represent the instrumental component of uncertainty. Whilst the instrumental component

will also be present in the uncertainty of the shell materials, other substantial and unaccounted for effects will also be reflected in the final precision estimate of biomineral amino acids.

For this reason, precision estimates derived from standard solution analysis should not be used as an estimate of precision in a solid matrix.

From the D/L values in Table 4.8, it would appear that the youngest, least racemised material was ILC-A, followed by ILC-B with the oldest material being ILC-C. This can also be seen in the concentration data (Table 4.9), with concentrations in L isomers generally decreasing and concentrations of D-isomers increasing with time. Exceptions to this can also be seen such as D-Asx, D-Ser and D-Arg, where additional degradation processes interact.

From the figures of confidence intervals for standard solutions (Figure 4.26 –Figure 4.28), the (b) charts show that RSD% values have a tendency to be wide at low D/L values and narrow as the D/L value approaches its fully racemic value of 1, i.e., with increasing age. Evaluation of data from Table 4.8, doesn't appear to follow this pattern, since the RSD% values might be expected to be widest in the youngest samples, (ILC-A), which isn't reflected in the data. However, it can be seen that generally, RSD% values for ILC-B are larger than those of ILC-C (the oldest material), as expected. These differences in ILC-A may be due to differences in genus between *Saxidomus* (ILC-A) and *Mercenaria* (ILC-B and ILC-C) and specific differences in biomineral protein composition, folding and interaction with the biomineral crystalline structure, or age effects (Collins and Riley, 2000). "*The Venerids have a complex ultrastructure with an outer prismatic layer, underlain by cross-lamella, then homogeneous and complex layers*" (Collins, pers.coms.) It may be that heterogeneity of the inter-crystalline proteins may be an issue in younger shells as the biomineral develops. Differences in D/L values between species has been previously reported (Penkman, 2005; Penkman *et al.*, 2008) and believed to be due to the variations in the ordering and binding of individual amino acid residues in the protein affecting their rates of racemisation. After hydrolysis, matrix molecules will remain in solution and complex interactions between matrix constituents and amino acids will continue to affect the availability and detection of individual isomers, affecting precision estimates. Such interactions could also affect the recovery of individual isomers and affect their accurate quantification. These effects could also contribute to differences in observed D/L values.

Whilst it is not possible to determine the individual contributions to analyte loss as a result of preparation, extraction, analysis and matrix effects, significant, unrecoverable losses

are a major concern for analysts. If significant and left uncorrected, this can result in a substantial systematic error in the final measurement result. The effect of bias will be looked at in more detail later in the chapter.

Table 4.8: Means and standard deviations for D/L values in ILC materials

Amino Acid D/L Value	Inter-Laboratory Calibration Standard											
	ILC-A				ILC-B				ILC-C			
	mean	p	std dev	RSD%	mean	p	std dev	RSD%	mean	p	std dev	RSD%
Asx D/L	0.39	5	0.008	2.16	0.70	6	0.094	13.41	0.88	5	0.041	4.70
Glx D/L	0.21	5	0.010	4.93	0.53	6	0.182	34.66	0.87	5	0.094	10.73
Ser D/L	0.50	5	0.089	17.62	0.43	6	0.171	39.52				
Arg D/L	0.65	4	0.213	32.93	0.66	3	0.314	47.90	3.04	3	1.463	48.05
Ala D/L	0.36	5	0.035	9.70	0.72	6	0.133	18.52	0.88	5	0.029	3.27
Val D/L	0.18	5	0.020	11.24	0.41	6	0.064	15.73	0.83	5	0.053	6.31
Phe D/L	0.25	5	0.029	11.39	0.55	6	0.122	22.47	0.77	5	0.109	14.22
Leu D/L	0.21	3	0.027	12.82	0.48	3	0.101	21.17	0.78	3	0.056	7.16
D-Aile/L-Ile	0.26	5	0.14	52.33	0.63	5	0.20	31.80	1.33	5	0.42	31.47

p = number of independent samples

Table 4.9: Means and standard deviations for concentrations in ILC materials

Amino Acid isomer conc.	Inter-Laboratory Calibration Standard											
	ILC-A				ILC-B				ILC-C			
pmol/mg	mean	p	std dev	RSD%	mean	p	std dev	RSD%	mean	p	std dev	RSD%
L-Asx	905	5	188.87	20.87	370	6	67.99	18.36	182	5	74.80	41.17
D-Asx	353	5	72.94	20.67	258	6	43.73	16.98	158	5	60.17	38.03
L-Glx	352	5	78.63	22.32	213	6	111.87	52.53	154	5	85.80	55.54
D-Glu	73	5	15.73	21.64	99	6	32.98	33.35	132	5	69.07	52.17
L-Ser	99	5	22.94	23.24	60	6	63.70	106.52	16	5	16.76	106.31
D-Ser	49	5	10.53	21.51	17	6	7.98	45.92	-	-	-	-
L-Arg	76	5	61.27	80.67	66	5	71.18	108.15	11	5	6.09	57.12
D-Arg	52	4	32.10	61.75	56	3	13.88	24.81	35	3	7.15	20.38
L-Ala	253	5	53.58	21.15	210	6	99.48	47.37	187	5	102.62	54.99
D-Ala	91	5	22.15	24.29	141	6	46.86	33.19	164	5	90.03	54.95
L-Val	181	5	59.08	32.59	134	6	93.13	69.38	115	5	69.44	60.39
D-Val	32	5	9.81	30.56	50	6	26.14	51.91	97	5	60.87	62.76
L-Phe	119	5	39.68	33.22	86	6	78.28	91.35	58	5	45.72	78.32
D-Phe	30	5	8.48	28.42	40	6	25.19	63.27	47	5	40.23	84.76
L-Leu	119	5	62.89	52.87	101	6	109.83	108.39	83	5	74.35	89.97
D-Leu	33	5	9.97	30.10	81	5	31.19	38.40	97	3	51.84	53.50
L-Ile	95	5	49.96	52.79	84	6	90.78	108.03	58	5	49.36	84.71
D-Aile	20	5	5.10	24.97	43	6	32.71	75.98	68	5	55.56	81.33

p = number of independent samples

4.4.2 Precision estimates from Proficiency Test (PT) data

Whilst the ILC data does not provide repeatability estimates, other biomineral data is available which does.

As part of this research, an inter-laboratory proficiency test was designed and carried out (see Chapter 5 and Chpt 5: Appendix 1). During the preparation of the test materials used in the study, samples were analysed and tested for homogeneity under repeatability conditions. That is, 10 individual vials from the bulk of measured samples, were taken and each split to give 2 sub-samples each, and a total of 20. All samples were run in random order in the same analytical run. In order to ensure analytical conditions remained as constant as possible, including the same batch of buffer and elution solutions were used, all samples had to be run within a maximum of three days. Due to the time required to analyse each injection, the most samples that can be analysed in a single day is 12. Therefore the maximum number of individual measurements in three days would be 36. In order for 20 individual samples to be run, with blanks and standard solutions as carried out routinely, this meant that only single measurements could be made for each sample.

Nonetheless, evaluation of the data by ANOVA provides precision estimates between pairs of sub-samples, i.e. within-sample repeatability, $s_{r(s\text{-}w)}$, and also between-sample repeatability, $s_{r(s\text{-}b)}$,. When combined in the same way as s_{RW} was determined previously, an overall estimate of repeatability precision can be derived for the whole method as applied to each specific matrix, at the relevant concentration / D/L value / age.

Further, because various members of the York BioArCh team were kind enough to carry out several sets of analysis, on different days over several months, using different individual PT samples (from the same bulk material), on different instruments, estimates of laboratory intermediate precision are now possible.

BioArCh team members performed the AAR measurement procedure by preparing single extracts from each biomineral sample and carrying out duplicate instrumental determinations. This provided two D/L values from a single run, but results were not independent. Precision estimates of the duplicate results therefore only represent injection or instrumental repeatability, $s_{r(i)}$, and not true sample repeatability. However, because the analysis of the individual samples incorporates the most variation in measurement conditions possible, (analyst, day, instrument), the between-sample precision estimate is equivalent to s_L in the evaluation of reproducibility (see section 4.3).

Because the data from both homogeneity assessment and proficiency test results are derived from the same homogeneous starting material, when taken together, these precision estimates provide us with a full-house! Table 4.10 shows the source of the various precision estimates and the abbreviations used to denote each one. The last column in Table 4.10 provides column references used in Table 4.11 - Table 4.19 for the precision estimate data, as determined for proficiency test materials.

Whilst six test materials were provided for the proficiency study, data from only three of them have been evaluated here; a standard solution, opercula and bleached mollusc shell (A).

Table 4.10: Key to Precision estimates derived from PT samples (Tables 4.11 – 4.19)

source	Precision component	Symbol	column
Homogeneity data	Within-sample repeatability	$s_{r(s\text{-}w)}$	A, B
Homogeneity data	Between-sample repeatability	$s_{r(s\text{-}b)}$	C, D
Homogeneity data	Sample repeatability	$s_r = \sqrt{(s_{r(s\text{-}w)}^2 + s_{r(s\text{-}b)}^2)}$	E, F
PT results	Injection repeatability	$s_{r(i)}$	G, H
PT results	Between-sample/run precision	s_L	I, J
PT results	Lab/method intermediate	$s_{RW(i)} = \sqrt{(s_{r(i)}^2 + s_L^2)}$	K, L
Homogeneity data + PT results	Intermediate reproducibility	$s_{RW} = \sqrt{(s_r^2 + s_L^2)}$	M, N

The standard solution was the same 0.5d evaluated earlier in this chapter, 20 μL sub-sampled and evaporated to dryness. Participants were required to rehydrate individual samples with 20 μL of rehydration fluid prior to analysis. Opercula test material was prepared from a 2 g bulk of individual Pleistocene opercula, taken from sediment collected at Funtham's Lane, Peterborough, UK, and the mollusc shell was the same ILC-A material described earlier, but bleached and air dried prior to measuring out into individual 20 mg test materials.

An overview of test materials, homogeneity evaluation and performance evaluation from an inter-laboratory perspective are given in the next chapter. Anonymous copies of all reports that were sent to participants can be found at www.neaar.co.uk, but are also included as separate Appendices.

However, precision data from the intra-laboratory perspective is given in the following tables. Table 4.11, Table 4.12 and Table 4.13, relate to the standard solution test

materials and show precision estimates derived from homogeneity data, PT results and overall intermediate reproducibility, respectively. Whilst these are not a biomineral matrix, D/L data can be compared to estimates derived earlier in the chapter (Table 4.6) and they also provide additional concentration data which isn't evaluated routinely from standard solutions.

Table 4.14, Table 4.15 and Table 4.16, are arranged similarly but relate to the opercula test materials. Whilst the method used at York can in principle be applied to many matrices, a large proportion of previous work has been carried out on Quaternary opercula. This data thus provides indicative values as might be expected in routine use for a mid-range D/L value material.

Finally, Table 4.17, Table 4.18 and Table 4.19, give data derived from bleached mollusc shell (A) test materials, previously referenced as ILC-A. This data has been included to provide a comparison with previous measurements of ILC materials shown in Table 4.8 and Table 4.9.

Any red text appearing in the tables indicates that a within-sample repeatability value has been used in place of a between-sample precision estimate. Occasionally, but particularly when dealing with smaller data sets, the ANOVA is unable to determine the between-sample variance, i.e., $s_L = \sqrt{(between\ MS - within\ MS)/n}$ as the within mean square is larger than the between mean square, and results in a negative value. This would suggest that the between-sample precision is in fact better than the within-sample precision, which wouldn't normally be expected. Therefore on these few occasions, ISO 5725 recommends s_L component be reduced to 0 and the $s_{RW} = s_r$.

Green text is separate from the evaluation of precision estimates but uses these values to determine repeatability limits. Data have been included in these tables to avoid repetition and are discussed in more detail below.

It should be noted that the ILC material used in the PT study was milled to a finer particle size than the originally provided material and may reduce observed imprecision. Also the opercula test materials were produced from a finely powdered bulk of many individual opercula. This may add additional uncertainty to the precision estimate than would be normally be expected for an individual operculum.

When comparing relative standard deviations (RSD%) between materials with different mean values, (that is D/L values or concentrations) care should be taken. Relative

percentage values are strongly influenced by the mean value, as seen from Figure 4.26 - Figure 4.28 earlier in the chapter, low values acting to elevate the RSD%. As such it is not always appropriate to directly compare RSD% values unless materials have equivalent mean values.

In Table 4.13, Table 4.16 and Table 4.19, RSD_{RW}% values are finally derived from the homogeneity data RSD_r% and the submitted PT data RSD_L%

4.4.2.1 Observations on D/L value precision estimates

1. For D/L values, the majority of the random error observed in the homogeneity data is generally accounted for by the within-sample repeatability precision estimate ($s_{r(s-w)}$) (columns A & B).
2. D/L value repeatability precision (s_r) (columns E & F) is smallest in standard solution (generally <1%), and larger in biominerals, reflecting the additional variability due to matrix effects and method preparation/extraction stages.
3. D/L value repeatability precision (s_r) (columns E & F) vary between matrices and amino acids.
4. Injection repeatability values, $s_{r(i)}$ (columns G & H) are similar in size to the s_r values (columns E & F). Homogeneity repeatability values represent true repeatability based on the analysis of independent samples. Therefore it might be expected to see slightly larger RSD_r% values compared to injection precision for biominerals. Whilst this was evident in some cases, the effect was not always observed.
5. Although there are exceptions, generally D/L value between-sample precision, s_L, (columns I & J), determined under reproducibility conditions, are wider than the repeatability estimates as might be expected.
6. Generally, opercula D/L value between-sample precision, s_L, (columns I & J) are wider than those of standard solutions, and the mollusc shell s_L estimates are wider than those from the opercula.
7. Overall, the D/L value intra-laboratory reproducibility, RSD_{RW}% (columns M & N) for standard solutions (Table 4.13) gave the tightest RSD_{RW}% values, in all cases (except Arg; 6.5%), these were ≤1%.

8. For opercula, (Table 4.16) the lowest RSD_{RW}% precision estimate was for Glx, 1.22% with Asx and Ser also being below 2%. Ala, Phe, Val D/Ls together with D-Aile/L-Ile gave values between 3-8%, Leu D/L nearly 18% and Arg gave 31%.
9. For mollusc shell D/L values (Table 4.19), Asx gave the tightest RSD_{RW}% value (1.8%) with Glx and Phe both following with <4%. Ser, Ala and Val D/L RSD_{RW}% values ranged between 7-9%, Arg and Leu were both around 17-18% and D-Aile/L-Ile had the widest precision or 25%.

4.4.2.2 Observations on isomer concentration precision estimates

1. RSDr% values are often much larger for individual isomer concentration data compared to the equivalent amino acids' D/L value.
2. Isomer concentration precision estimates appear to be far more consistent within a specified matrix, with certain exceptions, such as L-Arg.
3. In standard solution, apart from L-Arg, all RSD_r% values , for all isomers ranged between 1.3-1.7%.
4. In opercula, isomer RSD_r% values generally range between 5.5-7.5%, peaking at 11.3% for D-Aile and 10.5% for D-Leu.
5. For mollusc shell RSDr% values appear slightly lower, ranging generally between 3.5-5.5%, with 5.9% for D-Aile, 8.1% for D-Leu and 8.6% for L-Ser and D-Ala.
6. In standard solution and opercula, injection precision (RSDr(i)%) (columns G & H) are often larger than the repeatability precision, RSD_r% (columns E & F), which would not be expected. This effect is not observed with mollusc shell to the same extent.
7. Although there are exceptions, generally, concentration between-sample precision, s_L, (columns I & J), determined under reproducibility conditions, are wider than the repeatability estimates as might be expected.
8. Between-sample precision, RSD_L, (columns I & J) estimates for opercula isomer concentrations were remarkably consistent, ranging between 11.5-16.9%, (Table 4.15), but more varied for mollusc shell, giving values generally between 4-16% (Table 4.18). See comment below regarding s_L values for standard solution concentration data (Table 4.12).

9. Overall, for opercula concentrations, the intra-laboratory reproducibility (Table 4.16) RSD_{RW}% values ranged between 13-16% with a couple of exceptions; (D-Aile; 10.8% and D-Leu; 10.6%, L-Phe; 23% and D-Arg; 34%).
10. For mollusc shell concentrations (Table 4.19), RSD_{RW}% values ranged between 6-10% with the following exceptions; D-Ala; 11.4%, D and L-Glx; 14-15%, D-Leu 17.5%, D-Aile 23% and D-Arg 29%.

4.4.2.3 Further comments

D/L value repeatability precision estimates s_r in columns E and F, (Table 4.11), are comparable with those from Table 4.6. Whilst for the most part, with the exception of Arg, Ile and Leu, precision estimates in both tables are less than 1%, the homogeneity data estimates (Table 4.11) are generally tighter than those given in Table 4.6.

It could also be argued that the data set used for Table 4.6 is substantially larger and therefore more reliable. The preparation of proficiency test samples and analysis of individual vials for homogeneity evaluation were done under controlled conditions. However, the control of analytical conditions for samples of standard solution run over time cannot be assured. For these reasons, **repeatability precision** estimates derived from the homogeneity data are considered the more reliable and represent genuine within-sample differences rather than within-injection.

s_r (Table 4.11) and s_L (Table 4.12) then combine to give the overall estimate of total intra-laboratory or intermediate reproducibility precision, s_{RW}. Once again, a comparison of s_{RW} values in Table 4.13 with those for standard solutions (Table 4.6) show similar differences as observed for the repeatability precision. Data from Table 4.6 show larger values, usually between 1-2%, whereas data in Table 4.13, suggest s_{RW} values, once again, of less than 1%. However, since the s_{RW} is intended to reflect all the potential sources of variability encountered in routine analysis over time, perhaps in this case the Table 4.6 values should be the ones to rely on since results from the PT are more of a snap shot than a reflection of long term trends.

Table 4.11: ANOVA Precision estimates derived from PT homogeneity data for amino acid D/L values and concentrations in standard solution.

Proficiency Test Homogeneity Data: Repeatability Precision (S_r)					A	B	C	D	E	F	repeatability limit (2r) (cols E & F)	
matrix; 0.5 std sol	amino acid	mean	n1	p1	Sr(s-w)	RSDr(s-w)%	Sr(s-b)	RSDr(s-b)%	Sr	RSDr%	absolute (for Sr)	as % (for RSDr%)
D/L Value	Asx D/L	0.50	2	8	0.001	0.11	0.000	0.02	0.001	0.11	0.002	0.32
	Glx D/L	0.56	2	10	0.002	0.31	-	-	0.002	0.31	0.005	0.88
	Ser D/L	0.41	2	10	0.001	0.24	0.001	0.15	0.001	0.28	0.003	0.79
	Arg D/L	0.36	2	10	0.021	5.64	-	-	0.021	5.64	0.058	15.95
	Ala D/L	0.47	2	9	0.001	0.14	0.000	0.10	0.001	0.17	0.002	0.49
	Val D/L	0.59	2	10	0.003	0.46	-	-	0.003	0.46	0.008	1.30
	Phe D/L	0.48	2	10	0.001	0.13	0.000	0.05	0.001	0.14	0.002	0.39
	D-Aile/L-Ile	0.56	2	10	0.004	0.68	0.000	0.07	0.004	0.69	0.011	1.94
	Leu D/L	0.59	2	10	0.001	0.23	0.001	0.12	0.002	0.26	0.004	0.74
concentration (pmol/µL))	L-Asx	34.2	2	10	0.116	0.34	0.546	1.59	0.558	1.63	1.578	4.61
	D-Asx	17.1	2	10	0.102	0.60	0.245	1.43	0.265	1.55	0.751	4.38
	L-Glx	31.4	2	10	0.127	0.40	0.508	1.62	0.523	1.67	1.480	4.71
	D-Glx	17.5	2	10	0.084	0.48	0.271	1.55	0.284	1.63	0.803	4.60
	L-Ser	31.6	2	10	0.109	0.35	0.465	1.47	0.477	1.51	1.350	4.27
	D-Ser	12.8	2	10	0.038	0.30	0.184	1.43	0.188	1.46	0.531	4.14
	L-Arg	38.5	2	10	2.396	6.23	-	-	2.396	6.23	6.777	17.61
	D-Arg	14.0	2	10	0.112	0.80	0.205	1.46	0.233	1.66	0.659	4.71
	L-Ala	32.3	2	10	0.068	0.21	0.461	1.43	0.466	1.44	1.318	4.08
	D-Ala	15.2	2	10	0.036	0.24	0.224	1.48	0.227	1.50	0.642	4.23
	L-Val	33.0	2	10	0.087	0.26	0.459	1.39	0.467	1.41	1.320	4.00
	D-Val	13.8	2	10	0.048	0.35	0.173	1.25	0.180	1.30	0.509	3.68
	L-Phe	34.7	2	9	0.065	0.19	0.499	1.44	0.503	1.45	1.423	4.11
	D-Phe	16.8	2	10	0.055	0.33	0.245	1.45	0.251	1.49	0.709	4.21
	L-Ile	35.0	2	9	0.090	0.26	0.535	1.53	0.542	1.55	1.533	4.38
	D-Aile	19.7	2	10	0.129	0.66	0.322	1.64	0.347	1.76	0.981	4.99
	L-Leu	37.0	2	10	0.111	0.30	0.547	1.48	0.558	1.51	1.579	4.27
	D-Leu	21.7	2	10	0.095	0.44	0.338	1.56	0.352	1.62	0.995	4.59

where, n1 = final number of replicates after removal of Cochran's and Grubb's outliers and p1 = final number of samples after removal of Cochran's and Grubb's outliers

where, s = standard deviation and RSD% = relative standartd deviation expressed as a %, both being expressions of precision

Suffix; r(s-w) = within-sample repeatability, r(s-b) = between-sample repeatability and r = repeatability of material, calculated as; $\sqrt{((Sr(s\text{-}w))^2+(Sr(s\text{-}b))^2)}$

Table 4.12: ANOVA Precision estimates derived from BioArCh PT result data for amino acid D/L values and concentrations in standard solution.

Proficiency Test York Results: Intermediate Precision, $S_{RW(i)}$					G	H	I	J	K	L	repeatability limit (2r) (cols G & H)	
matrix; 0.5 std sol	amino acid	mean	n1	p1	Sr(i)	RSDr(i)%	SL	RSDL%	SRW(i)	RSDRW(i) %	absolute (for Sr(i))	as % (for RSDr(i)%)
D/L Value	Asx D/L	0.50	2	5	0.001	0.12	0.001	0.24	0.001	0.27	0.002	0.35
	Glx D/L	0.55	2	6	0.001	0.18	0.004	0.65	0.004	0.67	0.003	0.50
	Ser D/L	0.40	2	4	0.001	0.14	0.003	0.71	0.003	0.73	0.002	0.41
	Arg D/L	0.36	2	6	0.014	3.87	0.012	3.22	0.018	5.04	0.040	10.94
	Ala D/L	0.47	2	6	0.003	0.67	0.002	0.33	0.003	0.75	0.009	1.90
	Val D/L	0.41	2	6	0.001	0.32	0.002	0.50	0.002	0.60	0.004	0.91
	Phe D/L	0.49	2	6	0.001	0.21	0.003	0.59	0.003	0.62	0.003	0.59
	D-Aile/L-Ile	0.56	2	6	0.003	0.45	0.001	0.21	0.003	0.50	0.007	1.29
	Leu D/L	0.59	2	5	0.003	0.54	0.006	0.97	0.007	1.11	0.009	1.52
concentration (pmol/µL))	L-Asx	90.4	2	5	1.720	1.90	40.080	44.35	40.117	44.39	4.865	5.38
	D-Asx	45.0	2	5	0.793	1.76	19.972	44.38	19.988	44.42	2.243	4.99
	L-Glx	88.1	2	5	1.766	2.00	38.960	44.23	39.000	44.27	4.994	5.67
	D-Glx	48.8	2	5	1.009	2.07	21.746	44.54	21.769	44.59	2.855	5.85
	L-Ser	92.2	2	5	2.119	2.30	40.861	44.32	40.916	44.38	5.993	6.50
	D-Ser	36.6	2	5	0.881	2.40	16.335	44.59	16.358	44.65	2.491	6.80
	L-Arg	114.6	2	6	11.438	9.98	45.591	39.77	47.004	41.00	32.351	28.22
	D-Arg	39.1	2	5	0.946	2.42	17.219	44.01	17.245	44.08	2.676	6.84
	L-Ala	92.5	2	5	2.284	2.47	41.106	44.43	41.169	44.50	6.459	6.98
	D-Ala	42.9	2	5	0.866	2.02	18.946	44.12	18.966	44.16	2.449	5.70
	L-Val	93.1	2	5	1.843	1.98	41.099	44.13	41.141	44.18	5.213	5.60
	D-Val	38.4	2	5	0.796	2.08	16.764	43.68	16.783	43.73	2.253	5.87
	L-Phe	94.8	2	5	1.489	1.57	41.936	44.23	41.963	44.25	4.212	4.44
	D-Phe	46.6	2	5	0.723	1.55	20.472	43.90	20.485	43.93	2.045	4.39
	L-Ile	98.8	2	5	2.024	2.05	43.440	43.95	43.487	44.00	5.726	5.79
	D-Aile	54.8	2	5	1.128	2.06	24.032	43.82	24.059	43.86	3.190	5.82
	L-Leu	104.5	2	5	2.109	2.02	45.848	43.88	45.897	43.93	5.965	5.71
	D-Leu	66.6	2	4	1.491	2.24	29.261	43.92	29.299	43.98	4.217	6.33

where, n1 = final number of replicates after removal of Cochran's and Grubb's outliers and p1 = final number of samples after removal of Cochran's and Grubb's outliers

where, s = standard deviation and RSD% = relative standartd deviation expressed as a %, both being expressions of precision

Suffix; r(i) = instrumental or injection repeatability, L = between-sample/run variance and RW(i)= within or intra-lab reproducibility calculated as; $\sqrt{((Sr(i))^2+SL^2)}$

Note; SRW(i) is not quite a complete estimate of intermediate precision as the repeatability component should consist of independent samples, not replicated injections

Table 4.13: Intermediate reproducibility precision estimates for amino acid D/L values and concentrations in standard solution.

Intra-Lab Reproducibility or Intermediate Precision, S_{RW}			E (or A)	F (or B)	I (or G)	J (or H)	M	N
matrix; 0.5 std sol	amino acid	mean	Sr	RSDr%	SL	RSDL%	SRW	RSDRW%
D/L Value	Asx D/L	0.50	0.001	0.11	0.001	0.24	0.001	0.26
	Glx D/L	0.55	0.002	0.31	0.004	0.65	0.004	0.72
	Ser D/L	0.40	0.001	0.28	0.003	0.71	0.003	0.76
	Arg D/L	0.36	0.021	5.64	0.012	3.22	0.024	6.50
	Ala D/L	0.47	0.001	0.17	0.002	0.33	0.002	0.37
	Val D/L	0.50	0.003	0.46	0.002	0.50	0.003	0.68
	Phe D/L	0.49	0.001	0.14	0.003	0.59	0.003	0.61
	D-Aile/L-Ile	0.56	0.004	0.69	0.001	0.21	0.004	0.72
	Leu D/L	0.59	0.002	0.26	0.006	0.97	0.006	1.01
concentration (pmol/µL))	L-Asx	62.3	0.558	1.63	40.080	44.35	40.084	64.33
	D-Asx	31.1	0.265	1.55	19.972	44.38	19.974	64.30
	L-Glx	59.8	0.523	1.67	38.960	44.23	38.963	65.21
	D-Glx	33.1	0.284	1.63	21.746	44.54	21.748	65.61
	L-Ser	61.9	0.477	1.51	40.861	44.32	40.864	66.00
	D-Ser	24.7	0.188	1.46	16.335	44.59	16.336	66.07
	L-Arg	76.6	2.396	6.23	45.591	39.77	45.654	59.63
	D-Arg	26.6	0.233	1.66	17.219	44.01	17.220	64.83
	L-Ala	62.4	0.466	1.44	41.106	44.43	41.109	65.87
	D-Ala	29.1	0.227	1.50	18.946	44.12	18.948	65.20
	L-Val	63.1	0.467	1.41	41.099	44.13	41.102	65.16
	D-Val	26.1	0.180	1.30	16.764	43.68	16.765	64.21
	L-Phe	64.7	0.503	1.45	41.936	44.23	41.939	64.78
	D-Phe	31.7	0.251	1.49	20.472	43.90	20.474	64.53
	L-Ile	66.9	0.542	1.55	43.440	43.95	43.443	64.93
	D-Aile	37.3	0.347	1.76	24.032	43.82	24.035	64.50
	L-Leu	70.7	0.558	1.51	45.848	43.88	45.852	64.82
	D-Leu	44.1	0.352	1.62	29.261	43.92	29.2634	66.29

where, s = standard deviation and RSD% = relative standartd deviation expressed as a %, both being expressions of precision

Suffix; r = repeatability of material, L = between-sample/run variance and RW= within or intra-lab reproducibility, calculated as; $\sqrt{(Sr^2+SL^2)}$

Table 4.14: ANOVA Precision estimates derived from PT homogeneity data for amino acid D/L values and concentrations in opercula.

Proficiency Test Homogeneity Data: Repeatability Precision (S_r)					A	B	C	D	E	F	repeatability limit (2r) (cols E & F)	
matrix; opercula	amino acid	mean	n1	p1	Sr(s-w)	RSDr(s-w)%	Sr(s-b)	RSDr(s-b)%	Sr	RSDr%	absolute (for Sr)	as % (for RSDr%)
D/L Value	Asx D/L	0.58	2	9	0.009	1.49	-	-	0.009	1.49	0.024	4.21
	Glx D/L	0.17	2	9	0.001	0.70	0.001	0.80	0.002	1.06	0.005	3.01
	Ser D/L	0.66	2	9	0.008	1.15	0.006	0.88	0.010	1.45	0.027	4.10
	Arg D/L	0.82	2	10	0.199	24.27	-	-	0.199	24.27	0.563	68.64
	Ala D/L	0.26	2	10	0.021	8.07	-	-	0.021	8.07	0.060	22.81
	Val D/L	0.13	2	9	0.003	1.91	-	-	0.003	1.91	0.007	5.41
	Phe D/L	0.30	2	9	0.006	2.03	0.004	1.22	0.007	2.37	0.020	6.71
	D-Aile/L-Ile	0.17	2	7	0.004	2.51	0.006	3.53	0.007	4.33	0.021	12.25
	Leu D/L	0.25	2	10	0.031	12.05	-	-	0.031	12.05	0.086	34.07
concentration (pmol/mg))	L-Asx	1522.8	2	10	114.189	7.50	-	-	114.189	7.50	322.975	21.21
	D-Asx	874.3	2	10	57.271	6.55	-	-	57.271	6.55	161.986	18.53
	L-Glx	1201.3	2	9	82.896	6.90	-	-	82.896	6.90	234.465	19.52
	D-Glx	203.2	2	10	15.508	7.63	-	-	15.508	7.63	43.865	21.59
	L-Ser	394.5	2	9	26.670	6.76	-	-	26.670	6.76	75.433	19.12
	D-Ser	261.5	2	10	17.696	6.77	-	-	17.696	6.77	50.052	19.14
	L-Arg	309.1	2	8	11.920	3.86	-	-	11.920	3.86	33.715	10.91
	D-Arg	259.7	2	10	56.629	21.81	-	-	56.629	21.81	160.172	61.68
	L-Ala	1220.7	2	10	98.509	8.07	-	-	98.509	8.07	278.627	22.83
	D-Ala	320.8	2	10	18.649	5.81	2.503	0.78	18.816	5.86	53.219	16.59
	L-Val	924.0	2	9	51.804	5.61	-	-	51.804	5.61	146.524	15.86
	D-Val	122.9	2	10	6.885	5.60	-	-	6.885	5.60	19.474	15.85
	L-Phe	299.9	2	9	17.667	5.89	-	-	17.667	5.89	49.969	16.66
	D-Phe	89.6	2	10	5.762	6.43	-	-	5.762	6.43	16.297	18.20
	L-Ile	361.6	2	9	19.820	5.48	-	-	19.820	5.48	56.060	15.50
	D-Aile	59.3	2	10	6.694	11.28	-	-	6.694	11.28	18.935	31.91
	L-Leu	688.2	2	9	42.269	6.14	-	-	42.269	6.14	119.556	17.37
	D-Leu	177.0	2	10	18.144	10.25	3.846	2.17	18.547	10.48	52.458	29.63

where, n1 = final number of replicates after removal of Cochran's and Grubb's outliers and p1 = final number of samples after removal of Cochran's and Grubb's outliers

where, s = standard deviation and RSD% = relative standartd deviation expressed as a %, both being expressions of precision

Suffix; r(s-w) = within-sample repeatability, r(s-b) = between-sample repeatability and r = repeatability of material, calculated as; $\sqrt{((Sr(s\text{-}w))^2+(Sr(s\text{-}b))^2)}$

Table 4.15: ANOVA Precision estimates derived from BioArCh PT result data for amino acid D/L values and concentrations in opercula.

Proficiency Test York Results: Intermediate Precision, S_{RW}'					G	H	I	J	K	L	repeatability limit (2r) (cols G & H)	
matrix; opercula	amino acid	mean	n1	p1	Sr(i)	RSDr(i)%	SL	RSDL%	SRW(i)	RSDRW(i) %	absolute (for Sr(i))	as % (for RSDr(i)%)
D/L Value	Asx D/L	0.57	2	6	0.002	0.42	0.003	0.45	0.004	0.62	0.007	1.17
	Glx D/L	0.17	2	7	0.000	0.27	0.001	0.59	0.001	0.65	0.001	0.77
	Ser D/L	0.66	2	7	0.009	1.40	0.006	0.87	0.011	1.64	0.026	3.95
	Arg D/L	0.74	2	7	0.162	21.98	0.148	20.12	0.219	29.80	0.458	62.17
	Ala D/L	0.26	2	7	0.007	2.65	0.003	1.21	0.007	2.91	0.019	7.49
	Val D/L	0.13	2	6	0.004	3.29	0.006	4.99	0.008	5.98	0.012	9.31
	Phe D/L	0.30	2	6	0.002	0.76	0.006	1.91	0.006	2.06	0.007	2.15
	D-Aile/L-Ile	0.24	2	7	0.041	17.49	0.011	4.80	0.043	18.13	0.116	49.46
	Leu D/L	0.28	2	6	0.017	6.15	0.036	13.04	0.040	14.41	0.048	17.39
concentration (pmol/mg))	L-Asx	1417.7	2	7	134.700	9.50	197.338	13.92	238.928	16.85	380.989	26.87
	D-Asx	813.7	2	7	73.541	9.04	114.904	14.12	136.423	16.76	208.006	25.56
	L-Glx	1191.7	2	7	109.343	9.18	140.929	11.83	178.373	14.97	309.268	25.95
	D-Glx	196.9	2	7	18.153	9.22	23.015	11.69	29.312	14.89	51.344	26.08
	L-Ser	396.2	2	7	33.752	8.52	47.556	12.00	58.316	14.72	95.466	24.09
	D-Ser	260.4	2	7	21.957	8.43	29.915	11.49	37.108	14.25	62.103	23.85
	L-Arg	294.0	2	7	20.714	7.04	37.144	12.63	42.529	14.46	58.589	19.93
	D-Arg	220.8	2	7	52.737	23.88	59.054	26.74	79.174	35.85	149.163	67.55
	L-Ala	1067.8	2	7	88.619	8.30	140.712	13.18	166.292	15.57	250.652	23.47
	D-Ala	275.0	2	7	16.371	5.95	40.234	14.63	43.437	15.80	46.303	16.84
	L-Val	745.6	2	7	69.744	9.35	125.820	16.88	143.857	19.29	197.265	26.46
	D-Val	97.1	2	7	4.637	4.77	15.556	16.01	16.232	16.71	13.116	13.50
	L-Phe	1417.7	2	7	134.700	9.50	197.338	13.92	238.928	16.85	380.989	26.87
	D-Phe	78.1	2	7	4.729	6.05	10.517	13.47	11.531	14.76	13.375	17.12
	L-Ile	302.2	2	7	26.401	8.74	48.219	15.96	54.973	18.19	74.672	24.71
	D-Aile	68.0	2	6	11.657	17.14	1.651	2.43	11.773	17.31	32.971	48.48
	L-Leu	781.9	2	7	64.846	8.29	113.564	14.52	130.774	16.73	183.413	23.46
	D-Leu	245.6	2	5	12.524	5.10	-	-	12.524	5.10	35.424	14.42

where, n1 = final number of replicates after removal of Cochran's and Grubb's outliers and p1 = final number of samples after removal of Cochran's and Grubb's outliers

where, s = standard deviation and RSD% = relative standartd deviation expressed as a %, both being expressions of precision

Suffix; r(i) = instrumental or injection repeatability, L = between-sample/run variance and RW(i)= within or intra-lab reproducibility calculated as; $\sqrt{((Sr(i))^2+(SL)^2)}$

Note; SRW(i) is not quite a complete estimate of intermediate precision as the repeatability component should consist of independent samples, not replicated injections

Table 4.16: Intermediate reproducibility precision estimates for amino acid D/L values and concentrations in opercula.

Intra-Lab Reproducibility or Intermediate Precision, S_{RW}			E (or A)	F (or B)	I (or G)	J (or H)	M	N
matrix; opercula	amino acid	mean	Sr	RSDr%	SL	RSDL%	SRW	RSDRW%
D/L Value	Asx D/L	0.58	0.009	1.49	0.003	0.45	0.009	1.56
	Glx D/L	0.17	0.002	1.06	0.001	0.59	0.002	1.22
	Ser D/L	0.66	0.010	1.45	0.006	0.87	0.011	1.69
	Arg D/L	0.78	0.199	24.27	0.148	20.12	0.248	31.88
	Ala D/L	0.26	0.021	8.07	0.003	1.21	0.021	8.25
	Val D/L	0.13	0.003	1.91	0.006	4.99	0.007	5.28
	Phe D/L	0.30	0.007	2.37	0.006	1.91	0.009	3.04
	D-Aile/L-Ile	0.20	0.007	4.33	0.011	4.80	0.013	6.65
	Leu D/L	0.27	0.031	12.05	0.036	13.04	0.047	17.84
concentration (pmol/mg))	L-Asx	1470.2	114.189	7.50	197.338	13.92	227.995	15.51
	D-Asx	844.0	57.271	6.55	114.904	14.12	128.386	15.21
	L-Glx	1196.5	82.896	6.90	140.929	11.83	163.502	13.66
	D-Glx	200.1	15.508	7.63	23.015	11.69	27.752	13.87
	L-Ser	395.4	26.670	6.76	47.556	12.00	54.524	13.79
	D-Ser	261.0	17.696	6.77	29.915	11.49	34.757	13.32
	L-Arg	301.6	11.920	3.86	37.144	12.63	39.010	12.94
	D-Arg	240.3	56.629	21.81	59.054	26.74	81.819	34.06
	L-Ala	1144.3	98.509	8.07	140.712	13.18	171.767	15.01
	D-Ala	297.9	18.816	5.86	40.234	14.63	44.416	14.91
	L-Val	834.8	51.804	5.61	125.820	16.88	136.068	16.30
	D-Val	110.0	6.885	5.60	15.556	16.01	17.011	15.46
	L-Phe	858.8	17.667	5.89	197.338	13.92	198.128	23.07
	D-Phe	83.8	5.762	6.43	10.517	13.47	11.992	14.31
	L-Ile	331.9	19.820	5.48	48.219	15.96	52.134	15.71
	D-Aile	63.7	6.694	11.28	1.651	2.43	6.895	10.83
	L-Leu	735.1	42.269	6.14	113.564	14.52	121.176	16.49
	D-Leu	211.3	18.547	10.48	12.524	5.10	22.379	10.59

where, s = standard deviation and RSD% = relative standartd deviation expressed as a %, both being expressions of precision

Suffix; r = repeatability of material, L = between-sample/run variance and RW= within or intra-lab reproducibility, calculated as; $\sqrt{(Sr^2+SL^2)}$

Table 4.17: ANOVA Precision estimates derived from PT homogeneity data for amino acid D/L values and concentrations in mollusc shell (A) (ILC-A).

Proficiency Test Homogeneity Data: Repeatability Precision (S_r)					A	B	C	D	E	F	repeatability limit (2r) (cols E & F)	
matrix; Mollusc A	amino acid	mean	n1	p1	Sr(s-w)	RSDr(s-w)%	Sr(s-b)	RSDr(s-b)%	Sr	RSDr%	absolute (for Sr)	as % (for RSDr%)
D/L Value	Asx D/L	0.42	2	8	0.004	0.88	-	-	0.004	0.88	0.011	2.50
	Glx D/L	0.23	2	9	0.009	3.72	-	-	0.009	3.72	0.024	10.52
	Ser D/L	0.53	2	8	0.038	7.24	-	-	0.038	7.24	0.108	20.47
	Arg D/L	0.83	2	9	0.096	11.49	0.034	4.06	0.101	12.19	0.287	34.48
	Ala D/L	0.45	2	8	0.026	5.86	-	-	0.026	5.86	0.075	16.57
	Val D/L	0.19	2	9	0.009	4.81	-	-	0.009	4.81	0.025	13.62
	Phe D/L	0.28	2	9	0.008	2.77	0.002	0.57	0.008	2.83	0.022	7.99
	D-Aile/L-Ile	0.25	2	9	0.012	4.63	0.007	2.75	0.014	5.38	0.039	15.23
	Leu D/L	0.33	2	9	0.030	9.11	0.016	4.94	0.034	10.36	0.097	29.31
concentration (pmol/mg))	L-Asx	712.1	2	8	24.62	3.46	16.28	2.29	29.52	4.15	83.50	11.73
	D-Asx	302.0	2	8	11.15	3.69	6.94	2.30	13.13	4.35	37.15	12.30
	L-Glx	277.6	2	8	9.91	3.57	-	-	9.91	3.57	28.04	10.10
	D-Glx	64.2	2	9	2.64	4.11	-	-	2.64	4.11	7.46	11.62
	L-Ser	72.2	2	7	6.19	8.57	-	-	6.19	8.57	17.51	24.23
	D-Ser	39.2	2	8	0.83	2.12	1.07	2.73	1.36	3.46	3.83	9.78
	L-Arg	90.6	2	9	4.58	5.05	2.02	2.23	5.01	5.52	14.16	15.63
	D-Arg	75.3	2	9	6.43	8.53	6.69	8.88	9.27	12.32	26.23	34.83
	L-Ala	214.3	2	9	7.10	3.31	5.37	2.51	8.90	4.15	25.17	11.74
	D-Ala	95.8	2	9	8.26	8.63	-	-	8.26	8.63	23.37	24.41
	L-Val	215.2	2	9	7.50	3.48	4.80	2.23	8.90	4.14	25.17	11.70
	D-Val	40.1	2	9	2.01	5.01	-	-	2.01	5.01	5.69	14.16
	L-Phe	121.7	2	8	4.49	3.69	1.06	0.87	4.62	3.79	13.06	10.73
	D-Phe	33.9	2	8	1.13	3.32	0.10	0.29	1.13	3.34	3.20	9.43
	L-Ile	131.4	2	9	4.78	3.64	2.88	2.19	5.58	4.24	15.77	12.00
	D-Aile	33.3	2	9	1.48	4.44	1.30	3.90	1.97	5.91	5.57	16.71
	L-Leu	104.6	2	7	1.96	1.87	3.75	3.59	4.24	4.05	11.98	11.46
	D-Leu	34.8	2	9	2.18	6.28	1.80	5.17	2.83	8.14	8.00	23.02

where, n1 = final number of replicates after removal of Cochran's and Grubb's outliers and p1 = final number of samples after removal of Cochran's and Grubb's outliers

where, s = standard deviation and RSD% = relative standartd deviation expressed as a %, both being expressions of precision

Suffix; r(s-w) = within-sample repeatability, r(s-b) = between-sample repeatability and r = repeatability of material, calculated as; $\sqrt{((Sr(s\text{-}w))^2+(Sr(s\text{-}b))^2)}$

Table 4.18: ANOVA Precision estimates derived from BioArCh PT result data for amino acid D/L values and concentrations in mollusc shell (A) (ILC-A).

Proficiency Test York Results: Intermediate Precision, S_{RW}'					G	H	I	J	K	L	repeatability limit (2r) (cols G & H)	
matrix; Mollusc A	amino acid	mean	n1	p1	Sr(i)	RSDr(i)%	SL	RSDL%	SRW'	RSDRW' %	absolute (for Sr(i))	as % (for RSDr(i)%)
D/L Value	Asx D/L	0.43	2	5	0.0014	0.33	0.0067	1.56	0.0068	1.59	0.004	0.92
	Glx D/L	0.23	2	5	0.0027	1.17	0.0029	1.23	0.0039	1.69	0.008	3.30
	Ser D/L	0.56	2	5	0.0141	2.53	0.0069	1.24	0.0157	2.82	0.040	7.16
	Arg D/L	0.62	2	4	0.1967	31.73	0.0648	10.45	0.2071	33.40	0.556	89.74
	Ala D/L	0.41	2	5	0.0080	1.97	0.0251	6.18	0.0263	6.49	0.023	5.58
	Val D/L	0.19	2	5	0.0066	3.37	0.0141	7.26	0.0156	8.00	0.019	9.54
	Phe D/L	0.28	2	5	0.0040	1.44	0.0054	1.93	0.0067	2.41	0.011	4.06
	D-Aile/L-Ile	0.25	2	5	0.0378	15.07	0.0617	24.60	0.0724	28.85	0.107	42.63
	Leu D/L	0.29	2	4	0.0081	2.79	0.0443	15.25	0.0450	15.50	0.023	7.89
concentration (pmol/mg))	L-Asx	825.9	2	4	12.29	1.49	67.18	8.13	68.30	8.27	34.77	4.21
	D-Asx	355.0	2	4	5.65	1.59	26.01	7.33	26.62	7.50	15.97	4.50
	L-Glx	311.5	2	5	15.93	5.12	41.29	13.26	44.26	14.21	45.07	14.47
	D-Glx	72.6	2	5	4.18	5.76	10.02	13.80	10.86	14.95	11.82	16.29
	L-Ser	81.5	2	5	4.90	6.02	3.34	4.10	5.93	7.28	13.87	17.02
	D-Ser	45.5	2	5	2.32	5.11	3.06	6.73	3.84	8.45	6.56	14.44
	L-Arg	96.8	2	4	3.68	3.80	-	-	3.68	3.80	10.40	10.75
	D-Arg	62.1	2	4	17.77	28.63	-	-	17.77	28.63	50.26	80.97
	L-Ala	220.4	2	4	6.94	3.15	1.54	0.70	7.11	3.23	19.64	8.91
	D-Ala	92.9	2	5	2.37	2.55	6.83	7.35	7.23	7.78	6.71	7.22
	L-Val	207.8	2	4	5.90	2.84	-	-	5.90	2.84	16.68	8.03
	D-Val	41.9	2	5	2.05	4.90	2.66	6.36	3.36	8.03	5.80	13.86
	L-Phe	126.5	2	5	4.26	3.37	7.44	5.88	8.57	6.77	12.05	9.52
	D-Phe	35.3	2	5	1.50	4.26	1.99	5.62	2.49	7.05	4.26	12.04
	L-Ile	133.6	2	5	3.05	2.29	6.70	5.02	7.36	5.51	8.64	6.47
	D-Aile	33.3	2	5	4.53	13.62	7.46	22.42	8.72	26.23	12.82	38.53
	L-Leu	138.4	2	5	3.22	2.32	7.06	5.10	7.76	5.60	9.10	6.57
	D-Leu	40.6	2	4	2.24	5.51	5.94	14.64	6.35	15.64	6.32	15.58

where, n1 = final number of replicates after removal of Cochran's and Grubb's outliers and p1 = final number of samples after removal of Cochran's and Grubb's outliers

where, s = standard deviation and RSD% = relative standartd deviation expressed as a %, both being expressions of precision

Suffix; r(i) = instrumental or injection repeatability, L = between-sample/run variance and RW(i)= within or intra-lab reproducibility calculated as; $\sqrt{((Sr(i))^2+(SL)^2)}$

Note; SRW(i) is not quite a complete estimate of intermediate precision as the repeatability component should consist of independent samples, not replicated injections

Table 4.19: Intermediate reproducibility precision estimates for amino acid D/L values and concentrations in mollusc shell (A) (ILC-A).

Intra-Lab Reproducibility or Intermediate Precision, S_{RW}			E (or A)	F (or B)	I (or G)	J (or H)	M	N
matrix; Mollusc A	amino acid	mean	Sr	RSDr%	SL	RSDL%	SRW	RSDRW%
D/L Value	Asx D/L	0.43	0.004	0.88	0.007	1.56	0.008	1.79
	Glx D/L	0.23	0.009	3.72	0.003	1.23	0.009	3.89
	Ser D/L	0.54	0.038	7.24	0.007	1.24	0.039	7.15
	Arg D/L	0.73	0.101	12.19	0.065	10.45	0.120	16.58
	Ala D/L	0.43	0.026	5.86	0.025	6.18	0.036	8.50
	Val D/L	0.19	0.009	4.81	0.014	7.26	0.017	8.78
	Phe D/L	0.28	0.008	2.83	0.005	1.93	0.010	3.42
	D-Aile/L-Ile	0.25	0.014	5.38	0.062	24.60	0.063	25.04
	Leu D/L	0.31	0.034	10.36	0.044	15.25	0.056	18.04
concentration (pmol/mg))	L-Asx	769.0	29.52	4.15	67.18	8.13	73.38	9.54
	D-Asx	328.5	13.13	4.35	26.01	7.33	29.14	8.87
	L-Glx	294.5	9.91	3.57	41.29	13.26	42.46	14.42
	D-Glx	68.4	2.64	4.11	10.02	13.80	10.36	15.14
	L-Ser	76.8	6.19	8.57	3.34	4.10	7.03	9.15
	D-Ser	42.3	1.36	3.46	3.06	6.73	3.35	7.90
	L-Arg	93.7	5.01	5.52	3.68	3.80	6.21	6.63
	D-Arg	68.7	9.27	12.32	17.77	28.63	20.04	29.18
	L-Ala	217.3	8.90	4.15	1.54	0.70	9.03	4.16
	D-Ala	94.3	8.26	8.63	6.83	7.35	10.72	11.37
	L-Val	211.5	8.90	4.14	5.90	2.84	10.68	5.05
	D-Val	41.0	2.01	5.01	2.66	6.36	3.34	8.13
	L-Phe	124.1	4.62	3.79	7.44	5.88	8.75	7.05
	D-Phe	34.6	1.13	3.34	1.99	5.62	2.28	6.60
	L-Ile	132.5	5.58	4.24	6.70	5.02	8.72	6.58
	D-Aile	33.3	1.97	5.91	7.46	22.42	7.71	23.16
	L-Leu	121.5	4.24	4.05	7.06	5.10	8.23	6.78
	D-Leu	37.7	2.83	8.14	5.94	14.64	6.58	17.47

where, s = standard deviation and RSD% = relative standartd deviation expressed as a %, both being expressions of precision

Suffix; r = repeatability of material, L = between-sample/run variance and RW= within or intra-lab reproducibility, calculated as; $\sqrt{(Sr^2+SL^2)}$

One final observation concerns the standard solution concentration precision estimates reported in Table 4.12. Replicate injection precision estimates would appear to be reasonable, somewhere generally between 2-3%. However, the between-sample precision jumps to 44% in all cases and far greater than that of the biominerals. Closer inspection of the raw data revealed that data were divided with approximately half of the peak area values agreeing with those observed from the homogeneity data, but the other half were nearly double the size, for both L and D isomers, with no accompanying increase in size of the internal standard. All rehydration volumes quoted and calculations used to determine concentrations were the same in all cases and there is no obvious instrument or analyst bias.

There is no simple explanation for these observed differences in the peak area results. However, these observations do not appear to have affected the final D/L calculation, as the ratio cancels out this increased scaling. However, In terms of control of the measurement system, this gives rise for concern. It may be that the observations are due to differences in rehydration of the dried samples prior to analysis or possible stability issues, since dried standard solution samples had been kept at room temperature and not refrigerated to prevent condensation occurring. However one might then expect to see larger differences in D/L values as different isomers might be expected to exhibit different levels of stability.

Whatever the reason, RSD_{RW}% values given for standard solution concentrations in Table 4.13 should therefore not to be trusted.

4.4.3 Combined uncertainty and Covariance.

In a previous chapter, section 2.8 examined the way in which precision estimates for the respective L and D isomers, together with any contributions from the uncertainty due to bias, could (in principle) be combined to give a single overall estimate of uncertainty for an amino acid D/L value. Let the variable D/L = Y and the individual uncertainty contributions be X_1, X_2...X_n, then Y is related to the individual quantities by (JCGM 100, 2008, p8, 4.1);

$$Y = f(X_1, X_2, \ldots . X_n) \qquad \textbf{(4.7)}$$

The formula for the combined standard uncertainty is derived from a first-order Taylor series approximation, and is referred to in the GUM as the "*law of propagation of uncertainty"* ((JCGM 100, 2008, p19). Let $\mathrm{Y} = \mathrm{y}$ and $\mathrm{X} = \mathrm{x}$, therefore, the combined standard uncertainty is given by ((JCGM 100, 2008, p19, 5.1.2 & 5.1.3; EURACHEM / CITAC, 2000, p25, 8.2);

$$u_c^2(y(x_1 x_2 \dots x_n)) = \sum_{i=1}^{n} \left(\frac{\partial f}{\partial x_i}\right)^2 u^2(x_i) \quad \textbf{(4.8)}$$

$$and \quad u_c^2(y) = \sum_{i=1}^{n} [c_i u(x_i)]^2 \equiv \sum_{i=1}^{n} u_i^2(y) \quad \textbf{(4.9)}$$

Where $c_i \equiv \partial f / \partial x_i$ and $u_i(y) \equiv |c_i| u(x_i) \equiv u(y, x_i)$

The partial derivatives *ci* are termed sensitivity coefficients and describe how changes in *xi* affect the output y, or the uncertainty in y arising from the uncertainty in *x*. The GUM suggests that sensitivity coefficients may be derived mathematically (JCGM 100, 2008, p19, 5.1.3) and Eurachem provide an example of a spreadsheet derived approximation after Kragten (Kragten, 1994, cited in EURACHEM / CITAC, 2000, p104, E.2). However, it is acknowledged that whilst individual uncertainty components may be known, rarely is there information available on how the uncertainty of each input value affects the uncertainty of the end result (Thompson *et al.*, 2002, Appendix B1), and an experimental approach is suggested (Thompson *et al.*, 2002; JCGM 100, 2008).

However, Eurachem provide what might be seen as a "get-out clause";

> "....*However, when an uncertainty contribution is associated with the whole procedure, it is usually expressed as an effect on the final result. In such cases, or when an uncertainty on a parameter is expressed directly in terms iof its effect on y, the sensitivity coefficient* $\partial y/\partial x_i$ *is equal to 1.0*." (EURACHEM / CITAC, 2000, p26, B.2.4)

For independent variables, a general statement for combined uncertainty is given as (EURACHEM / CITAC, 2000, p25, 8.2.2);

$$u_c\big(y(x_1, x_2 \dots .)\big) = \sqrt{\sum_{i=1,n} c_i^2 u(x_i)^2} = \sqrt{\sum_{i=1,n} u(y, x_i)^2} \quad \textbf{(4.10)}$$

Which, assuming *ci* = 1, then equation 4.10 reduces to a much simpler form. (EURACHEM / CITAC, 2000, p26). (Rule 2) "*For models only involving a product or a quotient, e.g. y=(p x q x r x…) or y=p/(q x r x…), the combined standard uncertainty uc(y)....*" can be derived from each component's relative standard uncertainties, thus;

$$u_c(y) = y\sqrt{\left(\frac{u(p)}{p}\right)^2 + \left(\frac{u(q)}{q}\right)^2} \quad \textbf{(4.11)}$$

Therefore, since the D/L value is derived from the D and L isomer concentrations, then, assuming there are no significant bias contributions, the uncertainty of the D/L value should simply be;

$$\frac{u_c(D/L)}{D/L} = \sqrt{\left(\frac{u(D)}{[D]}\right)^2 + \left(\frac{u(L)}{[L]}\right)^2} \qquad \textbf{(4.12)}$$

Where, [D] and [L] are the concentrations of the D and L isomers and u(D) and u(L) are the intra-laboratory reproducibility standard deviations for the two isomers, $s_{RW(D)}$ and $s_{RW(L)}$. Also, since $(u(D)/[D])^2$ is the relative standard deviation, $RSD_{RW(D)}$, it follows that;

$$RSD_{RW\left(\frac{D}{L}\right)}(\%) = \sqrt{\left(RSD_{RW(D)}(\%)\right)^2 + \left(RSD_{RW(L)}(\%)\right)^2} \qquad \textbf{(4.13)}$$

From data given in Table 4.19 for Mollusc shell (A) (ILC-A), the D/L value for Asx is 0.43, with a RSD_{RW} of 1.79%. L-Asx has an RSD_{RW} of 9.54% and for D-Asx, a value of 8.87%.

Therefore, a combined uncertainty of individual isomers would give,

$$RSD_{RW\left(\frac{D}{L}\right)}(\%) = \sqrt{9.54^2 + 8.87^2} = 13.03\%$$

Clearly, 13.03% is not the same as a relative standard deviation of 1.79% reported as the precision estimate for the Asx D/L value of 0.43. Similar effects are observed for all amino acids in both the opercula and mollusc shell (standard solution data has been ignored here due to unaccounted for elevations of the concentration s_L precision estimates, previously discussed).

There are two possible causes of this effect. Either; i) there are either additional unaccounted for uncertainty components that are not included in the "top-down" precision estimation approach applied to D/L values, or ii) there are further substantial cancelling effects of uncertainty components common to both the L and D isomer concentrations.

In a single RP measurement, both L and D isomers are measured, for all amino acids, from the same chromatogram. If all amino acid isomers are determined during a single instrumental measurement and since the quantity of D is related to the quantity of L, they are not independent. Similarly, they are not independent of the other amino acids quantified in the same measurement. Consequently, the calculation of uncertainty based on individual contributions, becomes more complex requiring evaluation of sensitivity coefficients and covariances.

The general equation for combined uncertainty for variables that are not independent is given as (EURACHEM / CITAC, 2000, p26, 8.2.3),

$$u_c\big(y(x_1, x_2 \ldots.)\big) = \sqrt{\sum_{i=1,n} c_i^2 u(x_i)^2 \; + \sum_{\substack{i,k=1,n \\ i \neq k}} c_i c_k \cdot u(x_i, x_k)} \qquad \textbf{(4.14)}$$

Where $u(x_i, x_k)$ is the covariance between input quantities x_i and x_k.

Covariance describes the portion of variance shared by both variables and is given as (RSC Analytical Methods Committee, 2008a),

$$cov(x_i, x_k) = \sum\nolimits_i \frac{(x_i - \bar{x}_i)(x_k - \bar{x}_k)}{n-1} \qquad \textbf{(4.15)}$$

Derivation of covariance using a covariance matrix has been described (RSC Analytical Methods Committee, 2008a). Applied to the DL concentration data, it would like this;

Std dev, (s)	[L-Asx]	[D-Asx]
[L-Asx]	$[\text{L-Asx}]^2$	[L-Asx][D-Asx]
[D-Asx]	[D-Asx][L-Asx]	$[\text{D-Asx}]^2$

Because the $cov(x_i, x_k) = cov(x_k, x_i)$, the combined uncertainty for Asx D/L is determined as (Haesselbarth & Bremser, 2004);

$$\frac{u_c(D/L)}{D/L} = \sqrt{\left(\frac{u(D)}{[D]}\right)^2 + \left(\frac{u(L)}{[L]}\right)^2 - 2\frac{u(L,D)}{L \cdot D}} \qquad \textbf{(4.16)}$$

A screen shot of the Excel spreadsheet used for the calculations of covariance used in this example is shown in Figure 4.29. For Asx, given a D/L of 0.43, this results in an RSD_{RW}% value of 0.47%. Although this value is small compared to the derived value of 1.79% given in Table 4.19, it is of an appropriate order of magnitude and considerably closer than the 13% previously calculated. It is not therefore too difficult to appreciate that other small contributions perhaps due to interactions between Asx and other amino acids, or sampling, analytical or other matrix effects not accounted for here, could make up the difference. The above example illustrates the difficulty of applying the "bottom-up" GUM approach, even with a simplistic model such as combining two concentration uncertainty estimates. For this

reason, a "top-down" approach that incorporates all these elements in a single step, becomes very appealing.

Figure 4.29: Excel screen shot showing calculation of covariance

Mollusc (A) homogeneity			Deviation (S)				Covariance
average conc (pmol/mg)			Difference; yi - ymean				
sample	[L-Asx]	[D-Asx]		(L diff)	(D diff)		(L diff)x(D diff)
1	710.56	300.87		-1.51	-1.15		1.74
2	667.48	283.39		-44.59	-18.63		830.81
3	728.51	308.06		16.45	6.03		99.24
4	691.55	291.98		-20.52	-10.04		206.06
5	730.74	311.77		18.67	9.75		181.98
6	702.32	297.95		-9.75	-4.08		39.72
7	733.70	313.43		21.63	11.41		246.77
8	731.69	308.74		19.62	6.71		131.68
mean	712.07	302.02	sum	0.00E+00	3.41E-13	sum	1737.99
stdev	23.84	10.51				df = n-1	7
variance	**568.33**	**110.37**				**covariance = sum / df**	**248.28**

covariance matrix		
	[L-Asx]	[D-Asx]
[L-Asx]	568.33	248.28
[D-Asx]	248.28	110.37

	var (S^2)	conc	$conc^2$	$S^2 / (D/L)^2$
[L-Asx]	568.33	712.07	507040	0.0011
[D-Asx]	110.37	302.02	91219	0.0012
2x[L][D]	496.57	-	215062	0.0023
u(Asx)/Asx D/L = SQRT((uL/L)^2+(uD/D)^2-2(uLD/LD))				
			RSD% = u(Asx)/Asx D/L	**0.47%**

4.5 Quality Control

4.5.1 Repeatability

Green text present at the end of the Tables 4.11, 4.12, 4.14, 4.15, 4.17 and 4.18, show repeatability limits, derived from the repeatability standard deviations. In Table 4.11, Table 4.14 and Table 4.17 which show repeatability precision estimates derived from homogeneity data, s_r values in column E have been used to calculate repeatability estimates for independent samples. In Table 4.12, Table 4.15 and Table 4.18, repeatability estimates are derived from duplicate injections using sr(i) data given in column G.

Repeatability, (*r)*, was mentioned in the previous chapter (section 3.4.2.4) and is the absolute permissible distance between replicate data.

$$r = t \times \sqrt{n} \times s_r \quad \textbf{(4.17)}$$

where *t* is the t-value for a normal distribution at 95% probability, i.e.; 1.96 , rounded to 2, and *n* is the number of replicates , i.e., 2, and s_r is the repeatability precision estimate (Horwitz, 1995; NMS, accessed 2009b).

Table 4.20 shows how repeatability limits can be used to control replicate precision during routine analysis. Data used, has been taken from the proficiency test opercula homogeneity assessment as it provides paired results. The precision value used to determine acceptable limits is the within-sample repeatability, $s_{r(s\text{-}w)}$ (Table 4.14, column A: Asx D/L $s_{r(s\text{-}w)}$ = 0.0086).

Thus; $r = 1 \times \sqrt{2} \times 0.0086 = 0.0122$

$2r = 2 \times \sqrt{2} \times 0.0086 = 0.0243$

$3r = 3 \times \sqrt{2} \times 0.0086 = 0.0365$

The difference between replicate values is the absolute difference, ignoring direction, hence the squaring and square rooting of the difference (diff). It can be seen that the difference between the pair of replicates for Sample 7, previously identified as a Cochran's outlier from the original homogeneity assessment, is also greater than the maximum permissible distance or 3 times the repeatability (3r). Replicates for Sample 8 also exceed the 2r limit, (95% probability level) and depending on the application may be unsuitable or flag up as a warning and possibly require a retest. Such controls become an essential element of laboratory QC which monitor measurement system stability. Thus, it can be appreciated how measuring replicates, can be used to monitor precision within an analytical run. Similar assessments could be applied using injection repeatability precision or applying reproducibility limits in the same way, between runs, depending on the method protocol and measurement requirements.

For comparison, repeatability limits have also been applied to L-Asx and D-Asx concentrations, and calculated in exactly the same way using data, once again, from Table 4.14, column A.

Table 4.21 shows the concentration difference between replicate pairs for each of the ten samples. Note how Samples 7 and 8 for both L-Asx and D-Asx are within the 2r limits.

Table 4.20: Use of Repeatability Limits; Asx D/L in Opercula Test Material

Asx D/L Precision estimate			Repeatability limits		
	mean D/L	Sr(w-s)	r	2r	3r
	0.578	0.0086	0.0122	0.0243	0.0365
Asx D/L Homogeneity data			**Difference acceptability**		
sample	rep 1	rep 2	diff	$\sqrt{(diff)^2}$	comment
1	0.582	0.584	-0.0021	0.00213	
2	0.570	0.584	-0.0138	0.01384	
3	0.584	0.573	0.01104	0.01104	
4	0.570	0.585	-0.0151	0.01511	
5	0.585	0.581	0.00389	0.00389	
6	0.579	0.580	-0.0008	0.00084	
7	0.522	0.571	-0.049	**0.04904**	>3r
8	0.554	0.580	-0.026	**0.02604**	>2r
9	0.570	0.579	-0.0093	0.00933	
10	0.580	0.578	0.00239	0.00239	

red text indicates sample was identified as a Cochrans outlier

Repeatability therefore looks at the closeness of agreement between two values. It monitors the effect of random error effects. It does not however assess whether the measured values or even the mean of measured values is acceptable. D/L values are determined from L-Asx and D-Asx concentrations, therefore the accuracy of each measured isomer concentration is also important for the accurate reporting of D/L values. If the L-isomer is too high or the D-isomer too low, the D/L will reduce, similarly, if the L-isomer is too low or the D-isomer too high, the D/L will increase.

For this reason, control of the measurement system, needs to be at the concentration level. If concentrations are determined accurately, the D/Ls will take care of themselves.

Within-run repeatability checks are often carried out on routine test samples, where a sample is chosen at random and duplicate samples are taken and worked up through the whole measurement procedure. Samples are then located at random positions in the run sequence. However, this assumes the method has been fully validated with precision values determined, and the test samples under investigation are within the scope of the validation with regard to matrix and concentration.

Table 4.21: Use of Repeatability Limits; D- & L-Asx Conc. in Opercula Test Material

[L-Asx]	Precision estimate		Repeatability limits		
	mean conc	Sr(w-s)	r	2r	3r
	1522.809	114.189	161.49	322.97	484.46
[L-Asx]	Homogeneity data		Difference acceptability		
sample	rep 1	rep 2	diff	√(diff)^2	comment
1	1779.582767	1484.486	295.10	295.10	
2	1556.192342	1448.646	107.55	107.55	
3	1474.280404	1472.633	1.65	1.65	
4	1490.806981	1636.481	-145.67	145.67	
5	1380.230918	1475.068	-94.84	94.84	
6	1640.811384	1460.532	180.28	180.28	
7	1788.416429	1475.227	313.19	313.19	
8	1475.289984	1479.757	-4.47	4.47	
9	1481.275252	1479.285	1.99	1.99	
10	1470.53306	1506.64	-36.10	36.10	
[D-Asx]	Precision estimate		Repeatability limits		
	mean conc	Sr(w-s)	r	2r	3r
	874.307	57.2706	80.99	161.99	242.98
[D-Asx]	Homogeneity data		Difference acceptability		
sample	rep 1	rep 2	diff	√(diff)^2	comment
1	1035.939399	867.3243	168.62	168.62	
2	887.5633324	846.2678	41.30	41.30	
3	860.6109307	843.3985	17.21	17.21	
4	849.4129691	957.1394	-107.73	107.73	
5	807.5691648	857.3168	-49.75	49.75	
6	950.0867091	846.9297	103.16	103.16	
7	933.4281461	842.3134	91.11	91.11	
8	817.7328634	858.7392	-41.01	41.01	
9	844.3751405	857.0441	-12.67	12.67	
10	852.800641	870.14	-17.34	17.34	

Under these conditions, knowledge of the acceptable repeatability limits are known or can be determined from the s_r of the method. However in situations such as AAR where method precision estimates are not known, in-house reference materials such as the ILCs that have sufficient stability, are in sufficient quantity and have some analytical history making it possible to derive a repeatability estimate, should be used.

4.5.2 Control Charts

As a general rule, Quality Control materials (QCMs) should be as similar to the test samples as possible, going through the whole measurement procedure. QCMs might be the

same as in-house RMs used for duplicate analyses, or they might be different if routine test samples are used for duplicate analyses instead. Where replicate analyses monitor random error effects, QCMs are designed to monitor both random and systematic changes in the measurement system, both during and between runs. QCMs are placed at intervals throughout the analytical run as a way of tracking measurement consistency. If used from run to run, matrix-matched QCMs can also provide a check on the intra-laboratory reproducibility over time. The frequency with which a QCM is used depends on the length of the run. However Nordtest suggest that the decision is based on the assumption that "*all measurements performed after the last approved sample in the quality control may have to be reanalysed*." (Hovind *et al.*, 2007, p22). It therefore becomes a matter of balancing measurement quality against measurement cost.

In situations where the stability of the matrix QCM is in doubt or for simply monitoring the stability of the instrumental analysis, a reference solution could be used. This might be either a CRM or in-house standard solution. However, where the repeatability precision of the solution is different from that of a routine sample, then inclusion of some form of matrix-matched QCM is recommended (Hovind *et al.*, 2007).

The number and type of QCMs can vary depending on available material and measurement requirements. For example; if a number of samples are to be analysed in a single run, and the range of concentrations of those samples varies, it may be appropriate to include a couple of QCMs that cover the expected concentration range of the test samples. Inclusion of CRMs if available will also check on measurement bias.

Having run all QCMs and duplicate samples, results are reviewed and assessed against statistical limits. This is often best achieved by plotting on control charts.

X-charts and X-bar charts are used to plot individual values and value means respectively. If the method protocol requires that results are determined from the average of replicate samples, then QCMs should be treated in the same way.

Statistical control limits are set based on method performance characteristics. These may have been derived from a collaborative trial or during method validation, but in essence are the repeatability precision estimate, s_r and intermediate or intra-laboratory reproducibility, s_{RW} precision estimates, multiplied by 2 or 3 for the required confidence level.

It is important to remember that both s_r and s_{RW} are precision estimates for single values. Where data are subject to averaging, the equivalent standard errors should be used, thus;

chart	results	Within-run 2 std dev CI	Between-run 2 std dev CI
X-Chart	A, B, C	mean ± 2 x s_r	mean ± 2 x s_{RW} = 2 x √(s_r^2+s_L^2)
X-bar Chart	A=B+C+.../n	mean ± 2 x s_r/√n	mean ± 2 x √(s_r^2/n+s_L^2)

Results of duplicate analyses can also be plotted on control charts, these are referred to as range charts or R-charts / r-charts. In its simplest form, the control limits can be set at 2 x repeatability or 3 x repeatability value, and the absolute difference plotted.

For illustration, L and D-Asx concentration data for opercula, previously given in Table 4.21, have been used here to plot range and X-charts, Figure 4.30. All data points are independent and have been plotted individually on the X-chart.

The R-charts (Figure 4.30) make interpreting the data given in Table 4.21 far easier to identify anomalous values. However, as the data originally suggested, all concentration values for Asx in opercula were within the 3r limit. When plotted individually on an X-chart, the precision of the replicates around the mean value can be observed. If the x-axis was in days or runs, rather than in sample number, and if plotted in run order rather than stacked as shown, instrumental drift could be observed over time.

However, what can be seen on the X-charts are the relative positioning of the individual values. Although the chart for D-Asx has slightly tighter control limits, the relative positioning of each pair of data points is approximately the same, except for rep 1 of Sample 7. In Sample 7, L-Asx rep 1 is positioned higher than the equivalent D-Asx value. A higher L isomer concentration compared to the D isomer value, will depress the D/L value and this is what is observed in the Asx D/L homogeneity data. Interestingly, rep 1 of Sample 1 is high in both the L and the D X-charts. However the D/L value determined from this is totally acceptable as the ratio is maintained.

Whilst the use of QCMs and duplicate analyses in routine use cannot measure accuracy of each and every single test sample analysed, regular use of control charts would go a long way to help monitor the stability of the measurement system over the course of each run and over time.

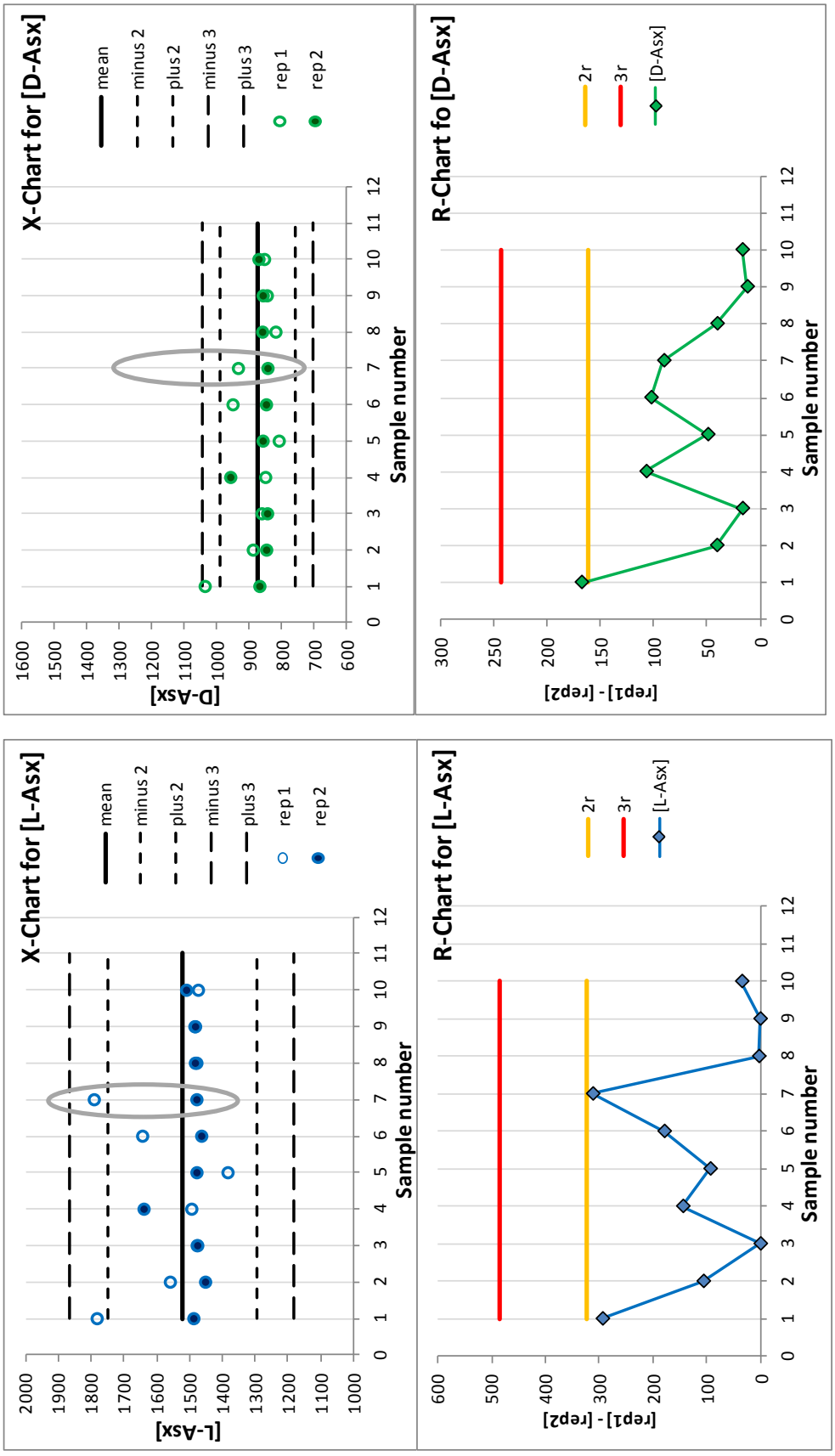


Figure 4.30: Control charts for L-Asx and D-Asx concentration values (pmol/mg) in opercula test materials

However, whilst good precision of results is always desirable, in terms of absolute accuracy, precision only goes part-way towards ensuring the accuracy of results. The evaluation of trueness of a method and the control of bias is of equal importance and frequently neglected. In AAR analysis, the absence of defined reference materials is a particular problem that needs urgent attention. For geochronological work, relative age differences are applied to AAR data, and usually calibrated by a different dating technique. Laboratories tend to work in isolation, unable to share AAR results due to specific effects resulting from laboratory and method bias, which they are currently unable to correct.

4.5.3 Bias Evaluation: Standard Solutions

The determination of bias requires a comparison of measurement results against a suitable comparator. This is most often a reference material with a known or reference value, such as CRMs or possibly the use of a reference method, defined through collaborative trial. In AAR analysis, no such matrix-matched reference materials for D/L values are commercially available, neither is there a method fully validated by collaborative trial. However, reference materials for the L and D isomers are commercially available and were previously acquired to make up the three original standard solutions (Penkman, 2005).

From the evaluation of precision estimates in standard solution given earlier in the chapter, (section 4.3.2), observed mean D/L values for each of the three levels of standard solution (y-axis), have been plotted in the following charts against the expected D/L values for each amino acid (x-axis). Further, the red dotted lines in each represent the predicted trendline, if each amino acid was present at its assumed level, i.e. 0.167, 0.5 and 0.91 D/L. The difference between the observed and expected lines, represent the theoretical bias.

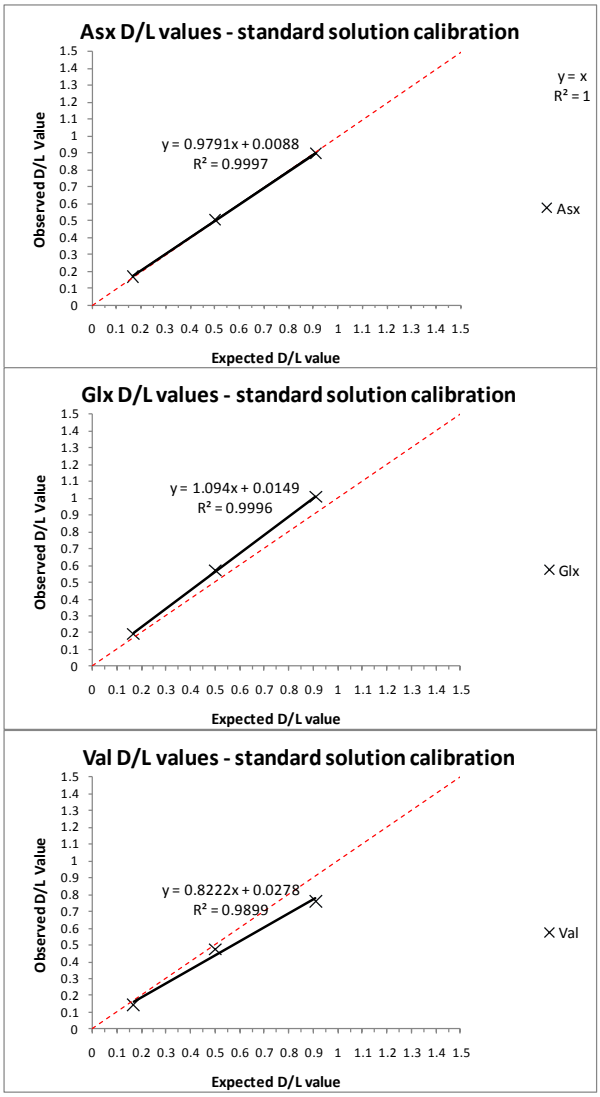
For example, Figure 4.31 shows that Asx D/L appears to be in alignment with the expected D/L values, Glx D/L would appear to determine D/L values slightly too high and valine, too low compared to the expected.

Because the D/L value of each standard solution is assumed constant, only the chromatogram peak areas are used routinely for QC during a run, or perhaps the normalised difference where the amino acid area is divided by the LhArg peak area. These can then be compared to previous values, as the ratios should be consistent. Because the concentrations are constant, there is no requirement to determine isomer concentrations and no correction factor is applied. However, the three charts shown above do suggest that bias exists which may be due to instrumental effects but may equally be due to differences in original preparations.

Figure 4.31: Examples of Observed D/L value in standard solutions against expected D/L value

Solid black line = observed D/L value

Red dotted line = expected D/L value



Data reporting the weights and volumes used in the preparation of the original bulk standard solutions, has subsequently been acquired. Consequently, it has been possible to determine the precise molarity of the D-isomers in each standard solution and get a more accurate estimate of the D/L value present in each solution for each amino acid. All L-isomers were present in a single reference solution, supplied by Sigma, and used at a molarity of 0.001M. All D-isomers were obtained as dry powders requiring weighing and dissolution. Weights, volumes and molar concentrations are summarised in Table 4.22 and the resulting D/L values given in Table 4.23.

Table 4.22: Amino acid constituents and concentrations present in standard solutions; 0.167d, 0.5d and 0.91d.

RM	Amino acid isomer	Prep of stock sols. MW	D-isomer mg in 30mL	molarity	Concentrations of isomers in 0.167d D/L std sol. volumes used (µL)	Total volume (µL)	conc in 0.167 D/L std sol µmole/µL	x10 dilution conc in 0.167d µmole/µL	Concentrations of isomers in 0.5d D/L std sol. volumes used (µL)	Total volume (µL)	conc in 0.5 D/L std sol µmole/µL	x10 dilution conc in 0.5d µmole/µL	Concentrations of isomers in 0.91d D/L std sol. volumes used (µL)	Total volume (µL)	conc in 0.91 D/L std sol µmole/µL	x10 dilution conc in 0.91d µmole/µL
L-aa	L-Asx			0.001			8.216E-04	8.216E-05			6.061E-04	6.061E-05			4.581E-04	4.581E-05
solution	L-Glu			0.001			8.216E-04	8.216E-05			6.061E-04	6.061E-05			4.581E-04	4.581E-05
	L-Ser			0.001			8.216E-04	8.216E-05			6.061E-04	6.061E-05			4.581E-04	4.581E-05
	L-Thr			0.001			8.216E-04	8.216E-05			6.061E-04	6.061E-05			4.581E-04	4.581E-05
	L-His			0.001			8.216E-04	8.216E-05			6.061E-04	6.061E-05			4.581E-04	4.581E-05
	Gly			0.001			8.216E-04	8.216E-05			6.061E-04	6.061E-05			4.581E-04	4.581E-05
	L-Arg			0.001			8.216E-04	8.216E-05			6.061E-04	6.061E-05			4.581E-04	4.581E-05
	L-Ala			0.001	1000		8.216E-04	8.216E-05	1000		6.061E-04	6.061E-05	1000		4.581E-04	4.581E-05
	L-h-Arg			0.001			8.216E-04	8.216E-05			6.061E-04	6.061E-05			4.581E-04	4.581E-05
	L-Tyr			0.001			8.216E-04	8.216E-05			6.061E-04	6.061E-05			4.581E-04	4.581E-05
	L-Val			0.001			8.216E-04	8.216E-05			6.061E-04	6.061E-05			4.581E-04	4.581E-05
	L-Met			0.001			8.216E-04	8.216E-05			6.061E-04	6.061E-05			4.581E-04	4.581E-05
	L-Phe			0.001			8.216E-04	8.216E-05			6.061E-04	6.061E-05			4.581E-04	4.581E-05
	L-Ile			0.001		1217.1	8.216E-04	8.216E-05		1650	6.061E-04	6.061E-05		2183	4.581E-04	4.581E-05
	L-Leu			0.001			8.216E-04	8.216E-05			6.061E-04	6.061E-05			4.581E-04	4.581E-05
D-aa	D-Ala	89.09	2.42	0.0091	16.7		1.242E-04	1.242E-05	50		2.744E-04	2.744E-05	91		3.774E-04	3.774E-05
Powders	D-Arg	174.2	5.11	0.0098	16.7		1.342E-04	1.342E-05	50		2.963E-04	2.963E-05	91		4.076E-04	4.076E-05
	D-Asp	133.1	3.31	0.0083	16.7		1.137E-04	1.137E-05	50		2.512E-04	2.512E-05	91		3.456E-04	3.456E-05
	D-Glu	147.1	4.39	0.0099	16.7		1.365E-04	1.365E-05	50		3.015E-04	3.015E-05	91		4.147E-04	4.147E-05
	D-His	155.2	4.03	0.0087	16.7		1.188E-04	1.188E-05	50		2.623E-04	2.623E-05	91		3.608E-04	3.608E-05
	D-aile	131.2	3.67	0.0093	16.7		1.279E-04	1.279E-05	50		2.826E-04	2.826E-05	91		3.887E-04	3.887E-05
	D-Leu	131.2	3.47	0.0088	16.7		1.210E-04	1.210E-05	50		2.672E-04	2.672E-05	91		3.675E-04	3.675E-05
	D-Met	149.2	3.8	0.0085	16.7		1.165E-04	1.165E-05	50		2.573E-04	2.573E-05	91		3.539E-04	3.539E-05
	D-Phe	165.2	5.01	0.0101	16.7		1.387E-04	1.387E-05	50		3.063E-04	3.063E-05	91		4.214E-04	4.214E-05
	D-Pro	115.1	3.04	0.0088	16.7		1.208E-04	1.208E-05	50		2.668E-04	2.668E-05	91		3.670E-04	3.670E-05
	D-Ser	105.1	2.99	0.0095	16.7		1.301E-04	1.301E-05	50		2.874E-04	2.874E-05	91		3.953E-04	3.953E-05
	D-Thr	119.1	3.12	0.0087	16.7		1.198E-04	1.198E-05	50		2.646E-04	2.646E-05	91		3.640E-04	3.640E-05
	D-Val	117.1	3.57	0.0102	16.7		1.394E-04	1.394E-05	50		3.079E-04	3.079E-05	91		4.236E-04	4.236E-05
	LhArg						1.000E-05	1.000E-05			1.000E-05	1.000E-05			1.000E-05	1.000E-05

Table 4.23: Actual D/L values for each amino acid in standard solutions

amino acid D/L	std sol. 0.167d	amino acid D/L	std sol. 0.5d	amino acid D/L	std sol. 0.91d
D/L-Asx	0.1384	D/L-Asx	0.4145	D/L-Asx	0.7543
D/L-Glu	0.1661	D/L-Glu	0.4974	D/L-Glu	0.9053
D/L-Ser	0.1584	D/L-Ser	0.4742	D/L-Ser	0.8630
D/L-Thr	0.1458	D/L-Thr	0.4366	D/L-Thr	0.7946
D/L-His	0.1445	D/L-His	0.4328	D/L-His	0.7877
D/L-Arg	0.1633	D/L-Arg	0.4889	D/L-Arg	0.8898
D/L-Ala	0.1512	D/L-Ala	0.4527	D/L-Ala	0.8240
D/L-Val	0.1697	D/L-Val	0.5081	D/L-Val	0.9248
D/L-Met	0.1418	D/L-Met	0.4245	D/L-Met	0.7726
D/L-Phe	0.1688	D/L-Phe	0.5054	D/L-Phe	0.9199
D/L-Ile	0.1557	D/L-Ile	0.4662	D/L-Ile	0.8485
D/L-Leu	0.1472	D/L-Leu	0.4408	D/L-Leu	0.8023

These results show the differences in actual D/L values compared to those expected from each respective standard solution. These differences may therefore account for the biases observed above. As a check on this the observed D/L values are again plotted against expectation but this time the known D/L values are used as reference values and shown as the red dotted lines. These results are shown in Figure 4.32 and all trendlines fixed at the origin. Having now plotted the observed D/L values alongside the known D/L values, it would appear that the differences are if anything, slightly wider than before. For instance Asx D/L now shows a substantial bias when previously there was none.

Using the equations of the trendlines on each chart, pairs of y values can be calculated for stated x values. The difference between the two y values is the bias. This can then be divided by the y value for the observed D/L trendline and multiplied by 100 to give the relative percentage bias of the observed value compared to the known D/L value.

For example, for Asx D/L, trendline (Figure 4.32); y_1=0.9916x, and observed standard solution; y_2 = 0.828x.

If x = 0.5; y_1 = 0.4958 and y_2 = 0.4145. The difference (bias); y_1-y_2 = 0.0813.

The relative bias = $(y_1-y_2)/y1 \times 100$, therefore (0.0813/0.4958)x100 = 16.4%

Values for the bias of the observed data are given Table 4.24, with the sign included to indicate the direction of the bias.

As there are no method preparation effects for standard solutions, any differences observed can only be attributed to instrumental losses, detector sensitivity and possibly stability issues. Data used for the determination of the average observed values, plotted on the charts, have been derived from data collected over several years. A review of the normalised peak areas (Isomer area/ LhArg area) does not indicate any obvious stability issues affecting one isomer more than another that may impact on an amino acid's D/L value.

Table 4.24: Relative bias for amino acid D/L values in standard solutions

Amino acid D/L value	Relative Bias (%) of observed std sol data
Asx	16.40%
Glx	10.80%
Ser	-21.42%
Arg	-8.03%
Ala	13.36%
Val	-17.90%
Phe	-11.60%
A/I	-15.64%
Leu	25.00%

This leaves instrumental losses and detector sensitivity which are likely to affect all sample analytes in a similar way. In terms of calibration, this is known as the response factor. Response factors (F) were mentioned briefly in the previous chapter (section 3.2.3), and are a requirement for the correct use of internal standards in calibration. Response factors are determined from;

$$\frac{A_{L\,aa}}{C_{L\,aa}} = F \times \frac{A_{is}}{C_{is}} \qquad \text{therefore,} \qquad F = \frac{A_{L\,aa}/C_{L\,aa}}{A_{is}/C_{is}} \qquad \textbf{(4.18)}$$

From previously recorded data we now have chromatogram peak areas for each amino acid for all three standard solutions, (A_{Laa}), peaks area of the internal standard for each run (A_{is}), the concentration of the internal standard, LhArg, (C_{is}) which is assumed constant at 0.01mM, and now, information regarding the concentration of the amino acids used in the bulk standard solutions originally prepared (C_{Laa}). Thus, there is now sufficient information to calculate *F* directly from the standard solutions, which should correct for the observed biases.

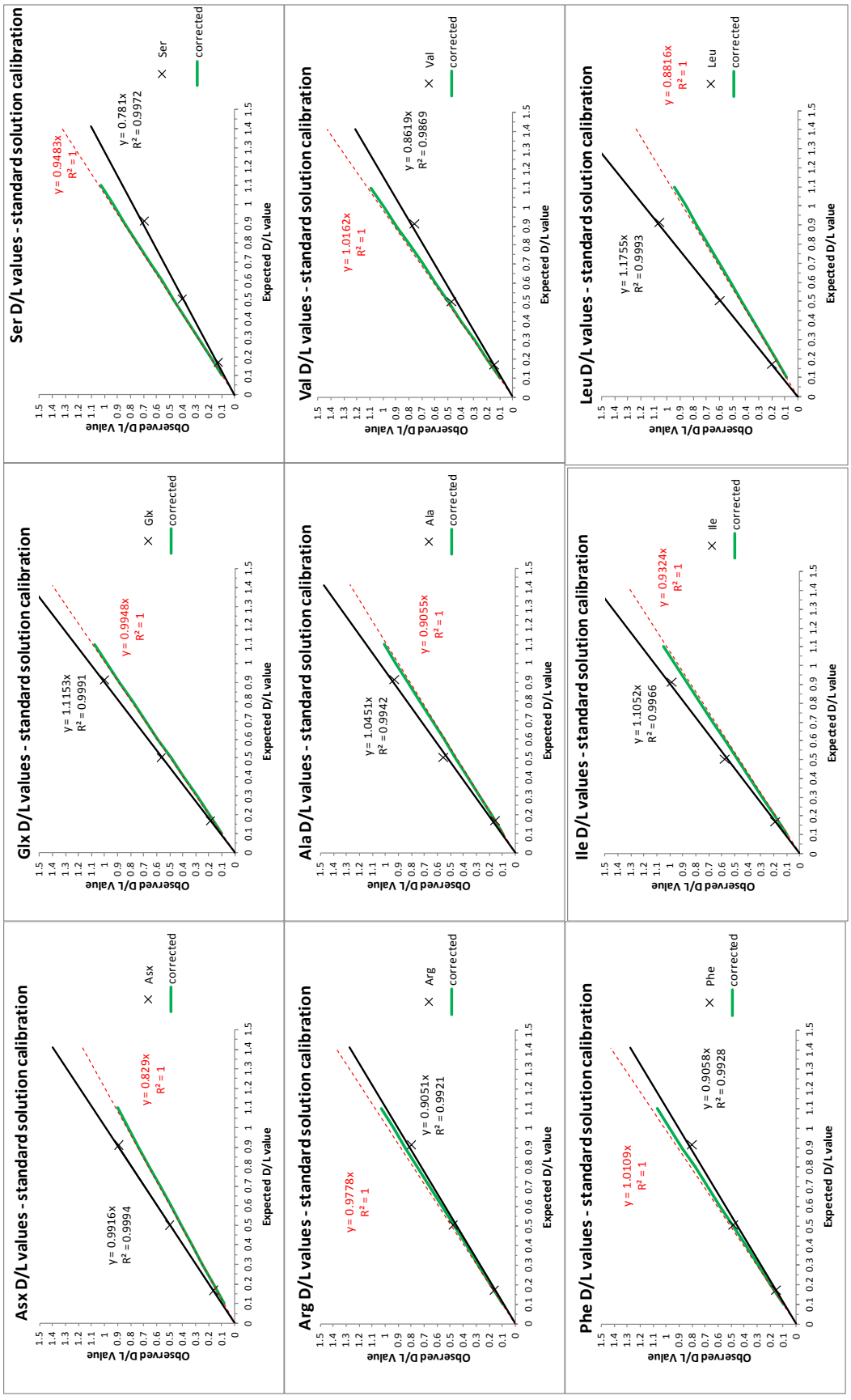


Figure 4.32: Observed and known D/L values for amino acids in standard solution

Solid black line = observed D/L value Red dotted line = expected D/L value Solid green line = corrected D/L value

Using standard solution and LhArg peak area data from every analytical run, individual response factors were calculated for each isomer for all amino acids in every standard solution. Response factors were then averaged to get a single representative value for a given L or D isomer for a particular amino acid in each solution (Table 4.25). Results indicated that response factors were not concentration dependent, therefore a single representative value could be derived as the average of the mean values for each isomer.

Table 4.25 shows the average response factors for all amino acid isomers, together with the 2 standard deviation uncertainty associated with the mean of these values (standard error). The error associated with these values is very small. It was noticed that response factors varied between amino acids but also between the L and D isomers of a specific amino acid. Further, it was also noticed that response factors showed no clear concentration dependence. Therefore F values for individual isomers could be averaged across the four standard solutions to give a single L or D isomer correction factor for each amino acid, Table 4.26.

These factors could potentially be used to correct for D and L isomer concentrations separately in future analyses, given by;

$$\mathrm{C_{L\,Val}} = \frac{A_{L\,Val}}{F \times (A_{is}/C_{is})} \qquad \textbf{(4.19)}$$

However a single correction factor for existing D/L values would be more practical. Therefore, in just the same way as a D/L value is obtained from [D]/[L], so a single Correction Factor, CF, was obtained by dividing the response factor for D by the response factor for L; CF = F_D/F_L. This data together with its uncertainty estimate is given in Table 4.27.

To evaluate the effectiveness of the correction factors, observed standard solution data, used for the charts in Figure 4.32, were corrected with the D/L correction values from Table 4.27. **Note; Response factors (F) are used in the denominator of the above equation**. **Therefore reported D/L values should be <u>divided</u> by the D/L correction values (CF), too. CF values greater than 1 indicate reported D/L results have been over-reported, whilst correction values less than 1, indicate measurement results have under-reported the D/L value.**

The results are encouraging. Figure 4.32, shows adjusted figures with solid green lines indicating the corrected values, which now line up along the known D/L trendlines (red dotted line).

Table 4.25: Response Factors (F) for amino acid isomers in standard solutions

amino acid isomer	Response Factors (F) for amino acids in standard solution											
	0.167dH2O standard solution						0.167d standard solution					
	conc M	F(Mean)	st dev	count	std u	RSU%	conc M	F(Mean)	st dev	count	std u	RSU %
L Asx	8.22E-05	**1.114**	0.0558	180	0.0042	0.37%	8.22E-05	**0.962**	0.0811	306	0.0046	0.48%
D Asx	1.14E-05	**1.356**	0.0728	180	0.0054	0.40%	1.14E-05	**1.195**	0.0792	291	0.0046	0.39%
L Glu	8.22E-05	**1.060**	0.0513	180	0.0038	0.36%	8.22E-05	**0.923**	0.0644	301	0.0037	0.40%
D Glu	1.36E-05	**1.226**	0.0639	177	0.0048	0.39%	1.36E-05	**1.038**	0.0979	305	0.0056	0.54%
L Ser	8.22E-05	**1.091**	0.0581	179	0.0043	0.40%	8.22E-05	**0.955**	0.0658	302	0.0038	0.40%
D Ser	1.3E-05	**0.908**	0.0515	181	0.0038	0.42%	1.3E-05	**0.803**	0.0503	299	0.0029	0.36%
L Arg	8.22E-05	**1.126**	0.0878	181	0.0065	0.58%	8.22E-05	**0.945**	0.0865	304	0.0050	0.52%
D Arg	1.34E-05	**1.088**	0.2415	181	0.0179	1.65%	1.34E-05	**1.013**	0.2465	305	0.0141	1.39%
L Ala	8.22E-05	**1.187**	0.0739	182	0.0055	0.46%	8.22E-05	**1.022**	0.0721	301	0.0042	0.41%
D Ala	1.24E-05	**1.242**	0.0823	179	0.0062	0.50%	1.24E-05	**1.169**	0.0642	298	0.0037	0.32%
L Val	8.22E-05	**1.275**	0.0750	182	0.0056	0.44%	8.22E-05	**1.095**	0.0821	277	0.0049	0.45%
D Val	1.39E-05	**1.090**	0.0717	180	0.0053	0.49%	1.39E-05	**0.949**	0.0669	275	0.0040	0.43%
L Met	8.22E-05	**1.087**	0.0698	182	0.0052	0.48%	8.22E-05	**0.907**	0.0856	275	0.0052	0.57%
D Met	1.16E-05	**1.530**	0.1132	182	0.0084	0.55%	1.16E-05	**1.296**	0.1232	275	0.0074	0.57%
L Phe	8.22E-05	**1.192**	0.0693	180	0.0052	0.43%	8.22E-05	**1.011**	0.0951	275	0.0057	0.57%
D Phe	1.39E-05	**1.114**	0.0761	181	0.0057	0.51%	1.39E-05	**0.954**	0.0834	277	0.0050	0.53%
L Ile	8.22E-05	**1.290**	0.0804	181	0.0060	0.46%	8.22E-05	**1.076**	0.1183	279	0.0071	0.66%
D Aile	1.28E-05	**1.581**	0.1199	171	0.0092	0.58%	1.28E-05	**1.358**	0.1455	277	0.0087	0.64%
L Leu	8.22E-05	**1.002**	0.0579	180	0.0043	0.43%	8.22E-05	**0.828**	0.1134	279	0.0068	0.82%
D Leu	1.21E-05	**1.383**	0.0946	179	0.0071	0.51%	1.21E-05	**1.151**	0.1825	275	0.0110	0.96%
isomer	0.5d standard solution						0.91d standard solution					
	conc M	F(Mean)	st dev	count	std u	RSU%	conc M	F(Mean)	st dev	count	std u	RSU%
L Asx	6.06E-05	**1.004**	0.177642	632	0.0071	0.70%	4.58E-05	**1.007**	0.1159	499	0.0052	0.52%
D Asx	2.51E-05	**1.191**	0.195202	558	0.0083	0.69%	3.46E-05	**1.203**	0.1414	497	0.0063	0.53%
L Glu	6.06E-05	**0.951**	0.170381	634	0.0068	0.71%	4.58E-05	**0.968**	0.1203	503	0.0054	0.55%
D Glu	3.01E-05	**1.091**	0.208014	635	0.0083	0.76%	4.15E-05	**1.078**	0.1433	503	0.0064	0.59%
L Ser	6.06E-05	**0.989**	0.173918	632	0.0069	0.70%	4.58E-05	**1.004**	0.1208	501	0.0054	0.54%
D Ser	2.87E-05	**0.852**	0.148909	632	0.0059	0.70%	3.95E-05	**0.812**	0.0962	498	0.0043	0.53%
L Arg	6.06E-05	**1.034**	0.213228	628	0.0085	0.82%	4.58E-05	**1.112**	0.1772	502	0.0079	0.71%
D Arg	2.96E-05	**1.003**	0.253115	628	0.0101	1.01%	4.08E-05	**0.968**	0.1575	496	0.0071	0.73%
L Ala	6.06E-05	**1.065**	0.18482	633	0.0073	0.69%	4.58E-05	**1.082**	0.1291	501	0.0058	0.53%
D Ala	2.74E-05	**1.309**	0.231027	631	0.0092	0.70%	3.77E-05	**1.225**	0.1389	499	0.0062	0.51%
L Val	6.06E-05	**1.155**	0.20162	604	0.0082	0.71%	4.58E-05	**1.166**	0.1303	468	0.0060	0.52%
D Val	3.08E-05	**1.079**	0.188095	603	0.0077	0.71%	4.24E-05	**0.959**	0.1091	468	0.0050	0.53%
L Met	6.06E-05	**0.959**	0.182619	602	0.0074	0.78%	4.58E-05	**0.975**	0.1279	469	0.0059	0.61%
D Met	2.57E-05	**1.096**	0.108733	283	0.0065	0.59%	3.54E-05	**1.288**	0.1654	470	0.0076	0.59%
L Phe	6.06E-05	**1.068**	0.2016	599	0.0082	0.77%	4.58E-05	**1.080**	0.1315	467	0.0061	0.56%
D Phe	3.06E-05	**1.030**	0.192982	603	0.0079	0.76%	4.21E-05	**0.946**	0.1110	469	0.0051	0.54%
L Ile	6.06E-05	**1.148**	0.216744	602	0.0088	0.77%	4.58E-05	**1.150**	0.1319	465	0.0061	0.53%
D Aile	2.83E-05	**1.136**	0.078127	192	0.0056	0.50%	3.89E-05	**1.341**	0.1500	464	0.0070	0.52%
L Leu	6.06E-05	**0.888**	0.174185	601	0.0071	0.80%	4.58E-05	**0.891**	0.1059	463	0.0049	0.55%
D Leu	2.67E-05	**1.213**	0.241212	599	0.0099	0.81%	3.68E-05	**1.185**	0.1546	463	0.0072	0.61%

RSU% =reative standard uncertainty expressed as a percentage

Table 4.26: Average Response Factors (F) for amino acid isomers in std sol.

amino acid isomer	Average Response Factors (F)				
	F(Mean)	st dev	count	std u	RSU%
L Asx	**1.022**	0.0649	4	0.032433	3.17%
D Asx	**1.236**	0.0798	4	0.039916	3.23%
L Glu	**0.975**	0.0596	4	0.029797	3.05%
D Glu	**1.108**	0.0817	4	0.040847	3.69%
L Ser	**1.009**	0.0581	4	0.029026	2.88%
D Ser	**0.844**	0.0479	4	0.023959	2.84%
L Arg	**1.054**	0.0835	4	0.041744	3.96%
D Arg	**1.018**	0.0505	4	0.025227	2.48%
L Ala	**1.089**	0.0701	4	0.035053	3.22%
D Ala	**1.236**	0.0574	4	0.028713	2.32%
L Val	**1.173**	0.0747	4	0.037359	3.19%
D Val	**1.019**	0.0757	4	0.037847	3.71%
L Met	**0.982**	0.0757	4	0.037863	3.86%
D Met	**1.302**	0.1780	4	0.089	6.83%
L Phe	**1.088**	0.0756	4	0.037816	3.48%
D Phe	**1.011**	0.0788	4	0.039384	3.90%
L Ile	**1.166**	0.0893	4	0.044664	3.83%
D Aile	**1.354**	0.1822	4	0.091077	6.73%
L Leu	**0.902**	0.0724	4	0.0362	4.01%
D Leu	**1.233**	0.1033	4	0.051645	4.19%

RSU% =reative standard uncertainty expressed as a percentage

Table 4.27: Single D/L Correction Factor ($F_{D/L}$) for each amino acid in std sol.

amino acid	single D/L correction factor CF	uncertainty with 2 and 3 std dev CLs					
		u(c)	RSU%	2xUCL	2xLCL	3xUCL	3xLCL
Asx	**1.210**	0.05478	4.53%	1.319	1.100	1.374	1.046
Glx	**1.136**	0.05439	4.79%	1.245	1.027	1.299	0.973
Ser	**0.836**	0.03378	4.04%	0.903	0.768	0.937	0.734
Arg	**0.966**	0.04511	4.67%	1.056	0.876	1.101	0.831
Ala	**1.135**	0.04505	3.97%	1.225	1.045	1.270	1.000
Val	**0.869**	0.04252	4.89%	0.954	0.784	0.997	0.742
Phe	**0.929**	0.04853	5.22%	1.026	0.832	1.075	0.784
A/I	**1.161**	0.08988	7.74%	1.341	0.981	1.431	0.891
Leu	**1.367**	0.07927	5.80%	1.525	1.208	1.604	1.129

u(c)=combined uncertainty

See section 4.5.3.1 for explanation

RSU% =reative standard uncertainty expressed as a percentage

UCL=upper confidence limit, LCL= Lower confidence limit

4.5.3.1 Does D-Aile/L-Ile really = 1.3?

Uncertainty estimates given in Table 4.26 ($u_{F(L\ or\ D)} = s_{F(L\ or\ D)}/\sqrt{n}$), are derived from the standard deviations of the four values from the four standard solutions used to obtain the average F values. Uncertainty estimates given in Table 4.27, are derived from the combination of the uncertainty estimates for both the L and D isomers (Table 4.26), using the simplified model for combining standard uncertainties where a quotient is involved, thus; $u_{CF} = CF \times \sqrt{[(u_{F(L)}/F_{(L)})^2+(u_{F(D)}/F_{(D)})^2]}$. Whilst these will represent the overall uncertainty of the mean CF values, other values are possible within the 95% and approx 99% probability range. Therefore confidence limits representing the limit of alternative values that the correction factor could take, can be determined. In Table 4.27, UCL and LCL represent the upper and lower confidence limits respectively.

Of particular interest is the effect of correction on the D-Aile/L-Ile (A/I) data. From Table 4.27, it can be seen that for A/I, the D/L correction value is 1.161, at 2 std dev, the UCL expands to 1.341 and at 3 std dev it becomes 1.431. If an observed D/L value of perhaps 1.341 is then corrected by being divided by the CF value, it can be seen that;

D/L / CF = 1.341 / 1.341 = 1

Therefore, if the correction factor wasn't in fact 1.161, but 1.341, an observed D/L for A/I of 1.341 would give a "corrected" ratio of 1 using this system.

However Ile has two stereogenic centres, and if the equilibria between the forms have different energies then the A/I ratio is not necessarily 1. Previous estimates of A/I have used both Gas Chromatography (Flame Ionisation Detection) and Ion Exchange Chromatography (ninhydrin detection), and both suggest an A/I value of > 1 (between 1.25 and 1.35).

It is therefore interesting to speculate whether the frequently discussed issue regarding the validity of A/I values of 1.3 or higher, might be simply explained by unaccounted for bias. If so, then the correct application of the internal standard and the use of an appropriate response / correction factor may be all that is required to correct ratios.

In principle, these correction factors could be used to correct all analytical data for instrumental effects. **HOWEVER**, they **do not correct** for analyte losses during preparation and extraction of biomineral matrices.

Correction factors are currently used in AAR analysis, derived historically from the analysis of collagen proteins. The recovered amino acid profiles were then compared to a

sequence based composition correction factor determined so that the observed profile matched the unknown composition (Collins, pers. coms.). In the absence of matrix-matched RMs, analysis of a well characterised natural compound would make sense. Correction factors thus determined should therefore correct for both losses during extraction and also instrumental and detector effects.

However, this assumes that;

- recovery of bone collagen is a suitable proxy for biomineral matrices,
- that the same extraction and analysis method has been applied, and
- that the correction is applied to the appropriate analyte, either the L-isomer, the D-isomer or the D/L value.

Currently, the correction factors applied are applied equally to the L and the D isomers. The problem with this is that when the corrected concentrations are used for determining the D/L value, because the correction has been applied to both sides, the correction cancels and gives exactly the same result as if no correction had been applied at all. There is the possibility that the correction should only be applied to the L-isomer as the collagen analysis was carried out on young material and likely to have very low amounts of the D form. However adjusting the L without being able to similarly adjust the D for losses and detection will unbalance the D to L ratio and produce inaccurate results. Alternatively, perhaps the adjustment should be applied to the final D/L value.

It is noted that the current “rt” corrections used by the laboratory are used as a multiplier. Equivalent values are easily obtained by taking the reciprocal of the CF value. Equivalent multiplication factors are compared in Table 4.28, to the previously derived CF correction factor and the current “rt” values used by the lab are also given alongside for comparison.

A comparison of corrected D/L values using the various correction options are provided in Table 4.29 at the end of the section. Data used has been based on a set of Asx data from a randomly selected analytical run for the 0.5d standard solution with known concentration values to give a known D/L value, and data from one of the opercula homogeneity evaluations discussed previously.

The results below indicate that use of the single estimated value for D/L correction produces the closest match to the known D/L in standard solution. Alternatively, for absolute accuracy, run-specific response factors could be determined if a known control material was placed in the analytical run with the unknown samples and applied to the

unknown concentrations or even final D/L values. It is also noted that the revised std sol correction values have a bigger impact on the final D/L than the current collagen derived values in the examples used above.

Because the new factors only represent instrumental correction, any analyte loss due to preparation and extraction, is likely to increase the final correction value even higher. It would therefore suggest that the current "rt" values may not be adequate for correction in biominerals.

Table 4.28: Current and alternative (std sol derived) D/L correction factors

amino acid	D/L correction factors		
	Derived from std sol		from lab
	÷ CF	x 1/CF	x rt*
Asx	1.210	0.827	0.929
Glx	1.136	0.880	0.970
Ser	0.836	1.196	0.964
Arg	0.966	1.035	0.949
Ala	1.135	0.881	0.896
Val	0.869	1.151	0.826
Phe	0.929	1.076	0.902
A/I	1.161	0.861	0.857
Leu	1.367	0.732	1.149

* note the "rt" factor is currently used to correct peak areas

4.5.4 Calibration Curves

Finally a brief mention about calibration curves to demonstrate how they could be used as an alternative to the internal standard approach described in the previous section.

Based on the collected mass of standard solution data, normalised peak areas of each L and D isomer for every amino acid was derived simply as Area(L or D-aa)/Area(LhArg). The individual values were then averaged to give single value estimates for each isomer. These values can then be plotted (y-axis) against the known concentrations from the preparation of the original solutions (x-axis).

The slope of the curve can then be used to determine the concentration of an unknown sample, if the normalised peak areas of the unknown are used to calculate the

predicted x value. Concentrations would then be adjusted subsequently, to take into account sample dilution and original sample weight, to return a final concentration in pmol/mg in the case of AAR.

In calibration, the uncertainties of the concentrations of the reference materials used are assumed to be minimal compared to the uncertainty of the instrument response, the concentrations are therefore plotted as the independent x variable and instrument response as the y. The use of reference or standard solutions in calibration will correct amino acid peak areas for instrumental effects and detector response but an independent check on analyte recovery from the matrix would usually be required by performing spiking experiments. If however matrix-matched reference materials were used as calibration standards (calibrants), then after having been taken through the entire extraction and measurement procedure, the resultant peak areas would reflect and correct for all unrecoverable analytical losses.

It should be noted that in Excel, the application of the least-squares regression, minimises the sum of the squared y variable residuals only. Therefore by convention, the line of regression is y on x, and the y variable is determined from the independent x variable, i.e.; $y = mx+c$ (where m is the gradient and c the y-axis intercept). The procedure applied by calibration determines x from y, such that $x = (y-c)/m$.

A procedure for determining the uncertainty of derived concentrations is described by the RSC Analytical Methods Committee (ref no22). This has been applied to the same data in the previous examples as a comparison. However, whilst this method enables the determination of uncertainty estimates for concentrations from unknown samples, this would normally be subsumed into the repeatability precision estimate, and therefore need not concern us. However it is informative to observe the effect graphically and compare differences between L and D isomers.

Using the opercula Asx homogeneity data and the 0.5d standard solution data used previously, D/L values have been determined from calibration curves for illustration. Calibration curves for the opercula L-Asx and D-Asx homogeneity data are shown in Figure 4.33 and Figure 4.34 for illustration. Both curves have been assumed linear and fixed at the origin. Data are also shown in Table 4.29.

If this technique was adopted in practice, then individual calibration curves would be required for every L and D isomer, and reference solutions would need to reflect adequately the full range of expected concentrations from routine samples.

Depending on the biomineral matrix concerned, this can be vary widely and a number of appropriate RMs would be required to cover the full range and avoid extrapolation. Calibration checks would be carried out by use of a single RM in each analytical run to ensure the system remains stable. Full calibration checks and recalibration (regression line adjustment) when necessary would be carried out at intervals, perhaps monthly and after any significant maintenance work, servicing and changes in batches of

Table 4.29: Comparison of correction methods on Asx D/L value in 0.5d std sol and the opercula PT test material.

0.5d std sol		**LhArg Peak Area**	957.9
D/L value	**0.414**	**LhArg Conc (M)**	0.00001
L-Asx Conc (M)	6.06E-05	**L-Asx Peak Area**	6440.6
D-Asx Conc (M)	2.512E-05	**D-Asx Peak Area**	3273.8
source	**correction factor**	**applied to**	**D/L value**
no correction			0.508
collagen (rt)	0.9286	both L & D isomers	0.508
	"	L-isomer only	0.547
	"	D/L value	0.472
std sol (F)	1.0217	L-isomer	0.420
	1.2362	D-isomer	
std sol (CF)	1.2099	D/L value	0.420
std sol (F)	1.1094	L-isomer	0.414
(run-specific)	1.3605	D-isomer	
std sol (CF)	1.2264	D/L value	0.414
calibration		L-isomer	0.416
		D-isomer	
red text = true D/L value			
opercula		**LhArg Peak Area**	1291.2
mass mg	4.58	**LhArg Conc (M)**	0.00001
vol rehyd (µL)	91.60	**L-Asx Peak Area**	12371.9
rehyd (µL/mg)	20	**D-Asx Peak Area**	7202.0
source	**correction factor**	**applied to**	**D/L value**
no correction			0.582
collagen (rt)	0.9286	both L & D isomers	0.582
	"	L-isomer only	0.627
	"	D/L value	0.541
std sol (F)	1.0217	L-isomer	0.481
	1.2362	D-isomer	
std sol (CF)	1.2099	D/L value	0.481
calibration		L-isomer	0.477
		D-isomer	

rehydration fluid as the normalised responses may shift. Each instrument would also require individual calibration. Whilst this may seem like a lot of work, calibration would normally be included in the initial method validation, prior to its introduction into routine use and determines whether the curves should be linear or not and whether they pass through the origin, or not. It then just becomes a matter of monitoring and making the occasional adjustment.

Figure 4.33: Calibration curve for L-Asx in 0.5d standard solution

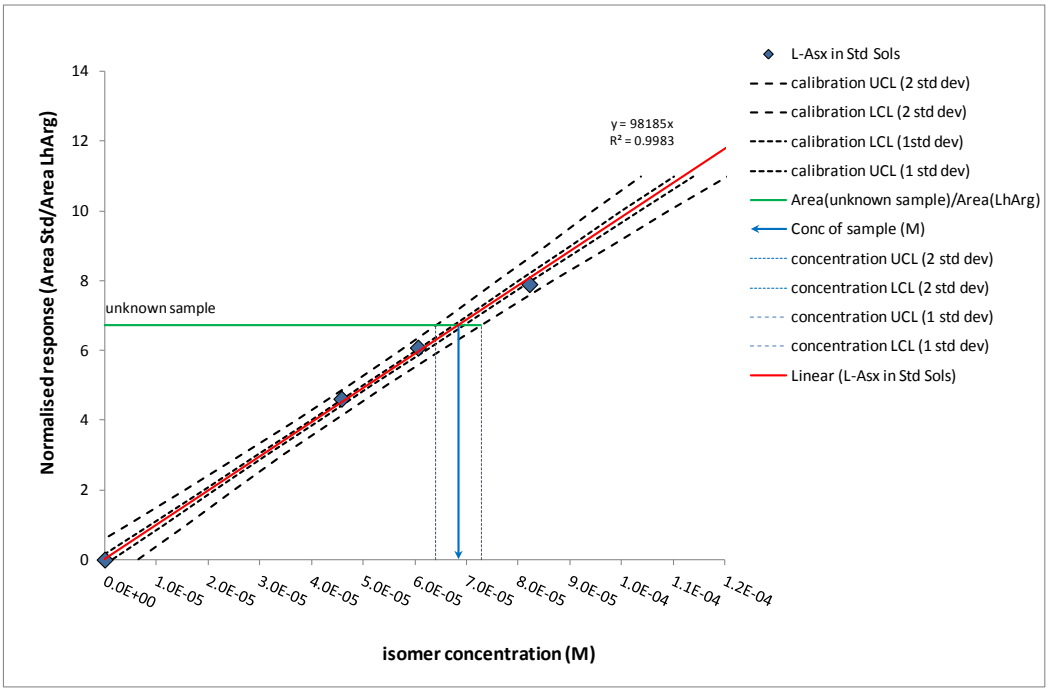
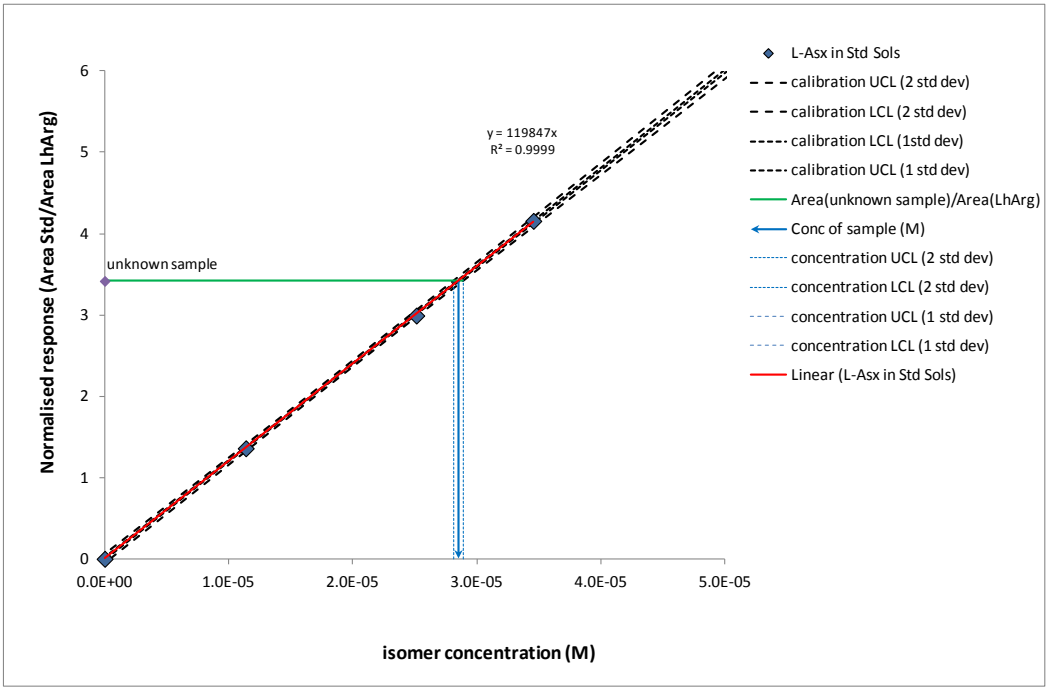


Figure 4.34: Calibration curve for D-Asx in 0.5d standard solution



4.6 Conclusion

A preoccupation with the measurement of D/L values has probably been a false economy. A slight shift in the determination of an L-concentration or a D-concentration will alter the D/L value. Therefore it would be far preferable to ensure that the L and D concentration values have been determined accurately, and by controlling the input values, the D/L output can take care of itself.

However, initial observations of the method in practice reveal that whilst standards and blanks are run in routine analysis, results are not used to monitor performance or correct for bias leading to the absence of suitable control measures necessary to ensure statistical control in routine application. In the absence of a validated method and no suitable CRMs, bias evaluation becomes problematic. Under these circumstances the only independent way of evaluating bias is by comparison against other laboratories or another method. Due to the same problem with the lack of RMs, comparison against other methods would be interesting but as these too are uncorrected for bias, may be equally systematically inaccurate, in different ways. The only remaining option would be comparison against other laboratories carrying out the same method and for this reason a proficiency study was designed and coordinated across as many AAR laboratories as could be included.

Long term, it will be necessary to extend this analysis to other biomineral matrices through the organization of inter-laboratory trials. Proposed matrices include standard solutions once again to compare the intra-laboratory variability with the inter-laboratory variability, artificially aged (through heat treating) ostrich egg shell, and existing mollusc shell inter-laboratory calibrants, previously prepared in bulk to aid comparability between AAR laboratories and help ensure some consistency. However, whilst this material has been used in a much earlier inter-laboratory study, techniques have been refined and GC and IE analysis have tended to be replaced by RP. Thus it is timely to reassess the material and provide reference values which can be subsequently used for validation, training and calibration.

Chapter 5.Inter-Laboratory Proficiency Study

Initially submitted to Quaternary Geochronology; November 2011

Author contribution approx 90%

Results from an Amino Acid Racemization Inter-Laboratory Proficiency Study; Design and Performance Evaluation

Joanne Powell[a], Matthew J. Collins[a], James Cussens[c], Norman MacLeod[d], Kirsty E.H. Penkman[b]

[a] BioArCh, Department of Archaeology, University of York, Heslington, York YO10 5DD, UK

[b] BioArCh, Department of Chemistry, University of York, Heslington, York YO10 5DD, UK

[c] YCCSA, Department of Computer Science, University of York, Heslington, York YO10 5GH, UK

[d] Department of Palaeontology, Natural History Museum, Cromwell Road, London SW7 5BD, UK

Corresponding author; Jo Powell; email jp588@york.ac.uk, tel; +44(0)1904 328806

5.1 Abstract

It is nearly thirty years since the last inter-laboratory study was carried out for amino acid racemization (AAR) analysis using powdered fossil material (Wehmiller, J. F. (1984) Interlaboratory Comparison of Amino Acid Enantiomeric Ratios in Fossil Pleistocene Mollusks. Quaternary Research, 22, 109-120). Since then there have been major changes in sample preparation and instrumentation, and it was considered timely to coordinate a new inter-laboratory study in support of current methodologies. In 2010, two such studies were undertaken. The first of these, coordinated by Wehmiller (2012; (this edition)), used

homogeneous hydrolysates of Pleistocene mollusc and eggshell materials and focused on the agreement of analytical measurements between laboratories, without interference from differing sample preparation procedures. The second (this study) was designed specifically as a proficiency test to compare the performance of laboratories carrying out their routine methods, including extraction. Participants were sent one dried sample of a mixed amino acid standards solution and five homogeneous powders: two Pleistocene mollusc test materials prepared from material (ILC-A) supplied and used by Wehmiller in previous inter-laboratory studies (1984; 2012 (this edition)), one Pleistocene opercula test material from the terrestrial gastropod, *Bithynia tentaculata*, and two heat-treated modern ostrich eggshell test materials. Previous AAR inter-laboratory evaluations have concentrated on comparisons of precision estimates, since it is the difference between amino acid enantiomeric/diastereomeric ratios which is utilized in aminostratigraphy and chronology building. However, inter-laboratory differences have been previously observed and preclude direct comparison of D/L data between laboratories, and therefore the wider application of the technology. Results from this study demonstrate that whilst individual laboratory precision may be excellent (often less than 1% for replicate measurements, suggesting good control of random error influences), agreement between methods, or even laboratories carrying out the same method, may be very different. Trueness evaluation (determined as the relative percentage bias) reveals the extent of the disagreement reflected by the inter-laboratory variability. Individual laboratory D/L value biases of 10-30% or more are not uncommon when compared to the consensus values. However, due to the limited number of laboratories submitting results, and because some methods are not sufficiently represented in this study, results should be seen as indicative and not absolute. No comment is made regarding the significance of any observed differences and no judgement is made as to which method may or may not be correct. Previously, AAR uncertainty estimates have been reported only as precision values, (i.e. the standard deviation of reported results expressed as the relative standard deviation, RSD% (or CV%)). However, bias is an essential component of measurement uncertainty. Here we demonstrate why bias contributions should also be included in uncertainty estimation and recommend that systematic error influences are controlled and corrected in the analytical system, where at all possible by the use of defined reference materials.

Keywords; Amino acid racemization, inter-laboratory comparison, proficiency test, accuracy, precision, bias, uncertainty, geochronology

5.2 Introduction

5.2.1 Amino Acid Racemization

Amino acid racemization (or epimerization[5.1] for molecules with two carbon centres) is a diagenetic process that occurs naturally following protein synthesis. The process involves the slow inter-conversion between the two chiral forms of amino acids, the building blocks of proteins, from the original *laevo* (L-form) in life to the *dextro* (D-form). Conversion of the L to D form continues until equilibrium is reached, which for most amino acids is usually equal to 1. This process can take many hundreds of thousands of years, thus the D to L ratio or D/L value can be used as an indicator of time. This technique has been particularly successful in dating Quaternary sediments using protein decomposition in fossil biominerals. The rates of racemization for the 20 or so naturally occurring amino acids and are highly temperature dependent, matrix and species specific (Wehmiller & Miller, 2000; Miller & Clarke, 2007). As the thermal history of a site is rarely known, it can be difficult to use AAR kinetic and temperature modeling to determine absolute age estimates (Clarke and Murray-Wallace, 2006; Kosnik *et al*., 2008). For this reason, much research tends to apply the technique as a relative stratigraphic tool (e.g. Miller *et al*., 1979; Miller & Hare, 1980; Bowen *et al*., 1989; Wehmiller *et al*., 2010; Penkman *et al*., 2011), with numerical ages only being assigned to samples within a defined locality using independently calibrated material (e.g. Hearty and Kaufman, 2009; Murray-Wallace *et al*., 2010; Demarchi *et al*., 2011), or by adopting a dual approach using both calibration and kinetic modelling (e.g. Wehmiller *et al*., 2010; Wehmiller *et al*., 2012a; Wehmiller *et al*., 2012b). The assumption is that if sites share the same temperature history, any observed D/L differences can be interpreted as relative age differences. Similarly, it becomes possible to use D/L values for palaeothermometry, (as indicators of relative temperature variation between same age sites), once independently dated using appropriate techniques (e.g. Kaufman, 2003; Owen *et al*., 2007; Bright *et al*., 2011; Reichert *et al*., 2011).

The last 30 years have seen significant changes in AAR analysis. Early research based on ion-exchange liquid chromatography (IEx) was able to separate L-isoleucine from its diastereomer D-alloisoleucine, yielding a D-Aile/L-Ile value, or often termed A/I value. As methods developed, it became possible to detect and measure increasing numbers of chiral

[5.1] Note; The more general term 'racemization' will be used hereafter to refer to both racemization and epimerization.

pairs of amino acids, from six or seven using gas chromatography (GC) to ten or more routinely determined today using reverse-phase HPLC (RP). These developments have continued to advance its application in routine analysis. AAR now requires mg sample sizes, is relatively fast and with inexpensive preparation and analytical costs, is a useful dating method with the potential to provide age estimates that cover the entire Quaternary (Wehmiller & Miller, 2000).

5.2.2 Accuracy or Precision ?

Clearly, the accuracy of age estimation relies heavily on the accuracy of the analytical data. Measurement accuracy is defined by the International Vocabulary of Metrology (otherwise known as the VIM) as "*the closeness of agreement between a measured quantity value and a true quantity value of a measurand*" (JCGM 200, 2008; p21, para. 2.13). However, accuracy is a concept and cannot be given a numerical value, although a measurement is said to be more accurate if it has a smaller measurement error. Measurement error is comprised of both random and systematic error components, determined as measurement (im)precision and measurement bias respectively. Where significant bias is detected, measurement results should always be corrected (EURACHEM/CITAC, 2000; JCGM 100, 2008). However any uncorrected bias, together with the uncertainty associated with that bias, plus precision estimates, reflect the overall doubt or the uncertainty associated with a measurement (Barwick and Ellison, 2000; EURACHEM/CITAC, 2000;). Precision can be determined through repeated measurements of the same or similar substance under repeatability or reproducibility conditions. Bias, however, requires evaluation against a true or reference value, which makes bias evaluation challenging in the absence of reference materials, as is currently the situation for AAR analysis. For this reason, in the absence of defined reference materials, most AAR uncertainty estimation focuses on precision evaluation.

For the majority of aminostratigraphic and geochronology studies, analyses are performed within a single laboratory, and therefore the most important factor is precision, as it is the differences between the D/L values which are used. Consequently, ensuring internal consistency within an individual laboratory is often all that is required, and the inability to correct for bias is not an issue. Nonetheless, precision estimates themselves will vary depending on sample type and analytical conditions. For example, measurements from several samples, perhaps individual shells, will likely show greater variability than estimates derived from subsamples of the same shell run within the same time frame, which in turn will

show greater variability than repeated measurements of the same subsample. Similarly, measurements taken during a single analytical run can be expected to show better precision than those obtained over several runs, and it would not be unreasonable to expect the precision of measurements from a simple solution of amino acid standards to be smaller than those which are matrix bound. These effects have been observed by several authors. For example, Wehmiller and Miller (2000) have reported intra-laboratory precision estimates of 2% for repeated instrumental determinations by gas chromatography (GC) of the same hydrolysate, between 3-5% for multiple analyses of different fragments of the same material, and between 5-10% for multiple samples from the same sample location. More recently, in an evaluation of marine molluscs from the North Carolina coastal plain (Wehmiller *et al*., 2010), analytical precision for most amino acids was reported as being better than 2% (based on D/L values from multiple chromatograms of the same derivative using GC). CV% values based on multiple shells approximated to about 6%, but the range varied for different amino acids, on a few occasions exceeding 30%. Uncertainty estimates for repeated analyses by RP of intra-laboratory reference solutions (approximate D/L of 0.5) carried out over several years have been reported as 1.5% for aspartic acid D/L values and 1.4% for glutamic acid D/L values (Kosnik *et al*., 2008). For a reference solution with a lower D/L ratio, (approximately 0.09), higher uncertainty estimates were obtained; 3.7% and 3.8% respectively, although an average of 1.4% is suggested as being representative of the analytical uncertainty for both aspartic acid and glutamic acid based on the mid-range D/L values.

Definition of the precision characteristics of an analytical system for target matrices and concentration / value range is a necessary and fundamental component of method validation. Knowledge of repeated measurements of in-house standard solutions is an important aspect of internal quality control. However, analysis of simple solutions free from matrix effects are not necessarily representative of the precision of solid matrix bound analytes, and their use to derive uncertainty values risks underestimation, which will then be carried forward to any subsequently derived numerical age confidence limits.

5.2.3 Previous AAR Inter-laboratory studies

In addition to the observed intra-laboratory matrix and sample variability discussed, several authors have observed important inter-laboratory and method related differences in D/L values from previous comparability studies (Bada *et al*., 1979; Kvenvolden, 1980; McCartan *et al*., 1982; Wehmiller, 1984; Hollin and Hearty, 1990; Bakeman, 2006; Wehmiller, (this edition)). Early inter-laboratory comparisons focused on GC method variations (Bada *et*

al., 1979) with ion exchange liquid chromatography (IEx) also being used for isoleucine determination (Kvenvolden, 1980, Wehmiller, 1984). In contrast, RP is the method more commonly used today. In Wehmiller's original study, eleven laboratories (using three GC methods and one IEx method) were each given six different materials to analyse: three marine mollusc shell powders (Inter-Laboratory Comparison materials or ILC A, B and C) and their respective desalted hydrolysates. Performance evaluation was carried out by a qualitative comparison of CV% values achieved by each laboratory. For example, for alanine, aspartic acid and glutamic acid, precision estimates ranged between 3-8%, for leucine and phenylalanine, 5-10% and for isoleucine, proline and valine, between 10-18%. Wehmiller (1984) reports that whilst CV%s for powders compared to the hydrolysates did not indicate significant differences, the median CV% from all the results of 9.6% for powdered samples and 6.5% for liquid samples, were higher than the 2-5% typically reported by an individual laboratory, and observed that significant differences between laboratories' results could lead to 25% differences in estimated age. As a result, Wehmiller called for the need for reference standards in routine analysis to ensure comparability more than twenty-five years ago. More recently Bakeman (2006; Bakeman and Wehmiller 2006), reported a 0.4% bias between GC and RP for aspartic acid, with RP giving the higher readings; a 6.8% higher systematic offset for isoleucine ratios by GC compared to RP and 1.9% compared to IEx are also reported. A further 4.6% difference was observed between GC and RP for glutamic acid D/L values, with as large as 25% for valine D/L values, with RP giving the higher readings in both cases (Bakeman, 2006; Bakeman and Wehmiller 2006).

Clearly there are noticeable discrepancies between the closeness of the intra-laboratory precision estimates achievable and comparability of data between different methods and/or laboratories, which inevitably could affect any subsequently derived numerical ages and their confidence intervals. This strongly suggests the presence of additional uncertainty contributions, due to unaccounted-for bias arising from analytical differences between methods and/or laboratories. For this reason, AAR dating is predominantly currently carried out by laboratories independently from each other and precludes direct comparison of D/L data.

Evaluation of trueness or bias estimation is an important component of measurement uncertainty determination. Every effort should be made to ensure systematic error influences are reduced to a minimum and any significant bias should always be corrected for, unless the method is empirical and by definition makes no correction (JCGM 100, 2008; EURACHEM/CITAC, 2000). Bias determination is usually carried out during

method validation and requires a comparison of the analytical result against a true or reference value, which by convention, usually involves the repeated analysis of a matrix-matched Certified Reference Material (CRM) or other suitably defined reference material (Thompson *et al*., 2002). As the use of CRMs eliminate laboratory, method and even run bias (Thompson, 2000), they are frequently used for calibration, to accurately transform instrument response into concentration units / values, and thus correct analytical results for bias. However, traceability back to standard materials with reference values with known uncertainty is currently impossible for AAR geochronology due to the absence of certified reference materials. The use of Wehmiller's original ILC powders has been suggested and are used routinely by some laboratories for internal quality control, but issues regarding method and laboratory bias have made defining reference values thus far problematic.

In the absence of a suitable CRM or reference material, spiking and recovery experiments with standard solutions might be used to determine losses during extraction and due to matrix effects (Thompson *et al*., 2002). However the validity of such data makes two important assumptions: i) that after extraction, the sample analyte is in the same chemical form as the spike; and ii) that the extracted analyte is as equally recoverable as the spike (Thompson, 2000). Nonetheless, accurate determinations of recovered concentrations / values may still require the use of calibration standards (Vanatta and Coleman, 2007), which again are not available for AAR analysis.

In the absence of comparable materials, comparability against other analytical data is the only remaining option. This may be an intra-laboratory comparison against data determined using a published or reference method, or an inter-laboratory comparison such as a collaborative trial, or results from proficiency tests (Thompson *et al*., 2002). A method specific inter-laboratory collaborative trial eliminates method bias, but incorporates laboratory bias into the between-laboratory precision estimate. It is thus designed to evaluate both repeatability and overall precision, expressed as the reproducibility of a method (Horwitz, 1995; ISO 21748, 2010). In contrast, a proficiency test is non-method specific and can evaluate a laboratory's routine method and individual laboratory bias by comparison against the assigned value, usually derived as the consensus of submitted results (Thompson *et al*., 2006). In addition, test materials left over after the end of a proficiency test (or even a collaborative trial) can later act as suitable matrix specific reference materials. As the value of the analyte has been determined by a consensus, there is minimal bias associated with it and it has a known uncertainty value.

The inability to evaluate laboratory and method bias routinely in AAR geochronology has important implications for the proper reporting of measurement uncertainty. Uncertainty estimates reported simply as an estimate of precision require further qualification to enable direct comparability. Precision is defined by the analytical conditions (JCGM 200, 2008):

- repeatability conditions: repeated measurements carried out during a single analytical run reflecting random error effects only, (JCGM 200, 2008; p23, para. 2.20).
- reproducibility conditions: repeated measurements carried out during multiple analytical runs, usually over a number of days or a period of time. Strictly speaking this involves an inter-laboratory element too, reflects random and systematic error effects (JCGM 200, 2008; p24, para. 2.24).
- intermediate conditions: (an intra-laboratory reproducibility equivalent) repeated measurements carried out during multiple analytical runs, usually over a number of days or a period of time but without the inter-laboratory element (JCGM 200, 2008; p24, para. 2.22).

Reproducibility precision estimates represent the overall imprecision of the measurement system. It may be determined by carrying out an analysis of variance (ANOVA) and combining the repeatability precision (within sample variance) with either the between-run precision estimates for intra-laboratory reproducibility (intermediate) precision, or the between-laboratory precision estimates for inter-laboratory reproducibility precision.

Examples of such definitions might include:

- instrumental repeatability precision (replicate analyses of the same subsample (i.e. single hydrolysate) analysed in a single run);
- sample repeatability precision (analyses of different subsamples (i.e. multiple hydrolysates) from the same sample analysed in a single run);
- site repeatability precision (analyses of different samples from the same horizon or site, analysed in a single run);
- instrumental (intermediate) reproducibility precision (replicate analyses of the same subsample (i.e. single hydrolysate) analysed in separate runs over several days by the same laboratory);

- sample (intermediate) reproducibility precision (analyses of different subsamples (i.e. multiple hydrolysates) from the same sample analysed in separate runs over several days by the same laboratory);
- site (intermediate) reproducibility precision (analyses of different samples from the same horizon or site, analysed in separate runs over several days by the same laboratory).

In the examples just given, the influence of error on the measurement result is gradually increasing and will be reflected in the final precision estimate. Such definitions therefore become important in order to ensure appropriate comparability of data.

Nonetheless, in the absence of bias determination, all quoted precision values risk underestimating the overall uncertainty to greater or lesser extents. So far, previous inter-laboratory comparisons have observed and reported on bias effects but have been unable to fully address the issue. A proficiency test was therefore coordinated to help laboratories observe the effect of their own overall bias when compared to other participant's data, and consider the implications for uncertainty estimation.

5.2.4 Proficiency Testing

It has long been widely appreciated that participation in inter-laboratory studies is a valuable tool enabling method comparisons and development (Thompson *et al*., 2006) and provides independent proof of competence (UKAS, 2004). Proficiency testing (PT) is a specific inter-laboratory assessment providing a formalized evaluation of accuracy against a consensus value and enabling an objective comparison with other laboratories' data, which is an important indicator of bias. Accuracy and by inference, performance, is characterized by elements of both precision and trueness. A laboratory may be inaccurate due to systematic bias effects, random error influencing poor repeatability, or both. In the absence of Certified Reference Materials (CRMs) for bias determination, participation in a proficiency test can provide a valuable alternative for laboratories.

Proficiency testing is commonly encountered in sectors that rely heavily on regulation and compliance, such as medicine and public health, forensic science, chemical and geochemical analytical services, manufacturing industries, calibration and engineering, food and feed industries. Today more than 1,300 PT schemes worldwide are listed on the EPTIS[5.2] website. Participation in such a scheme is also a requirement of analytical laboratories seeking accreditation to ISO 17025 (2005).

[5.2] European Proficiency Testing Information Service; http://www.eptis.bam.de/en/about/what_is_eptis/index.htm

The regular analysis of an independent quality control material forms a valuable part of external quality control (EQC), enabling comparability on a much wider scale with other laboratories, analysts and methods. As such, it is an essential element of any laboratory's Quality Assurance (QA) programme, together with the use of validated methods and internal quality control (IQC) procedures.

Proficiency testing acts as a snapshot of laboratory accuracy at any one moment. Whilst performance in individual rounds can identify unexpected error influences needing investigation, long-term trends are probably of greater value and can be observed using control charts (Thompson *et al*., 2006). The spread of results from a laboratory over a period of time should be comparable with that laboratory's own evaluation of uncertainty. Methods for the use of PT data in determining measurement uncertainty have also been described (EURACHEM/CITAC 2000; EUROLAB, 2006; EUROLAB, 2007; Magnusson *et al*., 2004).

5.3 2010-11 AAR Proficiency Test

5.3.1 Design and Organisation

Eight AAR laboratories from five countries agreed to take part in the 2010-11 study. Laboratories were sent six different test materials and were asked to use their routine method of analysis on each. Due to the low numbers of currently active laboratories routinely carrying out AAR analysis, laboratories possessing more than one instrument or having more than one member of staff competent to carry out the analysis, were asked to submit more than one set of results, i.e. one for every instrument / analyst combination, raising the number of potential sets of results to eighteen by RP, four by GC and 2 by IEx. Increasing the data set this way was important to help reduce the uncertainty of the assigned values, which otherwise may have been unreasonably large and would have had a significant effect on the evaluation of performance. It was understood that by increasing the number of submitted results from individual laboratories the risk of laboratory bias may also increase, which in-turn could bias the derived consensus values. However, the benefit gained from reducing the uncertainty of the assigned value was considered a priority, and individual laboratory effects could not be predicted. For a '*well-behaved*', unbiased data set, ideally results would be expected to be symmetrically distributed either side of the consensus value.

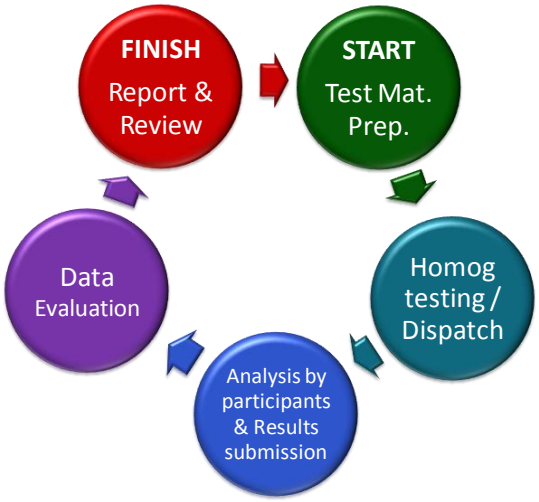


Figure 5.1: Schematic showing the general organisation of a proficiency test

The proficiency study was carried out according to documented guidelines (Thompson *et al*., 2006; ISO 13528, 2005). Figure 5.1 shows the general organization of the scheme. Test materials were prepared and dispensed into individual vials. Ten test materials from each bulk were randomly selected and assessed for homogeneity by RP using a modified version of the standard AAR method (Kaufman and Manley, 1998), prior to being dispatched to participants in July 2010. A total of fifteen sets of results were returned, each of which was given a unique identifying number. Participants were able to submit instrumental data such as peak heights, areas and concentrations, in addition to the D/L values requested for the study. Participants had the opportunity to submit both the free amino acid (FAA) and the total hydrolysable fractions (THAA). However, as only one laboratory submitted FAA data, this was not documented or assessed. Many laboratories chose to provide instrumental replicate values and these data were evaluated using normal summary statistics to derive precision estimates. Where relevant, the mean values of replicate data then underwent a statistical evaluation for performance. Finally an evaluation of measurement uncertainty was carried out to demonstrate how proficiency test data can be used and how the various precision and bias elements contribute to the overall uncertainty budget and affect confidence levels. Details of the statistical procedures used in the evaluation of performance are given in subsequent sections. Confidential reports for each test material were produced and sent electronically to participants. Anonymous copies of these reports can be accessed at; http://www.neaar.co.uk.

5.3.2 Test Materials

Six test materials were prepared and sent to participants. These comprised five dry powders and one mixed amino acid standards solution (with D/L values of approximately

0.5). The five powders included Pleistocene mollusc opercula, modern heated ostrich eggshell A and B, (OES (A) and OES (B)) and Pleistocene mollusc shell A and B, (Mollusc (A) and Mollusc (B)). The Opercula Test Material was prepared from a 2 g bulk of individual *Bithynia tentaculata* opercula, removed from sediment taken on 28 July 2005 from the mid-Pleistocene site of Funtham's Lane, approximately 5 km east of Peterborough, Cambridgeshire, UK (Langford *et al*., 2007; Penkman *et al*., 2007; Penkman *et al*., 2011).

Both ostrich eggshell materials were prepared from a single blown ostrich egg, obtained locally from an ostrich farm and subsequently heated to induce racemization. The mollusc shell materials were prepared from the ground bivalve *Saxidomus* bulk material, referred to as ILC-A in Wehmiller's inter-laboratory comparison studies (Wehmiller, 1984; Wehmiller, 2012; (this edition)). The opercula and broken ostrich eggshell pieces were initially cleaned. Large pieces of extraneous matter were removed and the bulk material was then repeatedly washed in ultrapure water using a sonicator until the water remained clear. The cleaned material was then lightly covered and left to air dry for 48 hours. The pieces of broken ostrich eggshell were then heated to 140^{o}C for 8 hours. This cleaned, dried material, together with the coarsely powdered mollusc shell were each ground using a sterile pestle and mortar and sieved (to ≤ 250 μm), then tumble-blended overnight on a roller mixer.

The powdered opercula, half of the ostrich shell and half of the mollusc shell material were then bleached, for 48 hours using 50 μL of 12% NaOCl per mg of powder. The bleach was removed and the powder rinsed with ultrapure water up to six times using a vortex mixer followed by centrifugation to pellet the solids in between washes. A final wash with methanol to remove any remaining water was carried out before the material was again lightly covered and left to air dry.

Individual 20 mg sub-samples of the cleaned, dried and bleached material were weighed into sterile glass vials, labelled as Opercula, OES (A) and Mollusc (A) Test Materials. The remaining unbleached materials were also weighed into sterile glass vials, and labelled as OES (B) and Mollusc (B) Test Materials.

Individual 20 μl sub-samples of an existing in-house standard solution (Penkman, 2005) were measured into sterile plastic 3 mL eppendorf tubes and labelled. Each aliquot was then dried over-night using a centrifugal evaporator and stored at room temperature to avoid condensation, prior to distribution.

The original standards solution was made up by the addition of thirteen D-amino acid powders (Ala, Arg, Asp, Glu, His, Aile, Leu, Met, Phe, Pro, Ser, Thr and Val), dissolved in HPLC

grade water, to a diluted liquid L-amino acid standard (Sigma; AA-S-18; containing L-amino acids at a concentration of 2.5 µmol/mL in 0.1 N HCl), diluted to 0.1 mM with HPLC grade water. D-amino acid powders were added to give final mid-range D/L values of ≈ 0.5.

All test materials were stored at room temperature prior to distribution. Participants were previously asked to notify the organizer with details of their proposed analytical method and were sent the appropriate number of individual test materials necessary to give sufficient bulk material required by the different methods. Those using RP were sent single individually numbered 20 mg test materials, those using ion-exchange HPLC (IEx) were sent three individual test materials (60 mg total) and those using gas chromatography (GC) were sent ten individual test materials (200 mg total), for each of the six materials. As homogeneity had been assessed and could be assured between the individual 20 mg sub-samples, participants receiving multiple test materials were asked to pool the contents to obtain the required quantity, rather than simply being sent a larger sample.

5.3.3 Homogeneity Evaluation

Ten randomly selected test materials were sub-sampled to give 10 duplicate samples (20 subsamples, i.e. 10 x a and b). These were then analyzed in a random order under repeatability conditions, for total hydrolysable amino acids (THAA) using a modified version of Kaufman and Manley's (1998) RP method. Asparagine (Asn) and glutamine (Gln) are known to undergo deamidation to aspartic acid (Asp) and glutamic acid (Glu) respectively during the hydrolysis extraction phase (Hill 1965). Because of this, chromatogram peaks are determined as Asx (representing the combined Asn plus Asp) and Glx (representing Gln plus Glu).

The purpose of carrying out homogeneity testing is to determine that any variation in composition between individual test materials is negligible compared to the variation in measurement determinations carried out by participants of the proficiency test. Due to the time and expense of preparing homogeneous test materials and carrying out the analysis, it is reasonable to start with the assumption that test materials are already homogeneous, and by carrying out homogeneity testing we are looking for evidence of heterogeneity, rather than vice versa. The procedure for the assessment of homogeneity follows that given in the international standard ISO 13528:2005 and has been described fully elsewhere (ISO 13528, 2005, Fearn and Thompson, 2001, Thompson *et al*., 2006).

Resulting data were initially scrutinized for obviously anomalous values, such as reporting errors giving values greater or less than 10 times the average. Plotting data in run

order helps to identify trends, stability issues or measurement problems. Data were then sorted and sub-samples re-paired prior to being assessed using a Cochran's outlier test and subsequent statistical evaluation.

The target value for standard deviation, σ_p (see section 5.3.4.3) used in the assessment of performance of participants' results is also used in the assessment of homogeneity and is usually derived from a data source external to the data under evaluation. However, due to the absence of a suitable value for σ_p (see section 5.3.4.4), in all cases, σ_h the target standard deviation for sufficient homogeneity, was set as the minimum value necessary to ensure fitness-for-purpose according to recommended criteria, i.e. that σ_h was at least twice the analytical precision (repeatability) and that the allowable sampling variance was sufficient to accommodate the observed between-sample differences (Fearn and Thompson, 2001, Thompson *et al*., 2006).

σ_h values thus define the level of observed homogeneity within a test material for each amino acid and can be used as a minimum value for σ_p in the performance assessment.

5.3.4 Performance Evaluation

The purpose of this evaluation is to provide a clear and independent statistical evaluation and comparison of participants' results. In routine analysis a laboratory's evaluation of analytical competence is often restricted to intra-laboratory precision evaluation of repeated analyses, or the evaluation of bias using certified reference materials (CRMs). However, in the absence of a suitable, matrix-matched CRM with a known value and uncertainty, evaluation of method and laboratory bias can be impossible without the cooperation of additional laboratories. Estimations of precision may be excellent when taken in isolation, but may give rise to unrealistically small uncertainties.

Participation in a proficiency test provides the opportunity to evaluate analytical bias by comparing an individual laboratory's result against the assigned value for the test material. Performance is traditionally determined by the calculation of a z-score, calculated using the submitted result, a reference or assigned value and the target value for standard deviation, using a procedure recommended in the IUPAC/ISO/AOAC International Harmonised Protocol for the Proficiency Testing of (Chemical) Analytical Laboratories (Thompson *et al*., 2006) and the international standard; ISO 13528: 2005, such that;

$$z = (\bar{x} - \hat{X})/\sigma_p$$

where; $\bar{x}$ = the mean of participant's reported replicate results (or simply x for a single reported result), $\hat{X}$ = the assigned value, and σ_p = the target standard deviation. (Note that; $(\bar{x} - \hat{X})$ is the calculation for bias.)

Satisfactory performance is indicated by achieving a z-score no greater than 2, i.e. $|z| \leq 2$. The results of a typical chemical analysis will be normally distributed about the mean with a known standard deviation (approximately 95% of data will be expected to lie within 2 standard deviations either side of the mean and 99.7% within ± 3 standard deviations). Thus, it is considered 'satisfactory' if a participant's z-score lies within this range. It follows that if a participant's z-score lies outside $|z| > 2$ there is about a 1 in 20 chance that their result is in fact an acceptable result from the extreme of the distribution. If a participant's z-score lies outside $|z| > 3$ the chance that their result is actually acceptable is only about 1 in 300 (Thompson *et al*., 2006; ISO 13528, 2005).

5.3.4.1 The Assigned Value, $\hat{X}$

The reference or assigned value, $\hat{X}$, is the best estimate of the true concentration of each analyte. Depending on the nature of a test material, this can be done in a number of different ways, for example the use of a reference value from a CRM, use of a reference method, a consensus of expert laboratories, or the consensus of submitted results.

In determining the assigned value for a specific analyte, a robust mean (Ellison, 2002a; RSC Analytical Methods Committee, 1989; RSC Analytical Methods Committee, 2001) is often used as the best estimate in a large data set (an iterative algorithm which minimizes the effect of outliers and gives a fairer estimate of central tendency). However, for small data sets such as here, whilst the robust mean may still be preferable to the standard mean, the influence of extreme values may still be significant. In such instances, the use of the median may be more suitable, or even the mode.

When determining the appropriate measure of central tendency, the effect of the uncertainty of the assigned value $u(\hat{X})$ on performance assessment also needs to be given consideration. If there is too much uncertainty associated with the assigned value, i.e. either the number of submitted results is too small or the distribution of results is too large, then this can have an adverse impact by exaggerating observed bias. For the robust mean and median the uncertainty of the assigned value is;

$$u(\hat{X}) = \hat{\sigma}/\sqrt{m}$$

Where: m = the number of laboratory results used to determine the consensus and $\hat{\sigma}$ = the standard deviation of the robust mean or median absolute deviation (sMAD), (Note this is not the same as σ_p the target standard deviation used for calculating z-scores).

For the mode, $u(\hat{X})$ is taken to be directly equivalent to the standard error of the mode, (SEM).

5.3.4.2 Derivation of the Assigned Values, $\hat{X}$

In this study all assigned values have been determined as the consensus of submitted data, which due to the low numbers of participants involved, also equates to the consensus from expert laboratories.

The consensus for each amino acid in each test material was determined as the most appropriate measure of central tendency. However, whilst assessing the data, in many cases it became clear that the robust mean was strongly influenced by extreme values, resulting in skewed distributions with a high or low-end tail. At times this appeared influenced by method and on other occasions by an individual laboratory where more than one result was submitted using the same method, but carried out using a different instrument or analyst. In addition, when assessing the mode (Ellison, 2002b; RSC Analytical Methods Committee, 2006; Lowthian and Thompson, 2002), it became clear that due to the low numbers of results, potentially false modes were identified due to only a couple of values and in some cases only a single data point.

In cases where there were two evenly matched modes or where a smaller second mode was dominated by data using a specific method such as GC, it would not be appropriate to penalise these laboratories by comparison against an assigned value determined from the primary or first mode. In this study, there is no judgment being made as to which set of results is '*correct*', therefore, it would not be appropriate to calculate performance for GC results using an assigned value determined from RP values if the GC data clustered differently. In situations such as this where the method may be empirical, the mode should not be used. Regrettably submitted results by GC were limited, making it difficult to know whether these observed differences were genuine method differences or simply values at the extreme of the distribution of results.

For these reasons, the median was used as the most appropriate measure of central tendency for all amino acids. The median could be thought of as the ultimate robust mean, as it ignores the effect of all outliers and is independent of distribution, placing data

symmetrically either side of the mid-point. This allows for any asymmetry arising from bimodality to be seen in the histograms, but makes no judgment as to which is the correct mode.

Proficiency tests in principle tend not to be method prescriptive, unless methods are known to be empirical and produce different results. In a comparison of GC with IEx and RP results from the recent inter-laboratory study (Wehmiller, 2012; (this edition)), observed RP D/L values for aspartic acid and glutamic acid were approximately 5% and 4% higher respectively than those by GC, and for valine, RP gave 20% larger D/L values than GC. Whilst these observations may represent genuine empirical differences in quantification by these two methods, as with this study, data reported by a single laboratory should only be seen as indicative and not be used for generalizations, which should preferably be based on a consensus of several.

For this reason, in this study GC data were initially included with HPLC values and evaluated against the same assigned value. However, RP results have also been evaluated separately for comparison. Insufficient data prevented a separate evaluation for GC or IEx methods individually.

5.3.4.3 The Target Standard Deviation; σ_p

The target standard deviation σ_p describes how the data is expected to perform for a given analyte and / or test material and determines the limits of satisfactory performance.

These values are often derived from the relative reproducibility standard deviation (RSD_R) determined by collaborative trials (method specific inter-laboratory studies designed to validate and assign performance characteristics to measurement methods). The RSD_R describes overall precision estimates under conditions of best practice for a specified method for a given matrix/analyte/concentration (Thompson *et al*., 2006). The RSD_R may then be used for the assessment of proficiency test data for the same or similar matrix/analyte/concentration combination, as it provides an external precision estimate describing how analytical results are typically expected to behave between laboratories.

$$\sigma_p = \frac{RSD_R}{100} \times c \qquad \textbf{(5.3)}$$

Where: RSD_R = Relative Standard Deviation of Reproducibility from collaborative trial data, expressed as % and c = concentration, i.e., the assigned value, $\hat{X}$.

In the absence of collaborative trial data, the Horwitz equation (Horwitz *et al*., 1980; Horwitz, 1982; RSC Analytical Methods Committee, 2004) is widely accepted as a suitable predictive measure for the target standard deviation in some sectors. However, the Horwitz function is not necessarily suited to every type of chemical analysis (Powell and Owen, 2002), and in the absence of a suitable alternative, the use of perception or fitness-for-purpose criteria may need to be employed, taking into consideration any uncertainty in homogeneity of test materials.

The distribution of submitted results and uncertainty of the assigned value $(\hat{X})$ should be small by comparison to the target standard deviation, σ_p This ensures that the data are sufficiently tight to give a measure of confidence in the assigned value, $(\hat{X})$ and that the target value is not overly restrictive (Thompson *et al*., 2006; ISO 13528, 2005).

5.3.4.4 Derivation of the target standard deviations, σ_p

To date, there has not been an inter-laboratory collaborative trial carried out to determine single method precision parameters for AAR on fossil material. The Horwitz equation requires the measurement units to be expressed as a mass fraction, i.e. mg/Kg = 10^{-6}, which is not appropriate in the current study as D/L results are expressed as a ratio. Therefore, in the absence of a suitable external value for target standard deviation, it was necessary to use experience and perception to determine fitness-for-purpose assessment criteria.

The target value derived during homogeneity evaluation, σ_h, is an excellent indication of the observed variation between test materials and reflects the uncertainty due to random error effects. The relative value of σ_h expressed as a percentage; i.e. the RSD%, is a more useful value and can be used to set the minimum permissible value for σ_h. Whilst an inter-laboratory collaborative trial reproducibility standard deviation (RSD_R%) would also reflect the additional laboratory component of variation, in the absence of such data, it nonetheless makes a good starting point for evaluating submitted results and provides a minimum fitness-for-purpose target value.

During the statistical evaluation of data, it was observed that for some amino acids in some test materials, the homogeneity target value was too wide compared to the submitted data for the test. Comparison of data between Table 5.1 and Table 5.2, shows that homogeneity precision values (Table 5.1) for serine in Opercula, OES (B) and Mollusc (B), valine in OES (B) and Mollusc (B), alanine in Mollusc (A) and leucine in Mollusc (B), were all wider than the RP inter-laboratory precision of submitted results (Table 5.2). This suggests

that the precision between laboratories in some instances was better than that observed between samples analyzed by a single laboratory under repeatability conditions for homogeneity!

5.3.4.5 Relative bias %

While these observations were surprising, it posed some difficulties in using objective fitness-for-purpose criteria for the determination of the target values for standard deviation and calculation of z-scores.

In order to overcome this problem and in the absence of independently determined performance criteria, it was decided to present the data as an assessment of relative bias (%) (Powell and Owen, 2002; Thompson and Wood, 1993), such that;

$$Relative\ bias\ \% = ((x - \hat{X}) \,/\, \hat{X}) \times 100 \qquad \textbf{(5.4)}$$

In this way it was possible to represent participants' results graphically as histograms in a similar way to z-score charts, but with the 2σ satisfactory range being given as a plus and minus relative percentage bias (%), rather than being expressed as a standard deviation.

When calculating z-scores, the use of a standard deviation, σ_p as the denominator acts to normalize results. This enables performance between different analytes or between different test materials to be compared on a common scale, but requires the target value (σ_p) to be scaled appropriately to the individual analyte or matrix. However, using the assigned value ($\hat{X}$) as the denominator, and calculating the relative percentage bias, still permits a comparison between analytes and test materials, but on a common percentage scale, thus providing perhaps a slightly more intuitive presentation of observed bias for individual results. It also uniquely presents the full extent of observed bias and allows for these differences (which was more significant for some amino acids and test materials than others) to be fully appreciated.

Therefore, for this study, performance was not determined by the calculation of z-scores but rather by an evaluation of relative bias. Satisfactory performance was assessed as plus or minus twice the standard deviation of the assigned value ($\hat{\sigma}$), representing 95% confidence limits, i.e. ± 2$\hat{\sigma}$

5.4 Results & Discussion

5.4.1 Homogeneity

D/L values from the homogeneity evaluation were plotted in sequence run order as a visual inspection and to pick up instrumental drift or other analytical inconsistencies. In addition, data were plotted as pair-wise duplicates as a demonstration of the level of agreement within and between the ten sets of samples for each amino acid and to identify potential Cochran's outliers. Target values of σ_h the standard deviation for sufficient homogeneity, were determined according to the within-sample and between-sample criteria previously described. σ_h is a measure of the imprecision observed for a specific analyte between different samples of test material analyzed under repeatability conditions. σ_h therefore represents the expected variability due to sampling, a smaller value indicating closer agreement than a wider one. In practice, additional variability due to method and between laboratory differences will make the imprecision observed between participants' results much larger. For this reason, any target standard deviation value, σ_p used in the subsequent evaluation of submitted results and calculation of z-scores will be at least equal to, but often larger than σ_h.

Table 5.1 shows the mean value for each analyte in each of the six test materials, together with values of σ_h given as a standard deviation and as a relative standard deviation, RSD (or CV) expressed as a percentage. Both values for σ_h have been given because for analytes with particularly low D/L values (e.g. the ostrich eggshell materials), the observed standard deviation may be very small, but when expressed as a value relative to the mean, the CV% value becomes elevated. This is because a small variability at low concentrations has a bigger influence on a low value mean than it would on a higher one.

Thus for comparisons between test materials with different analyte levels, the CV% can be misleading, so it is perhaps better to compare σ_h values more simply as a standard deviation.

For the **standard solution test material (D/L ~ 0.5)**, with the exception of D-alloisoleucine/L-isoleucine which has known reproducibility issues with RP, all amino acids were demonstrated to be homogeneous, with low variability between samples of less than 1% as might be expected from analytes in a solution. Even for D-alloisoleucine/L-isoleucine, homogeneity was achieved at 1.4%.

Table 5.1: Summary of homogeneity data showing the mean D/L and σ_h (the target standard deviation for sufficient homogeneity)

Amino acid by rpHPLC	Test Material								
	Opercula (bleached)			OES(A) (bleached)			OES(B) (unbleached)		
	Mean D/L	σ_h	σ_h as RSD%	Mean D/L	σ_h	σ_h as RSD%	Mean D/L	σ_h	σ_h as RSD%
Asx D/L	0.581	0.0077	1.3%	0.375	0.0086	2.3%	0.235	0.0084	3.6%
Glx D/L	0.167	0.0012	0.7%	0.094	0.0011	1.2%	0.070	0.009	1.3%
Ser D/L	0.662	0.0163	2.5%	0.325	0.0029	0.90%	0.119	0.0060	5.0%
Ala D/L	0.257	0.0096	3.8%	0.108	0.0041	3.8%	0.076	0.0040	5.3%
Val D/L	0.133	0.0046	3.5%	0.032	0.0024	7.5%	0.024	0.0036	15%
Phe D/L	0.296	0.0093	3.2%	0.083	0.0017	2.1%	0.062	0.0014	2.2%
D-Aile/L-Ile	0.167	0.0085	5.1%	0.036	0.0016	4.3%	0.033	0.0031	9.7%
Leu D/L	0.245	0.0103	4.2%	0.068	0.0011	1.6%	0.060	0.0045	7.5%
Amino acid by rpHPLC	Standard solution			Mollusc(A) (bleached)			Mollusc(B) (unbleached)		
	Mean D/L	σ_h	σ_h as RSD%	Mean D/L	σ_h	σ_h as RSD%	Mean D/L	σ_h	σ_h as RSD%
Asx D/L	0.501	0.0012	0.23%	0.424	0.0076	1.8%	0.408	0.0155	3.8%
Glx D/L	0.556	0.0035	0.63%	0.223	0.0172	7.5%	0.210	0.0100	4.6%
Ser D/L	0.405	0.0019	0.48%	0.527	0.0765	15%	0.423	0.0832	20%
Ala D/L	0.470	0.0014	0.29%	0.447	0.0540	12%	0.372	0.0377	10%
Val D/L	0.591	0.0054	0.92%	0.187	0.0181	9.7%	0.163	0.0235	14%
Phe D/L	0.485	0.0013	0.26%	0.278	0.0155	5.6%	0.255	0.0102	4.0%
D-Aile/L-Ile	0.562	0.0077	1.4%	0.254	0.0236	9.3%	0.235	0.0293	13%
Leu D/L	0.586	0.0028	0.47%	0.335	0.0452	14%	0.273	0.0572	21%

The D/L values for the amino acids in the **opercula test material** were generally lower than those of the standard solution and therefore the RSD% was proportionally higher. Nonetheless, the RSD% for sufficient homogeneity was less than 5% for all amino acids, again, with the exception of D-alloisoleucine/L-isoleucine which was 5.1%.

The two **ostrich eggshell test materials** had the lowest D/L values of all of the materials used. In all cases, the D/L values of the amino acids in the bleached OES (A) Test Material were higher than those from the unbleached OES (B) Test Material. Because the D/L values in both the OES (A) and OES (B) are of the same order of magnitude, it is appropriate to compare the RSD% values between these test materials. In all cases the RSD% was reduced in OES (A) to varying degrees (e.g.; the level of agreement between samples for glutamic acid was only marginally improved by bleaching, from 1.3% to 1.2%, whereas a greater level of agreement was achieved for valine, from 15% to 7.5%). It should also be noted that in both OES (A) and OES (B), D-alloisoleucine/L-isoleucine gave better levels of

agreement between samples than valine, which gave the highest variability of the amino acids listed in this matrix. For OES (B), with the exception of valine (which was 15%), the variability between samples for all amino acids was less than 10%. For OES (A), again with the exception of valine (having an RSD% for sufficient homogeneity determined as 7.5%), the variability between samples for all amino acids was less than 5%.

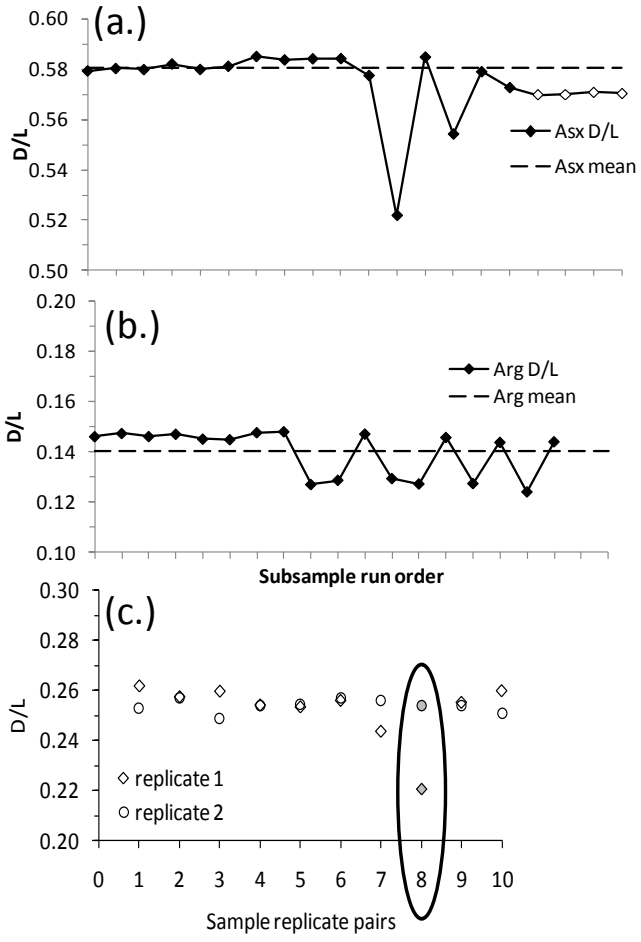
As for the ostrich eggshell test materials, bleaching increased all amino acid D/L values in the **mollusc shell test materials.** In the majority of cases, bleaching also improved agreement between samples, thus RSD% values for aspartic acid, serine, valine, D-alloisoleucine/L-isoleucine and leucine in the unbleached Mollusc (B) Test Material are all wider than their equivalent RSD% values in the bleached Mollusc (A) Test Material. However the converse was true for glutamic acid, alanine and phenylalanine, where increased variability between samples was observed on bleaching. Unlike the opercula and ostrich eggshell test materials, glutamic acid did not show the best agreement in mollusc shell matrices. Rather the closest level of agreement was found in aspartic acid, whilst the most variable data was observed in leucine and serine. In Mollusc (A), leucine and serine's RSD% values are 14% and 15% respectively; in Mollusc (B) these values widen to 21% and 20% giving the largest differences between samples for all the test materials. Interestingly, in mollusc shell matrix, D-alloisoleucine/L-isoleucine was remarkably well behaved and showed better agreement than valine, serine or leucine.

It should be noted that possibly one reason for the larger observed homogeneity standard deviations for the Mollusc test materials compared to the OES or the opercula, apart from their D/L level, may be due to the preparation of the Mollusc material from a collection of whole shells introducing additional variability into the bulk, rather than being prepared from a single shell as for the OES materials, or opercula which show consistent closed system behaviour (Penkman *et al*., 2008).

Figure 5.2 illustrates the advantages of plotting homogeneity results in sequence order and then as paired duplicates. During the analytical run for opercula homogeneity testing, instrumental errors caused the run to stop. After investigations the RP column was replaced and the run restarted. Figure 5.2a demonstrates the effect of this event on the analytical data. Pressure problems caused the data to fluctuate prior to system failure. The final samples to be run on the new column (shown as unfilled data points, Fig 5.2a) clearly demonstrate a systematic shift in measurement values.

Figure 5.2: Homogeneity data evaluation.

(a) Aspartic acid data for Opercula Test Material plotted in analytical sequence order, showing systematic bias following RP column change (unfilled data points). (b) Arginine data for OES (A) Test Material plotted in analytical sequence order, showing possible analyte instability and (c) Phenylalanine data for Mollusc (B) Test Material plotted as replicate pairs showing sample No. 8 as a Cochran's outlier.



This effect was observed to lesser or greater extents in all of the amino acids, with the effect being most noticeable in aspartic acid, alanine, phenylalanine, isoleucine and leucine. This may have been due to instability of the extracts over the duration of the investigation or issues such as the column needing time to 'bed down' before routine use. These issues are well-recognised within the laboratories, but it is clear that the regular monitoring of quality control materials for instrumental errors and system instability is essential in order to avoid and correct for systematic offsets such as this. The reproducibility

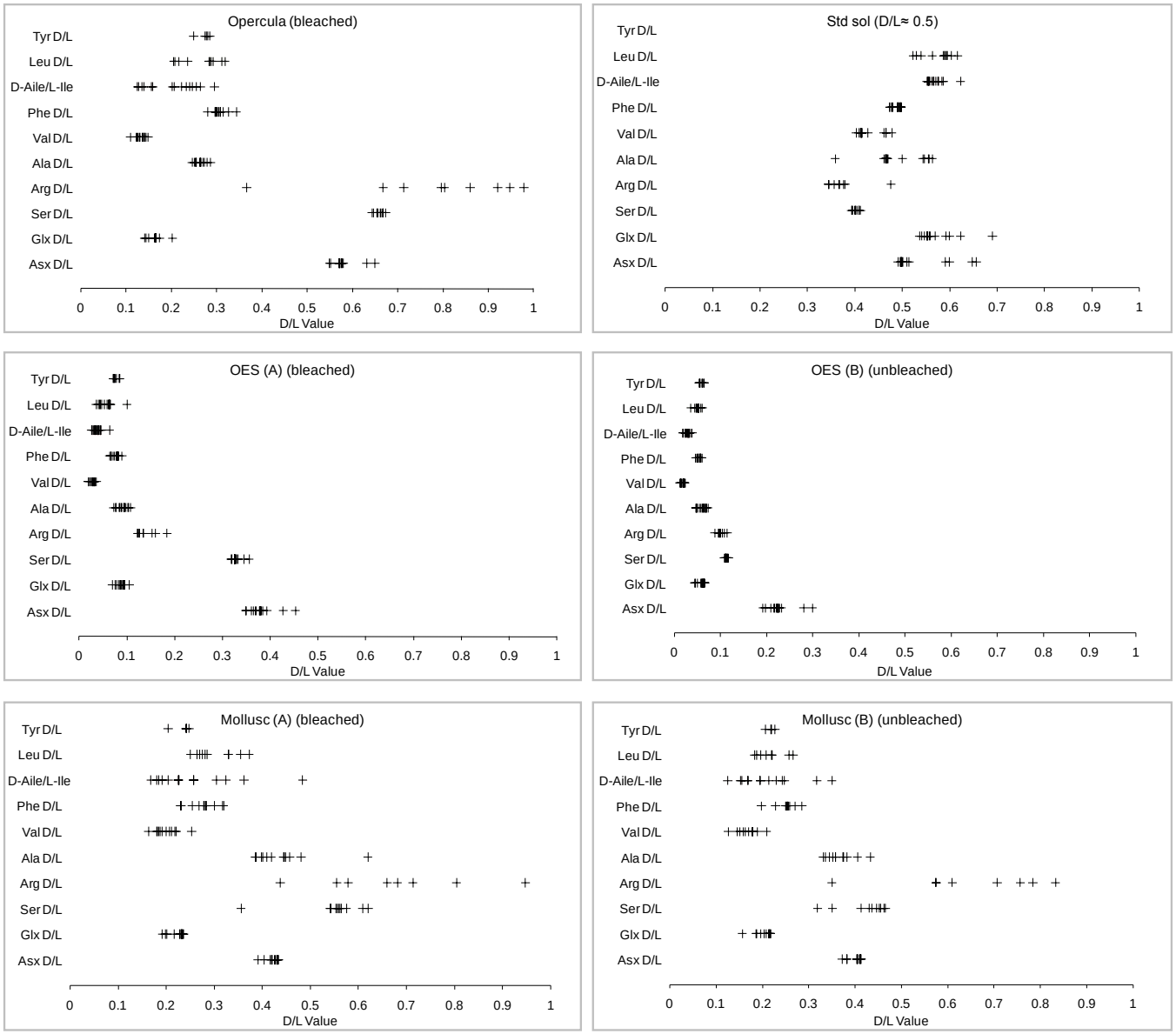
of D-arginine is a recognised problem for the RP method due to co-elution issues. Previous evaluations of standard solutions containing arginine have demonstrated bimodality and decreasing stability with time (Powell, (unpublished data)). This would appear to be supported by the homogeneity results in the bleached OES (A) Test Material (Figure 5.2b). It can be seen that for the first half of the analytical run, data remain consistent but then suddenly change giving randomly lower D/L values, although this effect is not observed in the other test materials. Whether this was due to instability and decomposition is not clear, but for this reason, arginine homogeneity data has not been included in Table 5.1. Plotting the re-paired subsamples can highlight where the within-sample variance is unacceptably wide compared to the other test materials, (Figure 5.2c). For example, subsamples 8a and 8b for phenylalanine in Mollusc (B) Test Material were identified as Cochran's outliers and removed from the data set so as not to unfairly influence the between-sample variance for the other nine pairs of data.

5.4.2 Intra- & Inter-Laboratory Precision (expressed as CV%)

For the proficiency test, participants were required to submit representative D/L values for all their determined amino acids in the test materials. However, in response to participant requests, participants were also invited to submit chromatogram peak information and L and D amino acid concentration data in addition to replicate D/L values. As a result, a substantial quantity of information was captured. Due to time constraints it was not possible to evaluate all of this additional chromatographic or concentration data, but it has been possible to carry out an additional evaluation of intra- and inter-laboratory D/L precision estimates.

Eight laboratories were sent test materials but three of these either reported instrumental problems or were unable to return results within the timeframe. In total, eleven sets of results were returned by four RP laboratories, two sets of results by one IEx laboratory and one set of results by one GC laboratory. In the majority of instances laboratories reported data for any given analyst/instrument combination, using a single test material hydrolysate. Where a laboratory submitted data derived from more than one hydrolysed subsample of a given test material by the same analyst/instrument, these values have been combined and averaged to provide a single representative D/L value for each amino acid in the test material. Figure 5.3 shows the distribution of reported results, given here as the mean of each participant's replicate D/L values, and clearly show the difference in levels of agreement between laboratories for the different matrices.

Figure 5.3: Distribution of participants' mean D/L values for amino acids in the six test materials.



Detailed evaluations of individual laboratory precision for each amino acid in every test material are provided in the individual study reports. However, a summary of the results is provided in Table 5.2. For every amino acid in each of the six test materials, the mean D/L value given is determined as the *mean of the means*; that is the average of all the participants' individual replicate D/L means.

For RP and IEx, data represented replicate injections of the same extract. For GC data, results were submitted as the mean of replicate values, a standard deviation and n, the number of replicates. For the proficiency test, GC D/L values derived from both chromatogram peak areas and peak heights were submitted. For the purpose of performance assessment, both sets of data were included in the assessments of relative bias for comparison. However, it was subsequently confirmed that only peak area derived D/L values would be used routinely for chronological purposes and it is these D/L values that have been used in the summary and comparison presented in Table 5.2. The laboratory carrying out GC analysis reported incomplete desalting for Mollusc (A) test material resulting in a poor derivative and low yields. However data have been included for completeness with the acknowledgement that due to problems D/L values may be inaccurate.

For consistency and comparison with previously published data, method specific intra-laboratory (repeatability) precision values have been determined as the relative standard deviation, expressed as the *CV% or RSD%* $= (s/\bar{x}) \times 100$, where s is the standard deviation and $\bar{x}$ is the mean of individual participant's replicate D/L values. The range of participants' *intra-laboratory* CV%s, for each amino acid in every test material are given, together with the average CV% based on the number of participants (m) who provided data. An estimate of the between–laboratory precision or the *inter-laboratory* CV% is then derived as the relative standard deviation of all the participants' D/L value means.

From the data given in Table 5.2, occasionally it can be seen that when m is greater than 1, no CV% range is given. This will be because only one of those participants will have provided replicate values and the other(s) will have submitted only a single D/L value, whose precision cannot be determined. Therefore there is no CV% range to report, only the CV% from the participant providing replicate data.

Table 5.2: Summary of laboratory precision estimates (intra- & inter-laboratory, CV%s) derived from participants' submitted replicate results.

amino acid	Opercula Test Material					OES(A) (bleached)					OES(B) (unbleached)				
	m	[1]mean D/L	Intra-Lab CV% range	Intra-Lab CV% average	[2]Inter-Lab CV%	m	[1]mean D/L	Intra-Lab CV% range	Intra-Lab CV% average	[2]Inter-Lab CV%	m	[1]mean D/L	Intra-Lab CV% range	Intra-Lab CV% average	[2]Inter-Lab CV%
Asx D/L-rpHPLC	11	0.57	0.04 - 1.78	0.45	1.75	11	0.37	0.04 - 1.44	0.42	3.35	11	0.218	0.00 - 0.73	0.24	5.84
Asx D/L-GC	1	0.631	-	7.13	-	1	0.379	-	5.01	-	1	0.225	-	8.89	-
Glx D/L-rpHPLC	11	0.159	0.00 - 1.49	0.33	6.16	11	0.086	0.00 - 7.38	1.21	9.93	11	0.059	0.00 – 1.00	0.44	14.15
Glx D/L-GC	1	0.174	-	14.94	-	1	0.09	-	15.85	-	1	0.059	-	13.56	-
Ser D/L-rpHPLC	11	0.659	0.16 - 2.55	0.98	1.41	11	0.33	0.00 - 1.99	0.53	3.37	11	0.113	0.00 - 1.10	0.32	2.48
Arg D/L-rpHPLC	9	0.784	1.4 - 34.72	15.48	23.98	9	0.14	0.50 - 6.21	2.88	15.11	9	0.101	0.08 - 6.37	3.1	7.79
Ala D/L-rpHPLC	11	0.265	0.08 - 4.42	1.91	4.31	11	0.095	0.00 - 4.70	1.77	8.17	11	0.064	0.00 - 14.8	2.53	8.85
Ala D/L-GC	1	0.265	-	3.4	-	1	0.077	-	5.19	-	1	0.047	-	31.91	-
Val D/L-rpHPLC	11	0.135	0 - 13.27	3.17	6.68	11	0.03	1.02 - 10.64	3.55	11.7	11	0.02	1.08 - 10.62	5.22	10.13
Val D/L-GC	1	0.109	-	5.5	-	1	0.019	-	15.79	-	1	0.015	-	6.67	-
Phe D/L-rpHPLC	11	0.309	0.12 - 10.14	2.17	4.56	11	0.079	0.14 - 11.28	2.6	4.58	11	0.054	0.00 - 3.21	1.19	7.84
Phe D/L-GC	1	0.28	-	10.71	-	1	0.067	-	11.94	-	1	0.051	-	25.49	-
D-Aile/L-Ile-rpHPLC	11	0.222	0.45 - 26.21	12.48	21.92	11	0.04	0.88 - 10.1	4.64	24.96	11	0.03	0.00 - 11.03	3.4	20.46
D-Aile/L-Ile-GC	1	0.159	-	8.81	-	1	0.033	-	6.06	-	1	0.026	-	3.85	-
D-Aile/L-Ile-HPLC-IE	2	0.137	1.58 - 3.05	2.31	2.32	2	0.031	0.00 - 7.37	3.69	0.76	2	0.024	-	0	0
Leu D/L-rpHPLC	8	0.277	1.49 - 10.53	5.32	13.82	9	0.062	0.00 - 5.96	1.71	27.21	8	0.051	0.00 - 29.89	5.97	9.71
Leu D/L-GC	1	0.204	-	2.94	-	1	0.04	-	2.5	-	1	0.036	-	5.56	-
Tyr D/L-rpHPLC	5	0.273	0.07 - 3.57	1.68	5.19	7	0.077	0.13 - 1.85	0.77	6.92	7	0.059	0.46 - 6.74	2.09	7.5

m = no of sets of results [1] = mean of the participants' average D/L values

[2] = Due to results for IE being submitted by single a laboratory, Inter-Laboratory precision should be read as Intermediate precision.

Table 5.2: Summary of laboratory precision estimates (intra- & inter-laboratory, CV%s) derived from participants' submitted replicate results (continued).

amino acid	Standard solution Test Material					Mollusc (A) (bleached)					Mollusc (B) (unbleached)				
	m	[1]mean D/L	Intra-Lab CV% range	Intra-Lab CV% average	[2]Inter-Lab CV%	m	[1]mean D/L	Intra-Lab CV% range	Intra-Lab CV% average	[2]Inter-Lab CV%	m	[1]mean D/L	Intra-Lab CV% range	Intra-Lab CV% average	[2]Inter-Lab CV%
Asx D/L-rpHPLC	10	0.5	0.01- 1.44	0.36	1.38	11	0.421	0.07 - 1.68	0.52	3.1	10	0.399	0.18 - 1.18	0.54	3.65
Asx D/L-GC	1	0.652	-	7.98	-	1	0.433	-	7.62	-	1	0.413	-	5.81	-
Glx D/L-rpHPLC	10	0.552	0.06 - 1.14	0.32	1.67	11	0.224	0.31- 2.64	1.26	6.63	10	0.202	0.02 - 3.4	0.76	9.72
Glx D/L-GC	1	0.596	-	9.4	-	1	0.198	-	12.12	-	1	0.196	-	2.04	-
Ser D/L-rpHPLC	10	0.402	0.00 - 3.28	0.56	1.63	11	0.549	0.46 - 4.51	2.11	12.55	10	0.423	0.72 - 7.04	2.17	11.71
Arg D/L-rpHPLC	8	0.376	0.09 - 7.66	2.16	11.27	8	0.672	1.42 - 39.33	18.5	23.4	8	0.649	0.5 - 33.96	21.57	23.94
Ala D/L-rpHPLC	10	0.481	0.02 - 1.38	0.52	12.32	11	0.425	0.16 - 3.48	1.83	7.44	10	0.357	0.82 - 4.27	2.08	4.75
Ala D/L-GC	1	0.484	-	3.86	-	1	0.62	-	5.32	-	1	0.405	-	6.42	-
Val D/L-rpHPLC	10	0.435	0.00 - 0.64	0.25	6.61	11	0.203	0.28 - 7.61	2.84	10.83	10	0.172	0.47 - 7.71	3.32	10.92
Val D/L-GC	1	0.408	-	3.92	-	1	0.163	-	30.06	-	1	0.126	-	3.17	-
Phe D/L-rpHPLC	10	0.492	0.02 - 1.28	0.32	1.08	11	0.281	0.20 - 12.91	3.24	9.13	10	0.253	0.59 - 13.18	3.8	8.81
Phe D/L-GC	1	0.477	-	7.34	-	1	0.23	-	1.74	-	1	0.228	-	4.82	-
D-Aile/L-Ile -rpHPLC	10	0.561	0.05 - 1.23	0.44	1.54	11	0.274	0.00 - 23.76	9.84	33.06	10	0.228	2.6 - 23.7	11	29.26
D-Aile/L-Ile -GC	1	0.586	-	1.02	-	1	0.204	-	4.9	-	1	0.168	-	3.57	-
D-Aile/L-Ile -HPLC-IE	2	0.592	-	-	0.48	2	0.186	0.00 - 1.11	0.55	4.37	2	0.154	0.00 - 5.47	2.74	1.38
Leu D/L-rpHPLC	8	0.592	0.07 - 2.60	0.75	2.58	8	0.307	0.79 - 23.7	7.35	14.7	6	0.228	1.91- 26.87	9.58	12.25
Leu D/L-GC	1	0.526	-	1.33	-	1	0.27	-	11.58	-	1	0.185	-	4.85	-
Tyr D/L-rpHPLC	-	-	-	-	-	5	0.247	0.94 - 3.09	2.1	7.16	4	0.218	1.13 - 2.85	1.83	3.98

m = no of sets of results

[1] = mean of the participants' average D/L values

[2] = Due to results for IE being submitted by a single laboratory, Inter-Laboratory precision should be read as Intermediate precision.

5.4.2.1 Observations on D/L values

When designing the proficiency test, the original intention was to try to provide test materials with approximately similar D/L values, but unfortunately the heating time and conditions used on the OES materials were inadequate to induce the necessary level of comparable racemization. Nonetheless, some general observations can be made regarding method differences and observed D/L values across all the test materials (Table 5.2). However, it should be stressed that such observations are only indicative, due to limited representative datasets particularly for GC and IEx methods.

- In all solid test materials, Asx mean D/L values by GC were larger than those by RP by an average of 3.9% (range; 1.4 – 9.7%). This increased to 23.2% in the standard solution.
- For Glx and Ala, mean D/L values varied. Neither RP nor GC were consistently higher or lower than the other.
- In all test materials, Val, Phe and Leu mean D/L values by RP were larger than those by GC: For Val by an average of 22.3% (range; 6.2 – 36.7%); for Phe by an average of 8.1% (range; 3.3 – 15.2%); for Leu by an average of 22.1% (range; 11.15 – 35.5%).
- In all solid test materials, mean A/I values by RP were larger than those by GC and those by GC were larger than those by IEx, i.e. RP>GC>IEx. For RP, A/I values were on average 22.2% (range; 13.33 – 28.38%) higher compared to those by GC, and an average of 29.0% (range; 20.0 – 38.3%) higher than by IEx. GC A/I values were on average 8.9% (range; 6.1 – 13.8%) higher than by IEx.
- For the standards solution test material the trend described above was reversed. Mean A/I values by RP were smaller than those by GC and those by GC were smaller than those by IEx, i.e. RP<GC<IEx. For RP, the A/I value was 4.3% lower than that by GC, and 5.2% lower than by IEx. The GC A/I value was 1.0% lower than by IEx.
- For all amino acids, the effect of bleaching, both on the mollusc and OES test materials, raised the D/L values. For OES, D/L values increased by an average of 30% (range; 10-66%) and for mollusc shell by an average of 15% (range; 1.0-35%)

5.4.2.2 Observations on Precision

With regard to RP precision estimates, the inter-laboratory CV%s in all cases are wider than the intra-laboratory values. This indicates less agreement between participants due to method and laboratory bias influences compared to the intra-laboratory precision estimates, which represent the analytical imprecision arising from random error effects only during the analytical run.

Precision estimates are amino acid specific, and will vary considerably within a single matrix. For example, for Mollusc (B) Test Material, the average RP intra-laboratory CV%s ranged from 0.54%(Asx) – 11.0%(Ile) (or 21.6% for arginine) and inter-laboratory CV%s ranged between 3.65%(Asx) – 29.3%(Ile).

Because of the differences in the variability observed for different amino acids in any given sample, it is more informative to evaluate amino acids individually across the different test materials. However, as discussed previously, when comparing relative values between test materials, caution should be taken as this is strongly influenced by the mean D/L value; lower values acting to elevate the relative standard deviations thus widening the observed CV% range.

- A comparison of precision estimates between methods (Table 5.2), results in the following observations;
- For Asx, Glx and Ala, all average intra-lab CV%s for RP are smaller than those for the GC data.
- For Val, all average intra-lab CV%s for RP are smaller than those for the GC data with the exception of Mollusc (B).
- For Phe, all average intra-lab CV%s for RP are smaller than those for the GC data with the exception of Mollusc (A).
- For D-Aile/L-Ile and Leu, intra-lab CV%s varied. Neither RP nor GC were consistently higher or lower than the other.
- For D-Aile/L-Ile, all average intra-lab CV%s for IEx are smaller than those for either the RP or GC data.

These are very general observations made using the average RP CV% values. This therefore does not make comparisons with individual RP laboratories, whose individual intra-lab CV%s may not follow this pattern but may in fact be larger than that for GC or smaller than that for IEx for some amino acids / test materials.

For each amino acid, Table 5.3 compares the average CV% for Standard Solution test material with the range of average CV%s observed across the five biomineral matrices, separated by method.

For the Standard Solution test material by RP, the average intra-laboratory precision is often smaller than that for the solid matrices, the exceptions being for Asx in OES (B) test material and serine in both OES (A) and (B) test materials. Similarly for the inter-laboratory precision, where all Standard Solution CV%s, (with the exception of serine in Opercula test material and alanine), are all smaller than those observed for the solid matrices. For alanine, it is interesting to note that agreement between laboratories for all solid matrix test materials were better, giving tighter inter-laboratory CV%s, than the Standard Solution.

Table 5.3: A comparison of amino acid average precision estimates for Standard Solution with the solid matrix test materials.

Amino acid	RP Precision (CV%)		GC Precision (CV%)	IEx Precision (CV%)
D/L value	Intra-Laboratory Std sol, Biominerals	Inter-Laboratory Std sol, Biominerals	Intra-Laboratory Std sol, Biominerals	Intra-Laboratory Biominerals
Asx	0.36, 0.24 - 0.54	1.38, 1.75 - 5.84	7.85, 4.92 – 9.01	-
Glx	0.32, 0.33 - 1.26	1.67, 6.16 - 14.15	9.12, 2.18 - 15.36	-
Ser	0.56, 0.32 - 2.17	1.63, 1.41 - 12.55	-	-
Ala	0.52, 1.77 - 2.53	12.32, 4.31 - 8.85	3.81, 3.44 - 31.21	-
Phe	0.32, 1.19 - 3.80	1.08, 4.56 - 9.13	7.35, 1.54 - 25.08	-
Val	0.25, 2.84 - 5.22	6.61, 6.68 - 11.7	3.14, 2.78 – 30.14	-
Leu	0.75, 1.71 – 9.58	2.58, 9.71 - 27.21	1.36, 2.83 – 6.43	-
A/I	0.44, 3.40 - 12.48	1.54, 20.46 - 33.0	1.05, 3.58 - 8.88	0.00 – 3.69

However, by comparison, the single laboratory GC intra-laboratory CV%, with the exception of D-Aile/L-Ile and Leu, does not appear to experience the same general improvement with the analysis of the Standards Solution (Table 5.3). A similar observation was previously commented on by Wehmiller (1984) in the GC/IEx inter-laboratory comparison using fossil mollusk powders and their respective hydrolysates. A closer inspection of data in Table 5.2, indicates that although there are instances of smaller CV%s for some amino acids in Standard Solution compared to specific powdered test materials, the response is generally more varied.

Due to the absence of data from other GC laboratories, it is not possible to determine GC specific inter-laboratory precision data that would have given an indication of the extent of the GC laboratory bias. IEx data is not included in this evaluation as the inter-

laboratory CV given in Table 5.2 is actually only equivalent to an intermediate precision estimate; i.e. an intra-laboratory reproducibility.

5.4.2.3 Observations on the effect of bleaching solid matrix test materials

In all cases, the effect of bleaching the powdered test materials prior to hydrolysis, raised the D/L values, in OES by an average of 30% and for mollusc by an average of 15%, irrespective of method. From data in Table 5.2, the effect of bleaching on precision estimates was inconclusive. For some amino acids, CV%s of bleached materials were marginally larger, for others, marginally smaller. Bleaching might have been expected to significantly improve precision estimates by isolating the intra-crystalline protein fraction (Penkman *et al*., 2008). Table 5.4 summarizes this data taken from Table 5.1 and Table 5.2 and compares the unbleached OES (B) and Mollusc (B) Test Materials with their bleached equivalents; OES (A) and Mollusc (A). Results indicate the number of amino acids which have reduced CV%s in the bleached materials. For ostrich eggshell, bleached OES (A) had 4 out of 7 GC amino acid intra-laboratory precision estimates smaller than the unbleached ones in OES (B), half of the RP intra-laboratory precision estimates and half of the RP inter-laboratory values (5 out of 10 in both cases) were also smaller. It is noted that it was not always the same amino acid that improved in each case. For the bleached Mollusc (A), 2 out of 6 GC amino acids showed smaller intra-laboratory precision whilst for RP 8 out of 10 (intra-laboratory) and 4 out of 10 (inter-laboratory) amino acids showed smaller precision estimates than the unbleached Mollusc (B). In this evaluation, only the *average* RP intra-laboratory CV%s are being compared, which does not preclude the potential effect on individual laboratories. In comparison, results from the single laboratory homogeneity evaluation, (Table 5.1), show a definitive improvement with bleaching across many amino acids. In ostrich eggshell all 8 amino acids demonstrated tighter agreement and even for mollusc shell material, 5 out of 8 amino acids gave better homogeneity precision estimates for bleached material (the exceptions being glutamic acid, alanine and phenylalanine). This indicates closer agreement for the majority of amino acids between individual bleached test materials analysed by a single laboratory, suggesting that bleaching does appear to improve the precision of individual amino acids in a single analytical run (repeatability precision), but the extent of this is probably matrix and amino acid specific. Furthermore, it would also appear that individual laboratory biases assert a significant effect on precision estimates, in this instance, exceeding any gain to be had from bleaching.

Table 5.4: Effect of bleaching on precision estimates.

Data compares unbleached OES (B) with bleached OES (A) and unbleached Mollusc (B) with bleached Mollusc (A). Results show the number of amino acids whose precision estimates improved on bleaching, i.e. experienced a reduction in CV% (or RSD% for homogeneity data).

Precision source	Participants' results				%
	GC	RP	IEx	All	
OES (A); No of amino acids improved by bleaching					
Intra-Lab CV[a]	4/7	5/10	0/1	9/18	50%
Inter-lab CV	-	5/10	0/1	5/11	45%
Homog RSD	-	8/8	-	-	100%
Mollusc (A); No of amino acids improved by bleaching					
Intra-Lab CV[a]	2/6	8/10	1/1	11/17	65%
Inter-lab CV	-	4/10	0/1	4/11	36%
Homog RSD	-	5/8	-	-	63%

[a] = given as the laboratory average CV% from Table 2

5.4.3 Performance Analysis

Table 5.5 summarizes the range of relative percentage biases achieved by participants, the standard deviations of the assigned values, $\hat{\sigma}$ used in setting the satisfactory ranges, together with the number of participants achieving performance within $|z| \leq 2$. It should be understood that the $\pm 2\hat{\sigma}$ describes the 95% confidence limits of the assigned value (the median in this case) but not necessarily the spread of all the submitted results. For a normally distributed dataset this will also describe the distribution of the results, but skewed or bimodal data will generally fall outside of this region and will be indicated by a lower percentage satisfactory figure. Those amino acids showing 100% satisfactory performance are either (a) in excellent agreement with a low precision estimate, or (b) are bimodal, where there is too much doubt associated with the position of the assigned value compared to the distribution of results, and so the satisfactory range is uncharacteristically wide.

Due to the relatively small data set it can be tempting to read too much into them, especially for GC and IEx data. For every amino acid in each test material, relative percentage bias values have been determined using a consensus derived from data as submitted by participants, taken across all methods: RP, IEx, GC (peak area) and GC (peak height). Thus the GC and IEx bias has been assessed against an RP weighted value. Clearly, this is not ideal where there is any doubt about the agreement between different methods.

Table 5.5: Assigned Values (median), deviation of the Assigned Value (sMAD) and participants' range of relative percentage biases for each amino acid together with the percentage of results falling within ± 2 standard deviations of the Assigned Value.

amino acid	Opercula Test Material				OES(A) (bleached)				OES(B) (unbleached)			
	median $\hat{\mathbf{X}}$ (D/L)	sMAD $\hat{\sigma}$ (%)	Rel bias Range (%)	% within ± 2$\hat{\sigma}$	median $\hat{\mathbf{X}}$ (D/L)	sMAD $\hat{\sigma}$ (%)	Rel bias Range (%)	% within ± 2$\hat{\sigma}$	median $\hat{\mathbf{X}}$ (D/L)	sMAD $\hat{\sigma}$ (%)	Rel bias Range (%)	% within ± 2$\hat{\sigma}$
Asx D/L-all[a]	0.573	1.02	-4.1 - 13.5	69%	0.379	3.84	-7.8 – 19.6	80%	0.223	4.14	-14.0 – 33.9	73%
Asx D/L-RP	0.572	1.17	-4.0 - 1.5	82%	0.370	3.76	-5.6 – 4.0	100%	0.222	3.47	-13.6 – 4.5	82%
Glx D/L-all[a]	0.165	1.47	-15.0 - 22.5	62%	0.087	8.32	-20.0 – 20.9	87%	0.062	6.85	-28.9 – 7.7	77%
Glx D/L-RP	0.164	1.29	-14.6 – 1.4	73%	0.088	12.72	-21.5 – 7.2	100%	0.062	5.88	-29.0 – 7.4	73%
Ser D/L-RP	0.662	1.41	-2.7 – 1.7	100%	0.326	1.27	-2.5 – 9.0	73%	0.112	3.57	-3.3 – 4.7	100%
Arg D/L-RP	0.803	21.76	-54.4 – 21.9	89%	0.134	11.55	-8.9 – 37.3	89%	0.100	7.10	-12.0 – 14.7	89%
Ala D/L-all[a]	0.263	5.15	-6.6 – 8.6	100%	0.092	12.25	-21.8 – 17.1	100%	0.062	16.09	-24.2 – 18.0	100%
Ala D/L-RP	0.264	5.14	-5.2 – 8.2	100%	0.095	10.24	-11.4 – 13.3	100%	0.065	8.34	-15.6 – 12.4	100%
Val D/L-all[a]	0.137	6.99	-20.4 – 8.4	92%	0.029	14.13	-35.3 – 22.7	87%	0.020	11.49	-40.5 – 15.6	73%
Val D/L-RP	0.137	7.58	-10.7 – 8.4	100%	0.030	13.23	-21.4 – 21.4	100%	0.020	2.14	-23.5 – 13.9	55%
Phe D/L-all[a]	0.304	2.87	-7.9 – 13.0	77%	0.079	5.14	-19.0 – 13.9	67%	0.054	9.49	-14.1 – 10.9	100%
Phe D/L-RP	0.305	3.01	-2.6 – 12.6	100%	0.079	3.12	-11.0 – 4.4	82%	0.056	5.51	-17.3 – 6.8	82%
D-Aile/L-Ile-all[b]	0.206	35.21	-39.9 – 43.2	100%	0.039	20.37	-33.3 – 63.1	94%	0.030	14.82	-38.4 – 29.1	88%
D-Aile/L-Ile -RP	0.233	16.94	-46.2 – 26.8	91%	0.039	18.92	-33.8 – 61.7	91%	0.031	7.28	-40.9 – 23.9	73%
Leu D/L-all[a]	0.284	16.12	-27.8 – 12.1	100%	0.058	18.81	-39.2 – 72.1	85%	0.051	10.18	-30.1 – 17.3	92%
Leu D/L-RP	0.286	7.86	-28.4 – 11.2	75%	0.062	8.24	-43.0 – 61.2	78%	0.052	8.36	-14.6 – 17.1	88%
Tyr D/L-RP	0.277	1.99	-10.1 – 2.9	80%	0.076	6.89	-5.7 – 11.6	100%	0.060	10.20	-9.8 – 8.7	100%

[a] = RP and GC data [b] = RP, GC and IEx data

Table 5.5: Assigned Values (median), deviation of the Assigned Value (sMAD) and participants' range of relative percentage biases for each amino acid together with the percentage of results falling within ± 2 standard deviations of the Assigned Value.

amino acid	Standard solution Test Material				Mollusc(A) (bleached)				Mollusc(B) (unbleached)			
	median $\hat{\mathbf{X}}$ (D/L)	sMAD $\hat{\sigma}$ (%)	Rel bias Range (%)	% within ± 2$\hat{\sigma}$	median $\hat{\mathbf{X}}$ (D/L)	sMAD $\hat{\sigma}$ (%)	Rel bias Range (%)	% within ± 2$\hat{\sigma}$	median $\hat{\mathbf{X}}$ (D/L)	sMAD $\hat{\sigma}$ (%)	Rel bias Range (%)	% within ± 2$\hat{\sigma}$
Asx D/L-all[a]	0.500	2.02	-1.9 – 31.2	71%	0.426	2.24	-8.1 – 1.8	83%	0.404	2.58	-8.0 – 2.2	73%
Asx D/L-RP	0.499	0.48	-1.6 – 3.2	70%	0.426	2.24	-8.1 – 1.8	82%	0.404	2.55	-8.0 – 2.0	80%
Glx D/L-all[a]	0.556	2.94	-3.5 – 24.1	71%	0.230	2.57	-16.8 – 2.5	67%	0.208	6.00	-24.7 – 4.6	91%
Glx D/L-RP	0.552	1.50	-2.9 – 3.0	90%	0.232	2.39	-17.2 – 2.0	73%	0.211	4.16	-25.8 – 3.1	70%
Ser D/L-RP	0.401	2.38	-2.1 – 2.7	100%	0.559	4.38	-36.3 – 11.1	73%	0.442	5.60	-27.7 – 5.4	80%
Arg D/L-RP	0.367	4.76	-6.2 – 29.7	88%	0.671	22.93	-34.8 – 41.3	100%	0.658	20.49	-46.9 – 26.5	88%
Ala D/L-all[a]	0.469	6.12	-23.5 – 20.1	57%	0.432	9.52	-10.7 – 43.7	92%	0.358	6.98	-7.5 – 21.0	92%
Ala D/L-RP	0.467	1.28	-23.2 – 19.2	60%	0.419	9.35	-8.0 – 14.8	100%	0.358	5.97	-7.5 – 6.9	100%
Val D/L-all[a]	0.414	1.90	-2.6 – 15.5	69%	0.196	11.45	-16.8 – 29.1	92%	0.170	10.06	-25.7 – 23.1	82%
Val D/L-RP	0.414	1.49	-1.4 – 15.3	60%	0.200	11.85	-9.9 – 26.2	91%	0.171	11.01	-15.3 – 21.9	100%
Phe D/L-all[a]	0.491	1.53	-3.7 – 1.3	86%	0.281	8.27	-18.1 – 13.8	83%	0.254	2.51	-22.2 – 12.5	64%
Phe D/L-RP	0.493	0.85	-2.6 – 1.0	90%	0.282	7.04	-18.3 – 13.2	91%	0.254	1.91	-22.2 – 12.4	70%
D-Aile/L-Ile-all[b]	0.566	2.69	-2.4 – 10.2	93%	0.226	24.75	-25.5 – 114.2	86%	0.195	30.69	36.1 – 79.1	85%
D-Aile/L-Ile -RP	0.558	1.01	-1.0 – 3.4	80%	0.257	27.56	-34.5 – 88.3	91%	0.222	17.51	-43.7 – 58.0	70%
Leu D/L-all[a]	0.589	3.71	-11.2 – 4.7	73%	0.283	13.60	-11.4 – 31.7	90%	0.214	15.47	-14.3 – 24.5	100%
Leu D/L-RP	0.593	1.13	-5.1 – 3.9	75%	0.305	17.25	-17.9 – 22.1	100%	0.220	12.44	-11.1 – 20.8	100%
Tyr D/L-RP	-	-	-	-	0.241	0.57	-15.5 – 2.5	60%	0.219	3.10	-5.8 – 3.8	50%

[a] = RP and GC data

[b] = RP, GC and IEx data

In this case, GC data needs to be assessed against a GC weighted consensus, similarly for IEx. For these reasons, the 'all' data given in Table 5.5 may show a much wider bias range, reflecting any potential differences between the GC (or IEx) and RP data. Data from this and other studies have observed these effects between single laboratories, so in the absence of other evidence, 'all' bias ranges should be seen as indicative and suggestive of rather than absolute. Comparative values have been given for RP data, assessed separately using a RP specific consensus, but due to lack of additional data for GC and IEx this has not been possible for these two methods.

Space precludes presentation of the histograms of relative bias for each laboratory for every amino acid in each of the test materials in this paper, but Figures 5.4, 5.5 and 5.6 help to illustrate some of the issues involved in assessing and interpreting the data.

In an ideal situation, submitted results will be randomly and normally distributed, as indicated by glutamic acid in OES (A) Test Material, with approximately 95% of participants' results within the satisfactory range (Figure 5.4a). However this was not always the case and distribution patterns varied depending on the analyte or matrix concerned.

A low tail skew is observed for Glx in Mollusc (A) Test Material (Figure 5.4b), which may be method or even laboratory dependent. By comparison there is a distinctive high tail GC skew for Glx in the standard solution test material (Figure 5.4c), not seen in the other two test materials. Taken in isolation, one could draw different conclusions based on different matrices.

Figure 5.4: Histograms showing the distribution of participants' relative biases for glutamic acid. In (a) OES (A) Test Material, (b) Mollusc (A) Test Material, and (c) Standard Solution Test Material.

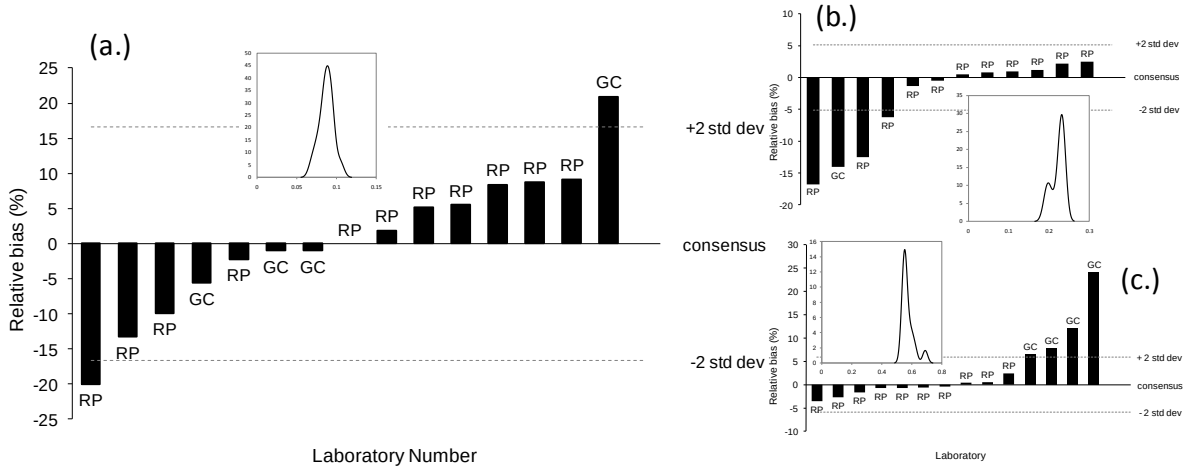


Figure 5.5: Histograms showing the distribution of participants' relative biases for valine. In (a) Standard Solution Test Material, and (b) OES (B) Test Material.

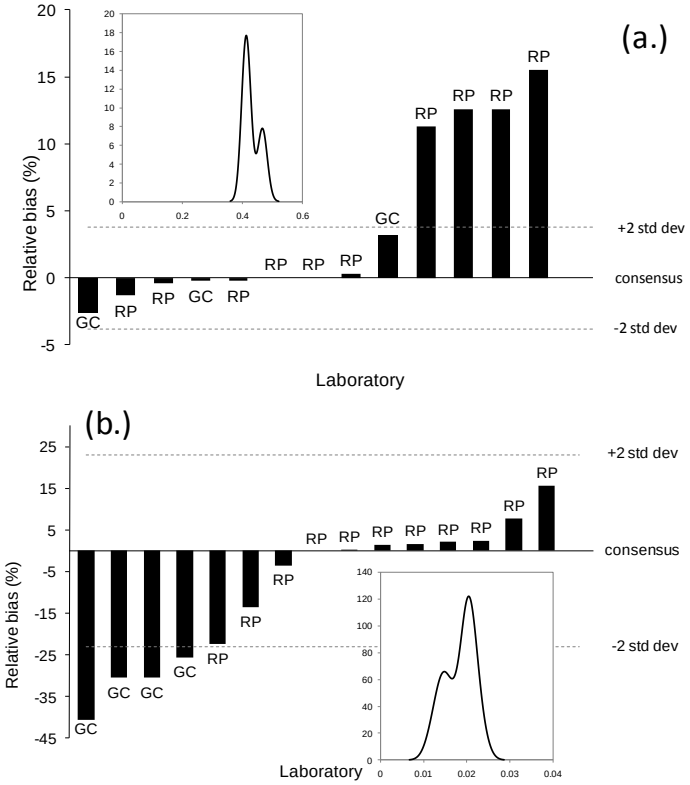
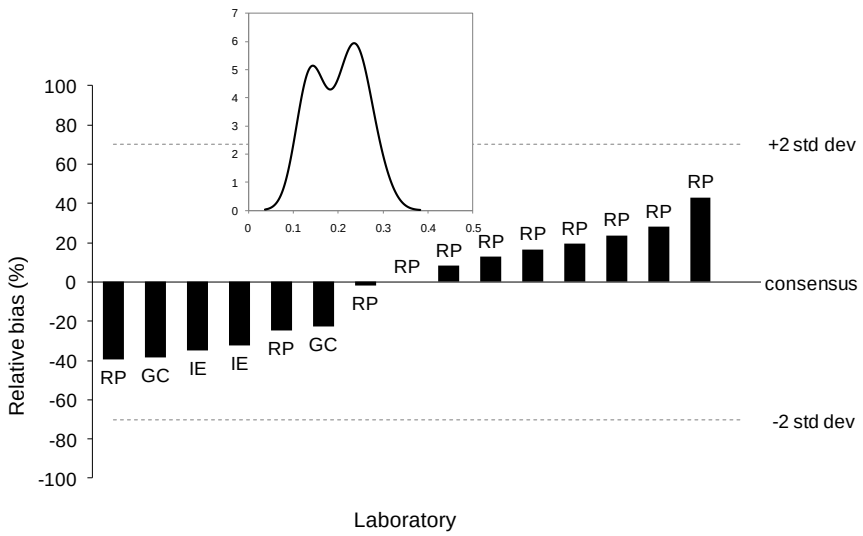


Figure 5.6: Histogram showing the distribution of participants'relative biases for isoleucine in Opercula Test Material.



Another problem sometimes encountered with small data sets and the use of the median is illustrated by valine data (Figure 5.5a & b). As for all amino acids in this study, the consensus was set as the median. In both the standard solution (Fig 5.5a) and ostrich eggshell (Fig 5.5b), data are evenly distributed either side of the mid-point. However, in there is a strong RP high-tail skew in standard solution (Figure 5.5a), whilst in ostrich eggshell there is an even larger GC low-tail skew (Figure 5.5b). Both sets of data are clearly bimodal, but the modes are unevenly balanced. Whether the skew is high or low simply depends on the positioning of the middle value and which camp it happens to fall in. In cases of bimodality, the primary mode is usually the one with the greatest number of data points, but when dealing with small data sets, judgement regarding which side is actually the correct one should be reserved in the absence of supporting evidence.

Finally, on rare occasions, two modes may be evenly matched and the median falls mid-way between the two (Figure 5.6). In this situation, there is no clear primary mode and the sMAD (median absolute deviation) increases to reflect the elevated uncertainty regarding the position of the consensus value. The ± 2 standard deviation satisfactory limits broaden to encompass the entire dataset and are clearly over generous for any formal performance evaluation. Both modes are fairly evenly populated and no judgement can be made about which one is correct.

The range of observed biases (Table 5.5), are dependent on the amino acid and matrix concerned; for example arginine and isoleucine are usually very wide compared to other amino acids. Generally though, it is difficult to see any clear patterns in this data, except perhaps to comment that in several cases it is not unusual to see relative percentage biases of up to 30% or more in either direction.

5.4.3.1 Average Relative Bias %

It is reasonable to assume that analytical/method/laboratory systematic bias might be expected to behave reasonably consistently for individual laboratories. Therefore it is helpful to compare behaviour for the same amino acid in different matrices. For each participant, an average relative percentage bias has been determined for each amino acid across all six test materials. It is expected that this should give a more balanced picture of the overall distribution of bias, giving due note to the direction of the bias values. No limits for satisfactory performance have been given, since every amino acid in each test material has its own specific satisfactory range (although an average pooled standard deviation for the assigned values could be determined).

However, the benefit is in identifying significant positive or negative bias, either as large individual biases or as a general trend, since a "*well-behaving*" laboratory will have evenly distributed values with minimal overall bias.

The calculation for the average relative bias was derived from the root mean square (RMS%) used to calculate the average overall bias (EUROLAB, 2007), where;

$$RMS_{bias} = \sqrt{\sum (bias_i)^2/n} \quad \textbf{(5.5)}$$

In this context, *n* is the number of proficiency test results submitted by an individual laboratory for a specific amino acid.

Therefore, the average relative bias, allowing for direction has been determined simply as;

$$average\ relative\ bias = \sum (rel\ bias_{\ i})/n \quad \textbf{(5.6)}$$

Figure 5.7 shows paired graphs (i and ii) for each amino acid separately (a-h). The first of these (Fig. 5.7i), shows relative percentage bias plotted against the mean of each participant's replicate results for each of the six test materials. Thus, data points extrapolated vertically onto the x-axis demonstrate the closeness of agreement between submitted results. A horizontal extrapolation illustrates each result's relative bias when compared to the assigned value. The position on the x-axis where each trendline crosses represents the consensus or assigned value, i.e. there is no observable bias. It can be seen that for all amino acids the spread of submitted results for OES test materials is most often the tightest, at times even better than the standard solution test material, with mollusc shell being generally the most variable.

The gradient of the line is a function of the assigned value and the y-axis intercept (-100) is a result of the relative bias values being expressed as percentages; i.e. the equation for a straight line is *y=mx+c* where *m* is the gradient and *c* the y-axis intercept. Equation 5.4 gives the function for the relative percentage bias, thus we now have:

$$y = relative\ bias\ \% = \hat{X}^{-1}(x - \hat{X}) \times 100 \quad \textbf{(5.7)}$$

After expanding the brackets and rearrangement we end up with it in the form for a straight line:

$$y = relative\ bias\ \% = x \cdot 100\hat{X}^{-1} - 100 \quad \textbf{(5.8)}$$

In equation 5.8, $100\hat{X}^{-1}$ describes the gradient as a reciprocal function, thus the smaller the value of the assigned value $\hat{X}$, (or D/L value), the steeper the gradient. These

diagrams therefore help to demonstrate that even small D/L differences can have a significant effect on bias estimates for materials with low D/L values (i.e., very young fossil samples, slow racemizing amino acids or samples with a cold temperature history). Conversely, it can also be said that a large uncorrected bias will have far greater implications for older samples, fast racemizing amino acids or those samples with a warm temperature history. For example, for a sample with a low D/L (e.g.; 0.1), a 10% bias would give an uncertainty contribution of ± 0.01 only, or an expanded uncertainty of ± 0.02, thus the D/L confidence limits would be 0.08-0.12, a difference of 0.04. For an older sample, perhaps with a D/L of 0.7, a 10% bias would give a standard uncertainty contribution of ± 0.07, which expands to ± 0.14, and confidence limits of 0.56-0.84, a difference of 0.28. This is one reason why observed uncertainty increases with age; not only will older samples have been exposed to environmental effects, re-working etc. for a longer time, which may increase sample or site (im)precision estimates, but any systematic value for laboratory/method bias will have a far greater influence. With the exclusion of gross errors and mistakes, precision due to random effects cannot be controlled. Bias however, once identified, should be controlled and where at all possible, significant bias needs either to be corrected for or included in the combined uncertainty estimate.

The second of the paired charts, shown in Figure 5.7ii, is simply a histogram of each participant's relative percentage biases for a given amino acid (equation 5.6), averaged across the six test materials. This therefore, removes matrix specific bias effects and looks for recurring systematic offsets for individual participants.

All RP data are shown in black, IEx are grey and GC white. Results indicate that for aspartic acid, all GC data have a strong positive bias compared to RP (Fig 5.7a.ii). This is also seen in glutamic acid (Fig 5.7b.ii) for GC data quantified using peak heights (laboratory numbers 6.2 and 7.2) rather than peak areas (6.1 and 7.1). For valine (Fig 5.7e.ii), phenylalanine (Fig 5.7f.ii) and leucine (Fig 5.7h.ii), GC data appear to give a negative bias compared to RP, but for both alanine (Fig 5.7d.ii) and isoleucine (Fig 5.7g.ii), data are more normally distributed, showing no clear evidence of any method bias at all. In the individual study reports, IEx data appear at the same position in all the biomineral matrix histograms, just left of centre. Although these data fit comfortably into the normal distribution for allo/isoleucine, there is clearly a systematic effect occurring, but just how significant this is has not been determined.

Figure 5.7: Relative bias distributions for each amino acid.

Amino acid specific, relative bias distributions are shown as line diagrams for each separate test material in Figures a-h (i). Here the gradient of the line is a function of the assigned value in each case; thus even slight variability at low D/L values can have a significant effect on the relative bias calculation and is the reason why precision needs to be determined carefully especially at low levels. These also illustrate how a fixed value for relative bias has increasing significance for samples with higher D/L values, adding to the uncertainty of a measurement result and estimated age.

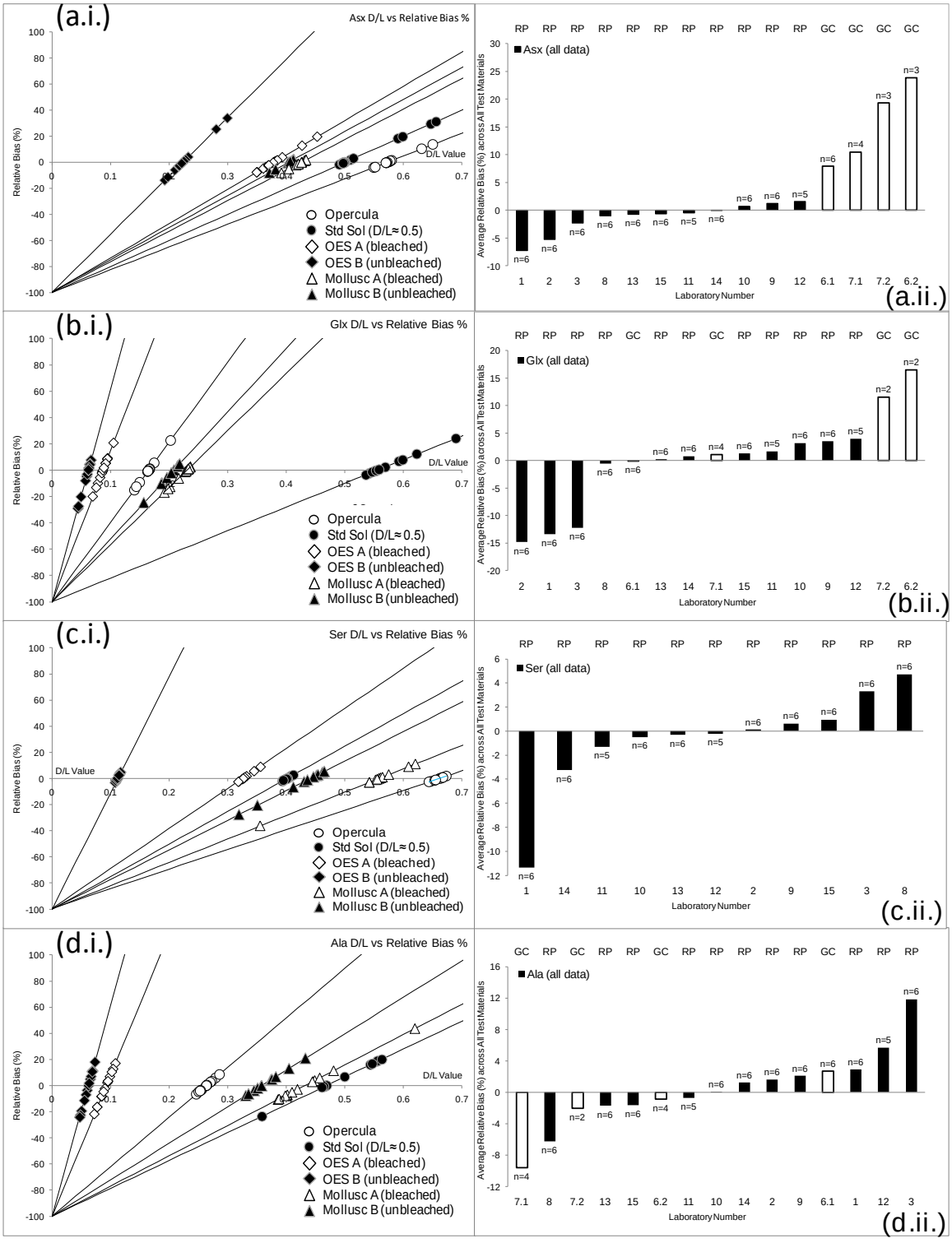
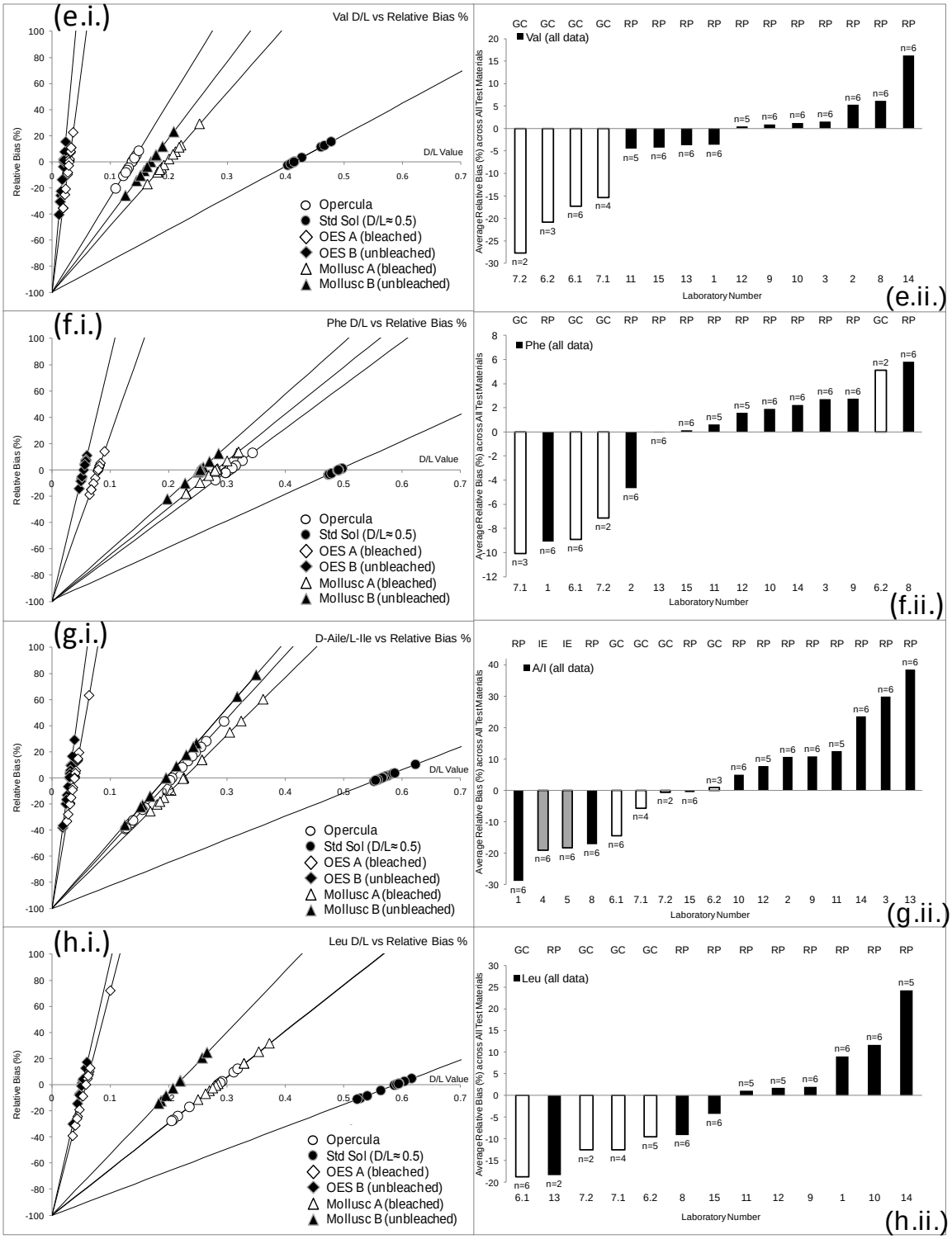


Figure 5.7: Relative bias distributions for each amino acid (continued).

Figures a-h (ii) show histograms of the average relative bias for each participant, averaged across their own submitted results for the 6 test materials. These charts help to identify laboratory specific bias trends. The individual amino acids shown are; a) aspartic acid, (b) glutamic acid, (c) serine, (d) alanine, (e) valine, (f) phenylalanine, (g) isoleucine, (h) leucine.



However, whilst the above data make for some interesting comparisons, it should always be borne in mind that the data represent a very small group of laboratories. Whether these observed biases are genuine methodological differences or simply laboratory biases that would lie at the edge of a normal distribution in a larger dataset, cannot be answered from this study.

5.5 CONCLUSIONS

This paper presents the concept of accuracy as being made up from both precision and bias components. Previous AAR studies have reported excellent precision estimates for repeated analyses and even between samples, although the exact level of agreement depends on the amino acid and matrix studied. In spite of this, on occasions unexplained differences in D/L values between different laboratories and subsequently derived numerical age estimates have been observed. Therefore, whilst the closeness of agreement between data may be tight, it would appear that the mean of the data may at times be slightly askew, or there are larger unaccounted for uncertainties that are not included in the intra-laboratory repeatability precision estimates. Such differences could be explained by method and laboratory bias.

A review of published studies indicates that many AAR uncertainty estimates are currently reported as the precision of analytical results. However, accuracy and therefore uncertainty of analytical data needs to consider elements of both precision and bias. The inability to evaluate and correct for bias is a serious issue and may lead to inaccuracies and an underestimation of uncertainties. On a larger scale, it precludes the direct comparison of AAR results between laboratories and prevents the wider application of the method, such as the development of extended regional or even global aminostratigraphies.

To date, it has not been possible to address bias within AAR geochronology due to the absence of certified reference materials. However, proficiency testing provides a unique opportunity to evaluate individual laboratory bias by comparing an analytical result against a consensus value (the best estimate of the true value of an analyte in a test material).

For comparison with previous AAR inter-laboratory studies, precision estimates were derived from participants' submitted D/L results and found to be amino acid and test material specific. In contrast, bias evaluation tends to be far more method and laboratory specific. By looking at the average relative bias for each individual participant, these trends become more identifiable. From the few GC and IEx results available, a comparison of mean

D/L values (Table 5.2) would suggest that for some amino acids (e.g. valine and A/I), there may be genuine empirical differences between RP, IEx and GC methods. The determination of the D-allo-isoleucine/L-isoleucine value by IEx and GC is historically important. This study demonstrates the close precision achievable for A/I by IEx and GC, which are often smaller than those obtained by RP with known resolution difficulties for this ratio. However, for several other amino acids the observed intra-laboratory CV%s are smaller for RP than for GC, even though agreement between different RP laboratories can be varied. Thus comparability between GC or IEx data with the increasingly popular RP presents some difficulties, as does comparison between different RP laboratories.

In order to address these issues, strategies to evaluate, monitor and correct for bias effects need to be employed. Such strategies might include bias evaluation as part of method validation, recovery correction and calibration in order to accurately interpret recovered chromatogram peak data (Vanatta and Coleman, 2007). However, these activities require the use of Certified or Standard Reference Materials (CRMs or SRMs) with known analyte levels and uncertainties, either as solutions or as matrix-matched substances. Whilst the list of CRMs available today is extensive, it was recognised very early on by the European Commission that supply didn't necessarily me*et a*l the analytical demands, and therefore it was proposed that a collaborative inter-laboratory approach may provide a practical solution to defining fit-for-purpose reference materials (Quevauviller, 1998). It is therefore recommended that a method specific AAR inter-laboratory collaborative trial should be conducted to formally validate instrument specific candidate reference methods, to derive performance precision parameters and define reference values for the analytes in the materials under evaluation. The relative reproducibility standard deviation (RSD_R%) thus derived, can then be used directly as an uncertainty estimate by laboratories, providing their own in-house repeatability estimates are in agreement with published values. Any remaining vials of material under evaluation that have been prepared in sufficient quantities, have been formally tested for homogeneity, can be stored appropriately and are stable for a sufficient period of time, can then be used as fit-for-purpose reference materials.

Although precision and bias are defined independently of each other, for example in the VIM (JCGM 200, 2008), the boundary is not always clear-cut. In this paper, it has been shown how a group of laboratories, each with their own individual method/laboratory bias, can expand the inter-laboratory precision estimate to reflect the additional between-laboratory variability. Therefore, for a single laboratory, the bias may be a fixed value, but

when viewed from a higher level, over several laboratories, the individual biases become a random variable that can be expressed as a standard deviation.

Figure 5.7 shows how for low D/L values, a small imprecision could result in a large relative bias, but for large D/L values, a small relative bias could result in a wide imprecision, both scenarios effectively increasing uncertainty estimates.

The relationships between bias, precision and uncertainty will be further considered in a subsequent paper, where, having determined participants' relative biases, the next stage is to incorporate this information into an uncertainty estimate.

5.6 Acknowledgements

This work has been funded through a Collaborative Studentship provided by the UK Arts and Humanities Research Council (AHRC) with additional support from the Natural History Museum, London. Laboratory facilities were provided by the UK NERC-recognised NEaar Laboratory at the University of York and special thanks go to Richard Allen, Beatrice Demarchi and Molly Crisp for technical and analytical support. Appreciation and gratitude go particularly to John Wehmiller for his enthusiasm, support and provision of mollusc shell material, and also to Darrell Kaufman, Jordon Bright, Katherine Sides, Jose Ortiz, Trinidad de Torres, Colin Murray-Wallace, Terry Lachlan, Eric Oches, and others for their generosity and goodwill, support, advice and assistance in this project.

5.7 Publications

A second draft paper entitled "Results from an Amino Acid Racemization Inter-Laboratory Proficiency Study, Part 2; Measurement Uncertainty Evaluation" was prepared but has not been submitted for publication. This manuscript can be found as Chpt5: Appendix 1.

Anonymous copies of each of the six Proficiency test reports are available on the NEaar website; http://www.neaar.co.uk/reports and have been given as individual appendices to this chapter; Chpt5: Appendices 2-7.

The final version of this manuscript was accepted for publication 01 November 2012 and subsequently published in Quaternary Geochronology 16 (2013) p183-197. Available online at; http://dx.doi.org/10.1016/j.quageo.2012.11.001 (Powell *et al.*, 2013).

5.8 References

Bada, J. L., Hoopes, E., Darling, D., Dungworth, G., Kessels, H. J., Kvenvolden, K. A. & Blunt, D. J. (1979) Amino Acid Racemization Dating of Fossil Bones, I. Inter-Laboratory Comparison of Racemization Measurements. Earth and Planetary Science Letters, 43**,** 265-268.

Bakeman, V. R. (2006) Pacific and Atlantic Coast Mollusc Shells: Chromatographic Amino Acid Racemization Kinetics and Interlaboratory Comparisons. MS thesis. Dept of Geology. University of Delaware, Newark.

Bakeman, V. R. and Wehmiller, J. F. (2006) Analytical Comparisons using Alloisoleucine/Isoleucine for Amino Acid Geochronology. Geological Society of America Abstracts with Programs, 38, No 7, 482.

Barwick, V. J. & Ellison, S. L. R. (2000) Development and Harmonisation of Measurement Uncertainty Principles Part (D): Protocol for Uncertainty Evaluation from Validation Data. Vam Technical Report, Lgc/Vam/1998/088.

Bowen, D.Q. Hughes, S. Sykes, G.A. & Miller G.H. (1989) Land-sea correlations in the Pleistocene based on isoleucine epimerization in non-marine molluscs. Nature, 340, 49–51.

Bright, J., Kaufman, D. S., forman, S. L., Mcintosh, W. C., Mead, J. I. & Baez, A. (2011) Comparative Dating of a Bison-Bearing Late-Pleistocene Deposit, Terapa, Sonora, Mexico. Quaternary Geochronology, 5**,** 631-643.

Clarke, S. J. & Murray-Wallace, C. V. (2006) Mathematical Expressions used in Amino Acid Racemisation Geochronology - A Review. Quaternary Geochronology, 1**,** 261-278.

Demarchi, B., Williams, M. G., Milner, N., Russell, N., Bailey, G. & Penkman, K. (2011) Amino Acid Racemization Dating of Marine Shells: A Mound of Possibilities. Quaternary International, 239**,** 114-124.

Ellison, S. (2002a) Robstat.Xla Version 1.0. Robust Statistics tool Kit based on RSC AMC Technical Brief No 6. Available from: http://www.Rsc.Org/Membership/Networking/Interestgroups/Analytical/AMC/Software/Index.Asp.

Ellison, S. (2002b) Kernal.Xla, Version1.0e. Kernal Density Estimation Based on RSC AMC Technical Brief No 4. Available from; http://www.Rsc.Org/Membership/Networking/Interestgroups/Analytical/Amc/Software/Index.Asp.

EURACHEM / CITAC (2000) Guide Cg 4: Quantifying Uncertainty in Analytical Measurements. S L R Ellison, S.L.R., Rosslein, M., Williams, A. (Eds.) 2nd Edition. Available from; http://www.Citac.cc/Quam2000-1.Pdf.

EUROLAB (2006) Technical Report No. 1/2006. Guide to the evaluation of measurement uncertainty for Quantitative test results. Available from; http://www.eurolab.org/docs/technical%20report/EL_11_01_06_387%20Technical%20report%20-%20Guide_Measurement_uncertainty.pdf.

EUROLAB (2007) Technical Report No. 1/2007. Measurement uncertainty revisited: Alternative approaches to uncertainty evaluation. Available from; http://www.eurolab.org/pub/i_pub.html.

Fearn, T. & Thompson, M. (2001) A New Test for 'Sufficient Homogeneity'. the Analyst, 126**,** 1414-1417.

Hearty, P. J. & Kaufman, D. S. (2009) A Cerion-based Chronostratigraphy and Age Model from the Central Bahama Islands: Amino Acid Racemization and 14C in Land Snails and Sediments. Quaternary Geochronology, 4**,** 148-159.

Hill, R.L. (1965) Hydrolysis of Proteins. Advances in Protein Chemistry, 20, 37-107.

Hollin, J. T. & Hearty, P. J. (1990) South Carolina Interglacial Sites and Stage 5 Sea Levels. Quaternary Research, 33**,** 1-17.

Horwitz, W. (1982) Evaluation of Analytical Methods Used for Regulation of Foods and Drugs. Analytical Chemistry, 54**,** 67a-76a.

Horwitz, W. (1995) IUPAC Protocol for the Design, Conduct and Interpretation of Method-Performance Studies. Available from; www.iupac.org/publications/pac/1995/pdf/6702x0331.pdf. Accessed 11/11/2011.

Horwitz, W., Kamps, L. R. & Bouyer, K. W. (1980) Quality Assurance in the Analysis of Foods and Trace Constituents. J.AOAC, 63**,** 1344-1354.

ISO 13528 (2005) Statistical Methods For Use In Proficiency Testing By Inter-Laboratory Comparisons. Available from the International Standards Organisation's website; http://www.iso.org/iso/home.html

ISO 21748 (2010) Guidance for the Use of Repeatability, Reproducibility and Trueness Estimates In Measurement Uncertainty Estimation. Available from the International Standards Organisation's website; http://www.iso.org/iso/home.html

ISO/IEC 17025 (2005) General Requirements for the Competence of Testing and Calibration Laboratories. Available from the International Standards Organisation's website.

JCGM 100 (2008) Evaluation of Measurement Data - Guide to the Expression of Uncertainty In Measurement (GUM). 1st Edition., Available from; http://www.BIPM.Org/Utils/Common/Documents/JCGM/JCGM_100_2008_E.Pdf.

JCGM 200 (2008) International Vocabulary of Metrology - Basic and General Concepts and Associated Terms (VIM). Available from; http://www.BIPM.Org/En/Publications/Guides/Vim.html

Kaufman, D. S. & Manley, W. F. (1998) A New Procedure For Determining DL Amino Acid Ratios in Fossils Using Reverse Phase Liquid Chromatography. Quaternary Geochronology, 17**,** 987-1000.

Kaufman, D. S. (2003) Amino Acid Paleothermometry of Quaternary Ostracodes from the Bonneville Basin, Utah. Quaternary Science Reviews, 22**,** 899-914.

Kosnik, M. A., Kaufman, D. S. & Hua, Q. (2008) Identifying Outliers and Assessing the Accuracy of Amino Acid Racemization Measurements for Geochronology: I. Age Calibration Curves. Quaternary Geochronology, 3**,** 308-327.

Kvenvolden, K. A. (1980) Interlaboratory Comparison of Amino Acid Racemization in Pleistocene Mollusc, *Saxidomus Giganteus*. In Hare, P. E., Hoering, T.C., and King, K. (Ed.) Biogeochemistry of Amino Acids. Usa, John Wiley & Sons Inc.

Langford, H. E., Bateman, M. D., Penkman, K. E. H., Boreham, S., Briant, R. M., Coope, G. R. & Keen, D. H. (2007) Age-Estimate Evidence For Middle-Late Pleistocene Aggradation of River Nene 1st Terrace Deposits At Whittlesey, Eastern England. Proceedings of the Geologists' Association, 118**,** 283-300.

Lowthian, P. J. & Thompson, M. (2002) Bump-Hunting for the Proficiency Tester - Searching for Multimodality. the Analyst, 127**,** 1359-1364.

Magnusson, B., Naykki, T., Hovind, H. & Krysell, M. (2004) NORDTEST Report TR 537. Handbook for calculation of measurement uncertainty in Environmental Laboratories, 2nd ed. *Available from; http://www.nordicinnovation.net/nordtestfiler/tec537.pdf*.

McCartan, L., Owens, J. P., Blackwelder, B. W., Szabo, B. J., Belknap, D. F., Kriausakul, N., Mitterer, R. M. & Wehmiller, J. F. (1982) Comparison of Amino Acid Racemization Geochronometry with Lithostratigraphy, Biostratigraphy, Uranium-Series Coral Dating and Magnetostratigraphy in the Atlantic Coastal Plain of the Southeastern United States. Quaternary Research, 18**,** 337-359.

Miller, G. H. & Clarke, S. J. (2007) Amino Acid Dating. In: Elias S.A. (Ed.), Encyclopedia of Quaternary Science. 2nd revised ed. Oxford. Elsevier. 41-52.

Miller, G.H. & Hare, P.E. (1980) Amino Acid Geochronology: Integrity of the Carbonate Matrix and potential of Molluscan Fossils. In: Hare, P.E., Hoering, T.C. & King, K. Jr. (Eds.), Biogeochemistry of Amino Acids, Wiley, New York, 415-443.

Miller, G.H., Hollin, J.T. Andrews J.T. (1979) Aminostratigraphy of UK Pleistocene deposits. Nature, 281, 539–543.

Murray-Wallace, C. V., Bourman, R. P., Prescott, J. R., Williams, F., Price, D. M. & Belperio, A. P. (2010) Aminostratigraphy and thermoluminescence Dating of Coastal Aeolianites and the Later Quaternary History of a Failed Delta: the River Murray Mouth Region, South Australia. Quaternary Geochronology, 5**,** 28-49.

Owen, L. A., Bright, J., Finkel, R. C., Jaiswal, M. K., Kaufman, D. S., Mahan, S., Radtke, U., Schneider, J. S., Sharp, W., Singhvi, A. K. & Warren, C. N. (2007) Numerical Dating of A Late Quaternary Spit-Shoreline Complex At the Northern End of Silver Lake Playa, Mojave Desert, California: A Comparison of the Applicability of Radiocarbon, Luminescence, Terrestrial Cosmogenic Nuclide, Electron Spin Resonance, U-Series and Amino Acid Racemization Methods. Quaternary International, 166**,** 87-110.

Penkman, K. E. H. (2005) Amino Acid Geochronology: A Closed System Approach to Test and Refine the UK Model. Unpublished thesis., University of Newcastle.

Penkman, K. E. H., Preece, R. C., Keen, D. H., Maddy, D., Schreve, D. C. & Collins, M. J. (2007) Testing the Aminostratigraphy of Fluvial Archives: the Evidence from Intra-Crystalline Proteins within Freshwater Shells. Quaternary Science Reviews, 26**,** 2958-2969.

Penkman, K. E. H., Kaufman, D. S., Maddy, D. & Collins, M. J. (2008) Closed-System Behaviour of the Intra-Crystalline Fraction of Amino Acids in Mollusc Shells. Quaternary Geochronology, 3**,** 2-25.

Penkman, K. E. H., Preece, R. C., Bridgland, D. R., Keen, D. H., Meijer, T., Parfitt, S. A., White, T. S. & Collins, M. J. (2011) A Chronological Framework for the British Quaternary Based on Bithynia Opercula. Nature**,** 1-4.

Powell, J. & Owen, L. (2002) Reliability of Food Measurements: The Application of Proficiency Testing to GMO Analysis. Accreditation and Quality Assurance: Journal for Quality, Comparability and Reliability in Chemical Measurement, 7**,** 392-402.

Quevauviller, P. (1998) Requirements for Production and use of Certified Reference Materials for Speciation Analysis: A European Commisssion Perspective. Spectrochimica Acta Part B:, 53**,** 1261-1279.

Reichert, K. L., Licciardi, J. M. & Kaufman, D. S. (2011) Amino Acid Racemization in Lacustrine Ostracodes, Part II: Paleothermometry in Pleistocene Sediments At Summer Lake, Oregon. Quaternary Geochronology, 6**,** 174-185.

RSC Analytical Methods Committee (1989) AMC Technical Briefs; Robust Statistics-how not to reject outliers: Part 1, Basic Concepts. the Analyst, 114**,** 1693-1697.

RSC Analytical Methods Committee (2001) AMC Technical Briefs No 6; Robust Statistics: A Method of Coping with Outliers. Available From; http://www.RSC.Org/Membership/Networking/Interestgroups/Analytical/AMC/Technicalbriefs.Asp.

RSC Analytical Methods Committee (2004) AMC Technical Briefs No 17; the Amazing Horwitz Function. Available from; http://www.RSC.Org/Membership/Networking/Interestgroups/Analytical/AMC/Technicalbriefs.asp.

RSC Analytical Methods Committee (2006) AMC Technical Briefs No 4; Representing Data Distributions with Kernal Density Estimates. Available From; http://www.RSC.Org/Membership/Networking/Interestgroups/Analytical/AMC/Technicalbriefs.asp.

Thompson, M. (2000) Towards a unified model of errors in analytical measurement. The Analyst*,* 125**,** 2020-2025.

Thompson, M. & Wood, R. (1993) The International Harmonized Protocol for the Proficiency Testing of (Chemical) Analytical Laboratories. Pure and Applied Chemistry, 65**,** 2123-2144.

Thompson, M., Ellison, S. L. R. & Wood, R. (2002) IUPAC Harmonized Guidelines for Single-Laboratory Validation of Methods of Analysis, (Technical Report). Pure and Applied Chemistry, 74**,** 835-855. Available from; http://www.iupac.org/objID/Article/pac7405x0835

Thompson, M., Ellison, S. L. R. & Wood, R. (2006) The International Harmonized Protocol for the Proficiency Testing of Analytical Chemistry Laboratories. Pure and Applied Chemistry, 78**,** 145-196. Available From; http://www.iupac.Org/Publications/Pac/2006/Pdf/7801x0145.Pdf.

UKAS (2004) TPS 47: UKAS Policy on Participation in Proficiency Testing. In Service, UKAS. Middlesex, UK. Available from; http://www.ukas.com/technical-information/publications-and-tech-articles/publications/Technical_Policy_Statements.asp. Accessed; 11/11/2011

Vanatta, L. E. & Coleman, D. E. (2007) Calibration, Uncertainty, and Recovery in the Chromatographic Sciences. Journal of Chromatography A, 1158**,** 47-60.

Wehmiller, J. F. (1984) Interlaboratory Comparison of Amino Acid Enantiomeric Ratios in Fossil Pleistocene Mollusks. Quaternary Research, 22**,** 109-120.

Wehmiller, J. F., York, L. L., Belknap, D. F. & Snyder, S. W. (1992) Theoretical correlations and lateral discontinuities in the Quaternary aminostratigraphic record of the U.S. Atlantic Coastal Plain. Quaternary Research, 38**,** 275-291.

Wehmiller, J. F. & Miller, G. H. (2000) Aminostratigraphic Dating Methods In Quaternary Geology. In : Noller, J. S., Sowers, J. M., Colman, S. M. & Pierce, K. L. (Eds.), Quaternary Geochronology: Methods and Applications. Washington Dc, American Geophysical Union, Reference Shelf Series 4. 187-222.

Wehmiller, J. F., Thieler, E. R., Miller, D., Pellerito, V., Bakeman Keeney, V., Riggs, S. R., Culver, S., Maillinson, D., Farrell, F. M., Tyork, L. L., Pierson, J. & Parham, P. R. (2010) Aminostratigraphy of Surface and Subsurface Quaternary Sediments, North Carolina Coastal Plain, USA. Quaternary Geochronology, 5**,** 459-492.

Wehmiller, J. F., Harris, W. B., Boutin, B. S. & Farrell, K. M. (2012a) Calibration Of Amino Acid Racemization (AAR) Kinetics in United States Mid-Atlantic Coastal Plain Quaternary Mollusks Using $^{87}Sr/^{86}Sr$ Analyses: Evaluation of Kinetic Models And Estimation of Regional Late Pleistocene Temperature History. Quaternary Geochronology, 7, 21-36.

Wehmiller, J. F., Harris, W. B., Boutin, B. S. & Farrell, K. M. (2012b) Corrigendum to "Calibration Of Amino Acid Racemization (AAR) Kinetics in United States Mid-Atlantic Coastal Plain Quaternary Mollusks Using $^{87}Sr/^{86}Sr$ Analyses: Evaluation of Kinetic Models And Estimation of Regional Late Pleistocene Temperature History." (Quaternary Geochronology, 7, (2012) 21-36). Quaternary Geochronology, 8, 52-53.

Chapter 6. An integrated approach to Site Uncertainty

6.1 Sampling Uncertainty

So far the emphasis of measurement uncertainty determination has been limited to the laboratory activities and the measurement system. Uncertainty estimates, reflecting analytical precision derived by ANOVA, have been determined for standard solutions of mixed amino acids and other stable and homogeneous biomineral matrices. For many commercial laboratories, their responsibilities start from receipt of material sent for analysis, and, but for perhaps some slight consideration towards the homogeneity of sub-samples, uncertainty from sampling plays no role in the uncertainty estimate that accompanies a measurement result.

Depending on the application, this may be perfectly adequate such as the nutritional composition of a product. However in terms of archaeology, we are not only interested in the level of the analyte, but also the level of the analyte for a specific horizon, site, location, depth, etc, of which our material is but a small representation. For geological, chronological and archaeological applications, the scope of any estimate of uncertainty must reach beyond the laboratory and encompass all other potential sources of variation that may impact on the interpretation of the measurement result.

An example of a cause and effect diagram for archaeological sampling may look something similar to Figure 6.1 (adapted from Grøn et al., (2006)), which reflects all possible uncertainty influences from physical and environmental factors acting on the in situ material, to considerations for the statistical representativeness of the sample(s) taken and the impact of sampling, storage and stability of the sample.

Since clearly it is not possible to account for all the individual uncertainty contributions that have acted on a geological or archaeological sample over time, these factors, once again have to be considered from a “top-down” perspective. Samples taken, have to be adequately representative of the material being studied. In an archaeological

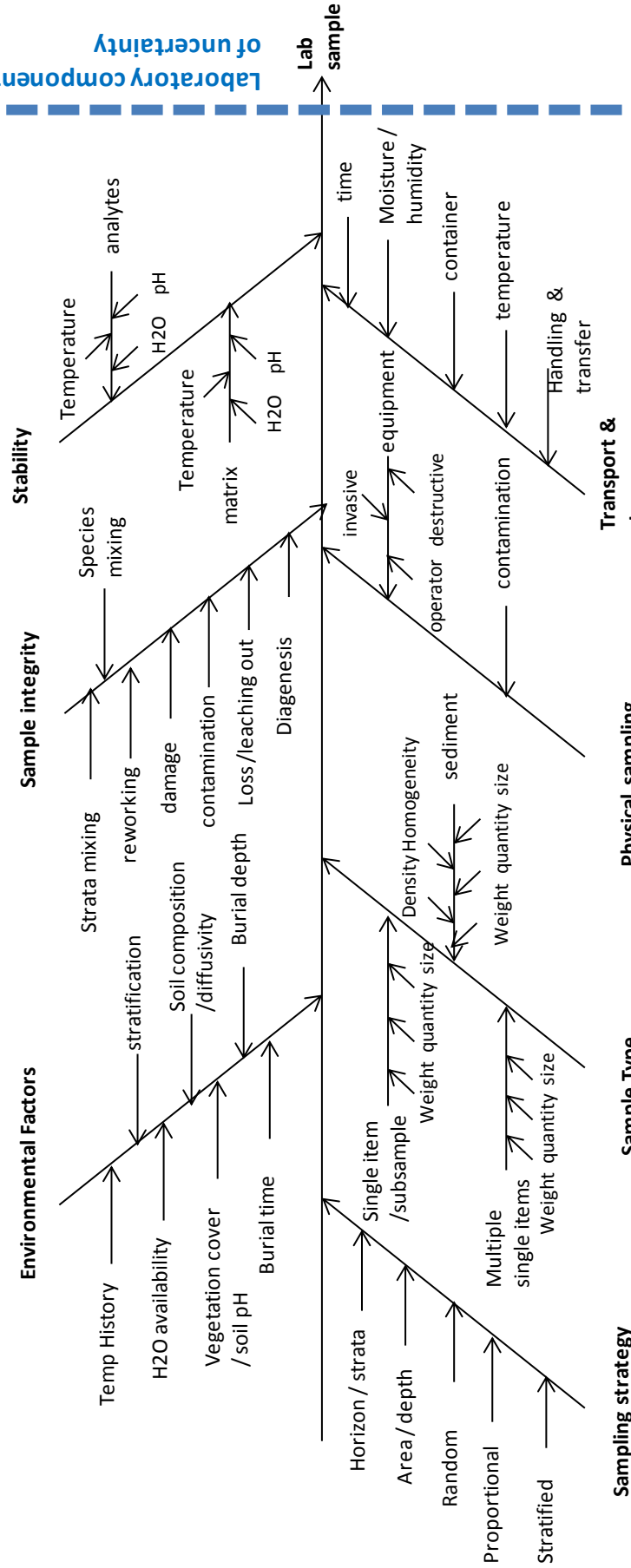


Figure 6.1 : Cause and effect diagram for sampling

context this may be challenging due to the small quantities of recoverable material, but less of a problem in a geological context due to the relative abundance of a sediment layer.

In recent years measurement uncertainty due to sampling has become an essential consideration. The Eurachem/CITAC Guide; Measurement Uncertainty arising from Sampling (EURACHEM/EUROLAB/CITAC/Nordtest/AMC, 2007), is a joint document contributed to by major European bodies involved in providing analytical expertise to the analytical community (Eurachem, EUROLAB, Nordtest and the RSC Analytical Methods Committee). The Guide compares the empirical or "top-down" approach to the modelling or "bottom-up" approach for determining uncertainty from sampling. As discussed in previous chapters, the "bottom-up" approach seriously risks under or over estimating the effects of individual uncertainty contributions. Figure 6.1 illustrates the complexity and near impossibility of accounting for individual uncertainty sources. For this reason the "top-down" approach is definitely to be favoured and the Guide provides a detailed discussion on the application of the duplicate or balanced design approach, based around ANOVA. However, whilst helpful examples are provided in the Guide's Appendices A1-6, the explanatory text (section 9.4) is poorly written with confusing terminology (although a list of definitions is provided in Appendix B) with inadequate statistical explanation. Alternatively the Nordtest Report 604 (Grøn *et al.*, 2006), provides a more user friendly account taking the reader from the simple differences between replicates to more complex ANOVA calculations.

However the principle of the replicate or duplicate design in both texts is fundamentally the same, that is *"...to apply the same sampling procedure two or more times on the same target or on different targets to estimate the random measurement error, preferentially at least 8 times for each calculation.*" (Grøn *et al.*, 2006, p17).

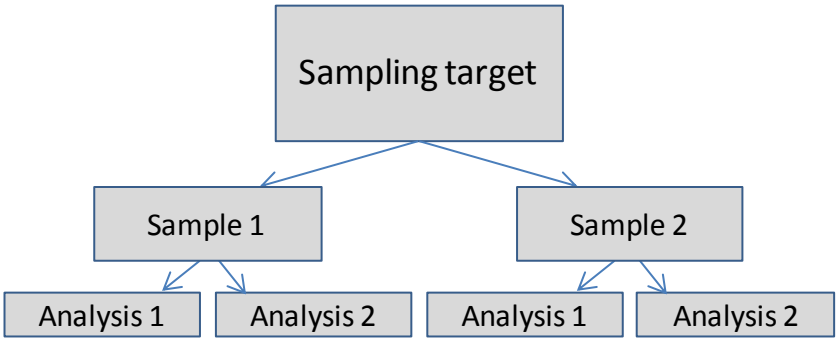
Sadly, interpreting the official Eurachem/CITAC guidance document isn't quite so simple. The sampling target is defined as "*Portion of material, at a particular time, that the sample is intended to represent*" and a primary sample is defined as "*The collection of one or more increments or units initially taken from a population*" (EURACHEM/EUROLAB/CITAC/Nordtest/AMC, 2007, Appendix B). The advice in the Eurachem/Citac sampling guide is that duplicate samples should be taken*...."(i.e. 10% but no less than eight targets) of the primary samples [Ramsey, 1998; Lyn et al.,2007] .*" (p17). For each sampling target, duplicate samples should be taken according to the sampling protocol. The duplicate samples are subjected to physical preparation to give two separate test samples. Each of these is then further split to give a further pair of test portions, which are

then analysed in duplicate (section 9.4.2, p17). The implication from the last sentence is that each of the test portions should be analysed twice, i.e.; two injections, which is not the case, as repeatability precision is dependent on independent analyses, that is, two separate samples, not separate injections of the same sample.

The technical guidance note written by the RSC Analytical Methods Committee isn't any clearer (RSC Analytical Methods Committee, 2009). "*Random duplicate primary samples are taken at 10% (n≥8) of sampling targets"*. Which would suggest that it is the primary samples that are taken at *n≥8,* (but shouldn't that be sample targets, and what are they anyway?). Further in an example given for soil, it is suggested that primary samples are made up from 4 increments each. Perhaps these 4x2 samples represent the 8 sampling targets referred to above? In addition, the figure given to illustrate the sampling design (Figure 6.2), (EURACHEM/EUROLAB/CITAC/Nordtest/AMC, 2007, p18, Fig 2; RSC Analytical Methods Committee, 2009), is no clearer on whether the analysis samples are dependent or independent. It's all very confusing and little wonder sampling uncertainty tends to get ignored.

Figure 6.2: Analytical sampling design

Diagram taken from Eurachem/Citac Sampling Guide, Fig 2, p18.



However, Example A1 given in the Eurochem/Citac sampling guide, helps to shed some light on the issue using the example of lettuces commercially grown in a number of bays, each bay being considered a target.

Eight bays were selected at random to give **eight Sampling Targets (1-8)**, and two, ten lettuce head samples taken from each bay to give **eight pairs of Primary Samples (Sample 1.1 and Sample 1.2.....Sample 8.1 and Sample 8.2**). Portions of each of the 10 heads in each of the primary samples were taken and macerated (**physical preparation**) to

give **16 primary composite samples**. From each primary composite sample, two sub-samples were taken (**Analysis Sample 1.1a and 1.1b, 1.2a, 1.2b.....8.2a, 8.2b**) and worked up through the entire measurement procedure and analysed, to give **32 independent measurement results**. This process is illustrated in Figure 6.3, which shows a balanced, two split level design for precision estimation.

Analysis of replicate sampling in this way can be used to separate out the uncertainty contributions due to sampling and analysis. When carried out under repeatability conditions, the analytical contributions to the uncertainty estimate are due to random error effects only. Systematic error influences arising from the analysis can be determined using CRMs, spiking and recovery analyses, method validation data or even proficiency test results. However, sampling bias is generally ignored (RSC Analytical Methods Committee, 2009). Nordtest presents a simplified method of evaluation of the analytical results using range statistics based on the differences between replicates, (Grøn *et al.*, 2006). However since ANOVA is designed to separate out sources of uncertainty, it is ideally suited, giving more detailed information.

Based on the balanced, two split level scheme in Figure 6.3, different ANOVA designs can be used to determine the various precision estimates. Note that if carried out during the same analytical run, all precision estimates are repeatability determinations. Depending on the precision estimates required, different arrangements of analytical results will derive different precision values. Figure 6.4 demonstrates the arrangement of measurement results derived from the split two level design given in Figure 6.3. All calculations are based on the same use of within and between sample mean squares (MS_w and MS_b) previously discussed in Chapter Chapter 3, where; $(within)s_r^2 = MS_w$, $(between)s_L^2 = \frac{MS_b - MS_w}{n}$ $(total)s_R^2 = s_r^2 + s_L^2$.

Thus, with different ANOVA arrangements, it becomes possible to determine analytical precision, sampling precision, between-target precision and total precision (Figure 6.4). A robust version of ANOVA; RANOVA, minimises the effect of outliers for normally distributed data. Software is available to download from the AMC website (http://www.rsc.org/Membership/Networking/InterestGroups/Analytical/AMC/Software/ROBAN.asp). The program ROBAN (version 1.01) applies a hierarchical ANOVA to data and provides estimates of the total, between-target, sampling, analytical and measurement uncertainty estimates, where measurement precision is the combination of the sampling and analytical precision estimates combined (RSC Analytical Methods Committee, 2009).

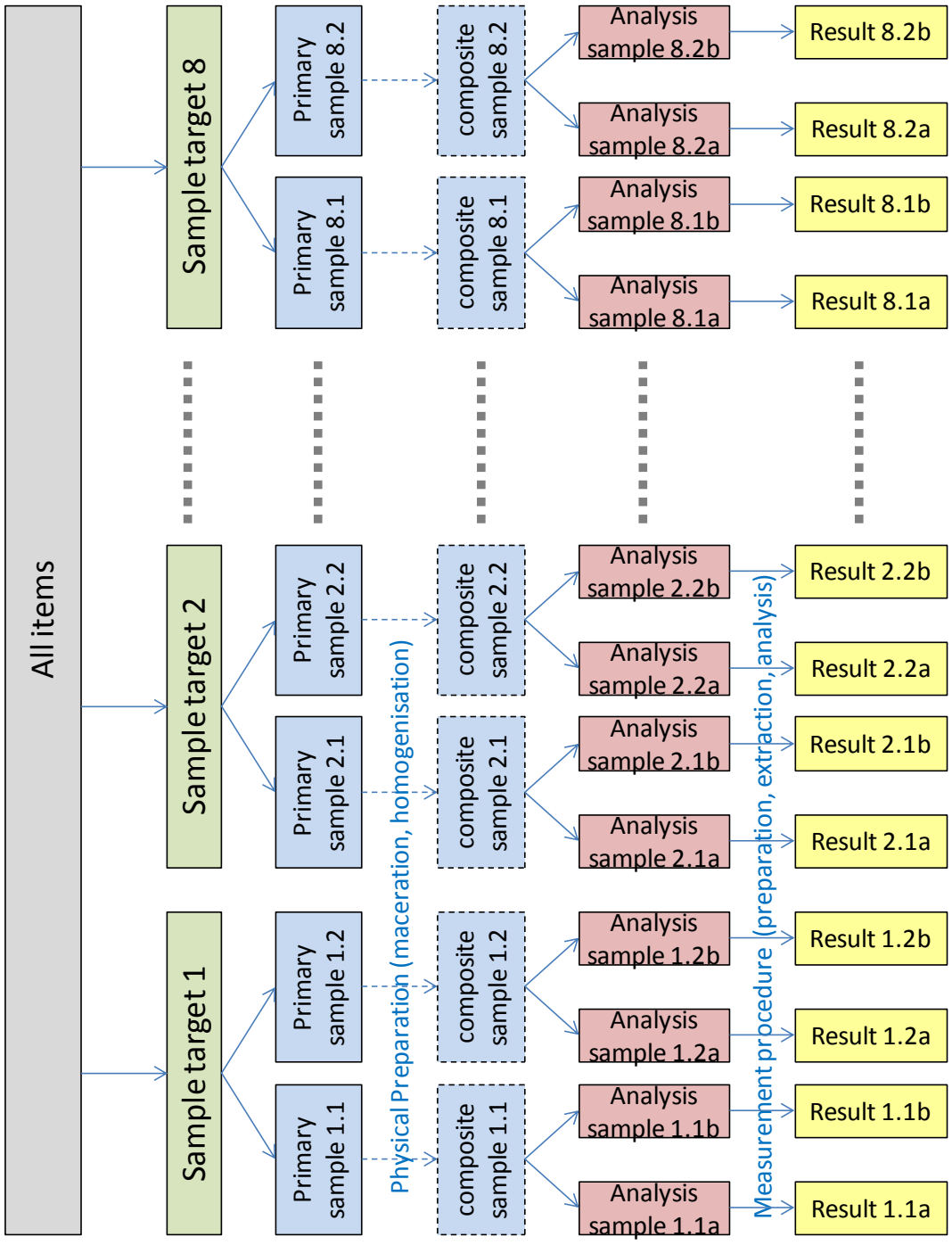


Figure 6.3: Strategy a balanced, two split level design for determining measurement uncertainty from sampling.

Figure 6.4: Arrangement of ANOVA data to derive different precision estimates

Analytical Variance $s_a^2 = MS_w$		
Primary sample	Analysis a	b
1.1	1.1a	1.1b
1.2	1.2a	1.2b
2.1	2.1a	2.1b
2.2	2.2a	2.2b
↓		
8.1	8.1a	8.1b
8.2	8.2a	8.2b

Sampling Variance $s_s^2 = MS_w$ Between-Target Variance $s_T^2 = (MS_b=MS_w)/n$		
Sample Target	Primary sample 1 Analysis mean (a+b/2)	Primary Sample 2 Analysis mean (a+b/2)
1	1.1	1.1
2	2.1	2.1
↓		
8	8.1	8.1

Total Variance (repeatability) $s_r^2 = MS_w + (MS_b=MS_w)/n$				
	Primary sample 1		Primary Sample 2	
Sample Target	Analysis a	Analysis b	Analysis a	Analysis b
1	1.1a	1.1b	1.2a	1.2b
2	2.1a	2.1b	2.2a	2.2b
↓				
8	8.1a	8.1b	8.1a	8.2b

6.1.1.1 Sampling for AAR

As far as sampling for AAR purposes, the chances of being able to isolate 8 sampling targets and acquire eight duplicate primary samples, seems remote, especially if working from a bag of sediment previously collected for opercula, or perhaps a small collection of shells. Whilst it is possible to image eight potential sampling target regions from a specific stratum running across a cliff face or perhaps the exposure of a shell bed during excavation, the exposed area is likely to be limited in size and would probably not warrant such detailed sampling.

Infact, the Eurachem/Citac sampling guide (p17) says that, "*If only one target exists, then all eight duplicates can be taken from it, but the uncertainty estimate will only be applicable to that one target.*" This would imply that all 16 samples should be taken from the same target region. Depending on the sampling area size, this might pose particular difficulty

in identifying the sample pairs. Especially if each sample is already made up of a composite, as it is unlikely there will be any difference between samples.

In this situation it is suggested that a level of sampling be omitted from the design and individual representative samples taken, either at intervals or as composites, which are then split to give two separate analytical portions for preparation, extraction and measurement (Figure 6.5). The number of individual primary samples taken will be a compromise between precision, instrumental run time, expense and available material. Previously, each sampling target was only sampled twice, whereas in this context, the bag of sediment is seen as a single sampling target. To only take two primary samples would be inadequate in the absence of other sampling targets. In contrast, the Eurachem/Citac sampling guide would seem to suggest that the number taken should be 16. If so, then with QC materials, the total number of samples to be analysed would be over 40 and take nearly 4 days to run by RP. This would seem excessive for one bag of sediment.

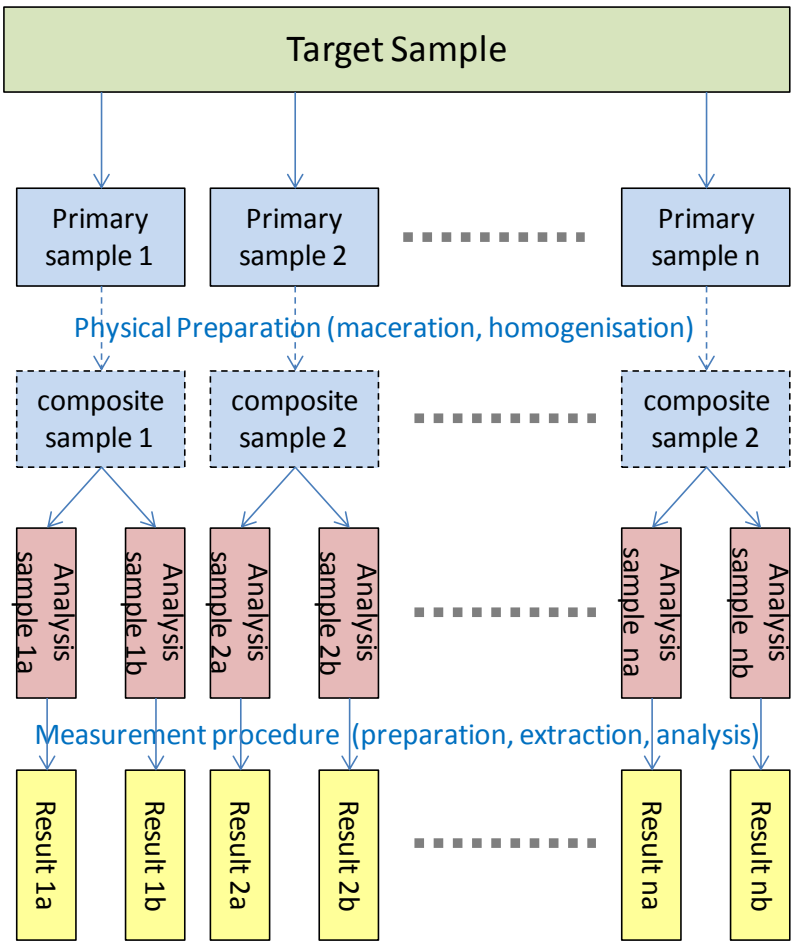
Remembering that each sample has to be further split to give two separate samples for preparation and analysis, it is suggested that at the very least 3 samples should be taken, better still, 5 and ideally 8 or more in a balanced, single split level design shown in Figure 6.6.

Thus the overall target repeatability precision estimate, i.e.; the uncertainty (due to precision) for the bag of sediment or shells etc, can be derived using the classical ANOVA design, where the total variance is the same calculation used for reproducibility precision, previously in Chapter Chapter 3.

Figure 6.5: Suggested Arrangement of ANOVA data for AAR analysis

Analytical Variance	$s_a^2 = MS_w$	
Sampling variance	$s_s^2 = (MS_b=MS_w)/n$	
Target Variance (repeatability)	$s_r^2 = MS_w + (MS_b=MS_w)/n$	
Primary sample	Analysis	
	a	b
1	1a	1b
2	2a	2b
↓		
n	na	nb

Figure 6.6: Suggested balanced, single split level design for determining measurement uncertainty from sampling in AAR.



6.1.1.2 Physical Preparation

Finally, a comment regarding the physical treatment of the primary samples. Routinely it would seem that individual opercula are taken for analysis. However, in order to provide a better estimate of the variation in the sample a homogeneous composite sample should be made from several opercula, perhaps 10 or more ground to a fine powder, and this should then be sub-sampled to give the two replicates used for the preparation , extraction and analysis. However it may be that for individual shells this would be impractical. In this case, several shells should be drawn from the bulk to make up each primary sample. If AAR analysis is carried out on a specific layer of the shell, then two separate shells would then be taken and worked up individually to give the two analytical replicates.

Note that in all cases, repeated instrumental injections can be carried out to give additional accuracy on the analytical result for each sample, but this is not necessary if

duplicate samples are run instead. If repeated injections are carried out then they should be averaged and the mean used as the value for the sample. The injection data itself is not directly evaluated as an estimate of precision, but is subsumed into the sample repeatability precision estimate.

Whilst it is assumed that all these estimates are derived from analytical results determined during a single run, (repeatability conditions), additional contributions would need to be included in the overall combined uncertainty estimate to reflect bias uncertainty components. However, additional levels of complexity can be built into the model to reflect between-run bias (intermediate reproducibility conditions) if repeated measurement procedures are carried out over multiple days. The evaluation of multi-level nested designs are covered in ISO 21749:2005 and are similarly based on outputs from ANOVA evaluations.

6.2 Determination of AAR uncertainty estimates for UK Archaeological Sites

In terms of evaluating existing data, once again a "top-down" perspective is adopted and ANOVA used to separate out the sources of uncertainty and determine overall precision. During the initial organisation of the data prior to evaluation, it was noticed that whilst for the majority of the time, replicate measurements that were reported related to repeated injections carried out on the same sample vial, in the same well position on the auto-sampler, although this was not always the case. Occasionally it was found that replicate samples with the same NEaar reference code, were carried out on different analytical runs. It is appreciated that from time to time it is necessary to stop a run, perhaps due to instrumental issues, and later start again with the next sequential run number. However this does not account for non-sequential run codes. It was also common for repeated measurements of the same material to be carried out on a different instrument, but it was not known whether this was simply a re-test of an existing extract or a new sub-sample from the original material worked up through the whole method preparation and extraction stages. It was also noticed that for the most part, replicates were carried out during the same run (repeatability conditions), but again this was not always the case and occasionally repeat measurements were taken, for the same material (NEaar reference), on different days, introducing an additional level of uncertainty into the data.

Fortunately, for the majority of site data, more than one physical sample had been taken from an individual site (multiple NEaar references), therefore in addition to repeated

injections, were replicated measurements on independent samples, which is ideal for the evaluation of precision estimates. However, whilst multiple samples were available, there was a general inconsistency as to whether these samples were analysed in the same run or across different ones and between instruments on different days. Thus there is a mixing of precision conditions for the repeated samples analysed.

Because of the size of the data set, (over 7000 samples for opercula alone), it was not possible to separate out repeatability and reproducibility data and results have been assessed using data exactly as recorded on the NEaar Excel spreadsheet. Data were evaluated based on the specific location recorded against each Quaternary site. For example, a specific site may have 3 or 4 different locations referenced to it, indicating samples had been taken from different areas, depths or trenches etc. As each location potentially represents a different set of environmental conditions, differences in D/L values and differences in age may also exist. Without additional information regarding the sampling details, each have therefore been assessed independently.

As there are eight well characterised amino acids to choose from, (Asx, Glx, Ser, Ala, Val, Phe, Leu and A/I), valine was selected as this is the slowest racemising amino acid and likely to cover the greatest time span. A faster racemising amino acid such as aspartic acid, would provide better resolution between younger site D/L values, however, depending on the age and temperature of the sites, it is possible that the amino acid would have reached a fully racemic state (D/L =1) fairly early in terms of geological time, and would therefore be unable to differentiate between older samples.

Therefore valine D/L values, previously determined using the laboratory's existing measurement procedure, have been assessed by ANOVA. The within-sample repeatability element (s_r), for the most part, represented repeated injections, (i.e.; a, b, c etc,) whilst the between-sample variance (s_L), for the most part, representing precision between samples from the same location. The total variance of a site was determined as the intermediate reproducibility precision estimate, (s_{RW}) using the equation; $s_{RW}^2 = s_r^2 + s_L^2$. However it should be emphasised that whether s_L and s_R represent the repeatability or an intermediate precision estimate is entirely dependent upon the analytical conditions under which the D/L data were originally acquired. Consequently, whilst the mean D/L values for all the sites are directly comparable with each other, the precision estimates may not necessarily be.

All valine D/L values and precision estimates were evaluated using an unbalanced design of ANOVA to accommodate uneven numbers of replicates (*n*) for site locations (see section 4.3).

Occasionally, there were locations where only single replicate values were available for samples and s_r values could not be determined, or, only single samples were available for a given location, in which case s_L values could not be determined. In both cases, the absence of either the s_r or s_L meant that the overall estimate of precision, s_R, could not be determined either. To overcome this, all s_r or s_L data were plotted against D/L value and approximations derived. Although these were only very rough approximations using the predictive curves derived from mixed condition data, it nonetheless permitted an informed approximation for an initial assessment.

Precision estimates derived by ANOVA for s_r, s_L and s_R, represent the observed distributions (within and between) of single values. However, it seems reasonable to express the measurement result from a number of independent measurement values, as the mean of those values. The larger the number of measurements made, the greater the confidence there is that the mean is in fact a reasonable approximation of the true value. Therefore, the larger the sample number, the smaller the uncertainty.

This relationship between sample number and uncertainty is reflected in the expression of the standard uncertainty derived from repeated measurements of independent samples (also known as the standard error of the mean or experimental standard deviation), thus; *std u = s/√n.* Similarly, repeatability and reproducibility can be adjusted for mean values too.

If, $s_R = \sqrt{s_r^2 + s_L^2}$ then it would normally be expected that if s_r was determined from several independent measurements (n), then $\bar{s}_R = \sqrt{\frac{s_r^2}{n} + s_L^2}$, note that s_L remains unaffected as ideally we need to know the uncertainty for a single measurement carried out at any time by the laboratory, by any analyst on any instrument. Therefore the estimate of the between-run precision in this context, s_L, remains unchanged.

However, with the current data, s_r represents repeated injections, not samples, and s_L represents the between-sample uncertainty, regardless of whether it was analysed during

the same analytical run or not. Therefore our estimate for single values must reflect the average number of samples (*p*), not just replicate injections (*n*).

Thus $\bar{s}_R = \sqrt{\frac{s_r^2}{n} + \frac{s_L^2}{p}}$ is required for the present data. In hind-sight, the 1/*n* could have been ignored since s_r from repeated injections contributes only a very small amount to the overall uncertainty compared to the influence of uncertainty between samples.

For those samples where either *n*=1 or *p* = 1, and estimates of s_r and s_L have been determined from plotted data as previously described, average values of n and p have been used, derived from the whole data set; giving *n*=2 and *p*=4.

A retrospective evaluation of all accumulated opercula AAR data has been carried out using ANOVA as described. An Excel spreadsheet showing these calculations and results is presented as Chpt 6: Appendix 1 and charts are shown in Figure 6.7. D/L values for site locations are plotted in order of ascending D/L value, thus low D/L values are the youngest sites and gradually increase in age, (assuming all sites share a common temperature history). Solid black symbols indicate observed s_R values from data, unfilled symbols are predicted s_R values when either n=1 or p=1. Charts also show the instrument used and the effective degrees of freedom as *df*.

Expanded confidence limits (2 std dev) have been determined for single values, assuming a normal distribution and a coverage factor of *k*=2 and are shown as the outer dotted lines in Figure 6.7. Thus;

$$D/L \pm 2 \times s_R = 2 \times \sqrt{s_r^2 + s_L^2} \qquad \textbf{(6.1)}$$

In addition, expanded confidence intervals for means have been given, using both *k*=2 and $k=t_{(\alpha=0.05,\ veff)}$, where t is the t-value at 95% probability level (≈2 std dev), and v_{eff} is the effective degrees of freedom determined by the Welch-Satterthwaite Equation (GUM, G.4.1 p73) using relevant values of n and p for v_i;

$$v_{eff} = \frac{u_c^4(y)}{\sum_{i=1}^{n} u_i^4(y)/v_i} \qquad \textbf{(6.2)}$$

therefore;

$$v_{eff} = \frac{(s_r^2 + s_L^2)^2}{\frac{s_r^4}{n-1} + \frac{s_L^4}{p-1}} \qquad \textbf{(6.3)}$$

Confidence intervals for means calculated as;

$$\frac{D}{L} \pm 2 \times s_R = 2 \times \sqrt{\frac{s_r^2}{n} + \frac{s_L^2}{p}} \qquad \text{and} \qquad \textbf{(6.4a)}$$

$$D/L \pm t_{(\alpha,veff)} \times s_R = t_{(\alpha,veff)} \times \sqrt{\frac{s_r^2}{n} + \frac{s_L^2}{p}} \qquad \textbf{(6.4b)}$$

The inner solid line represents expanded confidence limits for means using *k*=2, whilst the error bars indicate expanded confidence limits for means but with *k*=$t_{(\alpha=0.05,\ veff)}$.

From the charts it can be seen that very often, those locations with the widest confidence limits are those with the lowest effective degrees of freedom. Whilst this data reflects the uncertainty due to the precision of observed data and is location specific, it is sensitive to sample size. However, whilst this provides important information pertinent to the site location and should be retained, for geochronological purposes, it may be more instructive to be able to model the uncertainty around an observed D/L mean. This has been done using valine D/L data and is described in detail in the following section.

6.3 Modelling uncertainty using associations between amino acids

The use of closed system methodology (Penkman *et al.*, 2008) ensures that all amino acids released during hydrolysis originate from the intra-crystalline proteins within the biomineral matrix. Since all amino acids from a sample are measured from the same single injection volume taken for RP analysis, each measurement result is matched with equivalent measurement results from different amino acids. When individual D/L values for one amino acid are plotted against the D/L value of another, close associations between different amino acids can be seen and have long been recognised. Figure 6.8 illustrates how D/L values for many of the amino acids (y-axis) behave in a predictable way over time. As valine is one of the slowest racemising amino acids, data have been plotted against total hydrolysable valine D/L values as a measure of relative time along the x-axis. Whilst taken together, Figure 6.8 may look much like a piece of modern artwork, however when viewed separately, close associations based on differing patterns of protein degradation within the biomineral matrix, can be seen. In addition, each association has its own unique pattern of scatter which changes over time, and reflects the changing uncertainty in D/L values for different amino acids with increasing age.

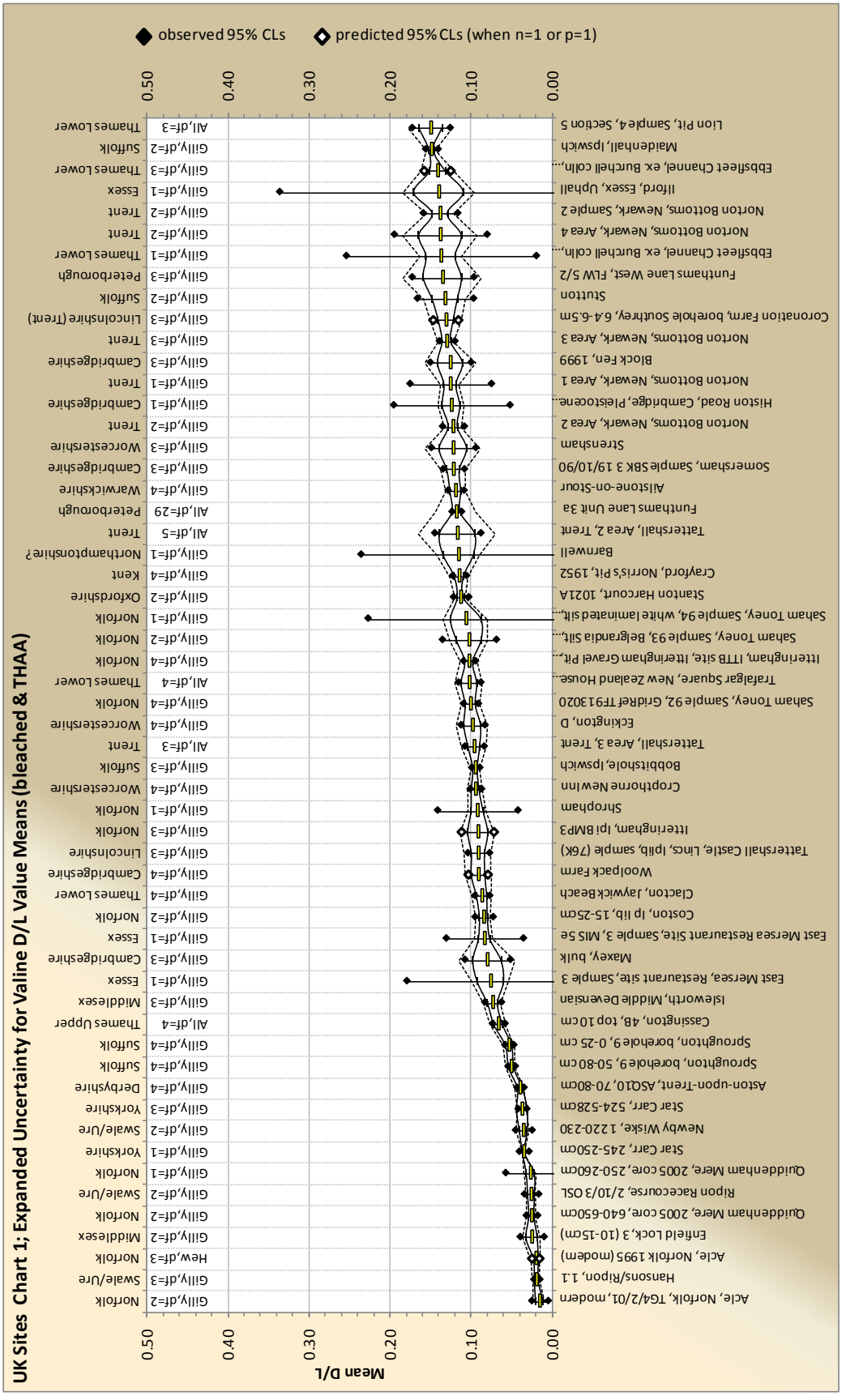


Figure 6.7: Retrospective evaluation of UK AAR site data derived from Bithynia opercula.

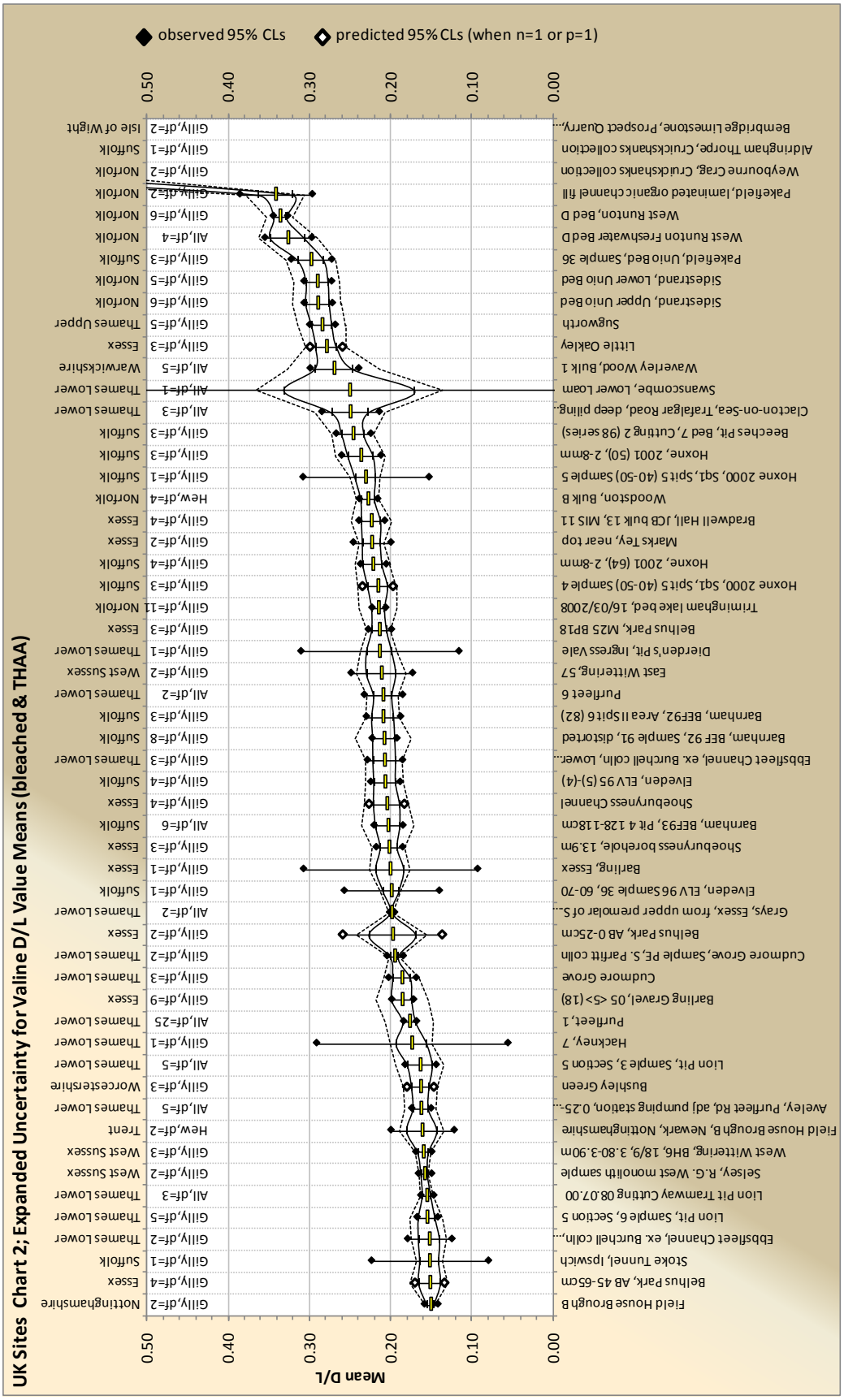
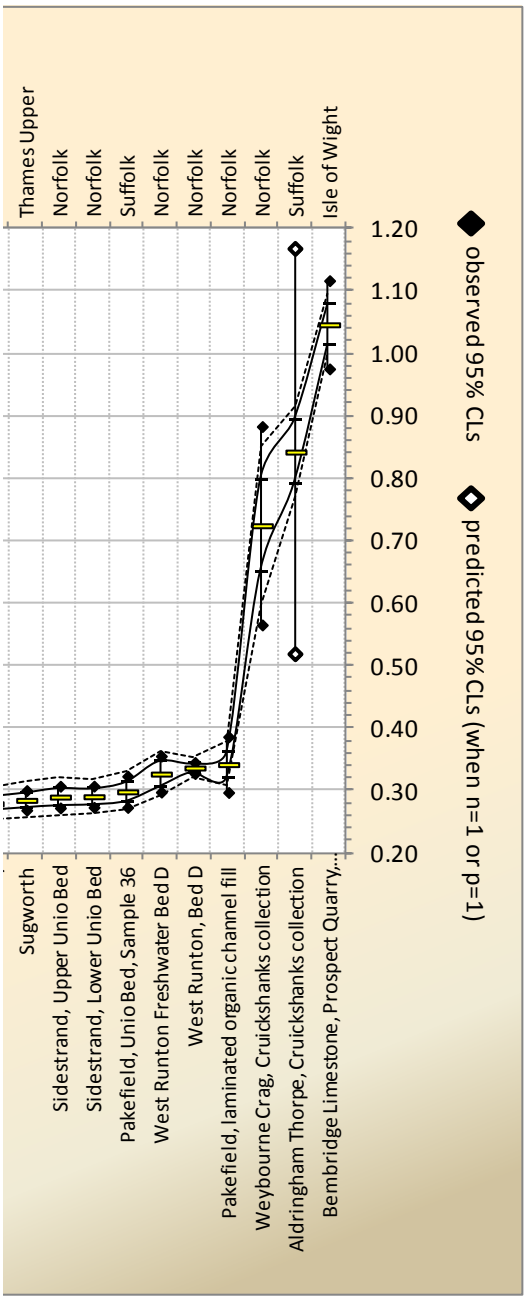


Figure 6.7: Retrospective evaluation of UK AAR site data derived from Bithynia opercula (continued).

Figure 6.7: Retrospective evaluation of UK AAR site data derived from Bithynia opercula (continued).



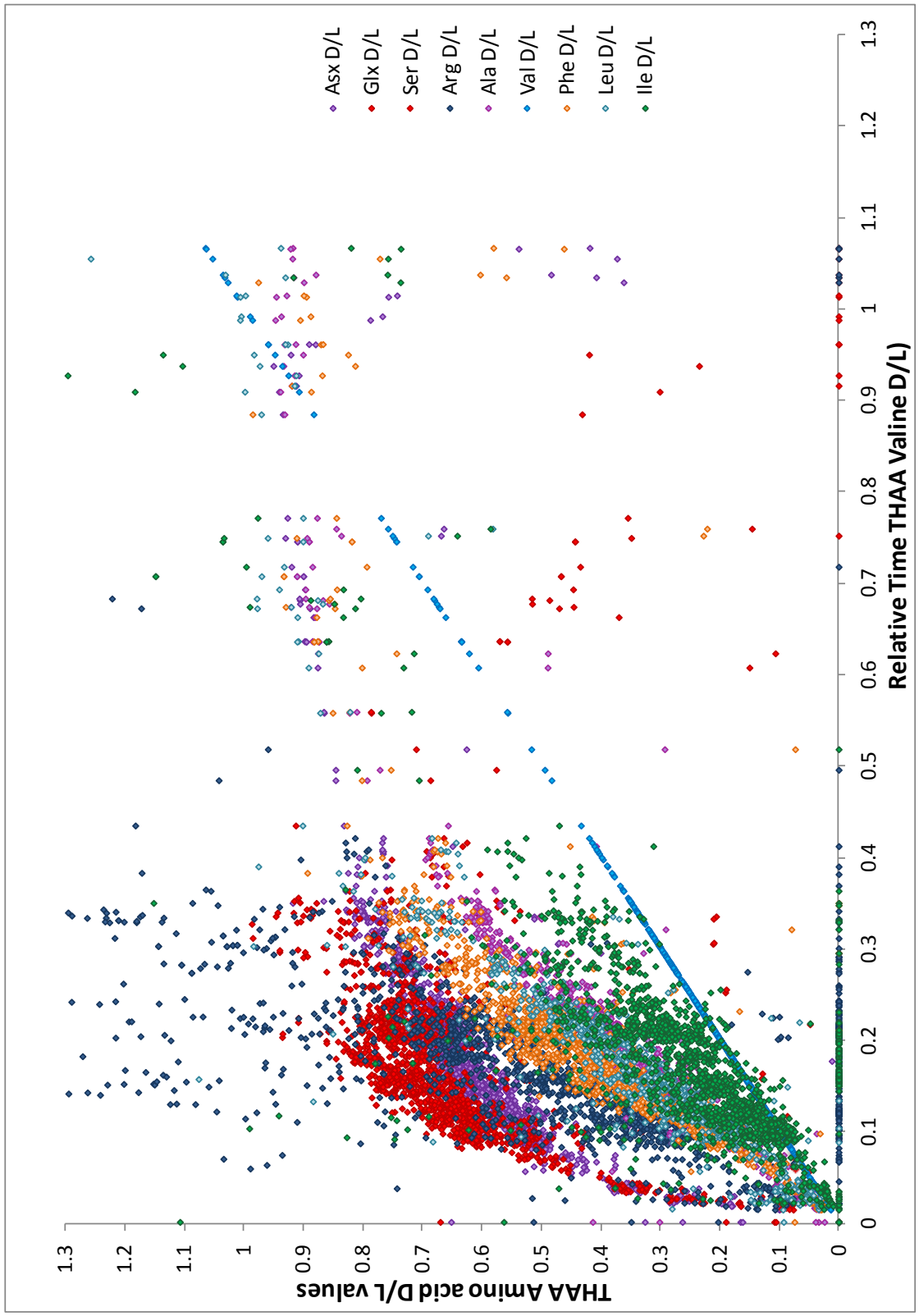


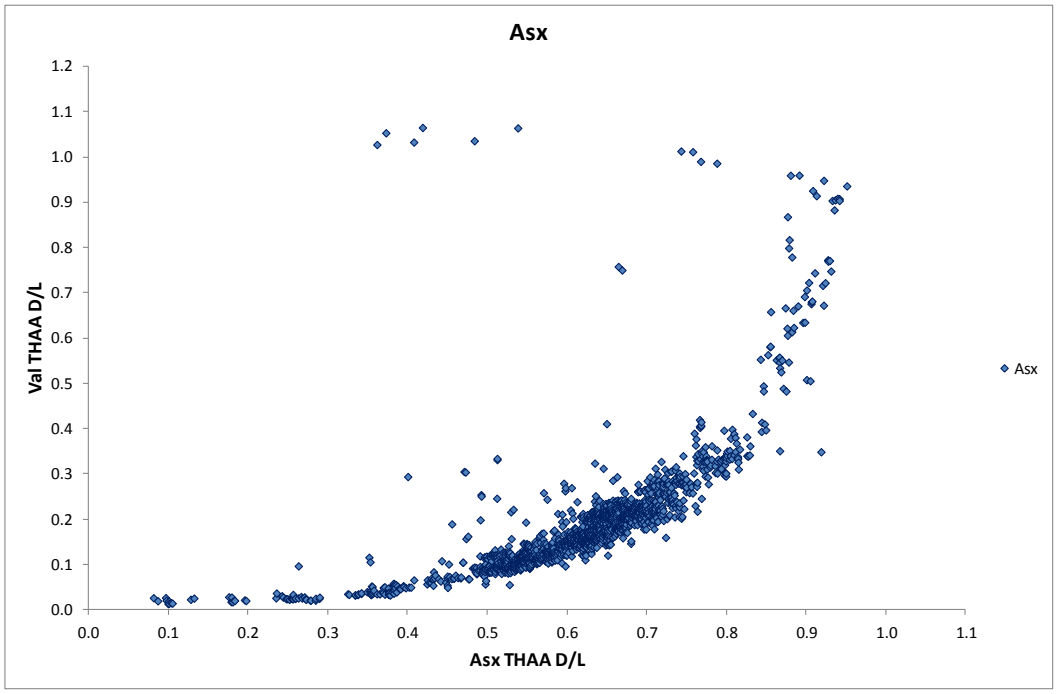
Figure 6.8: Correlations between amino acids (THAA) and valine (THAA) as an indicator of relative time

It would therefore seem to be a sensible idea to try to access this information and use these relationships to model the uncertainty, in this instance, of valine, using faster racemising amino acids to improve the resolution of valine data for younger samples, thus constraining the observed location uncertainty, whilst providing additional evidence to strengthen the confidence in the observed valine D/L value itself.

Figure 6.9 shows the relationship between Val and Asx D/L values, (with the axes swapped around). The initial rate of racemisation for Asx is much faster than for Val. By using the Asx D/L value to determine the equivalent Val D/L, especially at low levels, far better resolution of the Val D/L values is achievable. It can also be seen how tight the data points are early on, with very little dispersion which steadily increases for older samples (higher D/L).

Charts by convention, tend to plot the dependent variable along the y-axis and independent variable along the x-axis. Trendlines based on least squares regression, minimise the sum of the squared residuals of the dependent variable and provide equations y =*f*(x). Unknown values of y can then be determined using the equation, y off x. In this example with Asx, the dispersion of data and therefore the uncertainty in the y-direction early on, is particularly small. Therefore predicted Val D/L values derived from Asx D/L values (especially up to about 0.4), will have a high level of confidence associated with them and could be used to support and inform the observed site data.

Figure 6.9: Correlation between Val and Asx D/L (THAA) values



6.3.1 Model development

6.3.1.1 Trendline fitting

Using all site data, both FAA and THAA D/L values for each amino acid were plotted against valine THAA D/L data. As we are interested in the **association between amino acids, within the same biomineral extract,** it doesn't matter whether the data were acquired by repeat injections or from separate samples. Therefore all data points can be treated as independent. In all cases valine THAA D/L were plotted on the y-axis against a corresponding amino acid D/L value derived from the same analytical run.

The extraction processes for THAA and FAA fractions are different; THAA requiring hydrolysis and FAA requiring demineralisation. Therefore, two different sub-samples need to be taken and worked up through the respective extraction processes independently. For the most part, samples were analysed for both the THAA and the FAA, and very often both fractions were also analysed in duplicate. In this case, replicate (a) for the Val THAA was matched with replicate (a) for the FAA, and similarly (b) Val THAA with (b) FAA. However, since THAA and FAA are independent from each other it would have been equally valid to have matched (a) Val THAA with (b) FAA and vice versa. In hind-sight it would perhaps have been better therefore to have taken the average of THAA Val with the average FAA for a given location, but the difference in all probability, would likely to have been minimal due to the large dataset size used.

Having plotted Val THAA D/L on the y-axis against all other THAA and FAA variables on the x-axis, curves were fitted to each using Graphpad Prism software; (http://www.graphpad.com/scientific-software/prism/), designed originally for biologists and uses nonlinear regression to fit curves to nonlinear data. The exception to this was for A/I data, which for both the THAA and FAA data approximated closest to a linear model.

For the Prism functions selected, curve fits were based on minimising the sum of the squared y residuals, to find the best line with which to predict y from x. Consequently, the lines drawn by the software didn't necessarily "fit" with all the data as it ignores the x variable error values. However, what was needed was a line that allows us to describe how the two variables are related, that goes through the middle of the data and describes the association, allowing for variation in both the vertical and horizontal directions. This can readily done for linear regression, using alternative regression models that minimise the

combined deviations of the x and y variables, known as orthogonal, total least squares or reduced major axis (RMA) regression (MacLeod, 2004a), but not so easy for nonlinear data.

Therefore, having fitted a nonlinear regression model to the data, the function parameters given for the curve were *tweaked* and fitted by eye to the centre of the data as best as possible.

With Val THAA D/L on the y-axis, all associations were fitted using a power function of the form $y=A\cdot(x^B) + C\cdot(x^D)$. Specific values for the function parameters (A, B, C and D) are given in Table 6.1. Charts for individual associations are given in Chpt 6: Appendix 2.

6.3.1.2 Determining confidence limits

Having established the fitted association trendlines, these could then be used to predict valine THAA D/L values off the observed x-axis amino acid D/L values. The difference between the observed Val D/L (y_i) and the predicted value (y_p) gives a measure of the vertical bias (y_i-y_p). If the valine THAA D/L data approximates to normality, observed y values would be expected to be evenly distributed either side of the prediction curve, at given values of x.

Table 6.1: Correlation functions for amino acids when y=Val THAA D/L

y=Val THAA D/L	function		function parameters			
x-axis	type	formula	A	B	C	D
Asx THAA D/L	1	$y=A\cdot X^B + C\cdot X^D$	1.00	5.00	0.110	0.900
Glx THAA D/L	1	$y=A\cdot X^B + C\cdot X^D$	0.950	1.17	0.000	1.00
Ser THAA D/L	1	$y=A\cdot X^B + C\cdot X^D$	0.460	4.00	0.060	0.500
Ala THAA D/L	1	$y=A\cdot X^B + C\cdot X^D$	2.00	20.0	0.600	1.10
Phe THAA D/L	1	$y=A\cdot X^B + C\cdot X^D$	0.367	0.818	1.045	6.87
Leu THAA D/L	1	$y=A\cdot X^B + C\cdot X^D$	0.440	0.900	0.400	7.00
A/I THAA	1	$y=A\cdot X^B + C\cdot X^D$	0.420	1.05	0.312	0.750
Asx FAA D/L	1	$y=A\cdot X^B + C\cdot X^D$	1.00	10.0	0.200	1.60
Glx FAA D/L	1	$y=A\cdot X^B + C\cdot X^D$	0.850	1.30	0.050	1.00
Ser FAA D/L	1	$y=A\cdot X^B + C\cdot X^D$	0.550	1.40	0.400	0.900
Ala FAA D/L	1	$y=A\cdot X^B + C\cdot X^D$	0.530	1.20	1.20	17.0
Val FAA D/L	1	$y=A\cdot X^B + C\cdot X^D$	0.550	1.40	0.400	0.900
Phe FAA D/L	1	$y=A\cdot X^B + C\cdot X^D$	0.050	0.300	0.300	2.00
Leu FAA D/L	1	$y=A\cdot X^B + C\cdot X^D$	0.350	1.35	0.015	0.200
A/I FAA	1	$y=A\cdot X^B + C\cdot X^D$	0.520	1.350	0.090	0.550

Function type; 1 = power

Summing both the positive and negative bias values would therefore effectively cancel each other out and the overall deviation would end up equal to zero. Therefore, for least-squares regression, it is the square of the individual deviations (the residual sum of squares, $RSS=\sum(y_i-f(x_i))^2$) which is minimised, thereby removing the cancelling effect.

In nearly all cases, the size of the deviation increases with D/L value. (Most often concentrations are plotted in chemistry and this is then referred to as the concentration effect or concentration dependence).

If the deviations are proportional to *y*, then it should be possible to normalise the data by applying a transformation. In regression analysis, these larger distances tend to dominate and contribute more to the least squares total (Graphpad Prism http://www.graphpad.com/guides/prism/6/curve-fitting/). Therefore, weighting helps to even out the contributions to the RSS. In regression analysis, weighting by $1/y^2$ can be used to correct for concentration effects, however, rather than using squared deviations in this analysis it is the absolute deviations which describe the distribution of data above and below the association line. Therefore residuals were normalised using predicted values of Val D/L (y_p), to give estimates of the relative bias ($(y_i-y_p)/y_p$).

Data were first plotted sequentially, in order of increasing x-axis D/L value, (in this instance x=Asx THAA D/L), this provided a more even distribution of values (Figure 6.10a), and then plotted against predicted y (y_p), (Figure 6.10b). With the exception of a small number of values with larger residuals (later identified as outliers), the majority of data were evenly distributed within a fairly tight range either side of the central line (equivalent to zero deviation, i.e.; observed data lay on the predicted line).

Initially, the standard deviation of all the normalised residuals was taken to get a measure of the dispersion of the data. This standard deviation was then expanded to give 2 and 3 standard deviation confidence limits (std dev x 2 and x3 respectively), (Figure 6.10c). However, as can be seen, the expanded confidence limits are far too wide, are influenced by the minority of extreme values and don't adequately describe the majority of data. Therefore data were screened. Usually, a 2 standard deviation confidence interval is sufficient to describe the majority of data (data would be expected to lie within this range 95% of the time). To avoid rejecting acceptable data, 3 standard deviation confidence intervals could be used to increase the probability of data being acceptable to over 99%. However, to be absolutely certain and ensure that only outlying data were excluded (as far as reasonably possible), 5 standard deviations were used to set exclusion criteria.

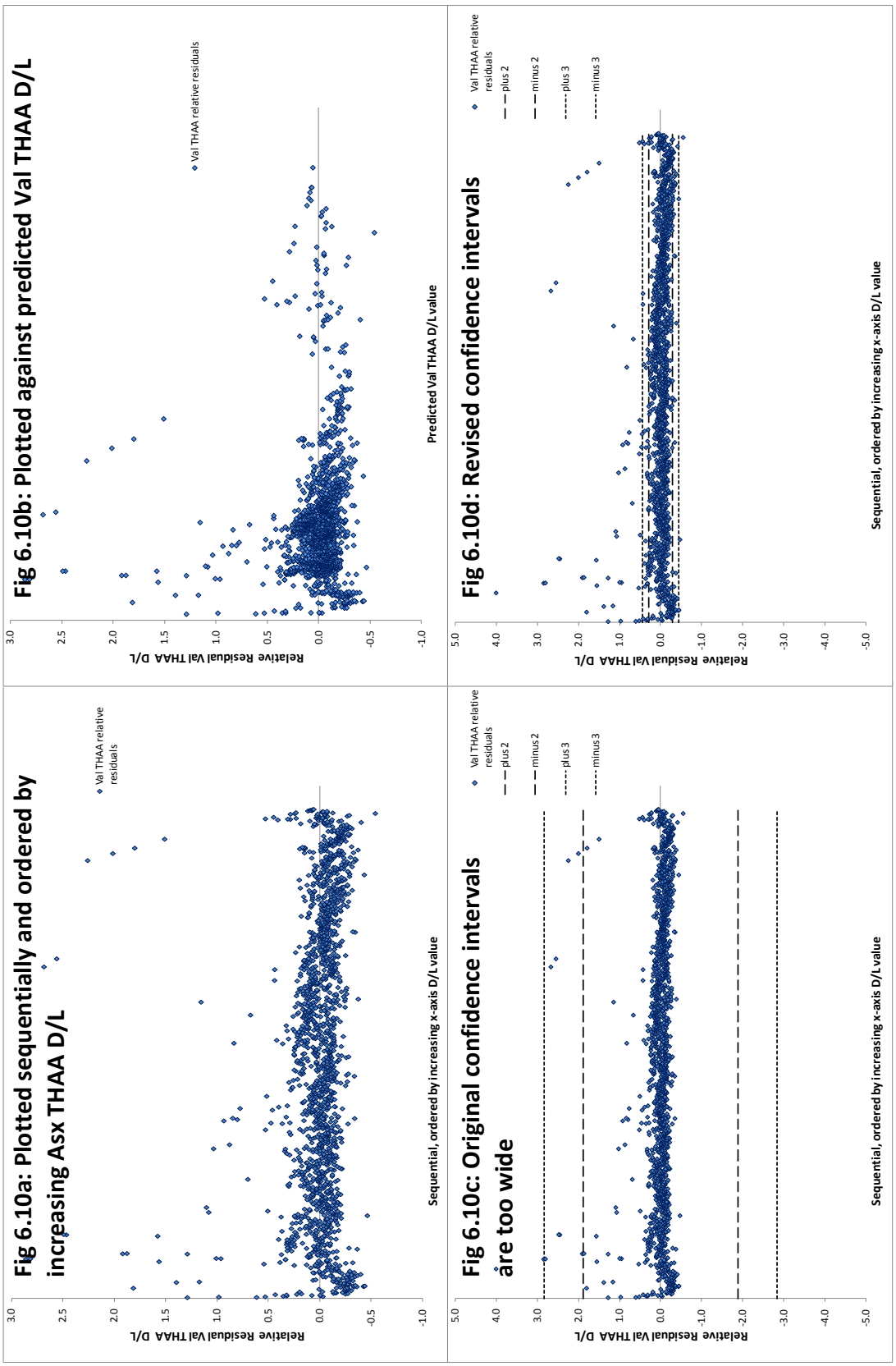
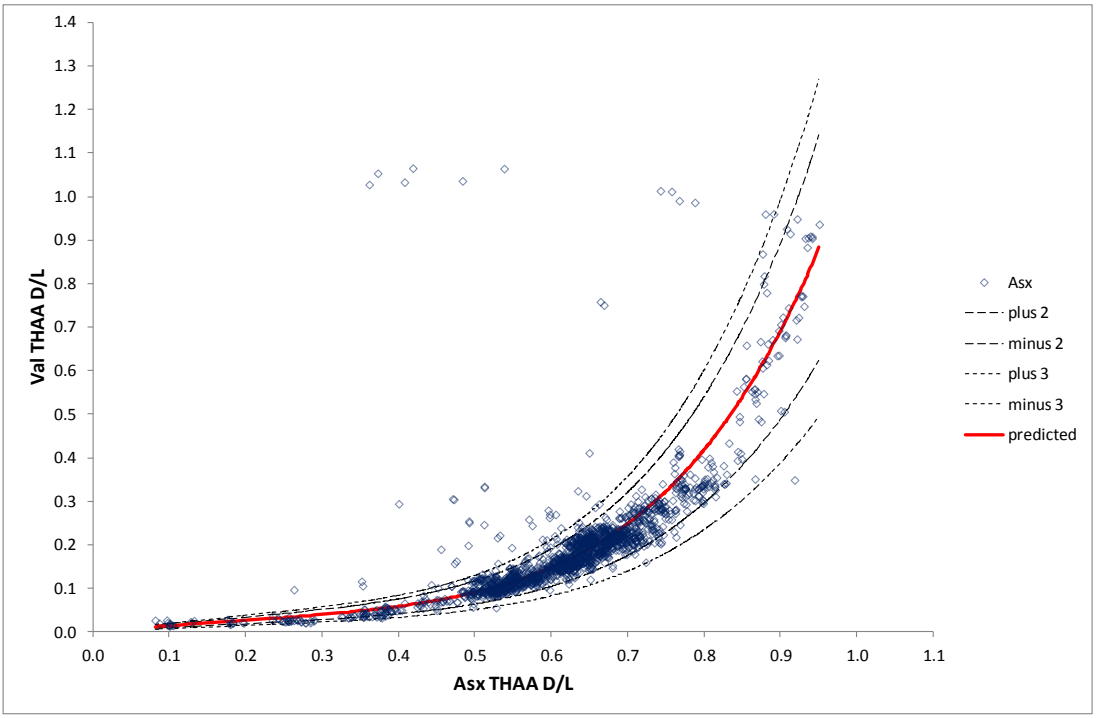


Figure 6.10: Distributions of normalised residuals for observed valine THAA D/L values

Thus any relative residual value, greater or less than 5 x std dev, were removed from the calculation. The average of the remaining deviations was taken and this was then used as the standard deviation to set the upper and lower confidence levels. Figure 6.10d shows how the revised confidence intervals fit snugly to the data and better describe the distribution for the majority of values.

Since the standard deviation here is actually a relative standard deviation (RSD) (since it was derived from the mean of relative residuals, and the difference between an observed data point and the true (predicted) value is equivalent to the deviation for a single value). Therefore, since the RSD=s/value (i.e.; D/L), then if the RSD is multiplied by the predicted y value, we can determine what the standard deviation should be for all points along the prediction curve. Expanding these standard deviations to 2 and 3 x the standard deviation now gives us the upper and lower confidence intervals surrounding the curve, Figure 6.11.

Figure 6.11: Confidence intervals for association between Val THAA D/L and Asx THAA D/L values



6.3.1.3 Identifying outliers

Having determined that the relative standard deviation is a constant, this can be used to calculate z-scores using the relative bias (residual) values determined previously.

$$z = \frac{x - \hat{X}}{s} \equiv \frac{x - \hat{X}}{1} \div \frac{s}{1} \; or \; \frac{x - \hat{X}}{1} \times \frac{1}{s} \qquad \textbf{(6.5a)}$$

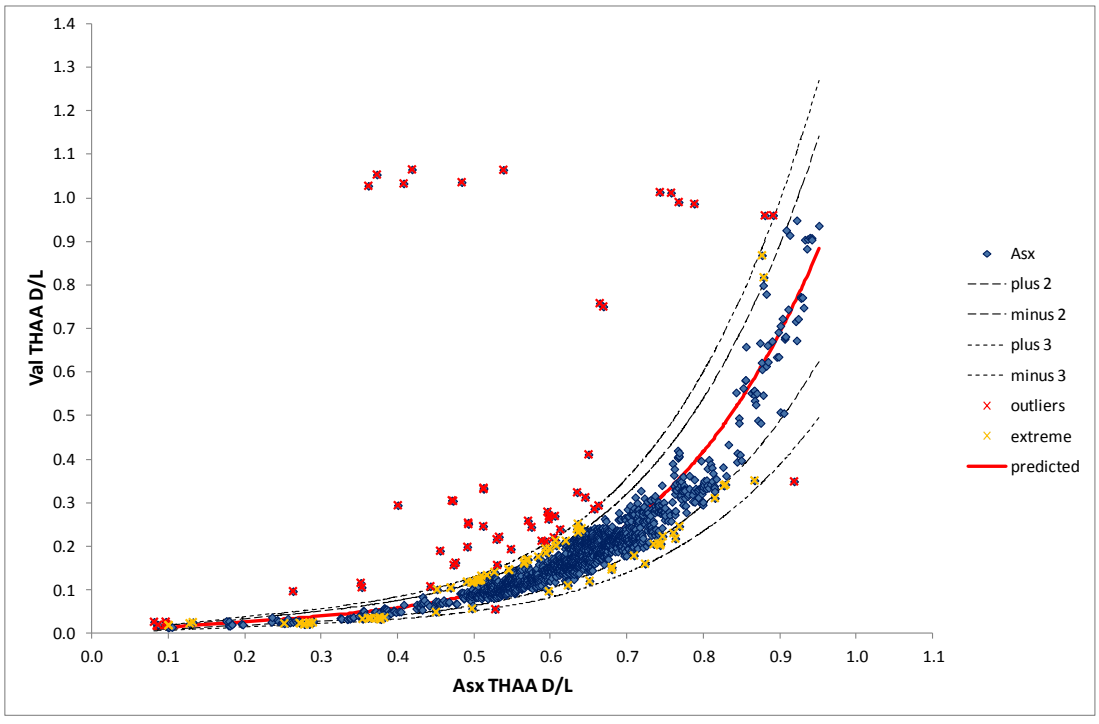
$$If, RSD = \frac{s}{X} \; and \; relative \; bias \; (residual) = \frac{x - \hat{X}}{\hat{X}}$$

Then,

$$\frac{x - \hat{X}}{\hat{X}} \div \frac{s}{\hat{X}} \equiv \frac{x - \hat{X}}{\cancel{\hat{X}}} \times \frac{\cancel{\hat{X}}}{s} \equiv \frac{x - \hat{X}}{s} = z \qquad \textbf{(6.5b)}$$

z-Scores can then be used to identify specific data points that fall between 2 and 3 standard deviations and therefore considered extreme values but are still probably acceptable and retained, and those that fall outside 3 standard deviations and are most likely to be beyond the normal distribution of the data. Figure 6.12 show z-scores greater than 2 but less than or equal to 3 standard deviations (2>|z|≤3) in yellow, whilst those considered outliers, such that z is greater than 3 (|z|>3) , are shown in red.

Figure 6.12: Confidence intervals for association between Val THAA D/L and Asx THAA D/L values showing extreme values and outliers



The data considered in this section relates to individual values plotted against each other, that is from every separate injection analysed by the instruments. Therefore, within this data are repeated measurements of replicate injections for each sample but also multiple samples for each location. When carrying out ANOVA as described in the previous section, individual extreme values will influence the precision estimates if not removed. To illustrate the effect that filtering the observed data using a z-score approach would have, individual expanded deviations (2 x std dev) have been determined. These have then been added to the predicted valine D/L value where data points lay above the prediction curve, and subtracted from the predicted valine value for those points falling below the prediction curve. This data is shown in Figure 6.13. Figure 6.14 then shows the same data but with those data points falling outside |z|=3, having been removed, and demonstrates the improvement made to the data set which would subsequently be reflected in the ANOVA precision estimates.

Figure 6.13: Expanded deviations of observed data from predicted Val D/L values

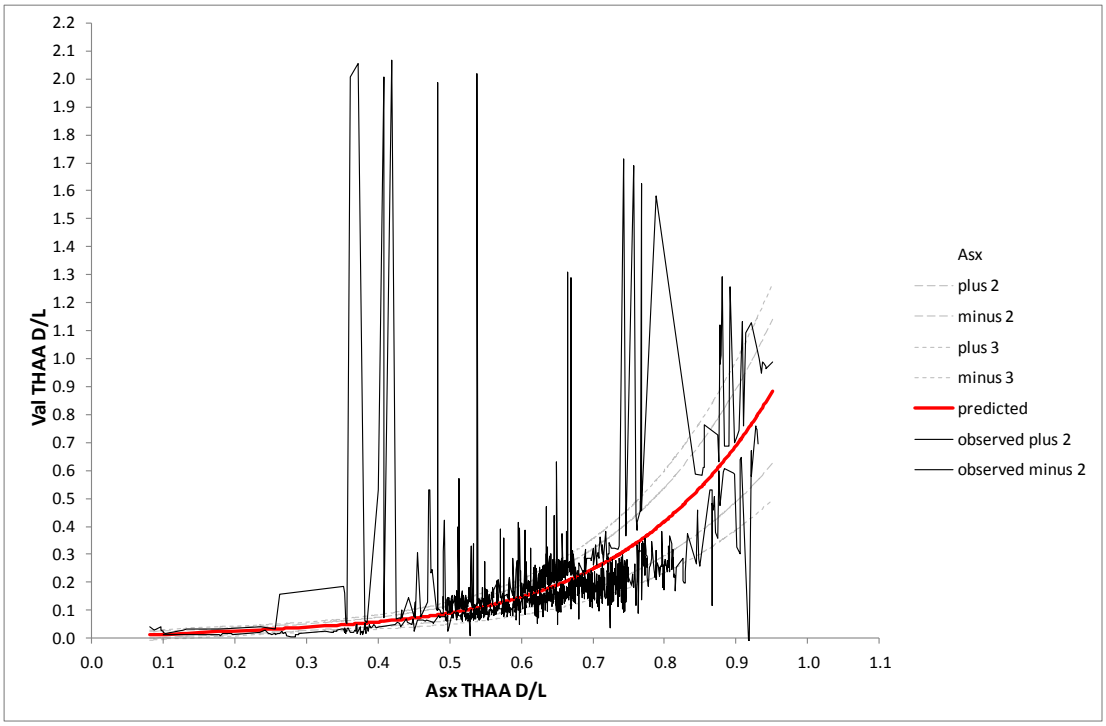
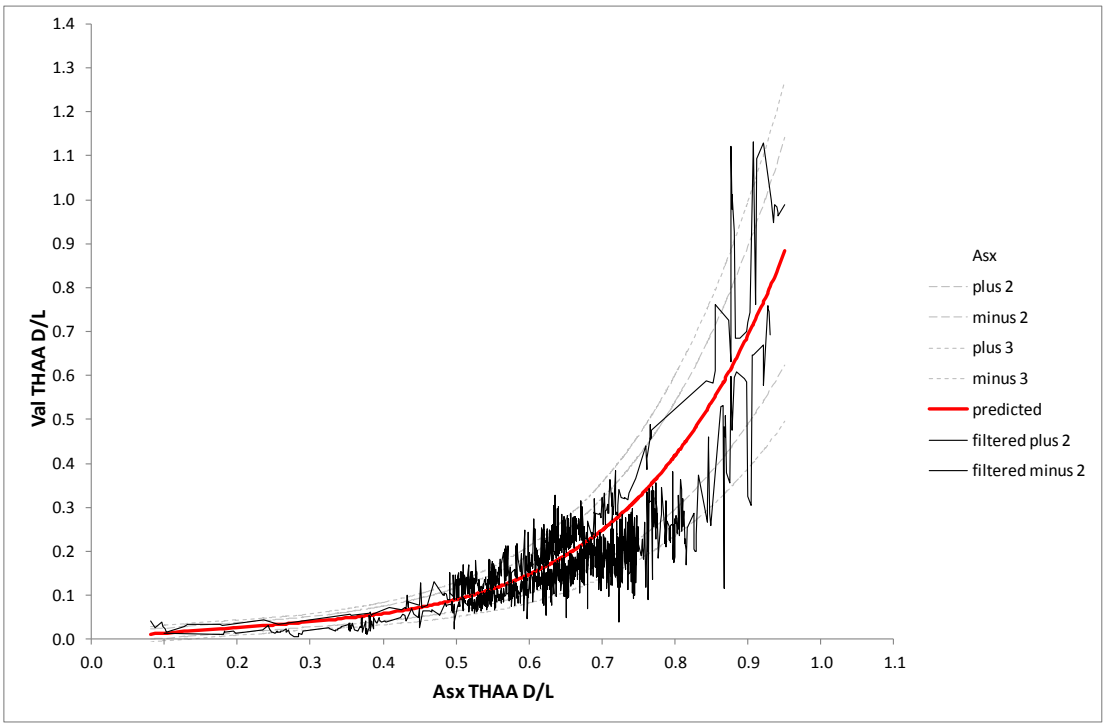


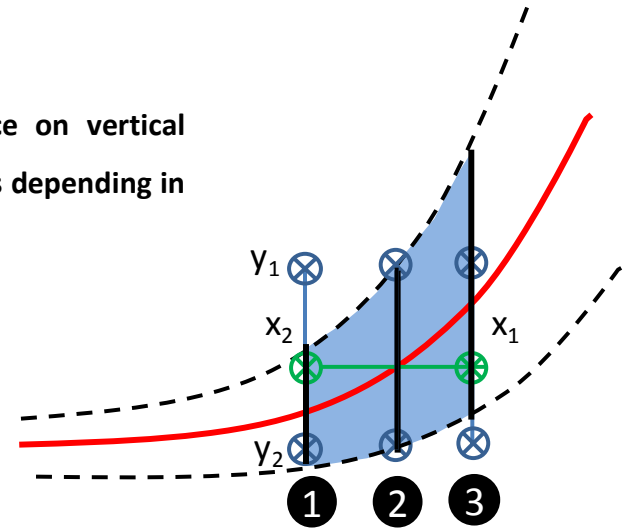
Figure 6.14: Expanded deviations of observed data from predicted Val D/L values after removal of outliers; |z|>3



6.3.1.4 Accommodating horizontal uncertainty

Having evaluated the variation in the vertical direction for the Val THAA D/L values, the variation in the horizontal direction also needs to be taken into consideration. This becomes particularly important as the curve becomes steeper as a small fluctuation to the left or the right of the observed amino acid D/L value could have a significant effect on the uncertainty of the valine in the vertical direction. Depending on the observed x-axis D/L values, this may increase or decrease the predicted uncertainty, depending on whether the true value of the x variable fell in position 1, 2 or 3 of Figure 6.15.

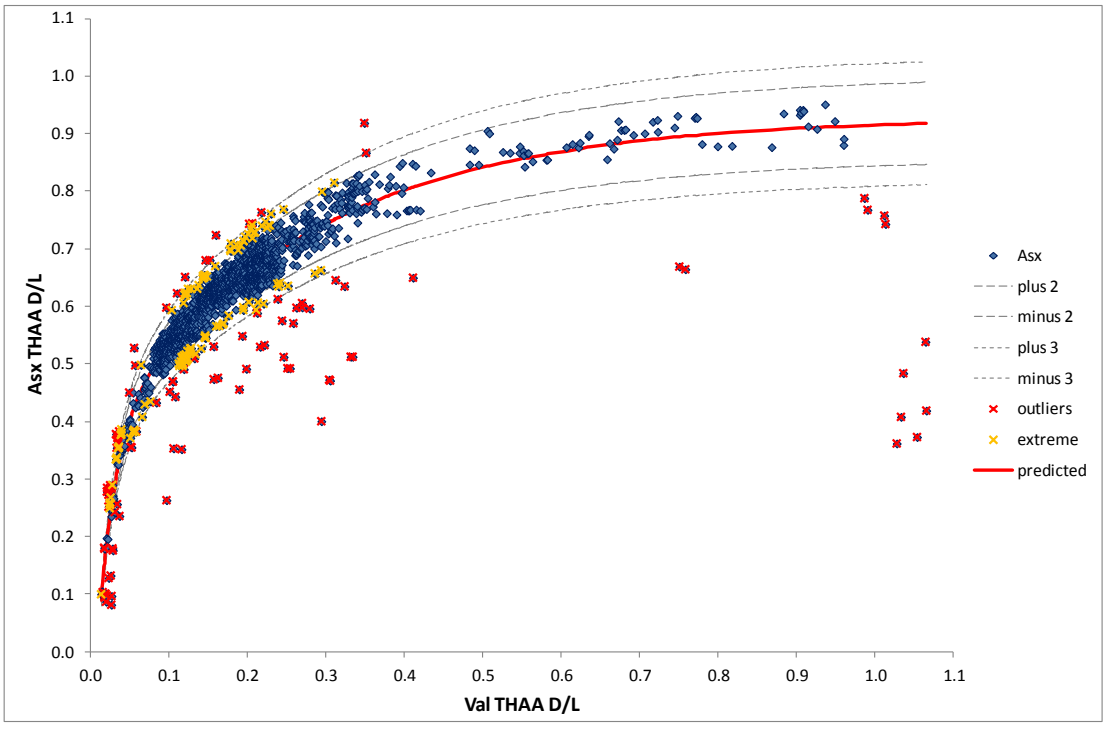
Figure 6.15: Influence on vertical uncertainty estimates depending in x-axis value



Whilst it would be possible to carry out a similar analysis using the vertical valine D/L values to predict x-axis values, using the equation for the line (Table 6.1) rearranged to solve for x. However, this isn't quite so straightforward for an equation with a double power function in it. Therefore a more practical solution was simply to swap over the variables on the two axes and repeat the whole process for the other amino acid, in this instance, measuring Asx THAA D/L on the vertical axis, off the valine THAA D/L on the horizontal axis. Figure 6.16 shows the distribution, association line, and 2 and 3 standard deviation confidence intervals for Asx THAA D/L against Val THAA D/L, together with data points identified as being at the extreme of the distribution (yellow) and those considered as outliers (red).

Table 6.2 gives the association functions for all combinations with Val THAA D/L this time as the x-axis variable. The function type used was derived from an initial template selected from a menu in Graphpad Prism, that appeared to best fit the data based on the initial non-linear regression. Function types such as the two-phase exponential association curve (type 2) and the dose response curve (type 3) are based on biological functions but have been adapted to best describe the associations between the amino acids. Further information on the individual curve functions in their biological context can be found on the Prism website. All individual associations can be found in Chpt 6: Appendix 2.

Figure 6.16: Confidence intervals for association between Asx THAA D/L and Val THAA D/L values showing extreme values and outliers



6.3.1.5 Further considerations and adjustments

For all associations, the distribution of data is generally well described by a fixed value for the relative standard deviation (RSD) and the standard deviation being proportional to the D/L value. Thus the standard deviation increases in a prescribed manner as the D/L value increases. However, this also means that the standard deviation also continues to reduce as the D/L approaches zero.

For many of the amino acids, this continual reduction at low levels appeared to be too tight for the data and is probably not realistic. After all, it would be unreasonable to expect the precision of analysis to continually reduce. At a certain point, the effect of the method and instrument sensitivity will limit the improvements in precision that can be achieved. This threshold will be related to the limit of quantification or possibly the limit of detection that would normally be determined during method validation.

Table 6.2: Correlation functions for amino acids when x=Val THAA D/L

x=Val FAA D/L	function		function parameters						
y-axis	type	formula	A	B	C	D	E	F	G
Asx THAA D/L	2	$y=(B+F\cdot(1-e^{-D\cdot X}))+(G\cdot(1-e^{-E\cdot X}))$	0.928	-0.201	49.3	47.6	3.77	0.557	0.573
Glx THAA D/L	1	$y=A\cdot x^{B}$	0.938	0.802					
Ser THAA D/L	2	$y=(B+F\cdot(1-e^{-D\cdot X}))+(G\cdot(1-e^{-E\cdot X}))$	0.900	-0.350	47.0	55.0	8.00	0.588	0.663
Ala THAA D/L	3	$y=A+(B-A)/(1+(10^{C-x})\cdot D)$	2.00	20.00	0.600	1.10			
Phe THAA D/L	3	$y=A+(B-A)/(1+(10^{C-x})\cdot D)$	-0.311	0.870	0.122	3.70			
Leu THAA D/L	3	$y=A+(B-A)/(1+(10^{C-x})\cdot D)$	-0.174	0.977	0.200	3.50			
A/I THAA	4	$y = A\cdot x$	1.37						
Asx FAA D/L	2	$y=(B+F\cdot(1-e^{-D\cdot X}))+(G\cdot(1-e^{-E\cdot X}))$	0.960	-0.200	65.0	30.0	3.50	0.754	0.406
Glx FAA D/L	3	$y=A+(B-A)/(1+(10^{C-x})\cdot D)$	-0.500	1.45	1.50	0.070			
Ser FAA D/L	2	$y=(B+F\cdot(1-e^{-D\cdot X}))+(G\cdot(1-e^{-E\cdot X}))$	2.30	0.020	105	0.550	0.000	2.39	-0.114
Ala FAA D/L	2	$y=(B+F\cdot(1-e^{-D\cdot X}))+(G\cdot(1-e^{-E\cdot X}))$	0.960	0.010	175	2.50	1.60	1.66	-0.713
Val FAA D/L	2	$y=(B+F\cdot(1-e^{-D\cdot X}))+(G\cdot(1-e^{-E\cdot X}))$	2.30	0.020	105	0.550	0.000	2.39	-0.114
Phe FAA D/L	2	$y=(B+F\cdot(1-e^{-D\cdot X}))+(G\cdot(1-e^{-E\cdot X}))$	0.900	-0.120	180	5.50	3.10	1.84	-0.816
Leu FAA D/L	2	$y=(B+F\cdot(1-e^{-D\cdot X}))+(G\cdot(1-e^{-E\cdot X}))$	1.15	-0.020	165	2.70	1.20	1.93	-0.761
A/I FAA	4	$y =A\cdot x + B$	1.42	0.125					

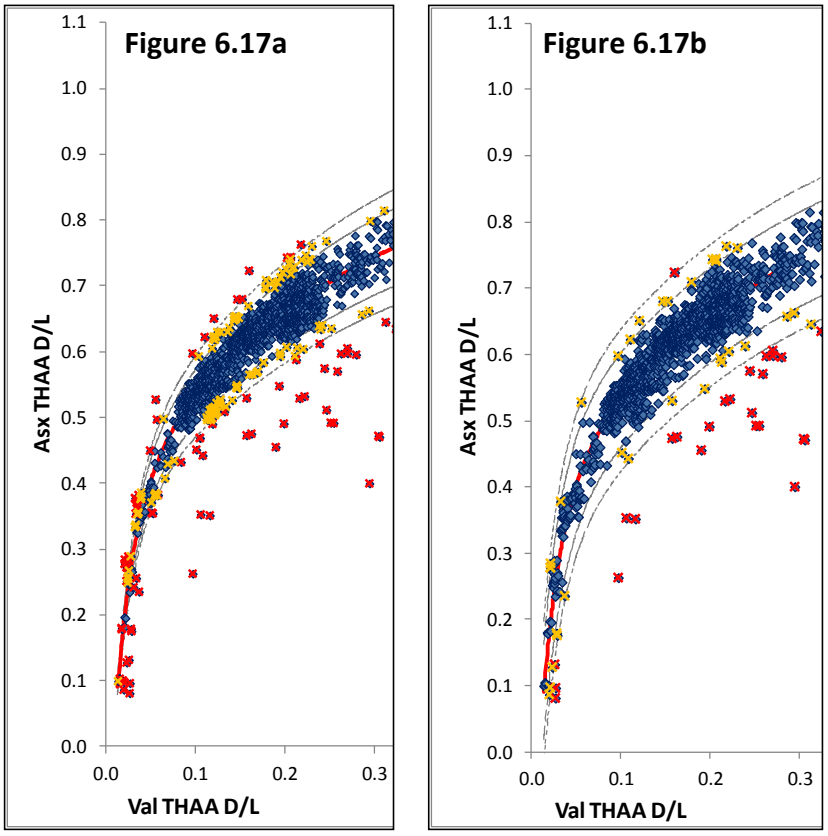
Fuction type; 1 = power; 2 = two phase exponential association curve; 3 = dose-response curve; 4 = linear

$F=(A-B)\cdot C\cdot 0.01$

$G=(A-B)\cdot(100-C)\cdot 0.01$

Using the distribution of data as a guide, a position along each x-axis was selected where it seemed reasonable to fix the vertical standard deviation, so no matter how small the D/L value became the standard deviation would not get any smaller. By effectively widening the confidence intervals at low values, this would also avoid rejecting data points that were probably acceptable but fell close to the limit of analytical capabilities. The effect of this can be seen in Figure 6.17b. which fixes the standard deviation, compared to the original data, Figure 6.17a, which doesn't.

Figure 6.17: Comparison between having a fixed (a) relative standard deviation and (b) having a fixed standard deviation



The effect of fixing the standard deviation means that the RSD now becomes proportional to 1/(D/L), and since RSD = s/(D/L). Therefore as D/L reduces, the RSD increases. This is shown graphically by the RSD confidence intervals in Figure 6.18 and Figure 6.19. A further check can be carried out on the data by plotting the observed deviations (*yi-yp*) against the predicted y-axis D/L values. Figure 6.20 and Figure 6.21 show the dispersion of residuals either side of the predicted lines. Figure 6.20 shows how the standard deviation for the majority of the data is proportional to the D/L value, except for the very lowest D/L values when the standard deviation becomes a fixed value; for Val THAA (y-axis) this was 0.04 D/L. For the swapped variables (Figure 6.21), there was no clear proportionality

between deviation and D/L value. The standard deviation applied to describe the association's confidence intervals was therefore fixed the whole way across the data set. For Asx THAA (y-axis) this was from 0.92 D/L.

It was also noticed that for some associations there were at times unbalanced distributions of data either side of the prediction curve due to other diagenetic effects occurring in the protein. Serine is an example of this since it is relatively unstable and known to decompose to alanine (Bada *et al.*, 1978). Where these effects were observed it was necessary to determine single sided RSDs that could then be applied to both sides and help identify extreme and outlying values.

For all associations, the fixed values for RSD, expressed as percentages, coordinates for the threshold levels at which the standard deviations were fixed and minimum permitted standard deviations, are given in Table 6.3.

An alternative approach to the evaluation of associated data was considered using the Matlab/Octave Gaussian Process (GP) software, (see http://www.gaussianprocess.org/gpml/code/matlab/doc/index.html (Rasmussen and Williams, 2006)). Gaussian processes are a powerful technique based on Bayesian modelling that utilise covariance and likelihood functions to make inferences from limited data sets (Garo Panikian, pers. comms.; Rasmussen and Williams, 2006).

Figure 6.18: RSD Confidence Intervals; Std dev fixed for Asx THAA (x) at 0.3 D/L, Val THAA (y) at 0.04 D/L

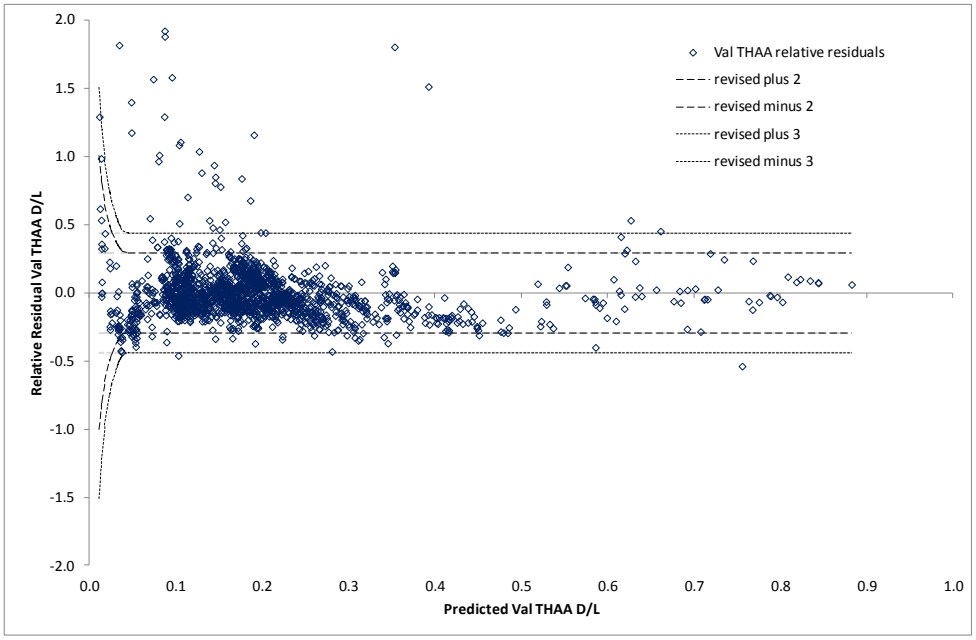


Figure 6.19: RSD Confidence Intervals; Std dev fixed for Val THAA (x) at 1.1 D/L, Asx THAA (y) at 0.92 D/L

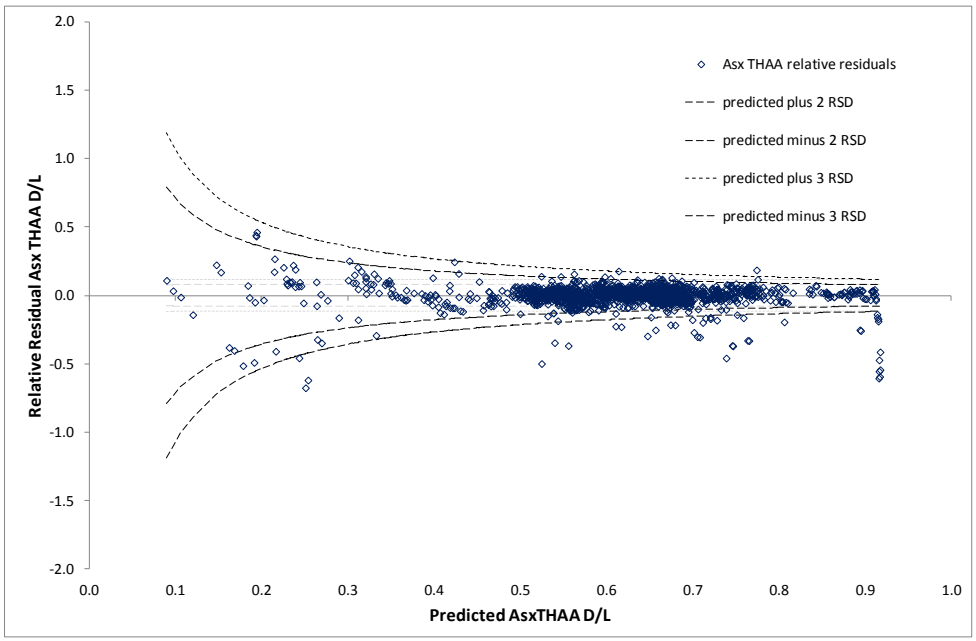


Figure 6.20: Standard Deviation Confidence Intervals; Std dev fixed for Asx THAA (x) at 0.3 D/L, Val THAA (y) at 0.04 D/L

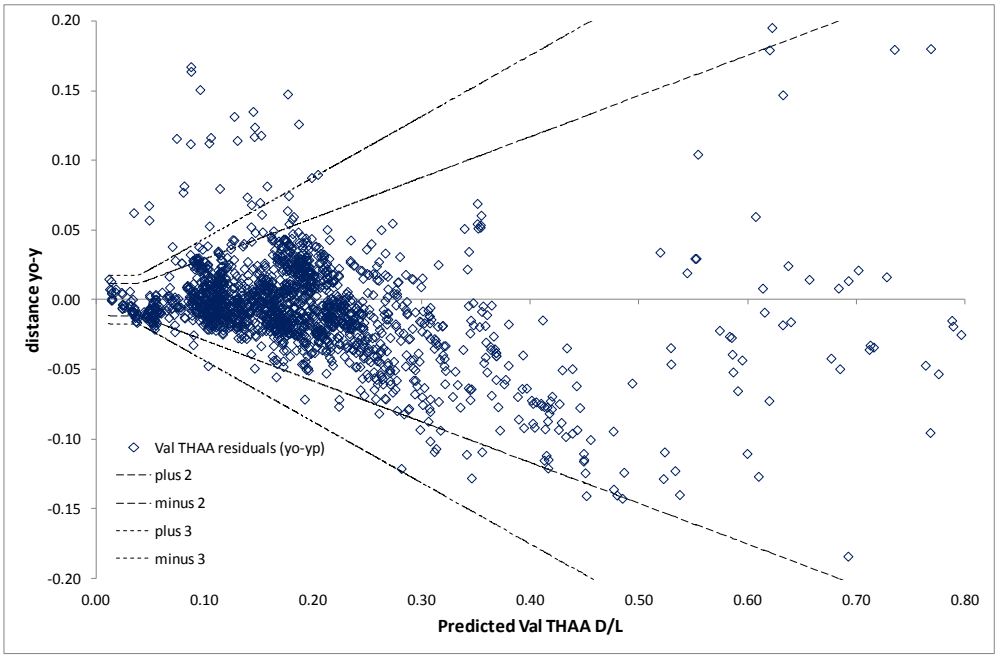
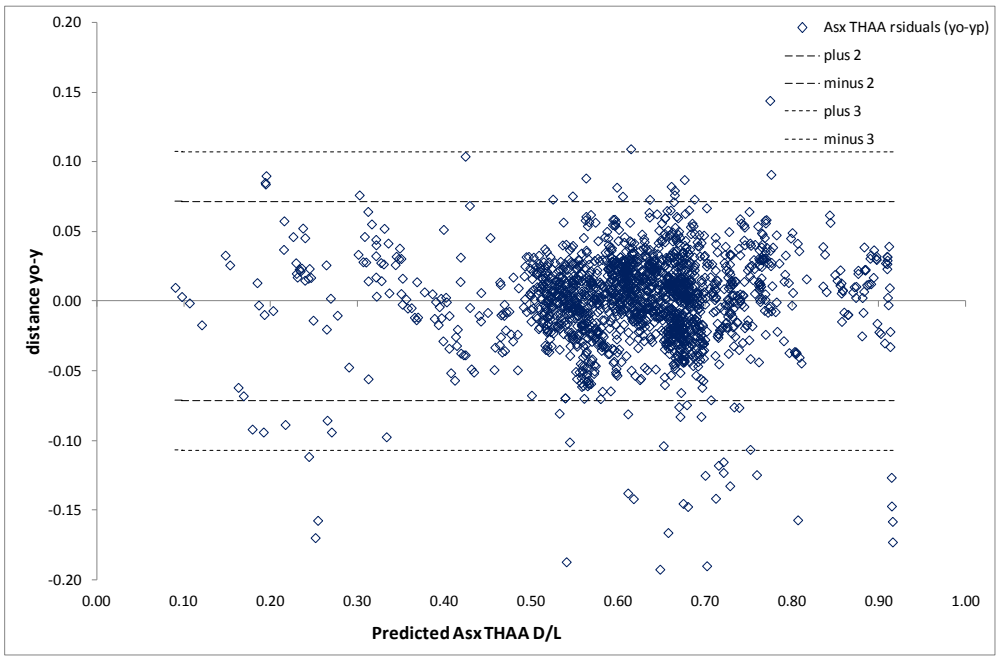


Figure 6.21: Standard Deviation Confidence Intervals; Std dev fixed for Val THAA (x) at 1.1 D/L, Asx THAA (y) at 0.92 D/L



An initial evaluation was carried out using a random subsample of 500 data points (from the original 1811), from across the same Val THAA, Asx THAA data set used in previous examples. Figure 6.22 shows the output of the Gaussian Process (GP) applied to a sample of 500 data points out of the possible 1811, and the inferred association and 2 standard deviation confidence interval for the joint density distribution.

If this is then superimposed over the whole data set, Figure 6.23, the GP credibility intervals (2 and 3 standard deviations), approximate to those previously modelled (see Figure 6.16). Although care would need to be taken with regard to appropriate sampling from the whole data set, this approach utilises complex statistical techniques and is probably far less labour intensive than the previous modelling description.

However, whilst the output closely follows the distribution of data, it was considered that for modelling purposes, it was too sensitive to the density of sampled data. Since even the entire data set does not reflect the whole population of possible values and is based only on the accumulated results acquired to date. Clustering of data is therefore purely coincidental.

Table 6.3: Summary of deviations for observed data from predicted values.

y=Val THAA D/L x-axis	constant RSD%	Fix point D/L Coordinates (x,y)	minimum SD
Asx THAA D/L	14.61	Asx = 0.30; Val = 0.04	0.006
Glx THAA D/L	10.02	Glx = 0.12; Val = 0.08	0.008
Ser THAA D/L	15.17	Ser = 0.43; Val = 0.06	0.008
Ala THAA D/L	8.16	Ala = 0.16; Val = 0.08	0.007
Phe THAA D/L	10.76	Phe = 0.19; Val = 0.09	0.010
Leu THAA D/L	15.56	Leu = 0.15; Val = 0.08	0.012
A/I THAA	21.22	A/I = 0.09; Val = 0.08	0.018
Asx FAA D/L	12.84	Asx = 0.50; Val = 0.07	0.009
Glx FAA D/L	19.89	Glx = 0.20; Val = 0.11	0.023
Ser FAA D/L	20.18	Ser = 0.40; Val = 0.03	0.006
Ala FAA D/L	10.34	Ala = 0.25; Val = 0.10	0.010
Val FAA D/L	12.94	Val F = 0.10; Val H = 0.07	0.009
Phe FAA D/L	10.99	Phe = 0.32; Val = 0.07	0.007
Leu FAA D/L	20.93	Leu = 0.25; Val = 0.07	0.014
A/I FAA	22.45	A/I = 0.20; Val = 0.10	0.022

x=Val THAA D/L y-axis	constant RSD%	Fix point D/L Coordinates (x,y)	minimum SD
Asx THAA D/L	3.88	Val = 1.1; Asx = 0.92	0.036
Glx THAA D/L	8.14	Val = 0.08; Glx = 0.12	0.010
Ser THAA D/L	6.38	Val = 0.90; Ser = 1.00	0.058
Ala THAA D/L	6.74	Val = 0.08; Ala = 0.15	0.155
Phe THAA D/L	8.68	Val = 0.08; Phe = 0.18	0.015
Leu THAA D/L	12.75	Val = 0.09; Leu = 0.16	0.021
A/I THAA	24.90	Val = 0; A/I = 0	-
Asx FAA D/L	3.24	Val = 1.1; Asx = 0.95	0.031
Glx FAA D/L	14.54	Val = 0.08; Glx = 0.19	0.027
Ser FAA D/L	5.01	Val = 1.0; Ser = 1.03	0.052
Ala FAA D/L	8.25	Val = 0.08; Ala = 0.23	0.019
Val FAA D/L	8.04	Val H = 0.08; Val F = 0.12	0.010
Phe FAA D/L	12.23	Val = 0.08; Phe = 0.35	0.043
Leu FAA D/L	11.31	Val = 0.08; Leu = 0.29	0.032
A/I FAA	25.93	Val = 0.08; A/I = 0.24	0.062

So, although the credibility interval reflects the distribution of data at any given point, based on the sample size available, it is unlikely to be true for the whole population. It would therefore seem unfair to use this model to predict uncertainty and exaggerate estimates for sites simply due to the scarcity of data available. For this reason, this approach was not taken any further.

Figure 6.22: Output of Gaussian Process applied to sampled data showing 2 std dev confidence interval (Garo Panikian, pers. comms.).

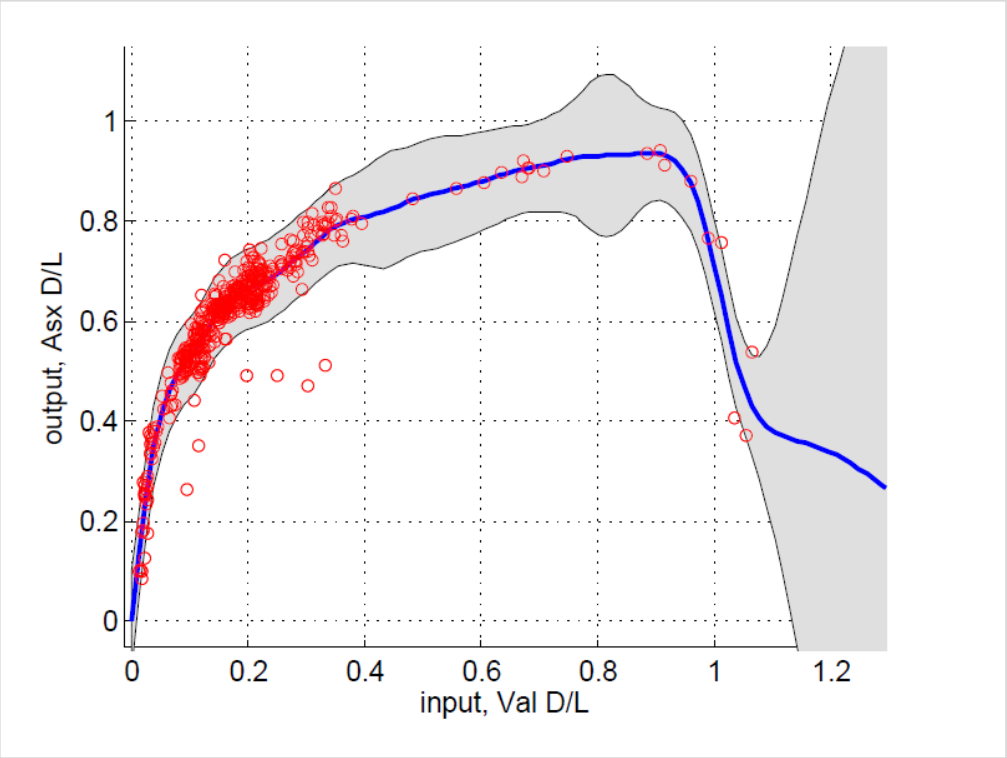
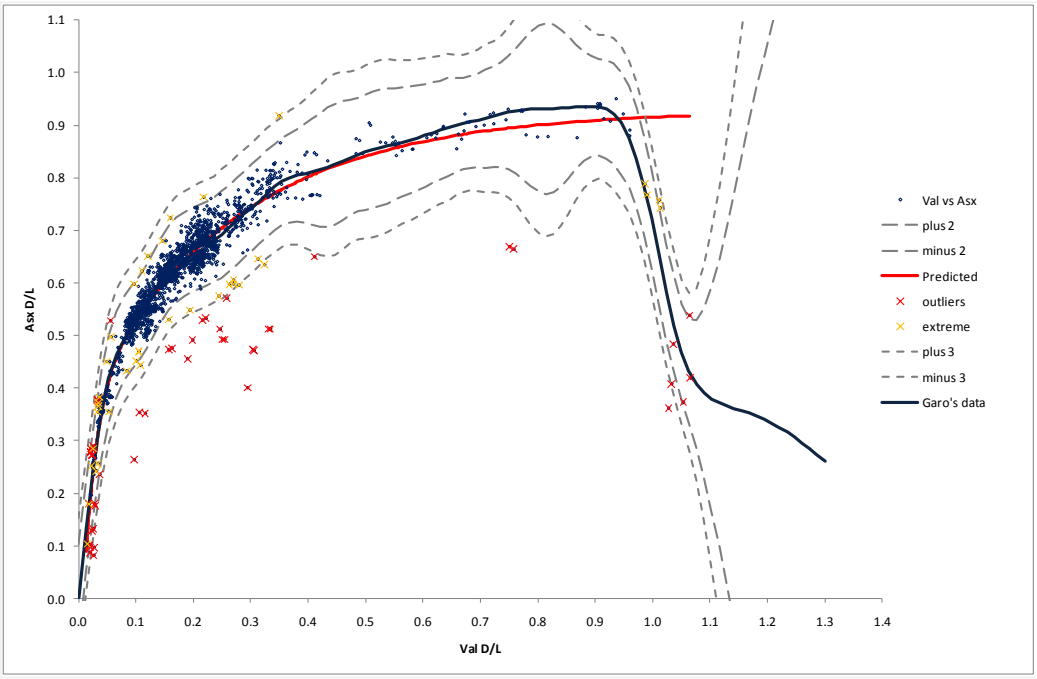


Figure 6.23: Gaussian process confidence intervals (2 and 3 std dev) superimposed over whole data set



6.3.2 Correlation, covariance and combined uncertainty.

6.3.2.1 Correlation v dependence

The frequent use of the r^2 function in Excel can be confusing. r^2 is the coefficient of dispersion and measures the goodness of fit between data and the fitted regression, it reflects the degree of dispersion of data round the trendline. r^2 values vary between 0-1; if r^2 = 0 there is no relationship between the x and y variables, if r^2=1, all data points lay on the regression line. r^2 values can be derived for both linear and curvilinear data. $\sqrt{r^2}$ is not necessarily the same as r, the Pearson product-moment correlation coefficient. r values range between -1 to 1 (MacLeod, 2004a; Miller and Miller, 2005), when r=-1 the data is said to be perfectly negatively correlated, or if r=1, positively correlated. Typically, correlation refers to the measure of linear dependence and the value of the coefficient, r, is derived from;

$$r = \frac{\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]}{\sqrt{\left[\sum_{i=1}^{n}(x_i - \bar{x})^2\right]\left[\sum_{i=1}^{n}(y_i - \bar{y})^2\right]}} \qquad \textbf{(8.6)}$$

Where $\bar{x}$ and $\bar{y}$ are the means of the x and y variables.

If r=0, the GUM says the variables are independent "...*a change in one does not imply an expected change in the other*." (EURACHEM / CITAC, 2000, p21). However Kirkup and Frenkel state that it is perfectly possible for variables to be uncorrelated but dependent. "....*independence implies zero correlation but zero correlation does not imply independence*." (2006, p78).

In the context of the current data, clearly there are strong correlations between the different amino acids, in this instance, with valine THAA D/L values. Variables are then said to have a high degree of dependence, to the extent that y can be predicted from x, however their linear correlation (with perhaps the exception of A/I data), is minimal.

6.3.2.2 Covariance

The correlation coefficient, r, is related to the covariance. Covariance is the proportion of variance shared by two variables, it describes the spread of values around their joint mean (MacLeod, 2004b).

From simple descriptive statistics we know that the standard deviation squared, gives an expression of the variance, and the square root of the variance gives us the standard deviation, thus s^2 = var and √var = s. Thus;

$$s(x) = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \text{ and } var(x) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})}{n-1}$$

Covariance is then given as;

$$covariance(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \quad \textbf{(6.7)}$$

Therefore r can also be expressed as;

$$r = \frac{covariance(x,y)}{\sqrt{var(x) \times var(y)}} \; or \; \frac{covariance(x,y)}{s(x) \times s(y)} \quad \textbf{(6.8)}$$

The above equations are often applied to data whose values lie within a discrete range. For example, repeated measurements of the same or similar item(s), i.e. sample data, which are all giving estimates of the true value plus random error. The mean of these estimates is then taken as the most representative value for the sample.

However, the associations discussed so far in this chapter evaluate data over a wide and continuous D/L range, i.e. multiple measurements for multiple samples each having different mean D/L values due to differences in age (or temperature history). Therefore for amino acids, the line of association fitted to the data, (which provides our predicted estimate of y), is equivalent to the x and y variable mean. Figure 6.24 shows the relationship between the x and y variables and the shared mean. Because of the difficulty in determining the deviations for the x variable, previously discussed, the axes have been swapped, so alternative vertical deviations can be determined for the second variable.

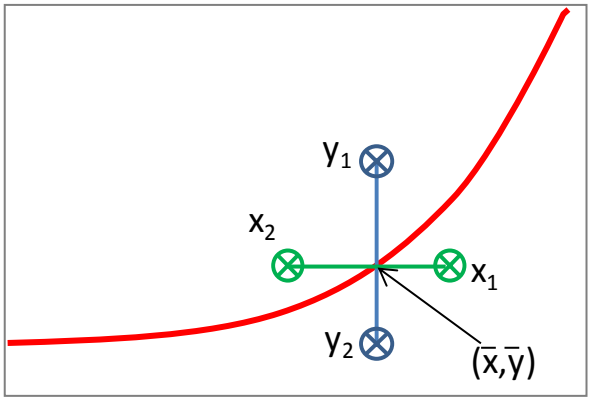


Figure 6.24: Covariant space for associated variables x and y

For a given location, the equation for the covariance becomes;

$$cov(y_{AsxH}, y_{ValH}) = \frac{\sum_{i=1}^{n}\left(y_{i(AsxH)} - y_{p(AsxH|ValH)}\right)\left(y_{i(ValH)} - y_{p(ValH|AsxH)}\right)}{n-1}$$

Where; $y_{i(AsxH)}$ = the observed Asx THAA D/L,

$y_{i(ValH)}$ = the observed Val THAA D/L,

$y_{p(AsxH|ValH)}$ = the predicted Asx THAA D/L given the Val THAA D/L,

$y_{p(ValH|AsxH)}$ = the predicted Val THAA D/L given the Asx THAA D/L.

6.3.2.3 Combined uncertainty

However, the purpose of evaluating the associations is to be able to use the predicted values of valine THAA D/L and its associated uncertainty to *inform* and *update* the mean and uncertainty estimate originally derived from the ANOVA performed on the raw site data.

Therefore, rather than wanting the covariance (that portion of variance shared between variables), what is required is the total variance (uncertainty) associated with the predicted valine D/L value (i.e. the uncertainty of Val THAA with another amino acid plus the uncertainty of that amino acid with Val THAA). Combining the two vertical deviations will therefore account for both uncertainty influences acting on the valine THAA from the association. An overview of the process used for determining the combined uncertainty is presented in Figure 6.25. In summary, for each location;

1. Associations between Val THAA and other THAA or FAA amino acid D/L values (aa), both in the x and y directions, are determined.
2. Using the average vertical relative deviations (y_i-y_p/y_p), determine uncertainty estimates (RSD and s) and confidence limits for y_p.
3. Compare the observed vertical deviation to the predicted vertical deviation for the same value of x, now allows us to identify extreme values and outliers using a z-score approach.
4. Using the criteria |z|>3, individual values are screened and potentially aberrant data removed.
5. Using screened data, improved mean D/L values can be determined for valine THAA and the associated amino acid, for each site.
6. Using the improved associated amino acid D/L mean, a single predicted uncertainty estimate can be determined for the Val THAA D/L using the curve function (valine THAA on the y-axis) from Table 6.1.

7. Using the improved valine THAA D/L mean, a single predicted uncertainty estimate can be determined for the associated amino acid using the curve function (valine THAA on the x-axis) from Table 6.2.
8. Both revised predicted uncertainty estimates can be combined in the usual way for the propagation of standard uncertainties.

For each location, predicted relative standard deviations for both amino acids can be combined to give an overall estimate of uncertainty. It is important to use relative values at this stage as different amino acids will give different D/Ls for the same location due to differences in racemisation rates, and individual deviations will not be comparable. Therefore we remove the "*concentration* (D/L value) effect" and work in relative values.

The two predicted RSDs can then be combined to give the combined standard uncertainty for valine THAA D/L, for a given site, based on the association with which ever amino acid it happens to have been associated with. Data have been combined using the conventional method for combining standard deviations described in the GUM (JCGM 100, 2008) and using values as originally recorded in the data archive, i.e. uncorrected, without any transformation or correction discussed previously in sections 4.5.3 and 4.5.4.

The risk of double counting uncertainty contributions when swapping axes was considered and alternative methods of combining values were evaluated. Using data given in Table 6.3 with Val THAA plotted against Asx THAA as an example,

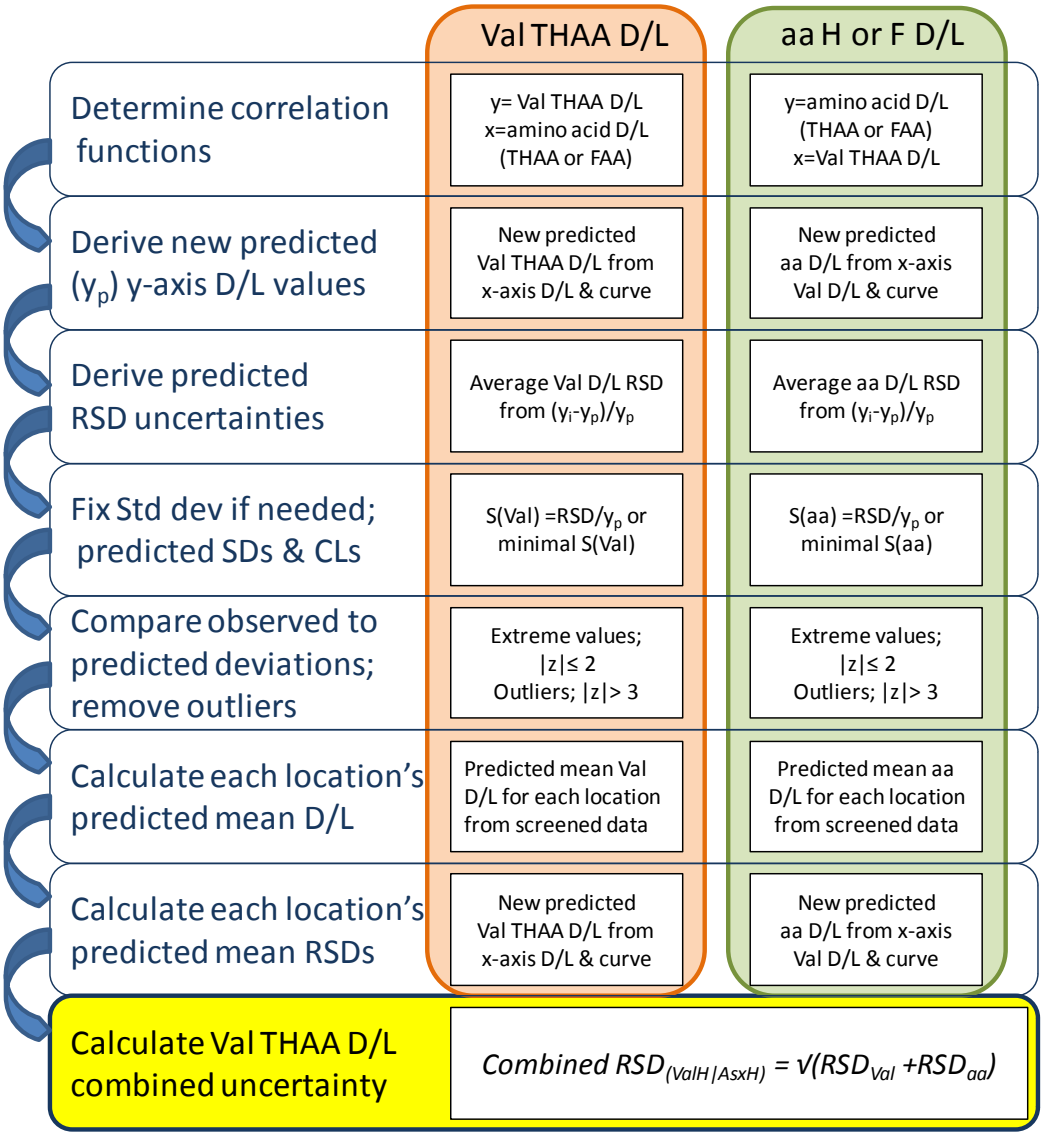
if a = u(Val) = 14.61 % (when x axis = Val THAA and y axis = Asx THAA), and

b = u(Asx) = 3.88 % (when x axis = Asx THAA and y axis = Val THAA),

i. $a+b = 14.61+3.88 = 18.49$ %
ii. $(a+b)/2 = (14.61+3.88)/2 = 9.25$ %
iii. $\sqrt{a^2 + b^2} = \sqrt{14.61^2 + 3.88^2} = 15.12$ %
iv. $\sqrt{(a^2 + b^2)/2} = \sqrt{(14.61^2 + 3.88^2)/2} = 10.69$ %

If u(Val) (14.61%) was simply added to u(Asx) (3.88%) then the combined value may well double count uncertainty contributions due to shared covariance resulting in (i) 18.48%. Averaged combined uncertainties (ii & iv) both result in final values less than that for valine on its own, so unlikely to be true (9.25 & 10.69 < 14.61). The most appropriate determination of the combined effect is (iii), the conventional approach, where the final result is slightly larger than that for valine on its own but less than the sum of the two components.

Figure 6.25: Schematic for determining combined uncertainty for associated amino acids (aa) with valine THAA D/L values



Thus;

$$u_{c(ValH|AsxH)} = D/L_{ValH} \times \sqrt{\left(\frac{u(ValH|AsxH)}{y_{p(ValH|AsxH)}}\right)^2 + \left(\frac{u(AsxH|ValH)}{y_{p(AsxH|ValH)}}\right)^2} \quad \textbf{(6.9a)}$$

$$\frac{u_{c(ValH|AsxH)}}{D/L_{ValH}} = combined\ RSD_{(ValH|AsxH)}$$

$$= \sqrt{RSD_{(ValH|AsxH)}{}^2 + RSD_{(AsxH|ValH)}{}^2} \quad \textbf{(6.9b)}$$

This combined valine uncertainty can be used to determine revised confidence intervals, which now allow for horizontal movement using the uncertainty of the x-axis amino acid. In effect this gives you the uncertainty on the uncertainty. Figure 6.26 shows the influence that the horizontal uncertainty has on the valine D/L confidence intervals. Because uncertainty is expressed both as the plus and minus standard deviation, both upper and lower confidence levels can be drawn around each of the original valine CLs. Using the revised outer CLs for each, extreme values and outliers can be reassessed (although the use of judgement for individual values should always be reserved for those closest to the boundaries). Figure 6.27 - Figure 6.40 Show revised ±3 standard deviation confidence intervals (solid lines) together with the original 2 and 3 standard deviation CIs (dotted lines) for valine THAA D/L for all THAA and FAA associations, together with revised extreme and outlier values. Table 6.4 provides the combined uncertainty estimates.

Table 6.4: Combined uncertainty estimates for Valine THAA D/L associated with different amino acids.

Val THAA D/L correlated with	constant RSD%	minimum SD	Val THAA D/L correlated with	constant RSD%	minimum SD
Asx THAA D/L	15.71	0.006	**Asx FAA D/L**	13.54	0.009
Glx THAA D/L	12.91	0.010	**Glx FAA D/L**	24.64	0.028
Ser THAA D/L	17.24	0.010	**Ser FAA D/L**	20.84	0.006
Ala THAA D/L	10.58	0.009	**Ala FAA D/L**	13.23	0.013
Val THAA D/L	-	-	**Val FAA D/L**	15.24	0.011
Phe THAA D/L	13.83	0.013	**Phe FAA D/L**	16.44	0.011
Leu THAA D/L	20.12	0.016	**Leu FAA D/L**	23.79	0.016
A/I THAA	32.71	0.028	**A/I FAA**	34.30	0.033

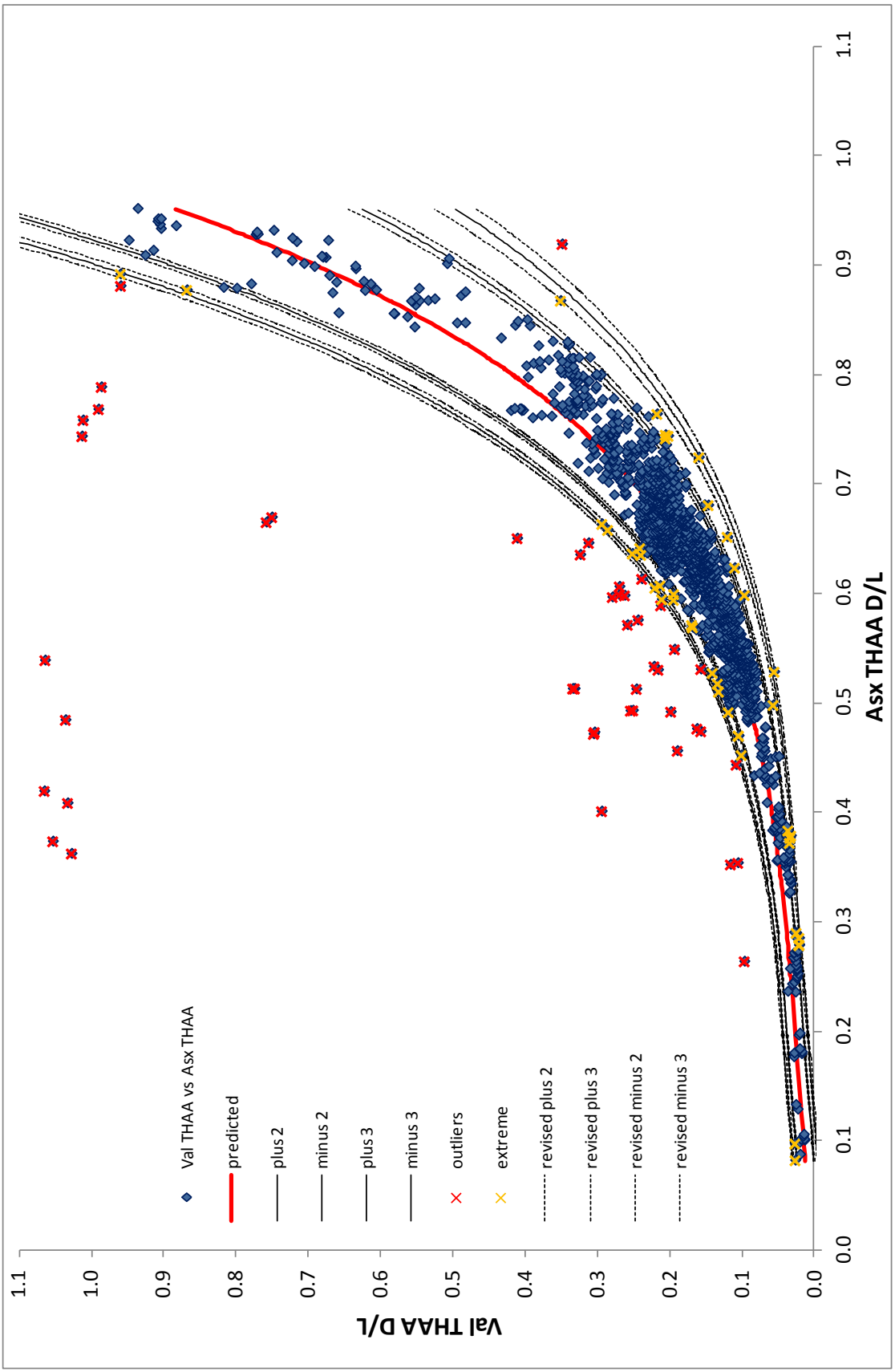


Figure 6.26: Influence of Asx THAA D/L uncertainty on Val THAA D/L 2 & 3 std.dev. Confidence Intervals

Figure 6.27: Revised ±3 std dev confidence limits for Val THAA associated with Asx FAA D/L

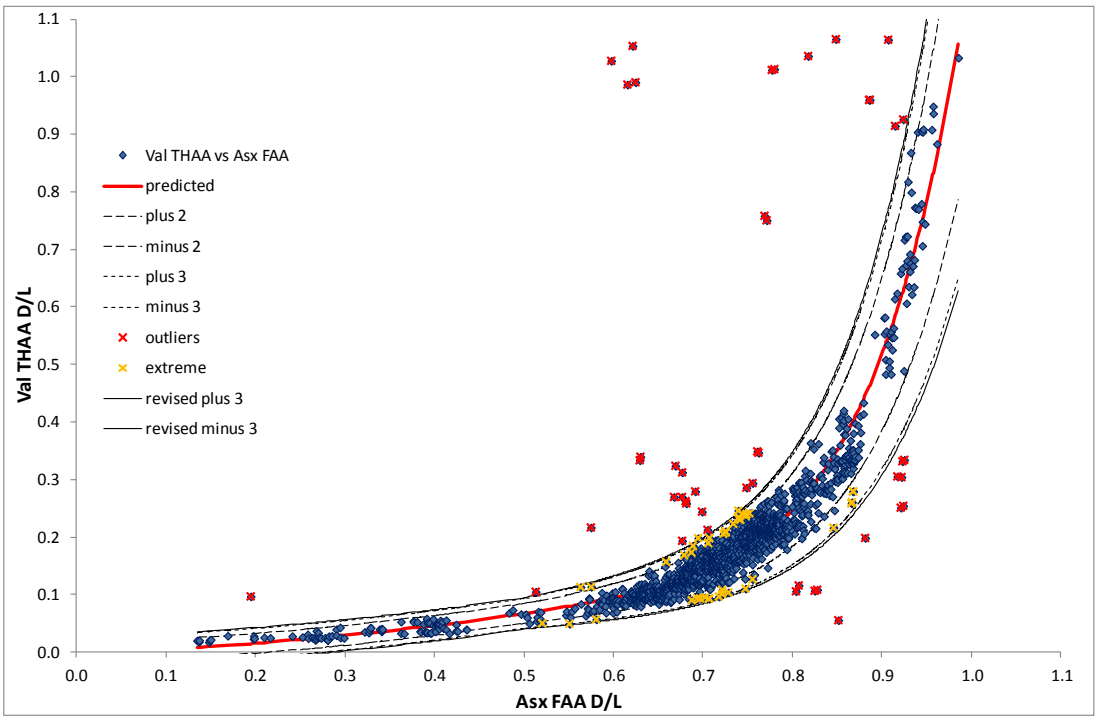


Figure 6.28: Revised ±3 std dev confidence limits for Val THAA associated with Glx THAA D/L

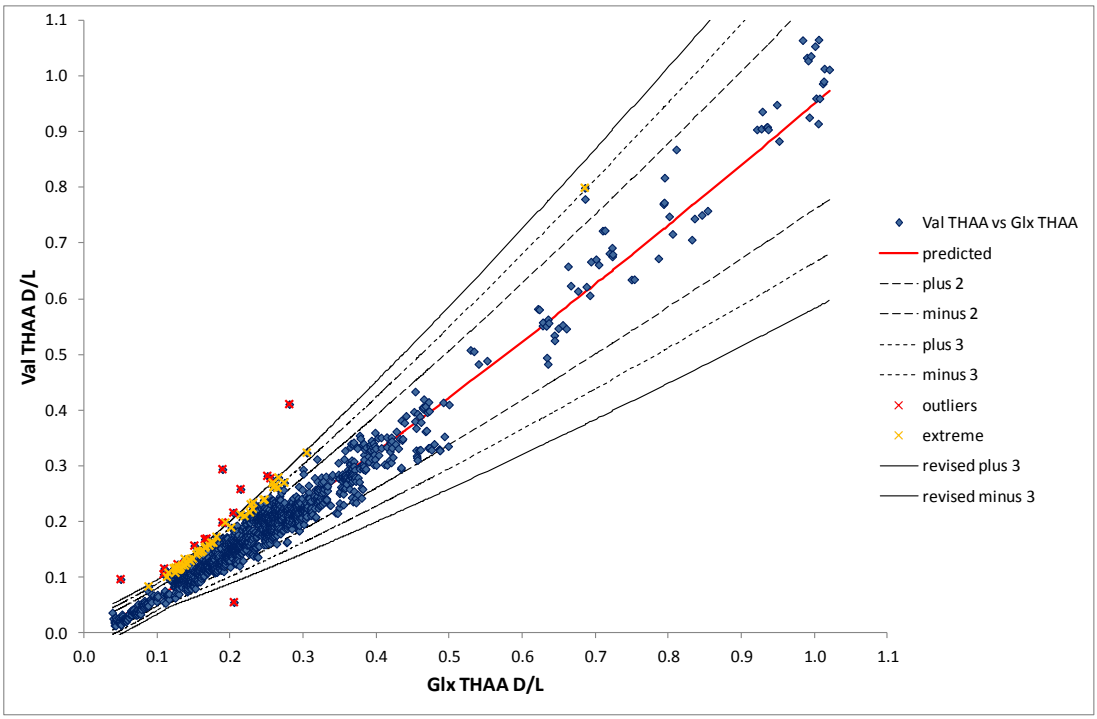


Figure 6.29: Revised ±3 std dev confidence limits for Val THAA associated with Glx FAA D/L

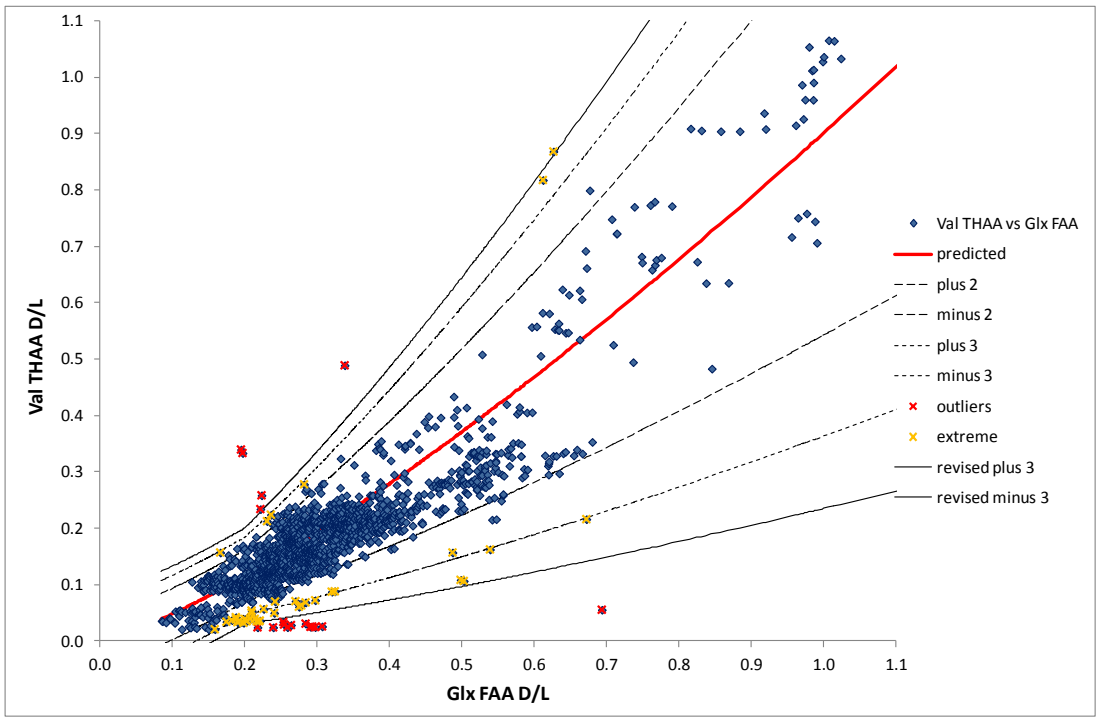


Figure 6.30: Revised ±3 std dev confidence limits for Val THAA associated with Ser THAA D/L

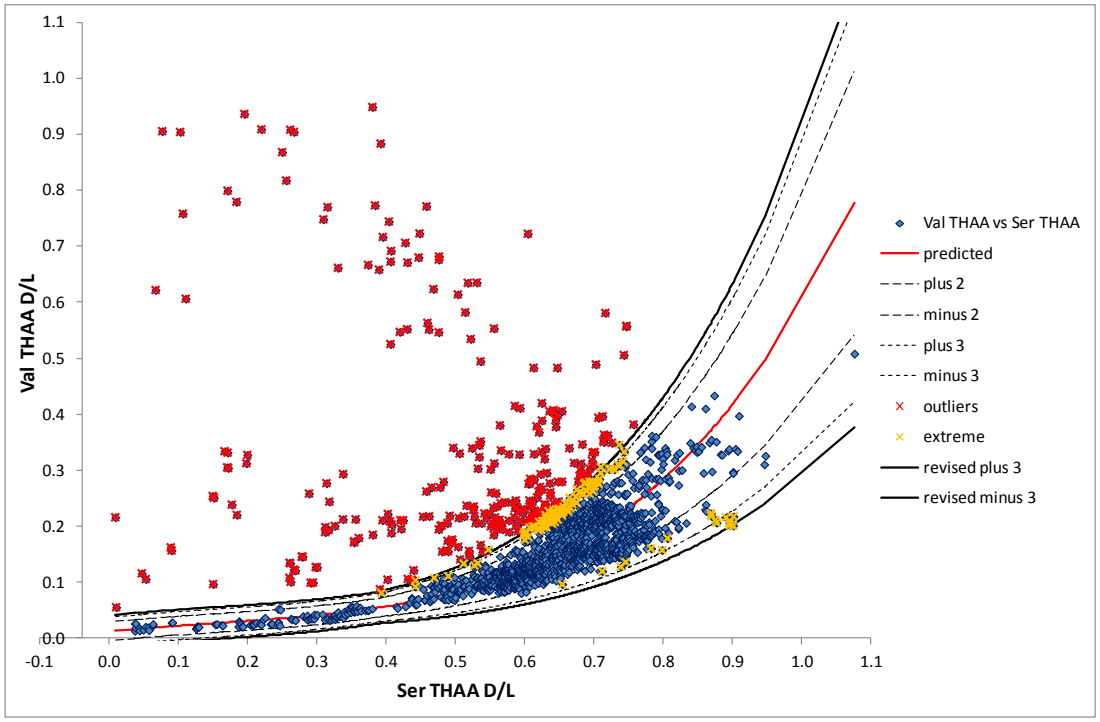


Figure 6.31: Revised ±3 std dev confidence limits for Val THAA associated with Ser FAA D/L

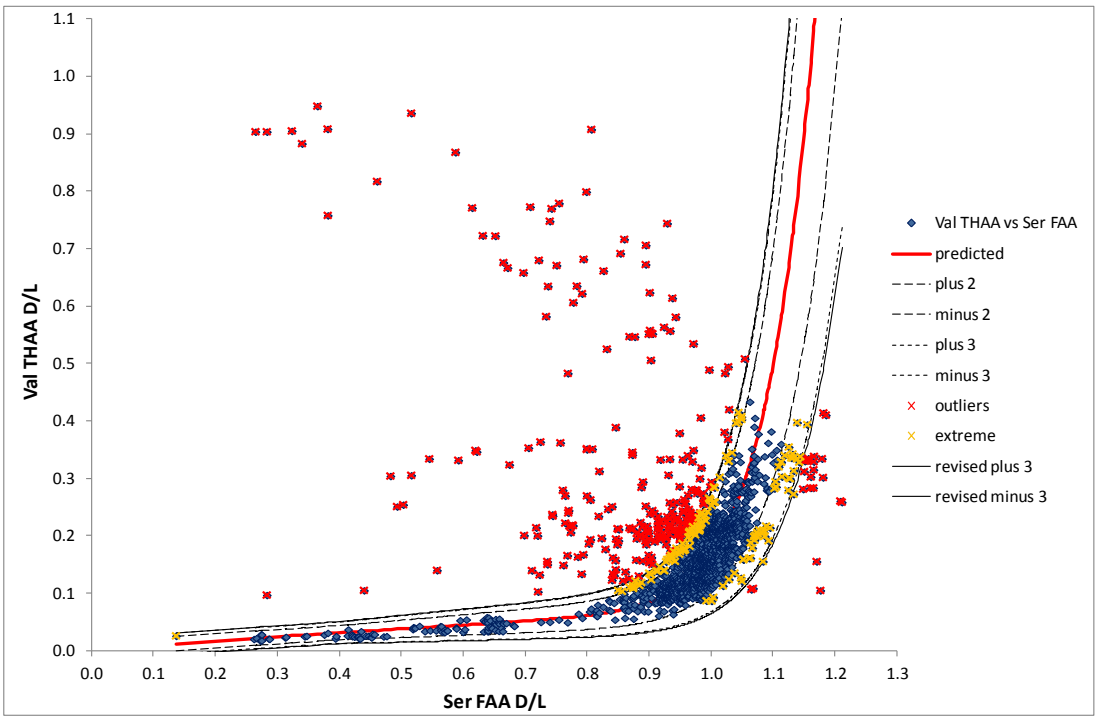


Figure 6.32: Revised ±3 std dev confidence limits for Val THAA associated with Ala THAA D/L

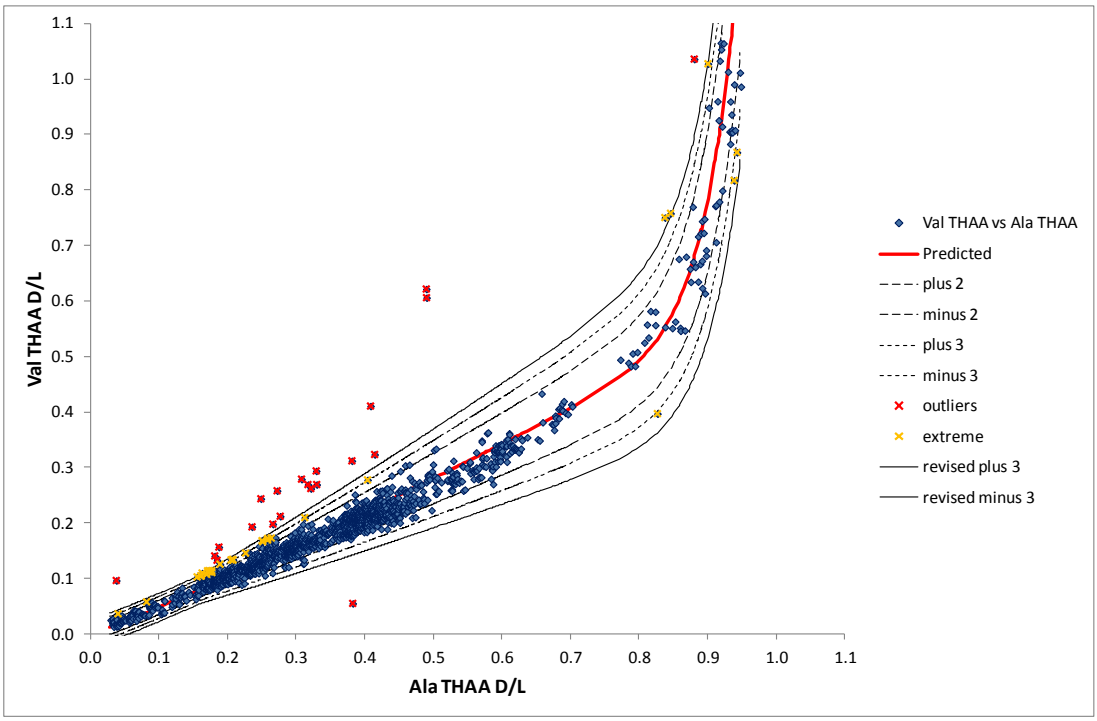


Figure 6.33: Revised ±3 std dev confidence limits for Val THAA associated with Ala FAA D/L

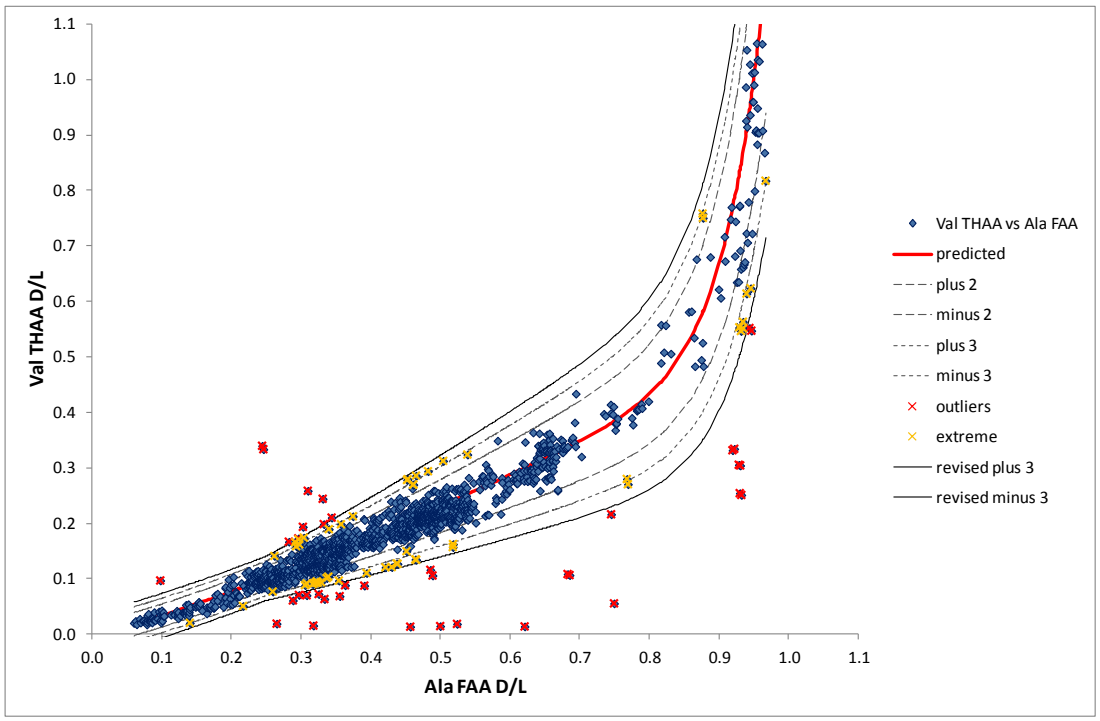


Figure 6.34: Revised ±3 std dev confidence limits for Val THAA associated with Phe THAA D/L

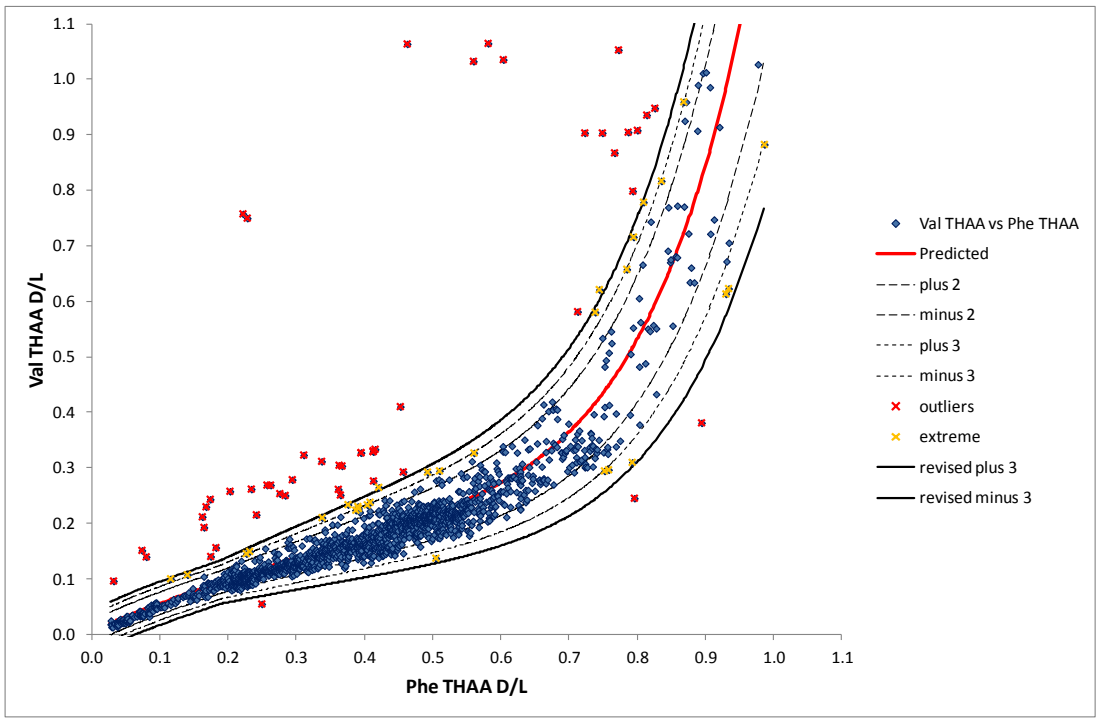


Figure 6.35: Revised ±3 std dev confidence limits for Val THAA associated with Phe FAA D/L

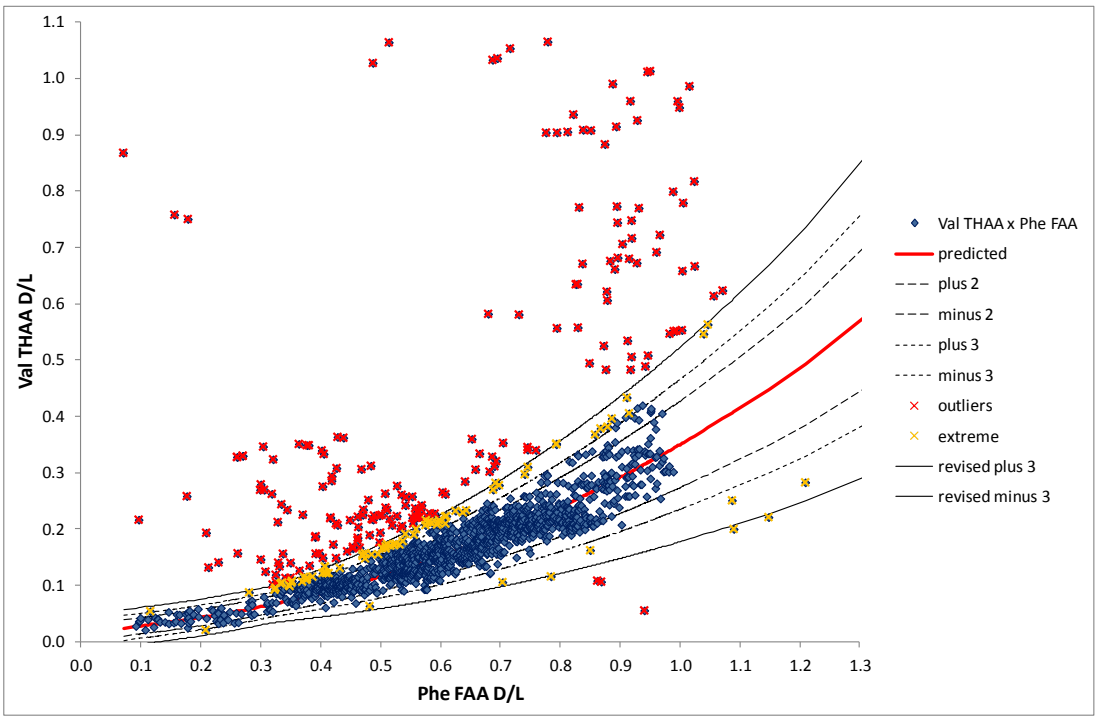


Figure 6.36: Revised ±3 std dev confidence limits for Val THAA associated with Leu THAA D/L

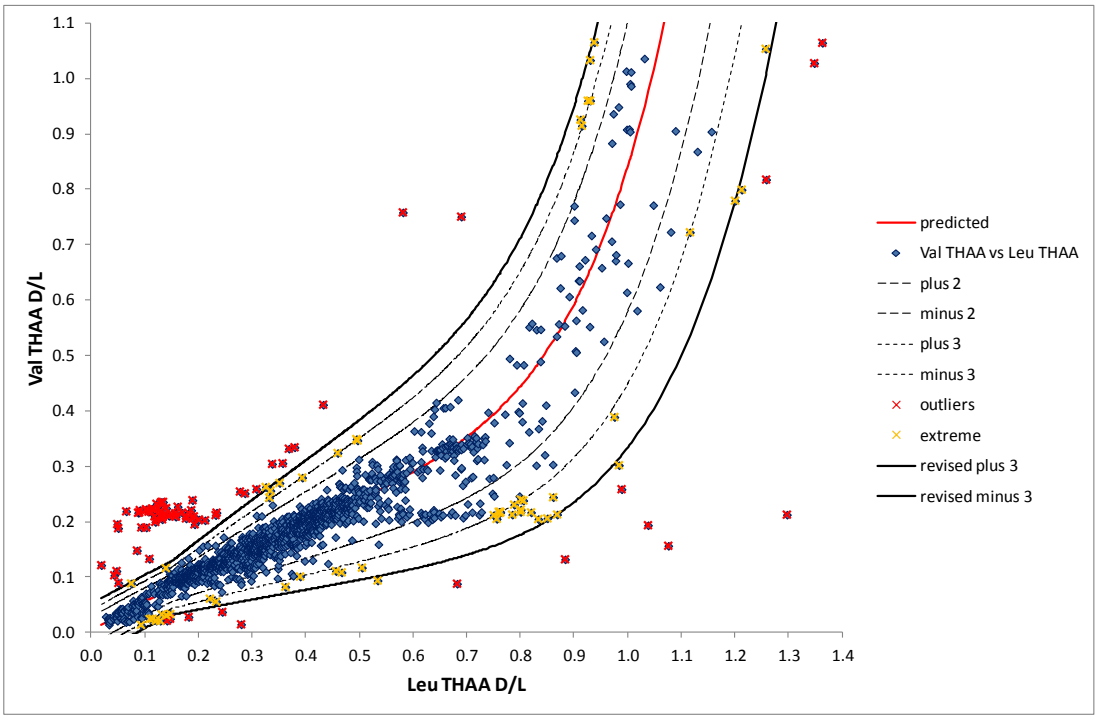


Figure 6.37: Revised ±3 std dev confidence limits for Val THAA associated with Leu FAA D/L

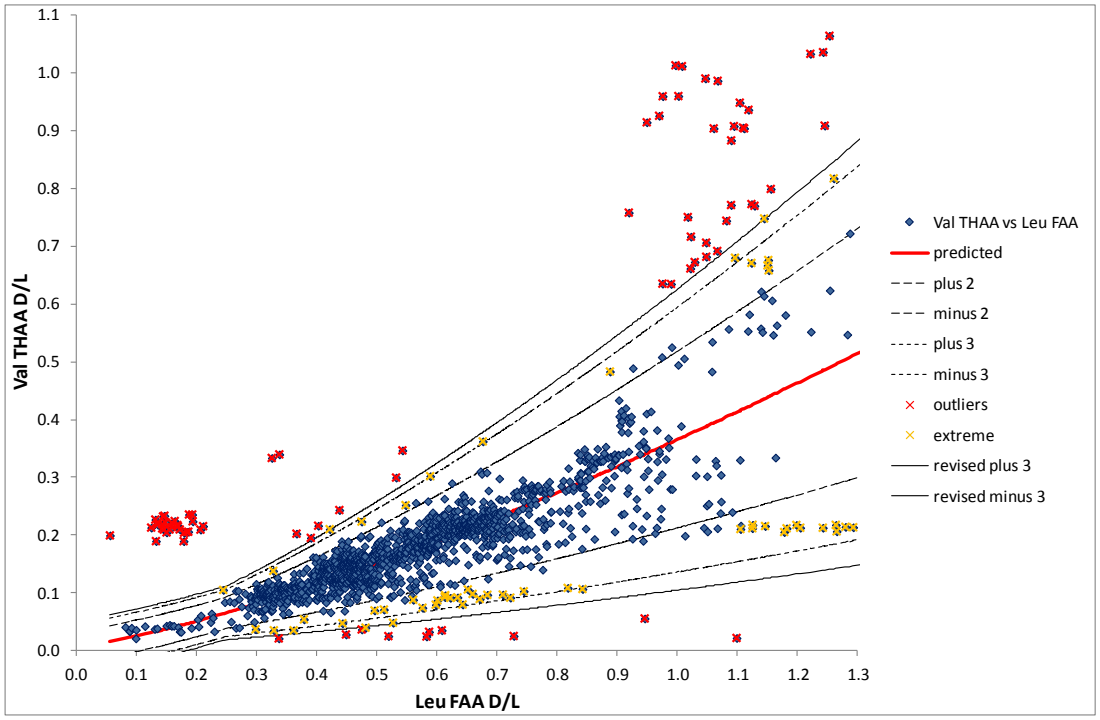


Figure 6.38: Revised ±3 std dev confidence limits for Val THAA associated with D-Aile/L-Ile THAA

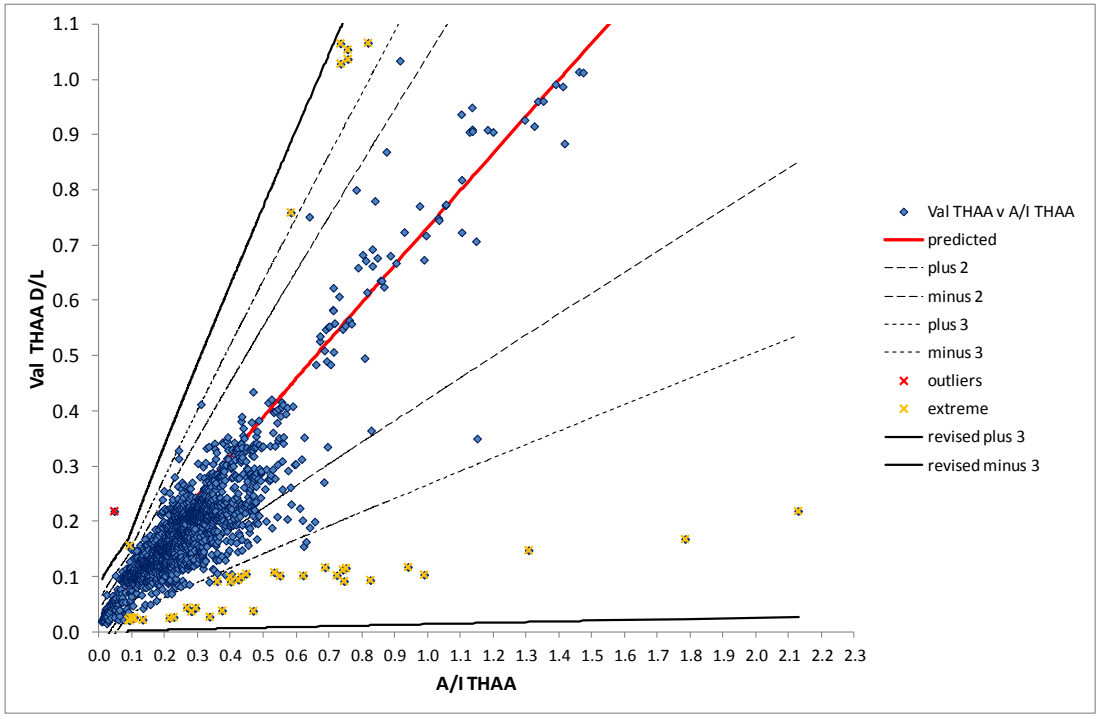


Figure 6.39: Revised ±3 std dev confidence limits for Val THAA associated with D-Aile/L-Ile FAA

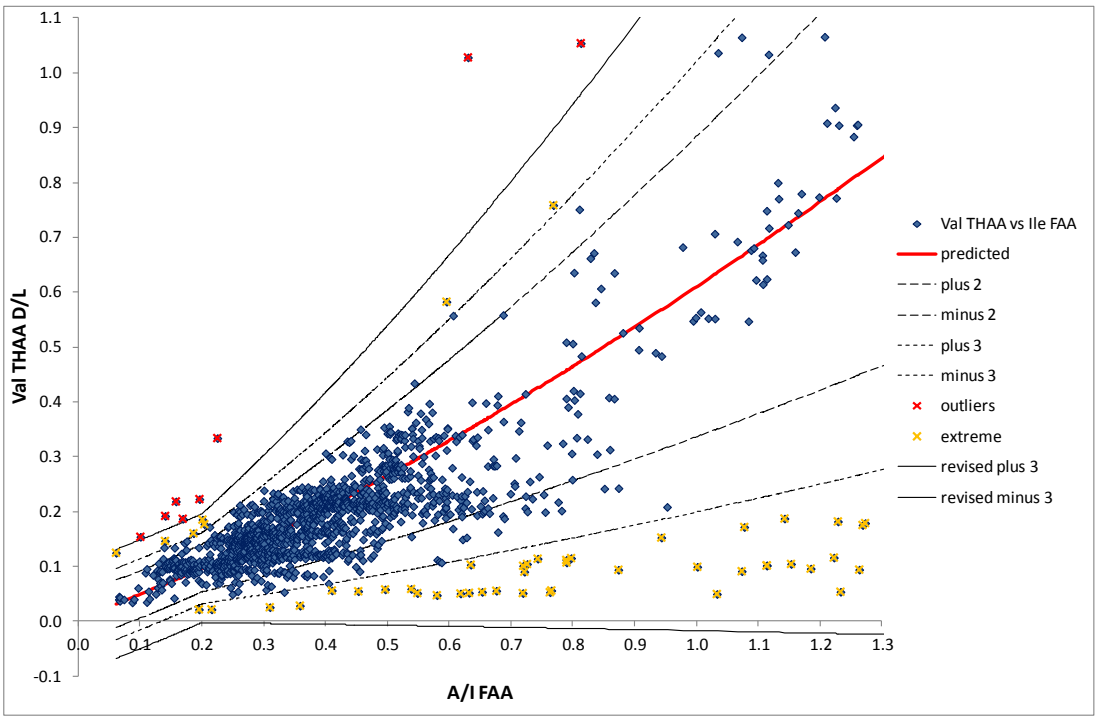
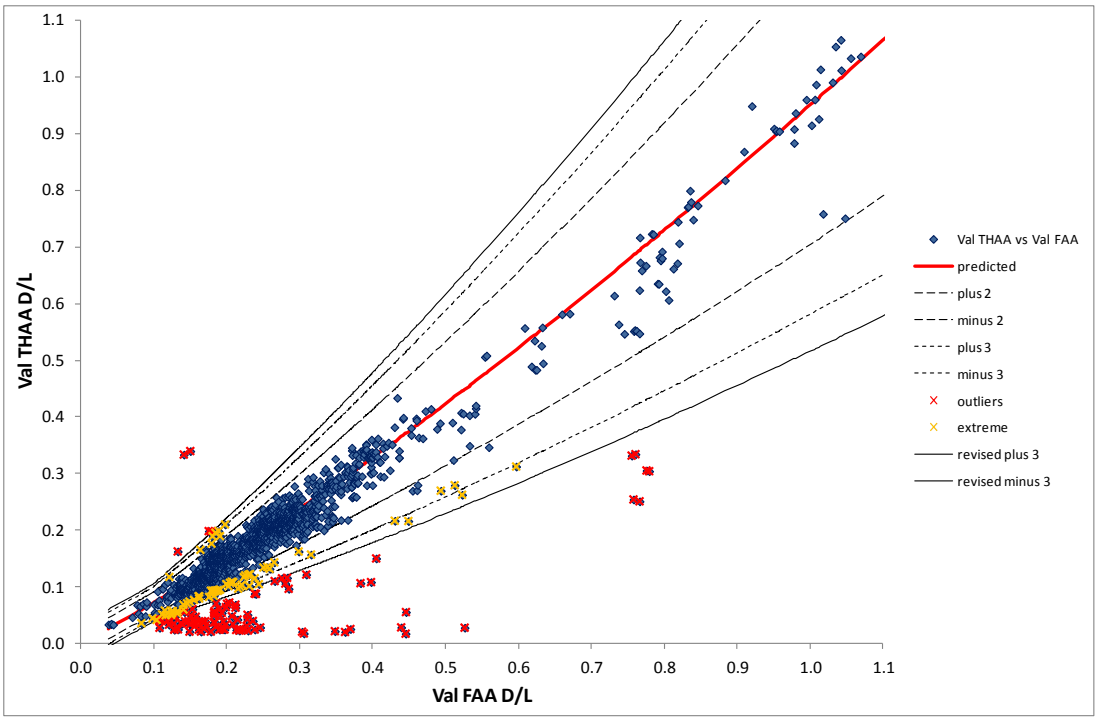


Figure 6.40: Revised ±3 std dev confidence limits for Val THAA associated with Val FAA D/L



6.3.3 Compromised or re-worked samples

Repeatability estimates given in section 4.5.1 provide a means of evaluating the precision of replicate samples, (see also (Grøn *et al.*, 2006)for calculations when n>2). Individual samples may give aberrant data because their integrity has been compromised in some way and amino acids have been lost or contamination has altered the D to L ratio. In situations such as this, the observed association between amino acids will probably shift and data points fall outside of the satisfactory region. Screening would therefore be expected to remove these individual values from further analysis. However there will also be situations where repeated measurements may still be wider than the expected repeatability limits but values fall within the acceptable region and are not identified as outliers. This provides a unique opportunity to identify potentially re-worked and mixed-aged samples suggesting that perhaps re-analysis or re-sampling is necessary.

6.4 A joint probability density model

The approach detailed in the previous section describes a more refined version of the original DMK model (decomposition model kinetic) (Penkman, 2005), later renamed the IcPD model (inter-crystalline protein decomposition) (Penkman *et al.*, 2007), which uses the average predicted valine D/L value as the basis of comparison between site locations.

The new approach correlates protein decomposition with more accurate curve fits for both THAA and FAA fractions to give a series of alternative, predicted valine D/L values with associated uncertainties which change with the valine D/L value. Such values can now be used to update the observed location uncertainty estimate previously derived by ANOVA in section 6.2. As discussed earlier in the chapter, the ANOVA, derived across all valine THAA D/L data for each site, can provide an overall uncertainty reflecting both the analytical and sampling contributions. However the extent to which this is true depends on the sampling and analysis strategy employed (section 6.1). In comparison, the associated predicted values assess the variability between the intra-crystalline amino acids. No distinction is made between analytical runs or samples or sites. Thus, in many ways the uncertainty derived by association for predicted valine D/L values, reflect uncertainty influences not only from the intra-crystalline protein, but also analysis, laboratory effects and sampling, over time.

The previous IcPD approach averaged predicted valine D/L values but gave no estimate of uncertainty. To determine the overall uncertainty for a given location, an

approximation based on the individual estimates is required. This could be achieved by simply taking the average of the uncertainties for each association, thus;

$$\bar{S}_{ValH} = \sqrt{\frac{S^2_{ValH|AsxH} + S^2_{ValH|AsxF} + S^2_{ValH|GlxH} + S^2_{ValH|GlxF} + \cdots etc}{n}}$$

Alternatively, data defined by a mean and standard deviation could be thought of in terms of probability density functions (pdfs). Whilst the GUM does not explicitly express uncertainty in terms of pdfs, it is acknowledged and forms the underlying principle behind Supplement 1 that utilises Monte Carlo simulation as an alternative model for uncertainty estimation. (JCGM 101:2008)

> *"....Thus a Type A standard uncertainty is obtained from a probability density function....derived from an observed frequency distribution...., while a Type B standard uncertainty is obtained from an assumed probability density function based on the degree of belief that an event will occur [often called subjective probability....]. Both approaches employ recognized interpretations of probability*." (JCGM 100:2008, 3.3.5, p7)

From Figure 6.27 - Figure 6.40, it was shown that, with the exception of outliers, relative data are generally evenly distributed either side of the association lines. For values of x, the greatest density of data can usually be found around the mid-point, and reflects the believability or probability that the corresponding value of y is in fact a close approximation of the true value (if it could be known). Given the large size of the data set, approaching 2000 data points, there are no reasons to suspect that data do not approximate to a normal or Gaussian distribution. The normal probability density function that gives the typical bell shaped curve, is defined by two parameters; μ (mu), the position of the true value and σ (sigma) the standard deviation or spread of data; N(μ,σ). Taken together, the parameters μ and σ determine the probability density of a value y, such that the smaller the spread, the tighter the data, the higher the peak and the greater the chances are that μ is the true value. The function is described by;

$$p(y|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{\left(-\frac{y-\mu^2}{2\sigma^2}\right)} = \frac{1}{\sqrt{2\pi\sigma}} exp\left(-\frac{1}{2}\left[\frac{y-\mu}{\sigma}\right]^2\right) \qquad \textbf{(6.10)}$$

Thus it can be appreciated that the pdfs for each association change with increasing D/L as the standard deviations widen with age. Given a location's observed amino acid D/L values, these can now be used to predict alternative values for valine plus a series of prior

predicted probability densities. Probability theory can now be used to derive a revised probability (uncertainty), based around a normal distribution model and observed data.

6.4.1 A Bayesian approach

Thomas Bayes (1702-1761) was a Presbyterian minister and mathematician, whose theorem was published posthumously in 1764 (Kruschke, 2011).

> Bayes rule *"...determine[s] the probability of a model when given a set of data. What the model itself provides is the probability of the data, given specific parameter values and the model structure. We use Bayes' rule to get from the probability of the data, given the model, to the probability of the model, given the data*." (Kruschke, 2011, p52).

> " *Bayesian inference is the process of fitting a probability model to a set of data and summarizing the result by a probability distribution on the parameters of the model and on unobserved quantities such as predictions for new observations*" (Gelman *et al*, 2004, p1).

Bayes theorem is based around the idea of conditional probability which looks at the probability of an event happening (y), given something else, (x), that is the probability of y is conditional upon x, written as; $p(y|x)$ (Currell and Dowman, 2005). Conditional probability is defined as;

$$\boldsymbol{p(y|x) = \frac{p(y,x)}{p(x)}} \qquad \textbf{(8.11)}$$

> In other words, "*..the probability of y given x is the same as the probability of x and y happening together, relative to the probability of x happening at all*." (Kruschke, 2011, p53).

The following derivations are all taken from Kruschke (2011) with page references where relevant, which is a strongly recommended text presenting Bayesian statistics for the non-mathematical!

When $p(x)>0$, Bayes rule is derived thus (ibid, p53);

1. From the above equation, multiply both sides by $p(x)$;

$$p(y|x)p(x) = p(y,x)$$

2. Because $p(x,y)=p(y,x)$, similarly we can derive;

$$p(x|y) = \frac{p(y,x)}{p(y)} \qquad [\times p(y)] \quad = \; p(x|y)p(y) = p(y,x)$$

3. So now there are two expressions that equal $p(y, x)$;

$$p(y,x) = p(x|y)p(y) = p(y|x)p(x)$$

4. Taking the last two expressions, divide them by $p(x)$;

$$\frac{p(x|y)p(y)}{p(x)} = \frac{p(y|x)\cancel{p(x)}}{\cancel{p(x)}}$$

This gives the basic Bayes formula;

$$\boldsymbol{p(y|x) = \frac{p(x|y)p(y)}{p(x)}} \qquad \textbf{(6.12)}$$

Now, the probability of x and y happening together is the same as y and x happening together; $p(x, y) = p(y, x)$. To determine the probability distribution of x on its own, $p(x, y)$ is summed across all values of y (ibid, p44), thus;

$$p(x) = \sum_y p(x,y)$$

Also, if;

$$p(x|y) = \frac{p(y,x) \; or \; p(x,y)}{p(y)} \; then \; p(x,y) = p(x|y)p(y)$$

Therefore;

$$p(x) = \sum_y p(x,y) = \sum_y p(x|y)p(y)$$

So now Bayes formula becomes;

$$\boldsymbol{p(y|x) = \frac{p(x|y)p(y)}{\sum_y p(x|y)p(y)}} \qquad \textbf{(6.13)}$$

Where the y in the numerator is a fixed value but the *"...y in the demoninator is a variable that takes on all possible values of y over the summation*." (ibid, p53).

When dealing with discrete variables, probabilities are expressed as probability masses (p44). However, when applied to continuous variables, probability masses become probability densities and the summation becomes an integral (ibid, p56). Thus Bayes formula changes to;

$$\boldsymbol{p(y|x) = \frac{p(x|y)p(y)}{\int dy \; p(x|y)p(y)}} \qquad \textbf{(6.14)}$$

In terms of the probability for a single value y, given parameters μ and σ, this becomes;

$$\boldsymbol{p(\mu,\sigma|y) = \frac{p(y|\mu,\sigma)p(\mu,\sigma)}{\iint d\mu\, d\sigma\, p(y|\mu,\sigma)p(\mu,\sigma)}} \quad \textbf{(6.15)}$$

Where; $p(\mu, \sigma)$ = the prior belief in μ,σ, before observed data are considered, $p(\mu, \sigma|y)$ = posterior belief in μ, σ, when data have been taken into account, $p(y|\mu, \sigma)$ = likelihood that the data could be derived from the model with parameters μ, σ, and $\int\int d\mu d\sigma p(y|\ \mu, \sigma)p(\mu, \sigma) = p(y)$ which is the probability of y given the model, referred to as the evidence or marginal likelihood (ibid, p57 & 58).

Prior information can be anything that helps to inform our prior belief about a distribution, before the data is observed. The likelihood is based on the observation. Since the posterior is proportional to the prior multiplied by the likelihood, equation 6.15 can be evaluated using a normal density likelihood to derive estimates of the posterior mean and uncertainty. The following derivation has again been taken from Kruschkel (2011), p392-393 and readers are recommended to refer to this text for a fuller explanation.

1. Let the parameters for the prior distribution on μ be normal with a mean M_μ and standard deviation s_μ, $(N(\mu|M_\mu, s_\mu))$ and the parameters for the likelihood also be normal and described by $N(y|\ \mu, s_y)$.
2. From the equation for a normal distribution we have;

$$p(y|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} exp\left(-\frac{1}{2}\left[\frac{y-\mu}{\sigma^2}\right]^2\right)$$

$$\propto exp\left(-\frac{1}{2}\left[\frac{y-\mu}{\sigma^2}\right]^2\right)$$

3. Therefore (likelihood x prior) ;

$$p(y|\mu,\sigma)p(\mu,\sigma)$$

$$\propto exp\left(-\frac{1}{2}\left[\frac{y-\mu}{S_y^2}\right]^2\right) exp\left(-\frac{1}{2}\left[\frac{\mu-M_\mu}{S_\mu^2}\right]^2\right)$$

$$= exp\left(-\frac{1}{2}\left[\frac{(y-\mu)^2}{S_y^2} + \frac{\left(\mu-M_\mu\right)^2}{S_\mu^2}\right]\right)$$

$$= exp\left(-\frac{1}{2}\left[\frac{S_\mu^2(y-\mu)^2 + S_y^2\left(\mu-M_\mu\right)^2}{S_y^2 S_\mu^2}\right]\right)$$

$$= exp\left(-\frac{1}{2}\left[\frac{S_y^2+S_\mu^2}{S_y^2 S_\mu^2}\left(\mu^2 - 2\frac{S_y^2 M_\mu + S_\mu^2 y}{S_y^2+S_\mu^2}\mu + \frac{S_y^2 M_\mu^2 + S_\mu^2 y^2}{S_y^2+S_\mu^2}\right)\right]\right)$$

$$= exp\left(-\frac{1}{2}\left[\frac{S_y^2+S_\mu^2}{S_y^2S_\mu^2}\left(\mu^2-2\frac{S_y^2M_\mu+S_\mu^2y}{S_y^2+S_\mu^2}\mu\right)\right]\right)$$

$$\times exp\left(-\frac{1}{2}\left[\frac{S_y^2+S_\mu^2}{S_y^2S_\mu^2}\left(\frac{S_y^2M_\mu^2+S_\mu^2y^2}{S_y^2+S_\mu^2}\right)\right]\right)$$

$$\propto exp\left(-\frac{1}{2}\left[\frac{S_y^2+S_\mu^2}{S_y^2S_\mu^2}\left(\mu^2-2\frac{S_y^2M_\mu+S_\mu^2y}{S_y^2+S_\mu^2}\mu\right)\right]\right)$$

$$\propto exp\left(-\frac{1}{2}\left[\frac{S_y^2+S_\mu^2}{S_y^2S_\mu^2}\left(\mu^2-2\frac{S_y^2M_\mu+S_\mu^2y}{S_y^2+S_\mu^2}\mu+\left(\frac{S_y^2M_\mu+S_\mu^2y}{S_y^2+S_\mu^2}\right)^2\right)\right]\right)$$

$$= \boldsymbol{exp\left(-\frac{1}{2}\left[\frac{S_y^2+S_\mu^2}{S_y^2S_\mu^2}\left(\mu-\frac{S_y^2M_\mu+S_\mu^2y}{S_y^2+S_\mu^2}\mu\right)^2\right]\right)}$$

This is the numerator of Bayes equation. "*When it is normalized by the evidence in the denominator, it becomes a probability density function*." (ibid, p393)

6.4.1.1 Posterior mean and standard deviation for single values

The above derived equation has the same structure as a normal distribution such that;

$$exp\left(-\frac{1}{2}\left[\frac{y-\mu}{\sigma^2}\right]^2\right) \cong exp\left(-\frac{1}{2}\left[\frac{S_y^2+S_\mu^2}{S_y^2S_\mu^2}\left(\mu-\frac{S_y^2M_\mu+S_\mu^2y}{S_y^2+S_\mu^2}\right)^2\right]\right)$$

Where;

$$\frac{S_y^2S_\mu^2}{S_{y+}^2S_\mu^2} \equiv posterior\ variance\ \sigma^2 \quad \text{and}$$

$$\sqrt{\frac{S_y^2S_\mu^2}{S_{y+}^2S_\mu^2}}$$

$$\equiv posterior\ standard\ deviation\ \sigma \quad \text{and}$$

$$\frac{S_y^2M_\mu+S_\mu^2y}{S_y^2+S_\mu^2} \equiv posterior\ mean$$

However, Kruschke observes that the reciprocal of the squared standard deviation, is in fact the precision of the normal. Thus, as the standard deviation goes up, so the precision goes down. When expressed as precision, this can be simplified to;

$$\frac{1}{\sigma^2} \equiv \frac{S_y^2+S_\mu^2}{S_y^2S_\mu^2} = \frac{1}{S_\mu^2}+\frac{1}{S_y^2} \qquad \textbf{(6.16a)}$$

The posterior precision is simply the sum of the prior and likelihood precisions (ibid, p292). The posterior standard deviation or uncertainty can therefore be simply expressed as;

$$Posterior\ standard\ deviation, s = \sqrt{\left(1/S_y^2 + 1/S_\mu^2\right)^{-1}} \quad \textbf{(6.16b)}$$

Or since s^2 = variance(RSC Analytical Methods Committee, 2003b), as;

$$Posterior\ variance = (v_y^{-1} + v_\mu^{-1})^{-1} \quad \textbf{(6.16c)}$$

Similarly the mean can be expressed in terms of precision. Thus, "*The posterior mean is a weighted average of the prior mean and the datum, with the weighting corresponding to the relative precisions of the prior and the likelihood.*" (Kruschke, 2011, p394).

$$Posterior\ mean = \frac{S_y^2 M_\mu + S_\mu^2 y}{S_y^2 + S_\mu^2} \quad \textbf{(6.17a)}$$

$$= \frac{1/S_\mu^2}{1/S_y^2 + 1/S_\mu^2} M_\mu + \frac{1/S_y^2}{1/S_y^2 + 1/S_\mu^2} y$$

Expressed as a variance (RSC Analytical Methods Committee, 2003b), this reduces to;

$$Posterior\ mean = \left(v_y^{-1} + v_\mu^{-1}\right)^{-1}(v_y^{-1} y + v_\mu^{-1} M_\mu) \quad \textbf{(6.17b)}$$

Weighting using precision estimates has important implications for the amino acid data. For example, a highly precise prior with a small standard deviation, will be weighted more than one that is less precise. Therefore the resulting posterior mean will fall closer towards the prior mean. This allows us to benefit from the closer precision of some amino acids, which will weight the predicted valine D/L towards those contributions with the best precision and down-weight predicted values from associated amino acids with poorer precision.

6.4.1.2 Posterior mean and standard deviation adjusted for means

These formulae also permit repeated measurements to be taken into account too, in just the same way that uncertainty estimates for a mean in classical descriptive statistics, reduce as the sample number goes up (uncertainty (mean) = s/√n).

However, rather than the distribution of y_i observed data (likelihood) being described by the parameters μ and σ, i.e.; $N(y_i|\mu,\sigma)$, the distribution of the single mean estimate for yi is described by $N(\bar{y}|\mu,\sigma/\sqrt{n})$ (Kruschke, 2011, p394). Thus for a normal likelihood and prior distribution as before, the posterior distribution on y has;

$$Posterior\ mean = \bar{\mu}_p = \frac{1/S_\mu^2}{n/S_y^2 + 1/S_\mu^2} M_\mu + \frac{n/S_y^2}{n/S_y^2 + 1/S_\mu^2} \bar{y} \qquad \textbf{(6.18a)}$$

Or

$$\bar{\mu}_p = \frac{\frac{1}{S_\mu^2} M_\mu + \frac{n}{S_y^2} \bar{y}}{\frac{1}{S_\mu^2} + \frac{n}{S_y^2}} \qquad \textbf{(6.18b)}$$

And

$$Posterior\ precision = \frac{1}{S_p^2} = \frac{1}{S_\mu^2} + \frac{n}{S_y^2} \qquad \textbf{(6.19)}$$

(Gelman *et al.*, 2004, p49; Kruschke, 2011, p394)

6.4.2 Application to Amino Acid data

6.4.2.1 Posterior mean D/L and uncertainty for single values

For each site location, ANOVA was carried out to give estimates of; the repeatability standard deviation, s_r, generally taken to represent the analytical uncertainty contribution; s_L, the between-sample variability, generally reflecting sampling uncertainty but which may have been determined under repeatability or reproducibility conditions; and s_{RW}, the intermediate reproducibility standard deviation, taken to be the overall estimate of uncertainty from the data available for any given location.

After screening the data and removing values exceeding |z|>3, amino acid mean D/L values were used to derive predicted valine THAA D/L values using the association functions described previously in section 6.3.1.

Predicted valine D/L values derived from each of the associations (Val^H_{aa},) and associated uncertainties (s_{aa}), were used as prior distribution parameters (θ_{prior}). The ANOVA mean valine D/L value (Val^H_{ANOVA}) and reproducibility standard deviation of observed data (s_R) were taken as the likelihood;

Prior distribution; $\theta_{prior} = N(\mu|Val^{H}_{aa}, s_{aa})$,

Where; $\theta_{prior} = (\theta_{AsxH}, \theta_{AsxF}, \theta_{GlxH}, \theta_{GlxF}, \theta_{SerH}, \theta_{SerF}, \theta_{AlaH}, \theta_{AlaF}, \theta_{PheH}, \theta_{PheF}, \theta_{ValF}, \theta_{LeuF}, \theta_{LeuH}, \theta_{A/I\ H}, \theta_{A/I\ F})$

and

$$\theta_{AsxH} = N(\mu|Val^{H}_{AsxH}, s_{AsxH})$$

$$\theta_{AsxF} = N(\mu|Val^{F}_{AsxF}, s_{AsxF})$$

$$\theta_{GlxH} = N(\mu|Val^{H}_{GlxH}, s_{GlxH})$$

.....etc

Likelihood; $\theta_{ValH} = N(Val^{H}_{ANOVA}|\mu, s_R)$.

Posterior mean (μ_{ValH});

(6.20)

$$\mu_{ValH} = \frac{\frac{1}{S_R^2}Val^H_{ANOVA} + \frac{1}{S^2_{AsxH}}Val^H_{AsxH} + \frac{1}{S^2_{AsxF}}Val^H_{AsxF} \ldots\ldots + \frac{1}{S^2_{aa}}Val^H_{aa}}{\frac{1}{S_R^2} + \frac{1}{S^2_{AsxH}} + \frac{1}{S^2_{AsxF}} \ldots\ldots + \frac{1}{S^2_{aa}}}$$

Posterior standard deviation (s_{ValH});

$$s_{ValH} = \sqrt{\left(\frac{1}{S_R^2} + \frac{1}{S^2_{AsxH}} + \frac{1}{S^2_{AsxF}} \ldots\ldots + \frac{1}{S^2_{aa}}\right)^{-1}} \quad \textbf{(6.21)}$$

Posterior Expanded uncertainty;

$$\mu_{ValH} \pm s_{ValH} \times k, \quad where\ k = 1.96 = t_{(\alpha=0.05, df=\infty)}$$

6.4.2.2 Posterior Valine THAA D/L and uncertainty adjusted for means

As the mean of each amino acid's D/L value is derived from a number of measurement results, rather than expressing the uncertainty as the standard deviation from individual values, it should perhaps be expressed as the standard deviation of means, where u=s/√n, (even though the requirement for independence may be questionable).

For ANOVA (section 6.2) this was achieved using $\bar{s}_{RW} = \sqrt{\frac{s_r^2}{n} + \frac{s_L^2}{p}}$. For predicted uncertainties, associations are derived from individual values.

Therefore;

Posterior mean ($\bar{\mu}_{ValH}$); **(6.22)**

$$\bar{\mu}_{ValH} = \frac{\left(\frac{\mathrm{n}}{s_r^2} + \frac{\mathrm{p}}{s_L^2}\right) Val_{ANOVA}^H + \frac{n_{AsxH}}{s_{AsxH}^2} Val_{AsxH}^H + \frac{n_{AsxF}}{s_{AsxF}^2} Val_{AsxF}^H \ldots\ldots + \frac{n_{aa}}{s_{aa}^2} Val_{aa}^H}{\left(\frac{\mathrm{n}}{s_r^2} + \frac{\mathrm{p}}{s_L^2}\right) + \frac{n_{AsxH}}{s_{AsxH}^2} + \frac{n_{AsxF}}{s_{AsxF}^2} \ldots\ldots + \frac{n_{aa}}{s_{aa}^2}}$$

Posterior standard deviation ($\bar{S}_{ValH}$);

$$\bar{s}_{ValH} = \sqrt{\left(\frac{\mathrm{n}}{s_r^2} + \frac{\mathrm{p}}{s_L^2} + \frac{n_{AsxH}}{s_{AsxH}^2} + \frac{n_{AsxF}}{s_{AsxF}^2} \ldots\ldots + \frac{n_{aa}}{s_{\mathrm{aa}}^2}\right)^{-1}} \quad \textbf{(6.23)}$$

Posterior Expanded uncertainty;

$$\bar{\mu}_{ValH} \pm \bar{s}_{ValH} \times k, \quad where\ k = 1.96 = t_{(\alpha=0.05, df=\infty)}$$

Alternatively, the effective degrees of freedom (v_{eff}) could be determined using the Welch-Satterthwaite equation (Equations 8.2 and 8.3). Therefore $k = t_{(\alpha=0.05, veff)}$.

6.4.3 Evaluating Results

Previously collected AAR data relating to the Thames Terrace sequence (e.g. Bridgland, 1994; Bridgland *et al.*, 2004a; Bridgland, 2006) were assessed using the standard ANOVA method which was then also combined with predicted values using Bayes. For each location, data are summarised in Table 6.5 and show mean D/L values, standard deviations (s) and relative standard deviations (RSD%) both for single values (to reflect the uncertainty on the distribution of results) and also adjusted for means, (to take into account the number of repeated measurements, (n). Figure 6.41 - Figure 6.44 show these data, associated against respective marine isotope stages, where independent evidence was available (Penkman *et al.*, 2011) and also presented as an Excel spreadsheet in Chpt 6: Appendix 3. In all cases, uncertainty estimates have been presented as expanded values, where the relevant standard deviation is multiplied by a coverage factor, *k.* In all cases *k* = 1.96, equivalent to 95% probability level or approximately 2 standard deviations, to allow for direct comparisons. Dotted lines represent confidence intervals for single values whereas solid lines represent confidence intervals for means, adjusted for *n*.

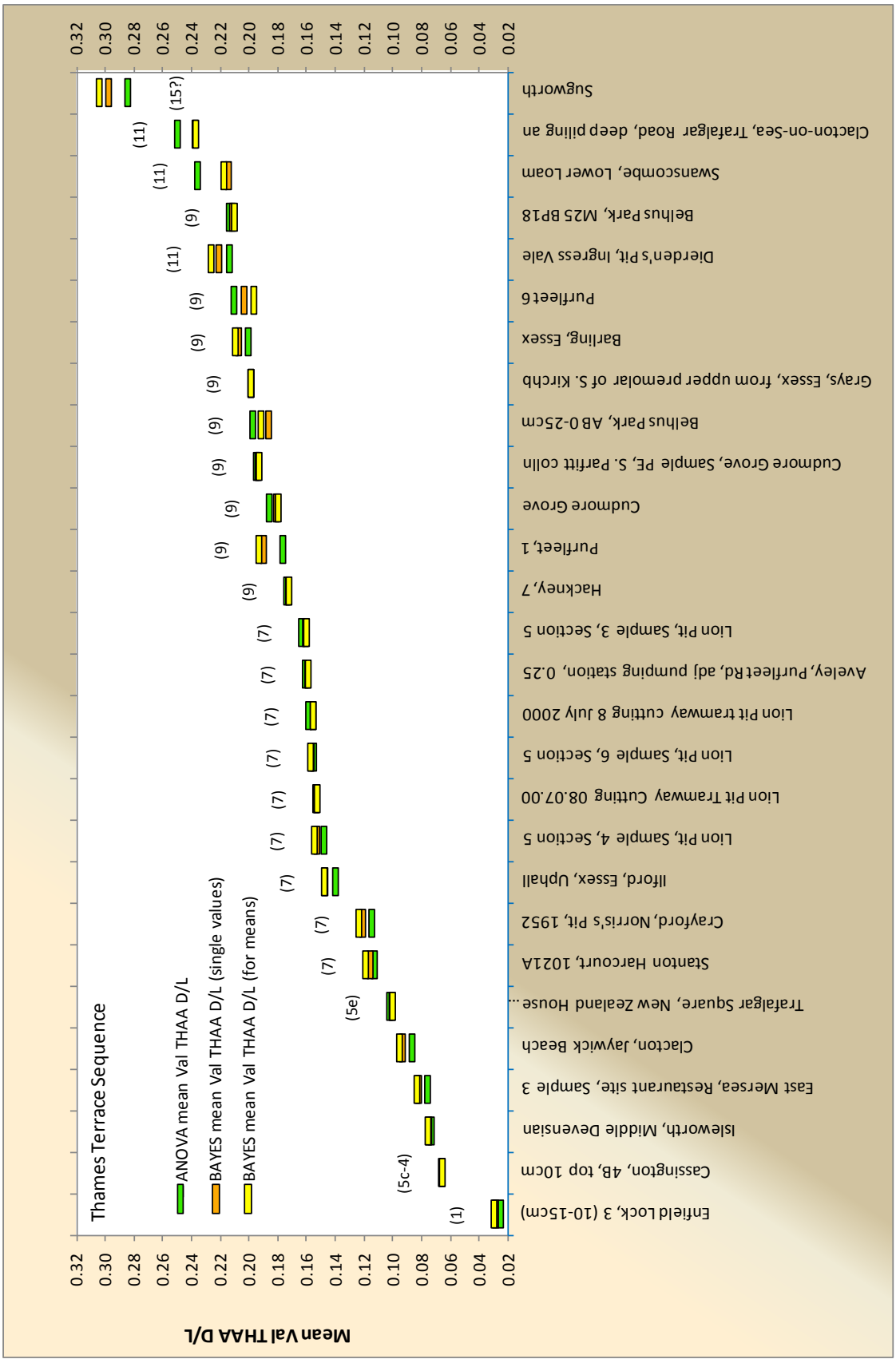


Figure 6.41: Thames Terrace Sequence. Comparison of Val THAA D/L values derived from ANOVA and as Bayesian posterior means (both for single values and adjusted for means). Independently derived Marine Isotope Stages are given in brackets if known.

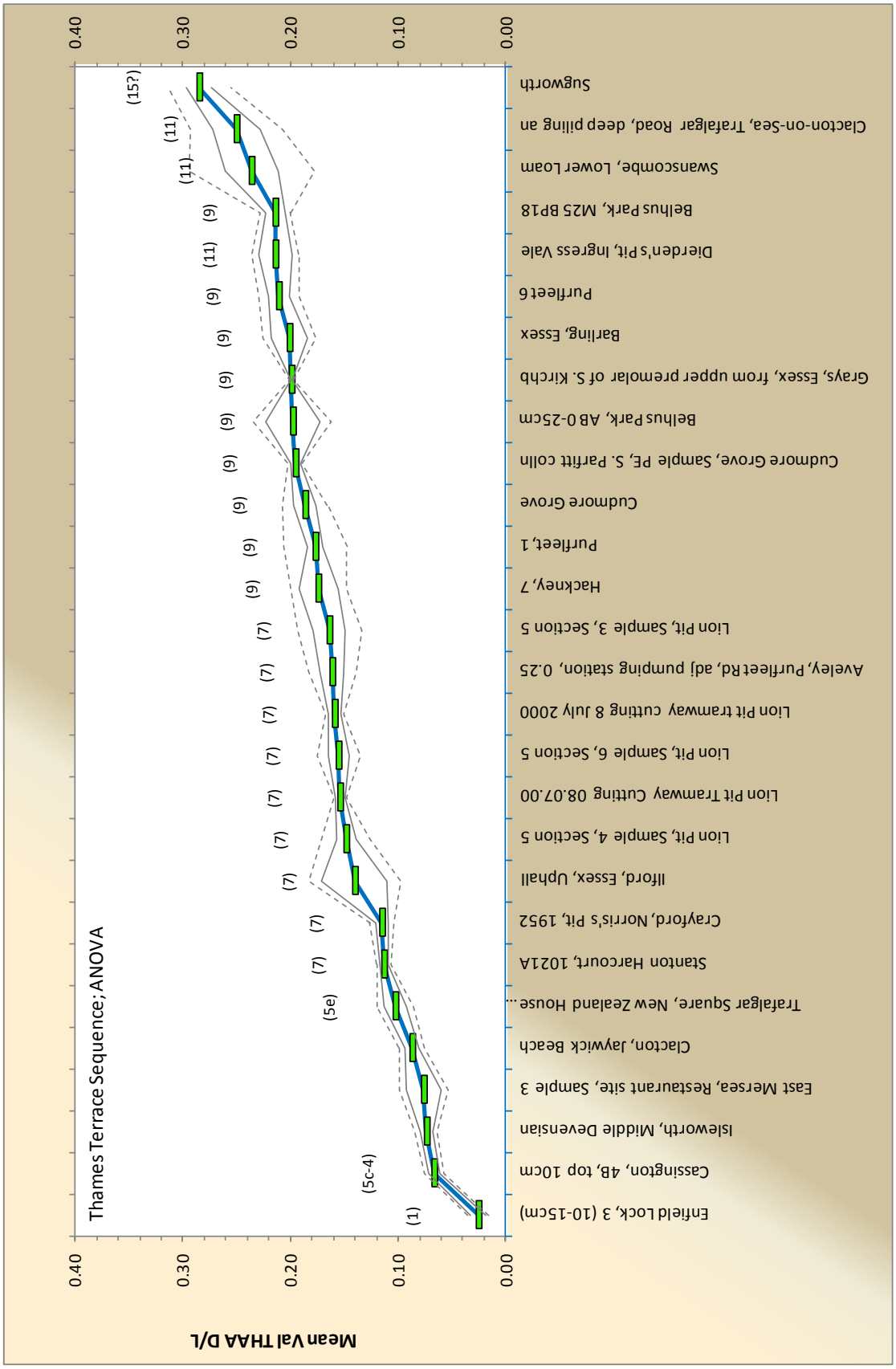


Figure 6.42: Thames Terrace Sequence. ANOVA derived Val THAA D/L values with expanded uncertainty based on s_R estimates and k=1.96 (95% probability). Dotted line = 1.96 x s_R for single values and solid line = 1.96 x s_R.for means

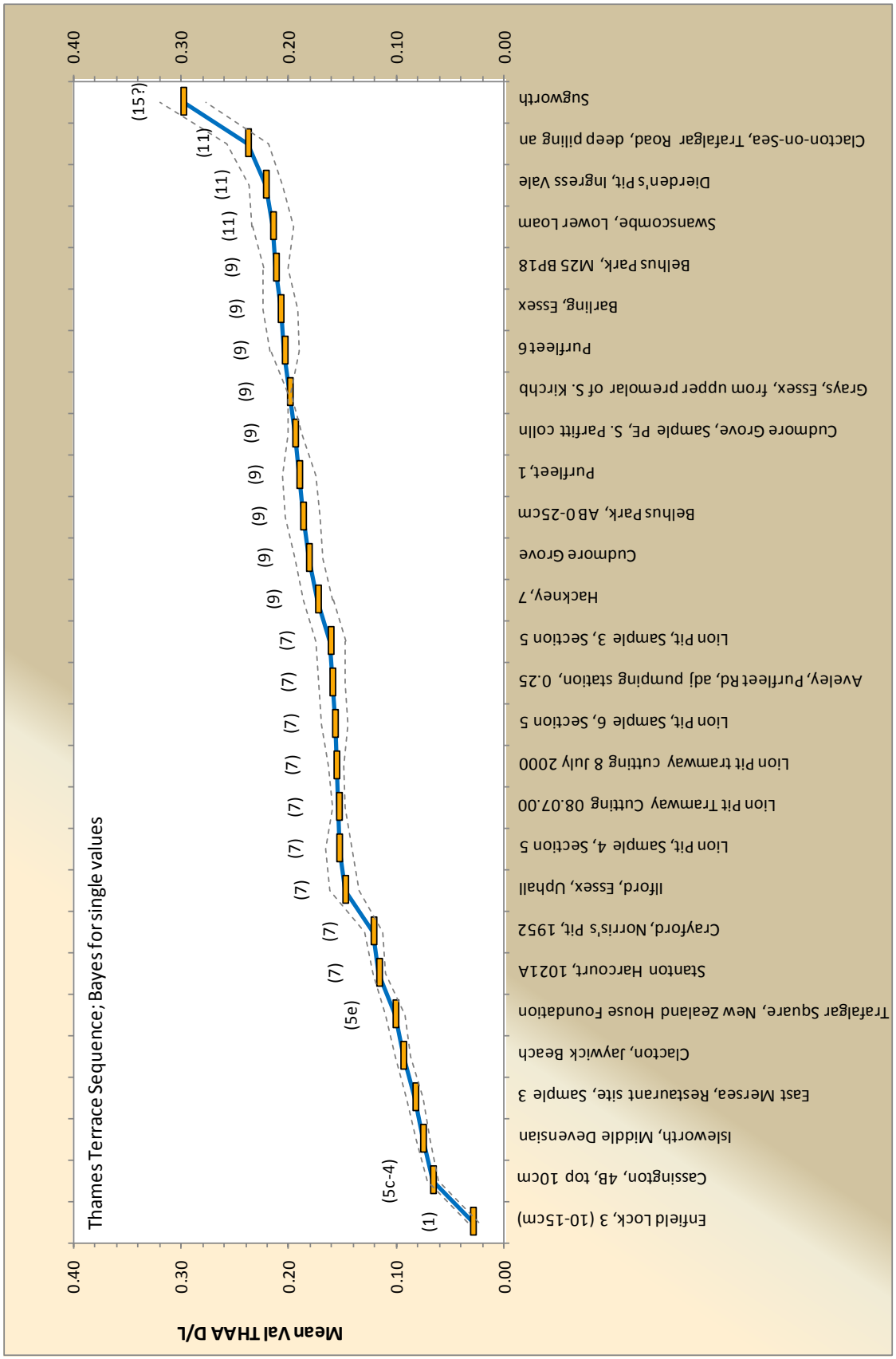


Figure 6.43: Thames Terrace Sequence. Bayesian derived posterior Val THAA D/L values with expanded uncertainty using k=1.96 (95% probability). Solid line = 1.96 x $s_{posterior}$ MIS given in brackets.

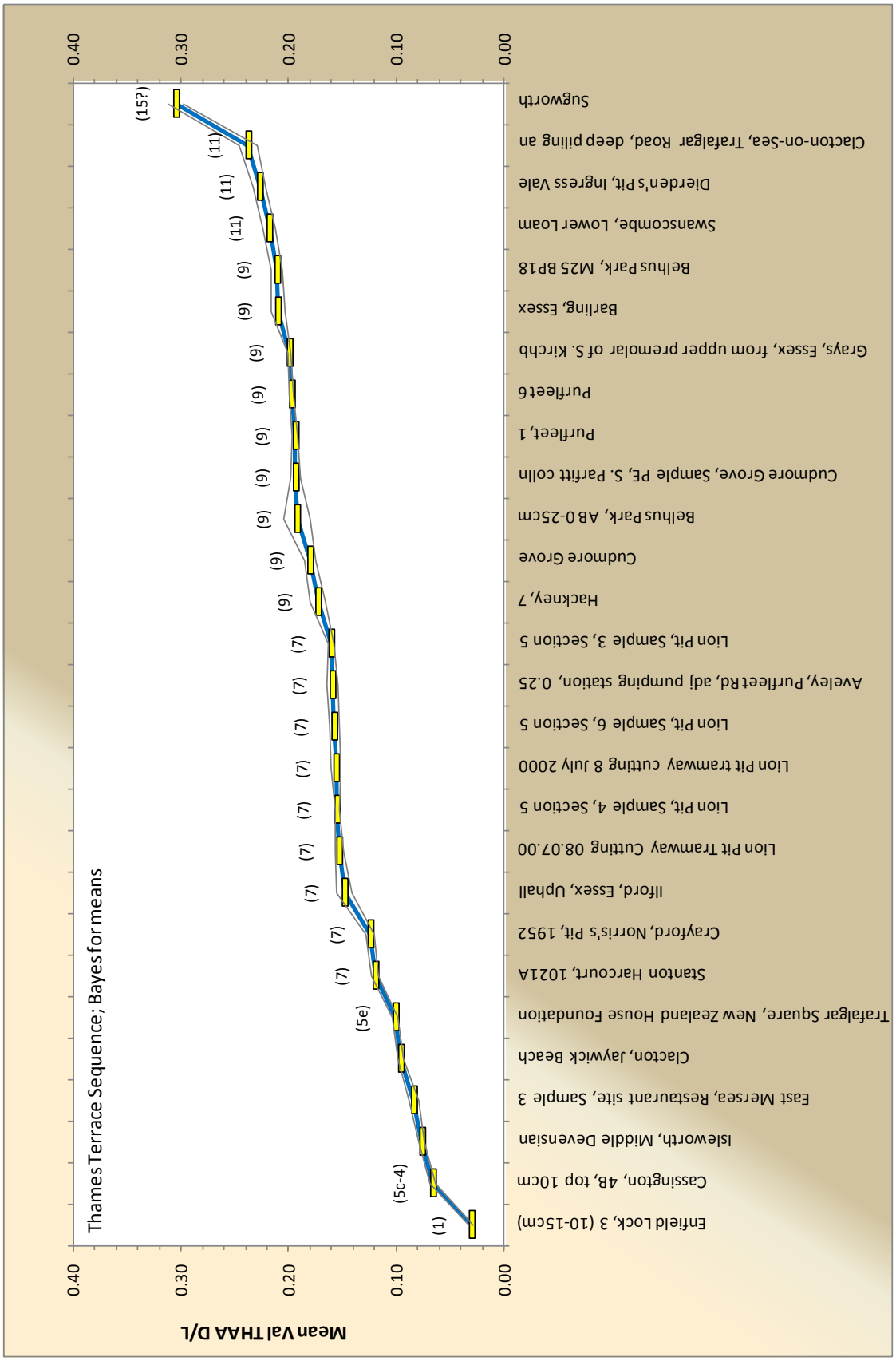


Figure 6.44: Thames Terrace Sequence, expanded scale. Bayesian derived posterior Val THAA D/L values with expanded uncertainty using k=1.96 (95% probability). Error bars = 1.96 x $s_{posterior}$ adjusted for means. MIS given in brackets.

Table 6.5: Comparison of D/L values and std dev s and RSDs for ANOVA and using Bayes

Thames Terrace				Classical method						Bayesian method					
Location reference		reps	samples	ANOVA: for single values			ANOVA: adjusted for means			Posterior: for single values			Posterior: adjusted for means		
	MIS	n	p	Val H D/L	S_R	RSD_R%	Val H D/L	S_R	RSD_R%	Val H D/L	S post.	RSD%	Val H D/L	S post.	RSD%
Enfield Lock, 3 (10-15cm)	1	1.5	5	0.025	0.0048	18.88	0.025	0.0036	14.08	0.029	0.0026	9.05	0.030	0.0010	3.36
Cassington, 4B, top 10cm	5c-4	2.0	4	0.067	0.0047	7.10	0.067	0.0028	4.16	0.066	0.0029	4.44	0.066	0.0012	1.83
Isleworth, Middle Devensian	-	2.0	4	0.074	0.0053	7.19	0.074	0.0033	4.42	0.075	0.0032	4.29	0.076	0.0013	1.74
East Mersea, Restaurant site, Sample 3	-	2.0	2	0.076	0.0116	15.17	0.076	0.0082	10.72	0.082	0.0039	4.77	0.084	0.0021	2.49
Clacton, Jaywick Beach	-	2.0	4	0.087	0.0062	7.09	0.087	0.0033	3.78	0.094	0.0038	4.05	0.096	0.0015	1.55
Trafalgar Square, New Zealand Hse Foundation 1957	5e	2.0	6	0.103	0.0087	8.48	0.103	0.0050	4.83	0.101	0.0043	4.25	0.101	0.0014	1.36
Stanton Harcourt, 1021A	7	2.0	4	0.113	0.0035	3.08	0.113	0.0022	1.95	0.116	0.0030	2.61	0.119	0.0015	1.30
Crayford, Norris's Pit, 1952	7	2.0	4	0.115	0.0058	5.06	0.115	0.0031	2.69	0.121	0.0043	3.53	0.124	0.0018	1.46
Ilford, Essex, Uphall	7	2.0	2	0.140	0.0219	15.62	0.140	0.0155	11.05	0.148	0.0069	4.68	0.148	0.0035	2.40
Lion Pit Tramway Cutting 08.07.00	7	2.0	2	0.154	0.0031	2.02	0.154	0.0022	1.43	0.153	0.0029	1.87	0.153	0.0019	1.24
Lion Pit, Sample 4, Section 5	7	1.9	25	0.148	0.0116	7.81	0.148	0.0045	3.03	0.153	0.0064	4.15	0.155	0.0011	0.69
Lion Pit tramway cutting 8 July 2000	7	2.0	1	0.159	0.0040	2.52	0.159	0.0028	1.78	0.156	0.0035	2.26	0.156	0.0025	1.60
Lion Pit, Sample 6, Section 5	7	2.0	5	0.156	0.0101	6.52	0.156	0.0050	3.19	0.157	0.0062	3.92	0.158	0.0022	1.40
Aveley, Purfleet Rd, adj pumping station, 0.25	7	2.0	4	0.161	0.0112	6.96	0.161	0.0057	3.51	0.160	0.0064	4.01	0.159	0.0025	1.56
Lion Pit, Sample 3, Section 5	7	2.0	25	0.164	0.0151	9.23	0.164	0.0075	4.57	0.161	0.0070	4.33	0.160	0.0011	0.68
Hackney, 7	9	2.0	2	0.174	0.0131	7.53	0.174	0.0093	5.32	0.173	0.0071	4.11	0.173	0.0038	2.23
Cudmore Grove	9	2.0	4	0.186	0.0106	5.68	0.186	0.0053	2.86	0.181	0.0067	3.71	0.180	0.0027	1.48
Belhus Park, AB 0-25cm	9	2.0	1	0.198	0.0185	9.35	0.198	0.0131	6.61	0.187	0.0085	4.53	0.192	0.0062	3.24
Cudmore Grove, Sample PE, S. Parfitt colln	9	2.0	2	0.195	0.0033	1.69	0.195	0.0023	1.20	0.194	0.0031	1.60	0.193	0.0021	1.08
Purfleet, 1	9	2.0	25	0.177	0.0148	8.38	0.177	0.0038	2.17	0.190	0.0081	4.24	0.194	0.0013	0.66
Purfleet 6	9	2.6	25	0.211	0.0094	4.44	0.211	0.0050	2.35	0.204	0.0067	3.29	0.197	0.0012	0.60
Grays, Essex, from upper premolar of S. Kirchb	9	2.0	2	0.199	0.0008	0.40	0.199	0.0006	0.28	0.199	0.0008	0.40	0.199	0.0006	0.28
Barling, Essex	9	2.0	4	0.201	0.0127	6.32	0.201	0.0085	4.23	0.208	0.0081	3.88	0.210	0.0034	1.61
Belhus Park, M25 BP18	9	2.0	4	0.214	0.0072	3.36	0.214	0.0045	2.10	0.212	0.0059	2.78	0.211	0.0028	1.32
Swanscombe, Lower Loam	11	1.8	6	0.236	0.0298	12.62	0.236	0.0127	5.37	0.215	0.0098	4.59	0.218	0.0031	1.43
Dierden's Pit, Ingress Vale	11	2.0	6	0.214	0.0112	5.22	0.214	0.0077	3.59	0.221	0.0079	3.58	0.227	0.0030	1.32
Clacton-on-Sea, Trafalgar Road, deep piling an	11	2.0	4	0.250	0.0215	8.61	0.250	0.0111	4.44	0.238	0.0102	4.27	0.238	0.0040	1.68
Sugworth	15?	1.8	6	0.285	0.0149	5.22	0.285	0.0061	2.14	0.298	0.0107	3.59	0.305	0.0037	1.20

If, p=1, then SR=Sr and RSDR%=RSDr%

Figure 6.41 shows a comparison of the mean Val THAA D/L values by each method. ANOVA uncertainty estimates for single values and for means share the same D/L value, however, the Val D/L values alter slightly using a weighted Bayesian approach.

ANOVA data has not been screened for outliers and represents the results for the raw data. Figure 6.42 is therefore comparable to Figure 6.7, data given earlier. However, two charts are given for the posterior D/Ls and confidence intervals; for the distribution of single values and Figure 6.43 for data adjusted for means. Figure 6.43 and Figure 6.44 show the same locations but uncertainty estimates have been determined using Bayes, for single values and then adjusted for means, respectively.

From Figure 6.42, and Figure 6.43, the effect of applying a Bayesian technique to assessing uncertainty data becomes clear. Because this approach is based on probability densities, using evidence derived from several sources, the confidence in the position of the mean value increases with the effect of reducing the posterior uncertainty compared to the classical ANOVA approach. The uncertainty estimates are reduced still further when they are adjusted to take into account sample numbers, n, Figure 6.44. However, it is noted that comparison of data in Table 6.5, shows that the RSD% values derived from ANOVA (for means) are only marginally different (higher or lower) from the Bayesian (single value) data, possibly because the number of data points used in the calculation of each were similar. Thus, in this context, the posterior probability density by Bayes provides a close approximation to the Classical uncertainty of the mean. For example, Table 6.5 site ref; "Cassington, 4B, top 10 cm", gives RSD% values, by ANOVA (for single values) = 7.10%; ANOVA (for means) = 4.16% and Bayes (for single values) = 4.44%. However, if the calculation is adjusted to account for sample number in the Bayesian estimate, this results in significantly reduced RSD% values, of 1.83% for this site.

From the ordering of D/L values, results of the ANOVA (Figure 6.44) indicate that Belhus Park: M25 BP18 (MIS 9) may in fact be MIS 11, whereas the stage 11 site Dierden's Pit: Ingress Vale, may in fact be MIS 9. Reordering of the posterior D/L values (Figure 6.45 & Figure 6.46) moves Belhus Park back next to the other MIS 9 sites, with Dierden's Pit aligning with the other MIS 11 data, confirming the stratigraphic positions based from independent dates (Penkman *et al.*, 2011).

The opercula, on which these analyses have been based, are derived from a freshwater, terrestrial gastropod *Bithynia*. Because *Bithynia* is rarely found during cold stages (Penkman *et al.*, 2011), its occurrence is a good indicator that it grew and died during

temperate climate conditions. As racemisation is a temperature dependent reaction, rates of racemisation in fossil opercula will increase during the warm interglacials (odd numbered MIS numbers) and slow down during the colder glacials. Therefore, the greatest differences in D/L value would be expected to occur between opercula found at the beginning and end of each warm stage, or perhaps the middle of one and the middle of the next interglacial. D/L values occurring part way through a warm marine isotope stage (MIS), may indicate the occurrence of a substage, a smaller oscillation in temperature. Plateaus seen in Figure 6.44, are not time related but simply reflect the occurrence of more than one site occurring at the same point in time (assuming a shared temperature history). This may be purely incidental based on biased sampling or reflect a genuine relative abundance of *Bithynia*, suggesting a warmer more stable phase in the palaeoclimate record. From Figure 6.44, a number of different levels can be seen within the MIS levels represented. The greater the number of results there are at each level, the more likely it is to be a genuine substage. For example, two separate levels can be seen in MIS 7, two or three in MIS 9, MIS 11 and MIS 5 are also suggested. However this is a limited data set and on its own does not provide sufficient or conclusive evidence of the occurrence of individual substages.

A spreadsheet has been developed as a tool for the calculation of valine D/L uncertainty estimates based on the amino acid associations presented in this chapter and can be found in Chpt 6: Appendix 4.

6.5 Conclusion

This chapter has looked at the importance of including sampling contributions in the overall uncertainty estimate for analysed samples in an archaeological / geological context. Suggested sampling schemes have been presented as given by the guidance documents (Grøn *et al.*, 2006; EURACHEM/EUROLAB/CITAC/Nordtest/AMC, 2007) but also from a more practical application with regard to samples collected for AAR analysis. Careful design of the sampling and analysis strategy can enable sources of uncertainty to be separated by ANOVA and overall precision estimates have been derived for existing quaternary site AAR data. However, such precision estimates are sensitive to sample number and a scheme for modelling uncertainty for valine THAA D/L data has been developed based on associations with other amino acids. Adopting a simple Bayesian approach, the associated data has been taken as prior information and combined with the observed ANOVA likelihood, to give much reduced uncertainty estimates. However, it should be noted that the resulting uncertainty

estimates are likely to be smaller than the true value due to assumed independence between the amino acids and unaccounted for covariance interactions which are likely to occur between all the various amino acids present in the intra-crystalline peptide chain, not accounted for here.

The Bayesian approach is a very different paradigm compared to the classical approach, which simply reports the observation. The concept of increasing the confidence in a measurement result with increasing evidence is understandably appealing, particularly to the archaeological community, where Monte Carlo simulation enables posterior distributions to be inferred from limited input quantities.

The principle is simple enough, for example, if it has been raining for five consecutive days, Bayesian statistics would infer that the chance that it rains on the sixth day, is highly likely. However, it could also be argued that living in the UK, the chances are that after five days of rain, the weather system bringing the rain will have passed, and in which case the chances of it raining again on the sixth are pretty slim. But this presupposes prior knowledge about UK weather patterns. The application of prior knowledge can therefore be very subjective and likely to result in different posterior results depending on who applies it. In this chapter, the modelled data has been taken as the prior information. If a different person had assessed the raw data, then they may have applied different curve fits and assigned different uncertainty estimates, but it's unlikely in this context that the results would change too much. Whether this is an appropriate perspective to take with archaeological / geological material is unclear. From the classical perspective, it makes intuitive sense that an uncertainty statement pertaining to a set of analytical results should be set so as to reflect the full extent of observed variability between samples. Thus we can be sure that any further samples would fit within the expanded confidence limits. To suggest that by taking a Bayesian approach the expanded uncertainty of the distribution might reduce, feels counterintuitive, unless it is argued that the improved confidence relates solely to the positioning of the mean, in which case this effect is also observed in classical statistics with the uncertainty being proportional to $1/\sqrt{n}$. Results from Table 6.5 would appear to agree with this distinction since RSD%s by ANOVA (for means) ≈ Bayes (for single values). If so, then the Bayes calculation for means may not a realistic estimation, in spite of being reasonably justifiable. Accessible Information regarding this distinction (at least in layman terms) appears limited. Clearly there are unresolved practical and philosophical implications that arise when comparing results by the two approaches which Buck *et al.* (1996) suggest are still being debated, and for now, shall be left to those with greater expertise.

Nonetheless, if the Bayesian approach for means is valid, this approach, using the protein decomposition of multiple amino acids and multiple samples to predict, in this case, valine THAA D/L values and uncertainty estimates, could potentially be a very powerful technique, providing better resolution of data than could be achieved by simple observation of valine data alone. If this same technique was applied to deriving posterior data for other amino acids in addition to valine, then multiple chronologies could be achieved and compared. Valine is the slowest racemising of the amino acids and therefore can be used to derive D/L values on some of the oldest samples. However, whilst faster racemising amino acids may reach equilibrium sooner, they can provide better resolution of younger samples, for example, aspartic acid (Asx) could be used for the youngest samples and alanine (Ala) for mid range values.

In section 6.4.3, it was suggested that the largest differences in D/L value are likely to occur between samples at the start and end of an interglacial. Similarly, the smallest differences will occur between samples at the end of one interglacial and the start of the next, since these samples are separated by a cold glacial phase when racemisation has all but stopped. One of the major difficulties facing AAR geochronology is trying to resolve the timescale between samples that may have very similar D/L values but could be separated by thousands of years and could fall in one of two warm stages.

Using kinetic models for racemisation and palaeoclimate reconstructions, it is possible to predict the probability of the age of samples with associated uncertainty. Depending on the age and temperature history of samples, this may give rise to multimodal distributions similar in appearance to those more commonly encountered in radiocarbon dating, except D/L estimates are calibrated against time and temperature. The next chapter tentatively presents such a model with due regard to unaccounted for uncertainties in the kinetic and temperature data.

Chapter 7. An integrated model for Quantitative AAR

7.1 Introduction

In Chapter 1, the use of AAR data as a relative dating technique was considered. Stratigraphic units based on D/L values are known as aminozones (Nelson, 1978) and the sequencing and correlation of amino zones is referred to as aminostratigraphy (Miller and Hare, 1980). The benefit of aminostratigraphy is that it avoids the complexity of kinetic and/or temperature modelling that is necessary to achieve quantitative dates. The focus of this chapter is therefore on the determination of quantitative AAR dates. We consider how the rate of racemisation is dependent on both time and temperature, and how the use of calibration, using an independently dated material, can avoid problems of an unknown temperature history. However, age uncertainty estimates are often very large and there is a need to develop a calibration free dating method. Using a kinetic model, we demonstrate a possible technique using a palaeoclimate reconstruction, to derive a new calibration curve based on predicted rates of racemisation. Then using the integrated model of racemisation and uncertainty estimates for the Thames terrace derived in the last chapter, consider its suitability and future potential.

7.2 Quantitative AAR dating

There are two main approaches to determining a quantitative age; calibration or time / temperature modelling. The calibration approach is the one most frequently used as it does not require detailed knowledge of the temperature history of a sample. Material is independently dated and used as reference values for a calibration curve. D/L values of the unknown samples are then interpolated or perhaps extrapolated from the calibration curve (Hearty and Kaufman, 2009; Murray-Wallace *et al.*, 2010; Wehmiller *et al.*, 2010; Demarchi *et al.*, 2011). Calibration materials need to be the same as the samples in question, or linked to them by association, perhaps occurring in the same sediment layer, and known to share a common temperature history. Thus any difference in D/L values can be interpreted as

difference in age. The second approach requires knowledge of the D/L value, rate kinetics and sample's temperature history. However Kaufman and Miller (1992) observe that most often a combination of the two approaches are most commonly used. Thus a kinetic model is used to extrapolate beyond the D/L range provided by the calibration curve, or used to adjust a calibration curve generated for one thermal regime to fit another (Wehmiller *et al.*, 2010; Wehmiller *et al.*, 2012). Furthermore, if the D/L value and the age of the sample is known, the average depositional temperature otherwise known as the effective diagenetic temperature (T_{eff}), can be derived and used to reconstruct palaeotemperatures (Kaufman, 2003; Owen *et al.*, 2007; Bright *et al.*, 2010).

However, reported uncertainties accompanying age estimates are often substantial due to issues of calibration and incorrect assumptions made regarding shared thermal histories. The use of A/I values to derive numerical dates was reported as " *no better than ±40-50% if the age equation is not calibrated locally and with a precision approaching 15% if appropriate calibration samples are available locally*." (McCoy 1987, p43). Following Wehmiller's inter-laboratory study (1984), differences of up to 25% in age estimates with Pleistocene molluscs are suggested if derived using calibrated results from other laboratories and taking both analytical uncertainty and the effect of 1.5^{o}C effective temperature uncertainty into account, a 25% uncertainty is further suggested (Wehmiller and Miller, 2000). An estimate of 30% is proposed when derived using different amino acids(Kosnik and Kaufman, 2008; Kosnik *et al.*, 2008) and Kaufman (2006) considers the effect of uncertainty in effective temperature and kinetic model choices and notes their effect on age uncertainties which range between 15 – 27% depending on amino acid with similarly large age uncertainties, up to about 40%, reported for museum specimens no more than 100 years old (Huntley *et al.*, 2012). It is interesting to note that in the absence of further explanatory text, all these uncertainty estimates should probably be interpreted as being equivalent to 1 standard deviation and therefore need to be expanded (x 2) to give approximately 95% confidence intervals! Age depth modelling of fossils from the Great Barrier Reef, Australia (Kosnik *et al.*, 2013), provides more recent ages again but observe skewed distributions of samples. Consequently ages are reported with a skewed 95% confidence range. For example Bramble Reef sediments are reported as being age homogeneous with a median age of 373 yrs but a 95% confidence range of 13-3491 yrs and Rib reef sediments have a median age of 326 years with a 95% confidence range from 4-2750 years (Ibid).

It should also be borne in mind that uncertainty estimates associated with the calibrated date itself should also be included in the overall estimate of uncertainty (Blaauw, 2012).

Therefore, before we look at an alternative model for quantitative AAR dating, it is necessary to consider the dependency of racemisation on time and temperature, to understand where some of the difficulties lie.

7.3 Time and Temperature dependency

Biomolecules (eg, DNA, collagen, intracrystalline amino acids), preserved over geological time, are unique in that they undergo time and temperature dependent degradation reactions. Consequently, they can act as miniature molecular time clocks, which, if correctly decoded can provide a direct measure of elapsed time since death. Therein lies the challenge and has been the subject of much research over the last forty years (Wehmiller and Miller, 2000; Miller and Clarke, 2007).

The temperature dependency of AAR and other diagenetic reactions (hydrolysis, condensation, deamination, decarboxylation) is critical and can be demonstrated by carrying out isothermal heating experiments. Linear, reversible first order kinetics (RFOK) were originally found to describe the behaviour of free amino acids heated at a constant high temperature in buffered solution (Bada, 1970; Kriausakul and Mitterer, 1978). Apparent RFOK was observed subsequently in experiments using fossil material heated over extended periods of time (usually days or weeks) (Wehmiller and Hare, 1971; Goodfriend and Meyer, 1991; Kaufman, 2000; Miller, 2000), with linearity being observed for A/I under environmental conditions up to a value of about 0.3 in foraminifera from marine sediments (Wehmiller and Hare 1971), slightly higher in mollusc shells up to about 0.55 (Mitterer and Kriausakul, 1989; Haugen and Sejrup, 1992), and up to 1.2 in ostrich eggshell (Miller *et al.*, 1992). The use of artificial heating experiments to mimic naturally occurring racemisation over geological time is often criticised as being inappropriate. Therefore other empirical approaches, (modelling fossil D/L values against independently dated samples using historical temperature records), to acquire an effective temperature have also been used (Miller, 1985; Hearty *et al.*, 1986; Wehmiller *et al.*, 1995; Ortiz *et al.*, 2004). The use of a hybrid approach therefore ensures that isothermally deduced reaction rates are constrained by those at ambient temperature, thus ensuring that subsequent temperature sensitivity modelling is also applicable to geological conditions (Miller, 1985; McCoy, 1987; Brooks *et al.*, 1990).

The **integrated rate equation** describes the relationship between a reversible reaction and time. For AAR, reversible first order kinetics (RFOK) can b expressed as;

$$L - isomer \rightleftharpoons D - isomer \quad \textbf{(7.1)}$$

Where k_1 is the forward rate constant an k_2 the reverse rate constant

The change in the abundance of the L-isomer with respect to time (*t*) is a function of the forward and reverse rate constants, thus;

$$-\frac{\partial L}{\partial t} = k_1 L - k_2 D \quad \textbf{(7.2)}$$

By integrating the above equation, k_1 can be related to *t* ; to give the integrated rate equation (Bada and Schroeder, 1972).

$$(1 + K) \cdot k_1 t + C = \ln\left(\frac{(1 + D/L)}{(1 - K \cdot D/L)}\right) \text{ where,} \quad \textbf{(7.3)}$$

$$C = \ln\left(\frac{(1 + D_0/L_0)}{(1 - D_0/L_0)}\right) at\ t = 0 \quad \textbf{(7.4)}$$

C represents the amount of laboratory induced racemisation measured in modern samples, where; D_0 and L_0 are values of samples at t=0.

For most amino acids $k_1 = k_2$, therefore for D/L values the k_1/k_2 ratio or $K = 1$. For A/I values k_1/k_2 = 1.0/1.3, thus K = 0.77 (Clarke and Murray-Wallace, 2006).

This the full expression of the equation becomes;

$$2 \cdot k_1 t = \ln\left(\frac{(1 + D/L)}{(1 - K \cdot D/L)}\right) - \ln\left(\frac{(1 + D_0/L_0)}{(1 - K \cdot D_0/L_0)}\right) \quad \textbf{(7.5)}$$

The gradient of $\ln(((1 + D/L))/((1 - D/L)))$ plotted against time provides the rate constants ($2k_1$).

For a very young sample (A/I<0.1), the reverse reaction can be ignored (due to the lack of D-isomer) and the integrated rate equation simplified (Bada *et al.*, 1970, Huntley *et al.*, 2012) to;

$$\ln(1 + D/L) = k_1 t + C \cong D/L \quad \textbf{(7.6)}$$

The measured rate constants can then be related to the temperatures of heating using the Arrhenius equation which describes the dependency or sensitivity of the rate to temperature.

The **Arrhenius equation** is given by;

$$k_1 = Ae^{(-Ea/RT)} \quad or \tag{7.7}$$

$$\ln(k_1) = \ln(A) - Ea/RT \tag{7.8}$$

Where A = pre-exponential constant (s^{-1}), Ea = the activation energy ($Jmol^{-1}$), R = the gas constant ($Jmol^{-1}K^{-1}$), T = integrated thermal history or effective diagenetic temperature (Kelvin)

Using several values of k_1 derived from different heating regimes, the natural log ($\ln(k_1)$) is then plotted against the temperature (Kelvin). The gradient of this line now gives our activation energy for racemisation and the y-axis intercept gives the pre-exponential constant (frequency or Arrhenius factor).

Substituting Equation (7.7) into (7.5), now provides a single expression linking time, temperature and D/L.

$$2 \cdot k_1 t = 2 \cdot Ae^{(-Ea/RT)} \cdot t = \ln\left(\frac{(1 + D/L)}{(1 - K \cdot D/L)}\right) - \ln\left(\frac{(1 + D_0/L_0)}{(1 - K \cdot D_0/L_0)}\right) \tag{7.9}$$

Thus it can be seen, that whilst isothermal experiments provide us with values of A and Ea, together with rate constants for defined temperatures, in the absence of knowledge about an unknown sample's temperature history or age, there remain two unknown factors, t and T.

If however, an independently dated sample can be correlated to a D/L value with an equivalent temperature history, a rate constant can be derived, allowing for the problems of integrating an unknown thermal record, to be circumvented. Using kinetic models, this can then be extrapolated to other samples within the same locality. However, the lack of knowledge about the temperature sensitivity prevents the calibration from being applied to samples with a different temperature history and places a strong reliance on the need for calibration.

7.3.1 Uncertainty

It is often assumed that samples sharing common temperatures from the same region today, also share the same effective temperatures (Wehmiller and Miller, 2000), which may not be the case. Miller and Clarke (2007) show that the rate of reaction increases exponentially with temperature and that the reaction rate approximately doubles with every 4°C increase at ambient temperature. Therefore, a fossil that has spent half its time at 20°C

and half at 0°C, has an effective diagenetic temperature of 16.7°C and not 10°C (Ibid). Thus the effect of even a slight error in temperature history, can have a significant impact on the rate and apparent age of the sample.

A further source of potential uncertainty lies in the kinetic model chosen. Whilst the original rationale to use the RFOK model was derived from early experiments of free amino acids in solution (Bada, 1970), kinetic experiments on fossil material identified discrepancies between dates derived kinetically compared to those derived from independent dates or stratigraphic information (Wehmiller and Hare, 1971; Wehmiller, 1981)

As a result, considerable effort has been spent in deriving modifications such as Wehmiller's (1981) model utilising a number of rate constants that changed with time, or alternative rate equations that are independent of kinetic theory. A review of mathematical expressions used in AAR is proved by (Clarke and Murray-Wallace, 2006). In brief, other models suggested include a logarithmic equation (Wehmiller 1988), parabola curve fitting (Hearty *et al.*, 1986; Mitterer and Kriausakul, 1989), simple linear equation, stepped linear model (Miller *et al.*, 1999) and power transformations (Goodfriend *et al.*, 1995; Kaufman, 2000; Manley and Miller, 2000; Kaufman, 2006). More recently a model-free approach has been developed free from kinetic theory (Demarchi *et al.*, 2013, In Press; Tomiak *et al.*, 2013, In Press). This approach is based on numerical optimisation and the determination of a scaling factor for time.

Regardless of the model chosen, there remains strong dependence on the use of calibrated samples. One way of overcoming this reliance was suggested by Miller *et al.*(1992), using the principle of protein diagenesis dating proposed by Hare (1969). Using knowledge of the differences in temperature sensitivities for hydrolysis and racemisation, values of t and T could be simultaneously calculated. However, with a reported uncertainty of ±10°C and with the implications for extrapolated age uncertainty estimates, this suggests that its application may be limited.

With this in mind, an alternative basic model has been developed for "calibration-free", quantitative AAR dating.

7.4 Its AAR dating – but not as we know it.

7.4.1 Calibration curve

The initial stage was to derive a calibration curve for predicted racemisation using a kinetic model and palaeoclimate reconstruction. From the previous discussion, it will be evident that the appropriateness of this quantitative approach is entirely dependent on the kinetic model and palaeoclimate record chosen. Having developed a quantitative AAR model prototype, the evaluation of the relevance of different models will therefore be a high priority for the future. However, we had to start somewhere!

The standard reversible first order kinetic (RFOK) model was used, against which other models can be later compared. However the use of RFOK is not entirely inappropriate. Valine is the slowest of all the amino acids to racemise (see Chapter 6). Some of the oldest samples in the archive have only achieved a valine D/L value of about 0.35. From the previous section, isoleucine D/Ls up to about 0.3 are considered to be consistent with RFOK. Therefore it is probably a fairly safe assumption, that the data considered for this model, based on an integrated valine D/L value (Chapter 6) will conform to a RFOK model.

The reconstructed palaeoclimate surface air temperature record used was provided by Richard Bintanja, Institute for Marine and Atmospheric Research Utrecht, Utrecht University, The Netherlands, who used "*...a coupled model of Northern Hemisphere ice sheets and ocean temperatures, forced to match an oxygen isotope record for the past million years compiled from 57 globally distributed sediment cores*..." (Bintanja *et al.*, 2005, p125). The temperature record used was provided as temperature differences from today, in 100 year intervals. The current mean annual temperature for the UK is in the range 8-11^{o}C, thus an initial modern day temperature approximation for the UK of 10^{o}C was assumed (taken from www.metoffice.gov.uk).

The model was developed using an adapted spreadsheet after Collins (unpublished work). The model utilises the concept of thermal age (Smith *et al.,* 2003). Thermal age is a concept of time equivalence. Using knowledge of the rate of a decay process for a biomolecule and its temperature history, the extent of decay (or preservation) can be determined. Thermal age is therefore the equivalent age of a biomolecule, based on the extent of undergone decay, assuming a fixed temperature. For comparative purposes it is convenient to use an arbitrary 10^{o}C (equivalent to current UK temperature average).

An activation energy (Ea) for valine THAA was derived using the model-free method (Demarche *et al*., 2013, In Press; Tomiak *et al*., 2013, In Press), to give a value of 119 kJ/mol (Penkman pers. comms.). The model-free approach does not provide a value for the pre-exponential constant (A), and so this was derived using a best-fit approach of the data to the temperature record. Thus using the Ea and A values and a temperature of 10°C (283.15K), a forward rate constant was determined ($k_{1@10}$).

This information was then used to determine the predicted rate of racemisation over the last million years (600ka shown in the illustrative figures). The process can be broken down into several stages;

- Step 1; Using the surface air temperature differences (Bintanja *et al*., 2005) a palaeoclimate record (100 yr average) for the UK was constructed using a current average UK temperature of 10°C.
- Step 2; using the Arrhenius equation (equation (7.7), calculate individual forward rate constants ($k_{1(1,2,3...n)}$) for each 100 year increment using the new average temperatures from Stage 1, and an Ea =119 $kJmol^{-1}$ and $A=1.89x10^{16}$ s^{-1} (or ln(A) = 37.38).
- Step 3; Using the individual forward rate constants, the yield was determined, representing the anticipated change of the L-isomer → D-isomer. Assuming at t=0, the yield at the start (Y_0) = 1 or 100% (all L). The revised yield after the first 100 years (Y_1) would be = Y_0-($k_{1(1)}$ x t) where t=100. The next value would be determined as; $Y_2 = Y_1$-($k_{1(2)}$ x t) and so on.
- Step 4; For each 100 year interval, calculate the thermal age by comparing the total change in yield over time ($1-Y_{1,2,3...n}$) (where yield was previously derived in Stage 3 using the specific rate constants). Then determine how long it would take to arrive at this amount of change if held at a constant 10°C, by dividing by $k_{1@10}$. Therefore T. $Age_{(k1@10)} = (1-Y)/ k_{1@10}$
- Step 5; For each 100 year interval, using the rate constant for 10°C ($k_{1@10}$) and the thermal age in Step 4, determine $2k_1t$ from the integrated rate equation (7.5); $2k_1t = 2(k_{1@}10 \text{ x T. } Age_{(k1@10)})$

  The value 2kt represents twice the D/L difference, where the difference is that between the observed D/L for a fossil and the D/L of a modern sample having undergone minimal racemisation (i.e. t=0) but may include some laboratory induced racemisation during extraction and analysis.

Samples previously collected from Acle, Norfolk, UK (Penkman, 2005), known not to be older than 100yrs provided a nominal D/L value at t=0 of 0.0173. Predicted D/L values could then be derived; $D/L = 2kt/2 + D/L_{(t=0)}$. These predicted D/L values are plotted against geologic time in Figure 7.1 and show the constructed climate record using the method as described previously. Rate can be seen to progress faster during the warmest phases (shown by those sections with the steepest gradients), and much slower during the colder stages, appearing to almost plateau at times when the D/L value barely changes.

7.4.1.1 Fitting the calibration curve

In Figure 7.1b, data from the archive have been superimposed to ensure consistency with the predicted model. Using Excel's lookup function, observed Valine D/L values were used to lookup the equivalent predicted D/L value and obtain corresponding ages. A slight discrepancy can be seen between the predicted valine rates derived from the integrated rate equation and the observed D/L values. This is because the RFOK model provides a close approximation for younger fossils, but other factors come into play within the matrix as time progresses, slowing down the reaction. It is thought that diminishing water availability within the intra-crystalline matrix may act as a limiting factor, slowing down hydrolysis of the peptide chain and reducing the availability of terminal amino acids for racemisation (Collins and Riley, 2000).

Transformation of the observed D/L values using ln((1+D/L)/(1-D/L)) into a kinetic equivalent value corrects for this discrepancy in RFOK (Figure 7.1c). Excel spreadsheets used for Figures 7.1b and 7.1c are provided as Chpt 7: Appendix 1 and 2 respectively.

From Figure 7.1c, a fairly even spread of D/L values can be seen along the length of the calibration curve. Many of the D/L values relate to samples taken from archaeological sites and can be correlated with discrete marine isotope stages (Penkman *et al.*, 2011). Thus whilst it might be expected to see many sites of human occupation associated with the warm stages (odd numbered MIS shown), it is highly unlikely that D/L values associated with archaeological sites will be found occurring during the colder glacial stages (Stringer, 2011).

For developmental purposes, the archive data were first correlated to appropriate isotopic stages based on the aminostratigraphy given by Penkman *et al*. (2011). A number of temperature flexibility features were built into the original spreadsheet (Collins, unpublished work), such as the ability to increase or decrease the temperature amplitude, the start and stop temperatures etc.

Figure 7.1: Predicted D/L (Val THAA) against time (krs)

The following three charts show the relationship between calibration curve (indicating predicted racemisation rate for valine), given as the dark red line, and temperature.

Figure 7.1a: Predicted D/L derived from 2kt/2 + D/L(t=0). The predicted rate is much faster during warm phases, but slows down, almost plateauing during cold phases

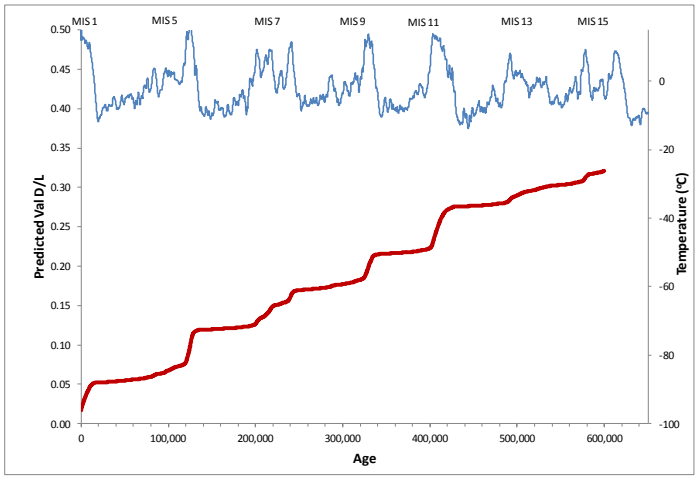


Figure 7.1b: Archive data have been plotted to determine the correlation between observed D/L value and predicted. The chart indicates that whilst the modelled rates agree well for younger samples, there is some disagreement in the RFOK model used towards the higher D/L values.

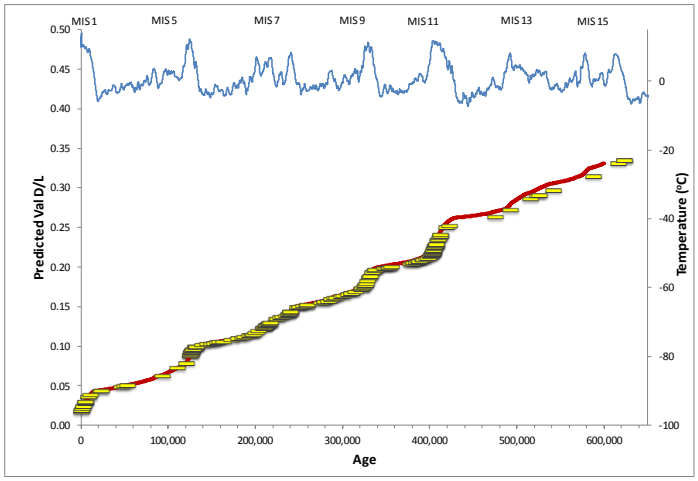


Figure 7.1c: Archive data have been adjusted using ln((1+D/L)/(1-D/L)), to derive equivalent kinetic D/L values and correct for mis-alignment due to deviation from RFOK

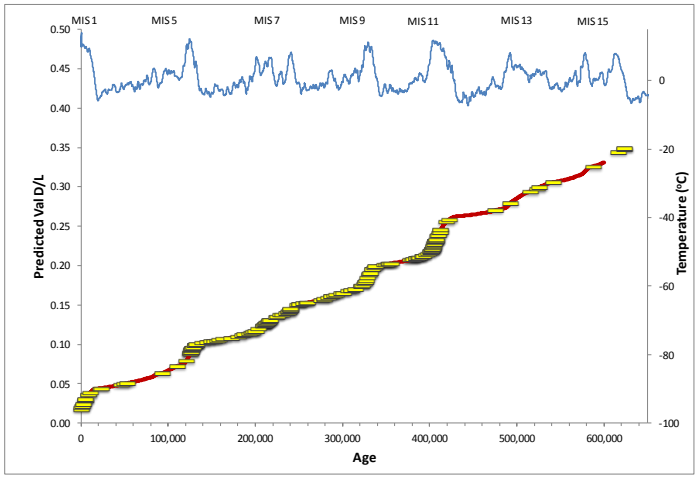


Table 7.1: Tie points used to fit calibration curve, showing independent dates and those derived using the AAR calibration curve.

Site name	MIS[1]	Integrated Valine D/L	Independent date (yrs)[2]	Predicted date (yrs)
Acle (modern) t=0	1	0.0173	<100	0
Cassington	5c-4	0.063	80,000	88,200
Trafalgar Square	5e	0.099	124,000	125,000
Strensham	7	0.125	2,000,000†	197,500
Swanscombe	11	0.220	471,000	400,900
West Runton	17/15?	0.335	500,000	648,900

[1] = Penkman *et al.*, 2011 [2] = Bowen 2000

† = reported as 2000 ka;in Bowen, 2000; p 255, Table 18-1. (Note, possibly a typographical error as 200 ka, has been predicted from the model)

Other features such as the ability to increase or decrease average temperatures across discrete isotope stages were also added to enhance the flexibility of the model. A set of discrete tie-points (see Table 7.1) were used to help fit the curve. These were derived from a revised aminostratigraphy (Bowen, 2000) and gave independently derived dates for some of the resampled Thames sites (Penkman *et al*., 2011).

Figures 7.2a-c, illustrate the curve fitting. a shows the predicted distribution using a starting temperature of 10°C, (equivalent to the UK's average modern day temperature). However, MIS 5e and 7 are not fast enough as they are taking too long to achieve the measured amount of racemisation (D/L value), compared to the temperature record. After some adjustment, some excellent separation was obtained by increasing the starting temperature by 2°C from 12 to 10 °C and more in line with Bintanja's reported present day surface air temperature differences and increasing the amplitude (from 1 to 1.5) (Figure 7.2b). Further separation between MIS 5 and 7 was achieved by adjusting the individual average temperatures of each stage, lowering the cold stage and raising slightly the temperature of the warm stages (Figure 7.2c), data provided in Chpt 7: Appendix 2.

Whilst the values selected were chosen based on goodness of fit and correlation with tie-points, they are nonetheless arbitrary but used to demonstrate the potential for the model. Further, more informed modelling will be required in the future based on independent dates and stratigraphies. For example, it has been established that the original AAR date for West Runton of 500,000 years (Bowen, 2000) is probably too low and the actual date is much older based on other stratigraphic information (for example, Rink *et al.*, 1996; Gibbard *et al.*, 2010; Maul and Parfitt, 2010; Penkman *et al.*, 2010).

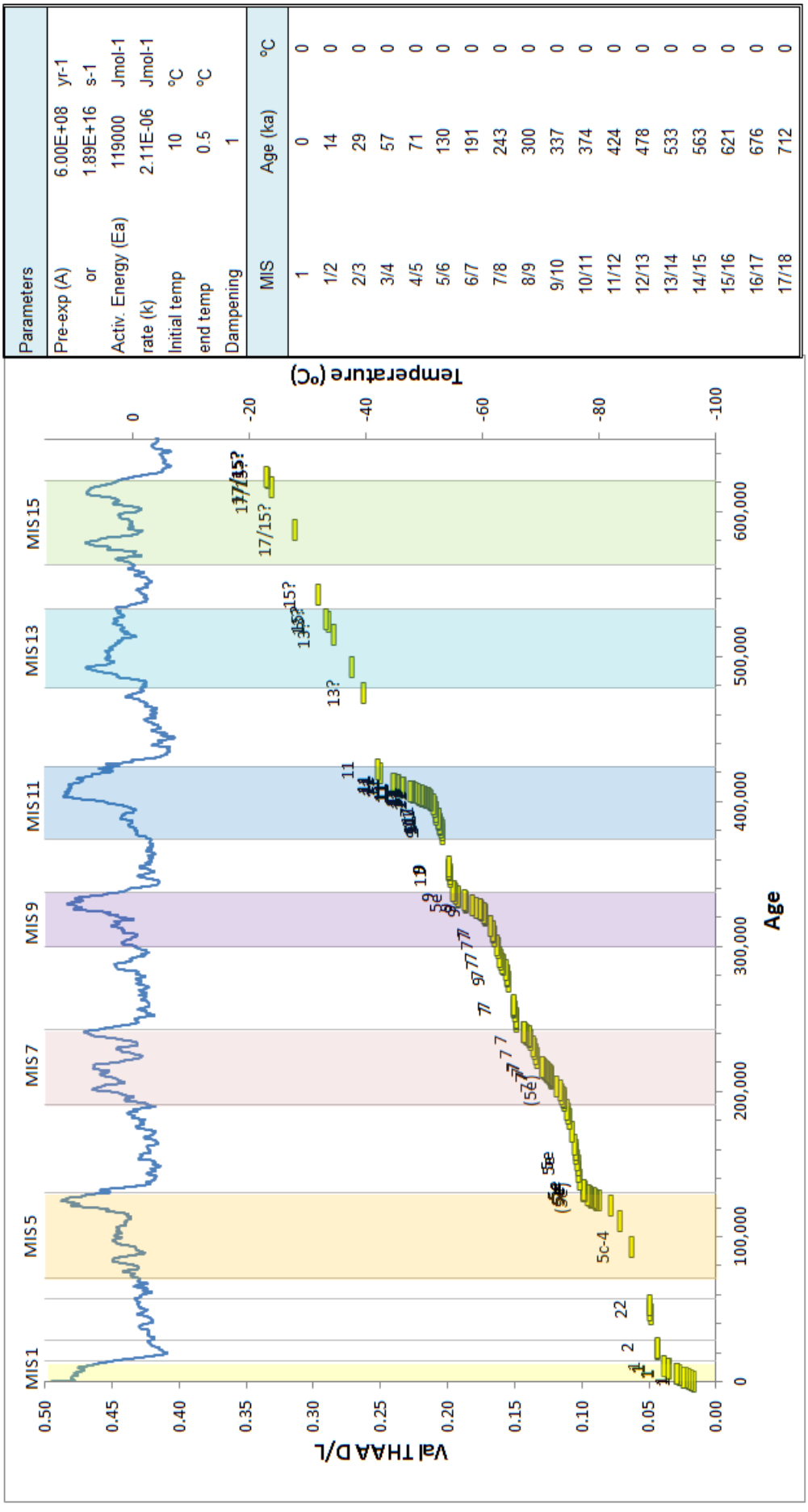


Figure 7.2: Improving the resolution of the calibration curve.

Figure 7.2a: Using default settings, data showed insufficient separation of sites compared to independent stratigraphic evidence

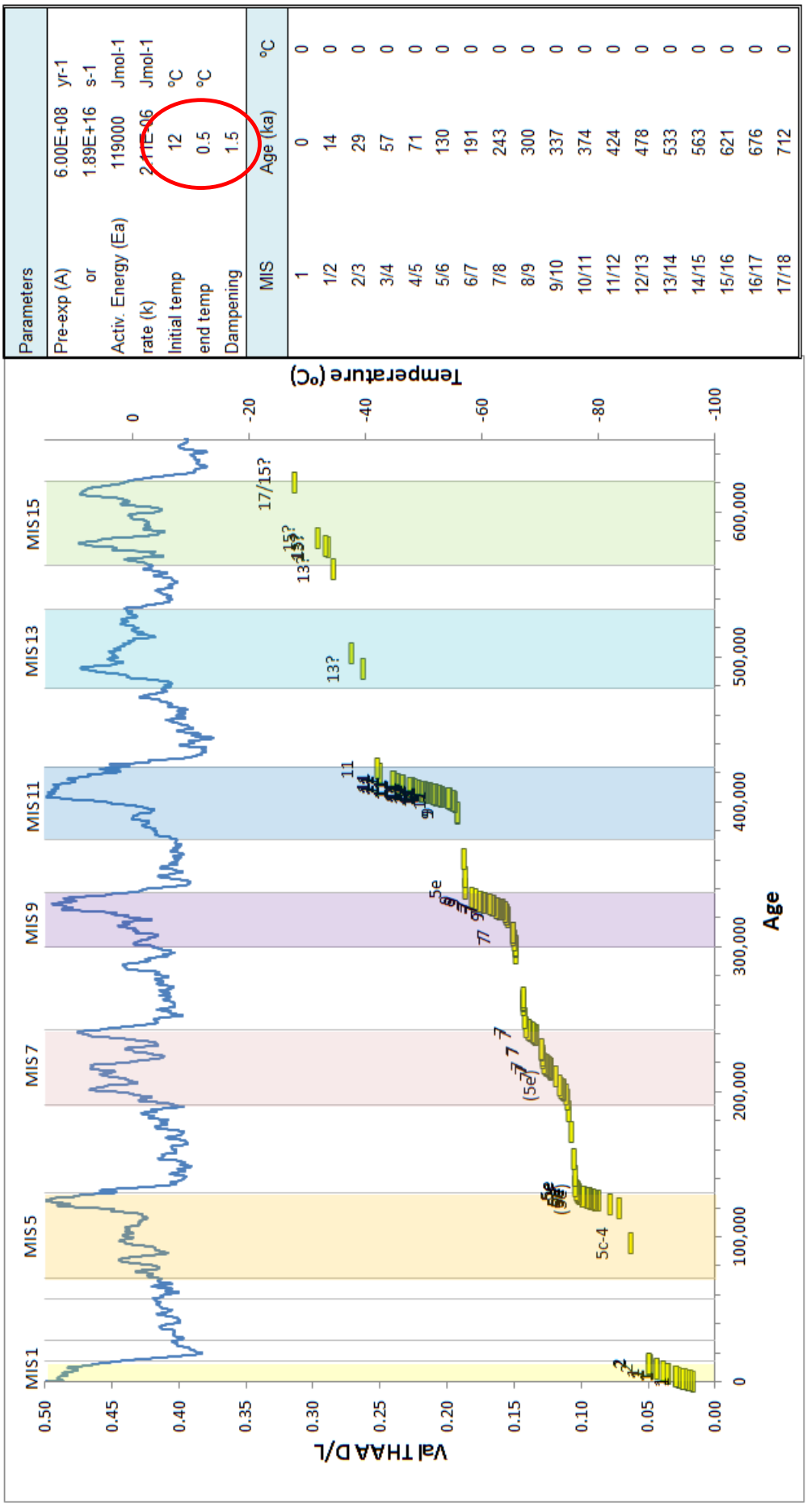


Figure 7.2b: The curve was adjusted by increasing the amplitude and start temperature

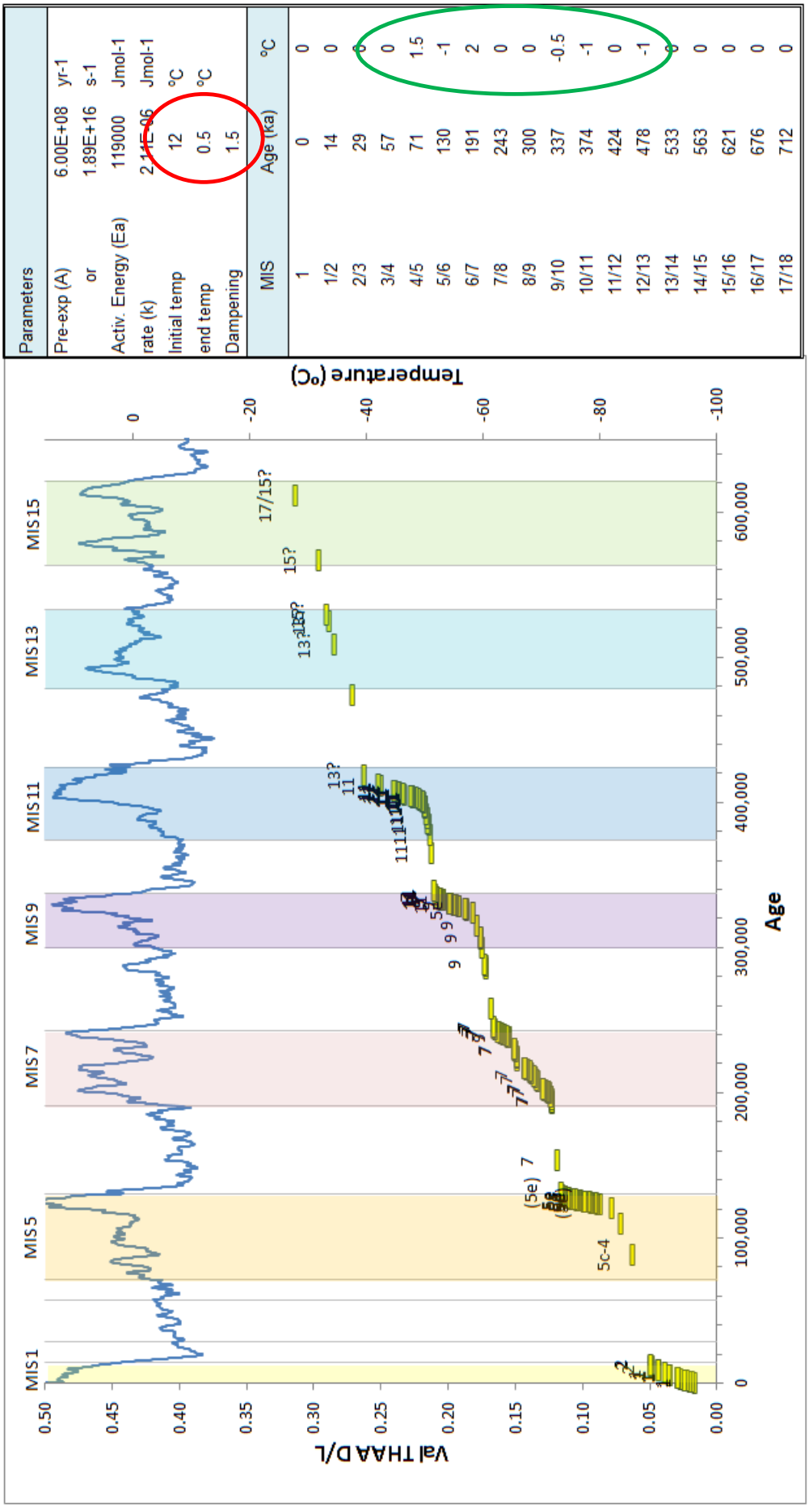


Figure 7.2c: Further separation was achieved between MIS 7 and 5 by adjusting the temperatures of individual stages

However it is interesting that by increasing the amplitude of the temperature range, data fell into a more expected distribution. Increasing the temperature highs and lows is therefore indicative of increased continentality than is perhaps otherwise indicated from the original temperature data set and warrants further investigation.

From Figure 7.2c site data appear nicely stacked against the appropriate warm stage MIS peaks but tail off towards the end of the stage (left-hand side), loosing resolution of sites at the end of one warm stage and the start of the next. During these cold glacial phases, mean air temperature drops and the rate of racemisation slows right down. It would be an interesting exercise to consider how a revised rate constant would compare based on an effective diagenetic temperature, determined across the sample's entire temperature history, rather than using discrete rate constants for fixed temperatures in each 100 year interval. Due to the exponential relationship between rate and temperature (see section 7.3.1), it is possible that the lower temperature record will have a minimal influence on the effective diagenetic temperature and overall rate constant, in effect speeding up the reaction rate to give a slightly higher D/L value which might provide the necessary resolution for these apparent cold stage sites in Figure 7.2c.

Furthermore, the initial temperature increase up to 12^{o}C on the model may also be explaned by the effective temperature. Temperature differences (Bintanja et al., 2005) are given as 100 year averages. When considering the determination of effective temperature, Smith et al. (2003, p214) use "*a simple sinusoidal model of soil temperature variation throughout the year, based upon ……[an] amplitude of 2.5^{o}C*." In this case, if the same approach was adopted and the daily/annual temperature fluctuations were taken into account rather than taking a 100 year average, it is possible that the effective temperature may be higher than the value used in this model and account for the need to increase temperature estimates.

7.4.2 Linking time, temperature and D/L values

From equation (7.9), expressions deriving estimates of time (age in yrs) or effective diagenetic temperature (T_{eff}), can be obtained. Thus a quantitative expression for age is given as (McCoy, 1987; Oches and McCoy, 2001);

$$t = \frac{ln[(1 + D/L)/(1 - K \cdot D/L)] - ln[(1 + D_0/L_0)/(1 - K \cdot D_0/L_O)]}{(1 + K)Ae^{-Ea/RT}} \quad \textbf{(7.10)}$$

Where definitions are as described previously, and T is the effective temperature (T_{eff}).

For Holocene samples, it is reasonably safe to assume the temperature history of fossils have been constant over the last 11.5 ka, and age estimates can be based on an assumed T_{eff} =10°C (283.15 K). However, for older Pleistocene samples, the integrated temperature history is more complex and less easy to determine. Therefore numerical age estimates for older samples are often less precise and open to larger uncertainties.

7.4.2.1 Palaeothermometry

However, if there is an independent date for the fossil, then an effective diagenetic temperature can be determined and used for palaeothermometry, comparing same age sites and deriving effective temperature differences (McCoy, 1987; Kaufman, 2003). Consequently, this broadens the range of potential applications of this new quantitative model once numerical ages can be determined. Thus;

$$T_{eff} = {-Ea}/{R} \cdot \left(\frac{ln([(1+D/L)/(1-K\cdot D/L)]-[(1+D_0/L_0)/(1-K\cdot D_0/L_O)])}{(1+K)At}\right) \quad \textbf{(7.11)}$$

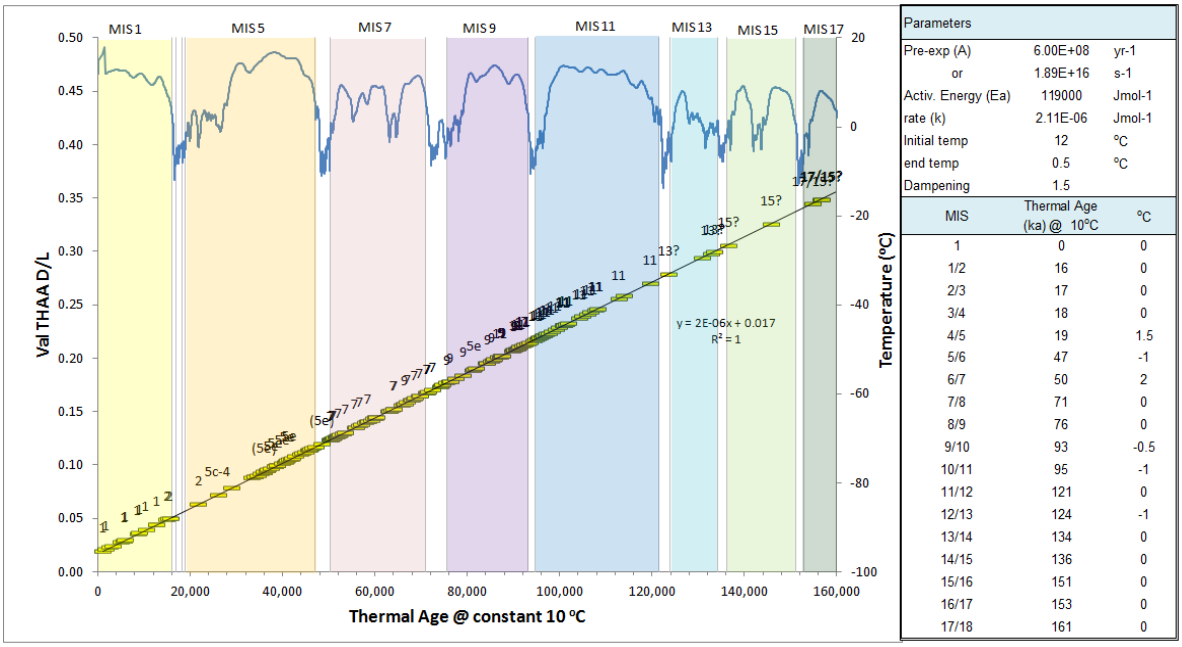
Clearly, the effectiveness of determining a quantitative age can be difficult in the absence of a palaeoclimate record. Therefore, to overcome this obstacle an alternative *Thermal age* approach has been proposed, which assumes a fixed diagenetic temperature of 10°C (Smith *et al.*, 2003).

7.4.2.2 Thermal Age

Thermal age is defined by Smith *et al* (2003, p204) as "...*the time taken to produce a given degree of DNA degradation when temperature is held at a constant 10°C*." Thus a fossil with a cold temperature history is more likely to have better biomolecule preservation than one with a temperate history, based on the temperature dependence of decay rate constants. The model developed, enables the average effective diagenetic temperatures to be determined, thus the extent of fossil preservation can be estimated from AAR D/L values. The potential recovery of amplifiable DNA has also been correlated to the extent of aspartic acid racemisation (Poinar *et al.*, 1996) and has important implications as a potential screening method for museums assessing requests for destructive sampling and analysis (see http://beta.thermal-age.eu/). Thermal-age.eu is a model developed out of the EU funded SYNTHESYS project by Matthew Collins and David Harker at BioArCh, as a tool to assist museum curator decision making.

The benefit of this quantitative AAR dating model is its ability to determine thermal age using the incremental changes in racemisation rate. Figure 7.3 (also given in Chpt 7: Appendix 2) shows the linearity between site D/L values plotted against their equivalent thermal age assuming a constant 10°C. This therefore becomes a far easier scale to interpret and use for comparisons when linked to DNA preservation and fragment length recovery (Poinar *et al.*, 1996; Deagle *et al.*, 2006).

Figure 7.3: Linear relationship between D/L value and thermal age (constant 10°C), showing significant age reductions based on biomolecule preservation.



Parameters		
Pre-exp (A)	6.00E+08	yr-1
or	1.89E+16	s-1
Activ. Energy (Ea)	119000	Jmol-1
rate (k)	2.11E-06	Jmol-1
Initial temp	12	°C
end temp	0.5	°C
Dampening	1.5	

MIS	Thermal Age (ka) @ 10°C	°C
1	0	0
1/2	16	0
2/3	17	0
3/4	18	0
4/5	19	1.5
5/6	47	-1
6/7	50	2
7/8	71	0
8/9	76	0
9/10	93	-0.5
10/11	95	-1
11/12	121	0
12/13	124	-1
13/14	134	0
14/15	136	0
15/16	151	0
16/17	153	0
17/18	161	0

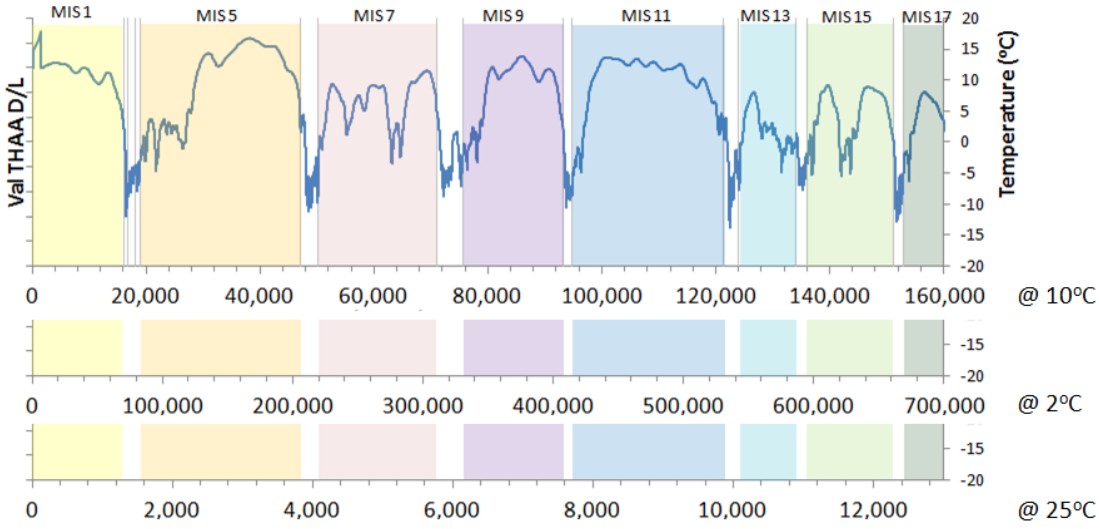
A thermal age can be calculated for any fixed temperature. For example, Figure 7.4 shows equivalent scales for thermal age based on different fixed temperatures, (2, 10 and 25°C). Because the rate of racemisation is much slower at 2°C than at 10°C, it takes much longer to arrive at the observed D/L value. Conversely, if held at a constant 25°C, the rate is much faster, so the fossil achieves its observed D/L value much faster.

7.4.2.3 Effect of temperature on geological age

It has been suggested that a 1°C increase in effective temperature would result in an age estimate 20% older (Miller *et al*., 2000). The effect of raising the effective temperature was assessed using the data from the tie-points shown previously in Table 7.1. For each tie-point (with the exception of Acle at t=0), the effective temperature was derived using equation (7.11).

Figure 7.4: Thermal Age Scales.

Shows the difference in time required to produce the same amount of racemisation when held at a constant 10, 2 or 25°C.



The ratio between $k_{(+1)}/k$ was used to multiply the original predicted age, to give the revised age estimate. In all cases, the observed increase in age represented an increase of approximately 20%, in agreement with Miller *et al*., (2000). This data is shown in Table 7.2 and calculated in Chpt 7: Appendix 1. Using each value of T_{eff}, rate constants (k) were calculated using the Arrhenius Equation (equation (7.8)). Revised effective temperatures were obtained by raising each by one degree Celsius (T_{eff} + 1), and revised faster rate constants determined ($k_{(+1)}$).

Table 7.2: Effect on age of raising the effective diagenetic temperature (T_{eff}), by 1°C

Site name	Predicted date (yrs)	T_{eff}	k	T_{eff} +1	$k_{(+1)}$	Revised date (yrs)	% difference
Acle (modern) t=0	0						
Cassington	88,200	2.8	5.65E-07	3.8	6.81E-07	106,366	20.6%
Trafalgar Square	125,000	4.3	7.42E-07	5.3	8.93E-07	150,450	20.4%
Strensham	197,500	3.4	6.32E-07	4.4	7.61E-07	237,987	20.5%
Swanscombe	400,900	3.6	6.60E-07	4.6	7.95E-07	482,931	20.5%
West Runton	648,900	4.3	7.50E-07	5.3	9.03E-07	780,955	20.4%

This has important implications for ensuring the burial environment is accurately represented and built into future models, although historical burial contexts may be difficult to come by. Deeper burial depths will act to buffer temperature fluctuations, whilst different soil matrices will have variable thermal diffusivity properties. These will need to be incorporated into uncertainty ranges where conditions are not known. Using the thermal-age.eu web tool, theoretical data were used to observe the effect of burial depth and soil type on effective temperatures. Data are purely illustrative as the kinetic parameters used in the thesis are based on estimates for valine and may not be comparable to values used for the web-tool. Nonetheless, some important differences can be seen (see Table 7.3), which suggests that the burial environment and possibly storage conditions may be equally as important to take account of as the palaeotemperature history.

Table 7.3: Effect of burial environment on effective temperature.

Site name	Burial depth (m)	T_{eff} [1]
Fresh sandy soil	0.01	9.4
Fresh sandy soil	0.5	8.6
Fresh sandy soil	1	8.0
Fresh sandy soil	2	6.9
Saturated sand	1	7.2
"Generic" rock	1	7.2
Sandy Clay, 10% moisture	1	4.1

[1] = all data derived using web tool thermal-age.eu

7.5 Uncertainty estimation

Having determined a calibration curve, it now becomes possible to use this to calibrate the D/L values and determine uncertainty ranges using the D/L uncertainty values derived using Bayes (for means) (see Table 6.5) in the previous Chapter.

The uncertainties derived, expressed as standard deviations, describe a normal probability density function. Thus, by dividing up the area beneath the normal curve into incremental bins, corresponding areas can be derived and plotted against calibration curve height. Three different approaches to this were attempted and results reviewed. Calculations carried out in both this and the following section and are given in Chpt 7: Appendix 2.

The first binned the area beneath the normal curve based on fixed values of z, thus each standard deviation was divided into perhaps 5 sections, expressed as D/L values, and corresponding areas determined. The equivalent age for each D/L value was then plotted and the height of the histogram bars adjusted based on required area. However because the areas varied between intervals, the heights of the resulting histogram were consequently affected not just by the steepness of the slope from the curve, but also the areas used, which could lead to problem with interpretation. For example a low histogram bar could be due to a slow rate of racemisation or because it occurred at the extreme of the D/L distribution.

The second version used a fixed area and calculated the intervals and D/L values from this. These were again plotted in a similar way as a histogram. This time the histogram bars areas now were evenly matched and the height determined only by the rate of racemisation and steepness of the calibration curve. However it was felt that the interpretation may be difficult with some very tight bars and others very wide. For Quaternary purposes regular time increments are more easily explained.

Therefore a third version was developed based on fixed time intervals and the corresponding D/L intervals and areas were determined from that in order to derive histogram bar heights. Examples of these three histograms are shown below.

Figure 7.5: Determining age probabilities calibrated against D/L uncertainty

Figure 7.5a: Calibration using fixed z values and variable area and age widths

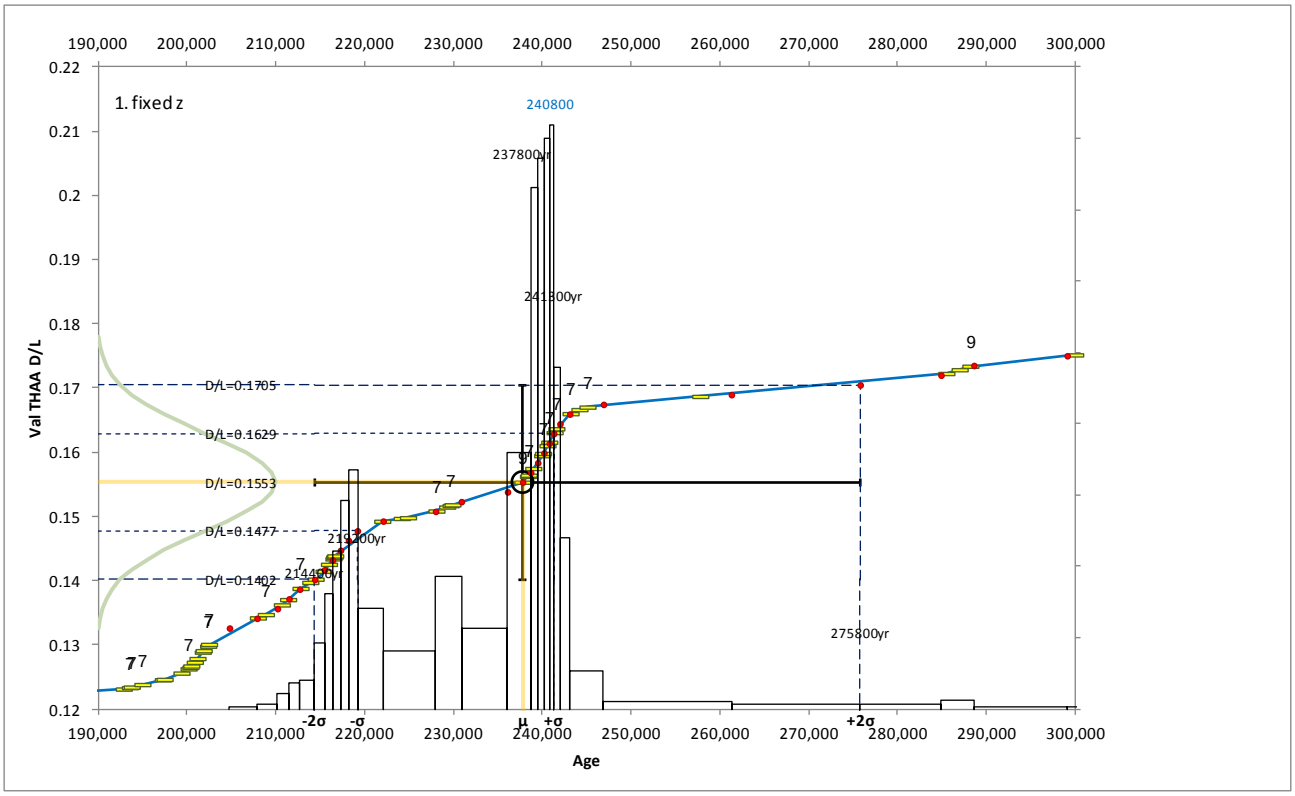


Figure 7.5b: Calibration using fixed areas with variable z values and age widths

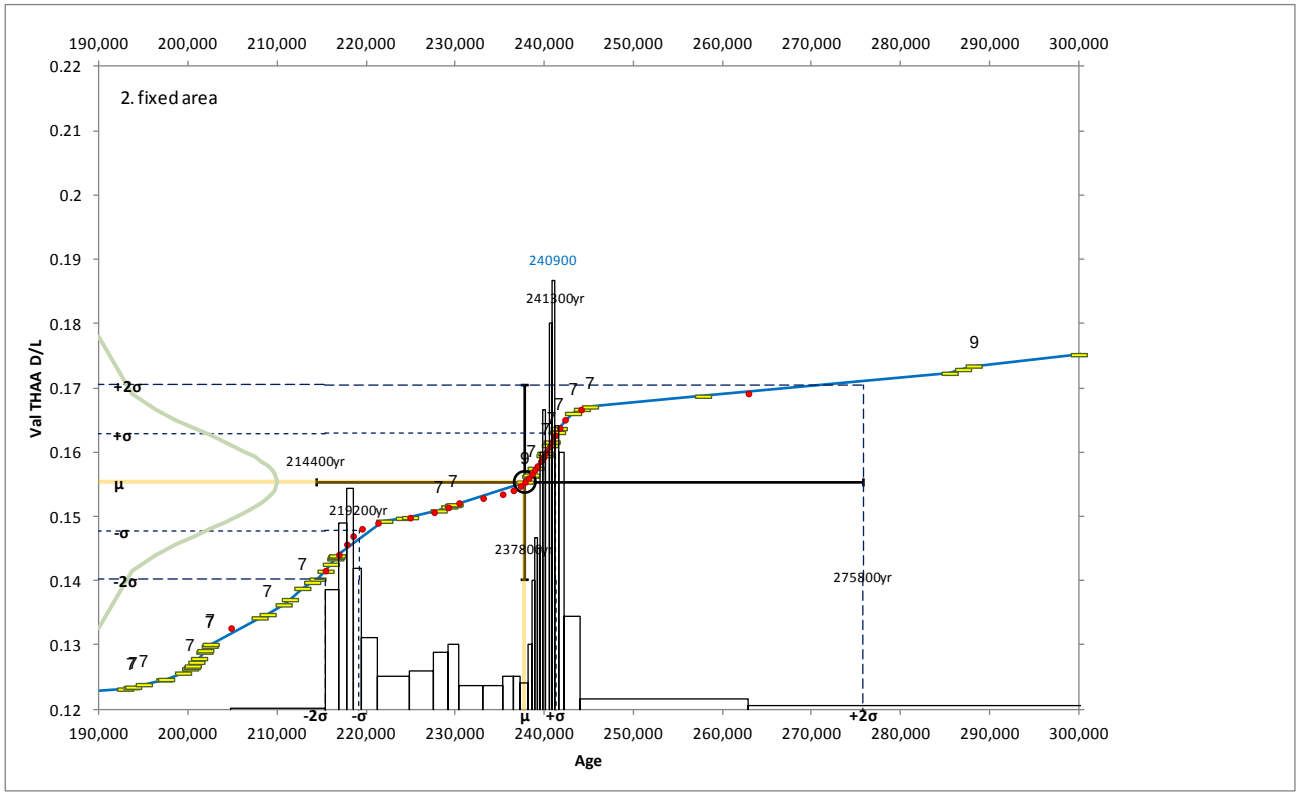
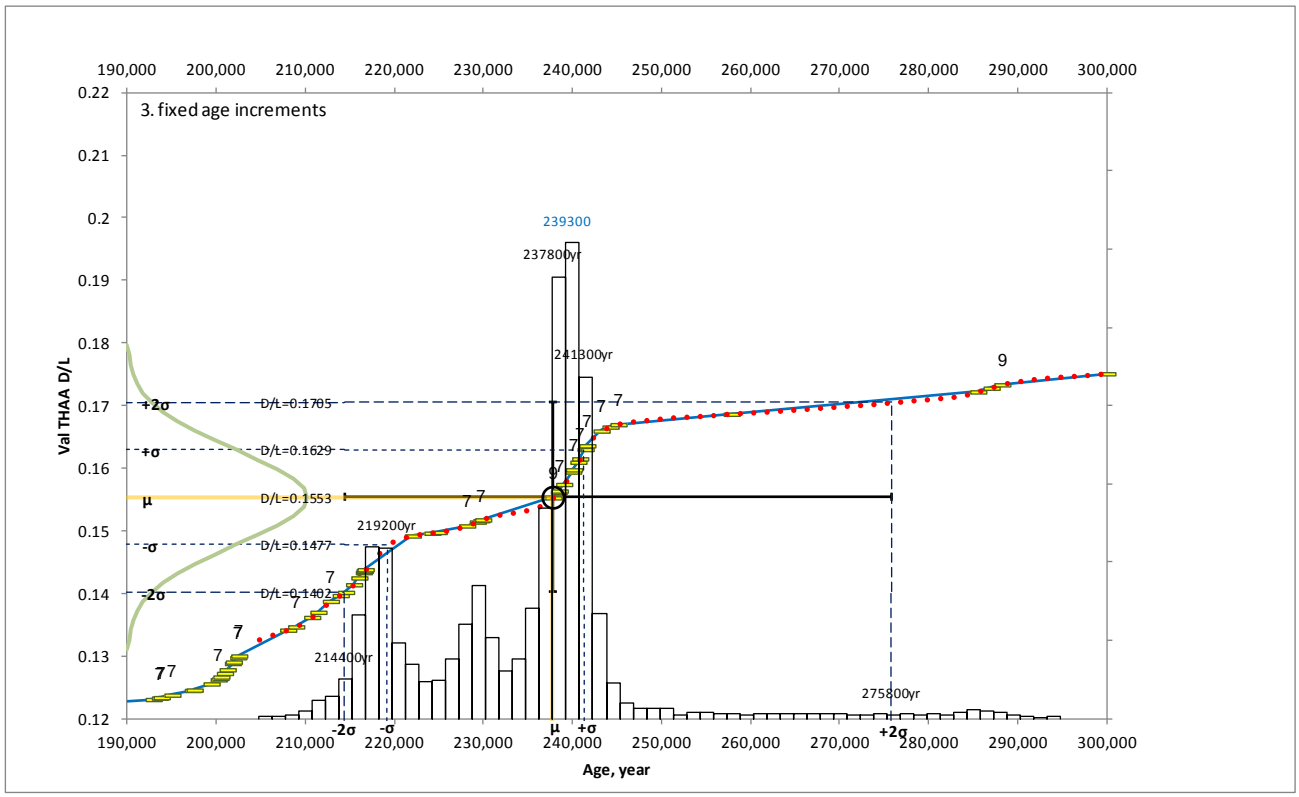


Figure 7.5c: Calibration using fixed age widths and variable z values and areas



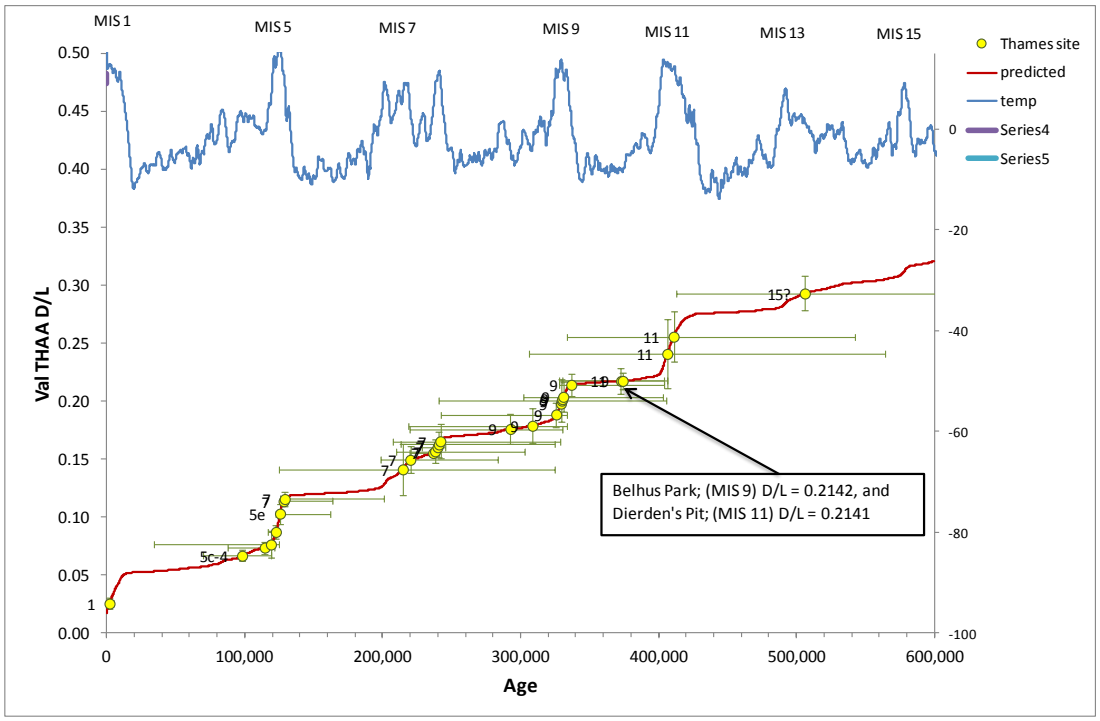
The temperature dependence of racemisation makes age determination during the end of one interglacial and the start of the next, potentially problematic. The D/L value will vary only marginally increase during a glacial stage, which could cover tens of thousands of years. Uncertainty ranges now based on probability distributions can be more easily interpreted, with peak heights indicating those regions that the true age of the sample is most likely to occur.

7.6 Testing the model

Although very much a prototype, it is helpful to look at some data in detail. Using the Thames terrace data from Chapter 6, Table 6.5, ±2 standard deviation uncertainty estimates for both D/L value and age (years) have been plotted against the predicted rate curve. Figure 7.6 shows D/L uncertainty estimates determined using ANOVA (for single values), and represents the worst case scenario.

The previous sequence chronologies (Figure 6.42) based only on ANOVA, indicated very little D/L value separation between the start of MIS 9 and the end of MIS 11. This is also illustrated in Figure 7.7 with both Belhus Park M25 (MIS 9) and Dierden's Pit (MIS 11) having a D/L of 0.214 (MIS assignment after Penkman, 2011).

Figure 7.6: Thames Terrace Sequence (uncertainty derived by ANOVA ±2 std dev).



Applying the calibration model, these values fall directly mid-way between MIS 11 and MIS 9. Figure 7.7a & b show the 2 std. dev. age confidence range for each site, both indicating increased probabilities that the true age is either in MIS 9 or MIS 11, and less likely to fall mid-way, which would be intuitively true due to glacial conditions during MIS 10.

Figure 7.7: Part of a Revised Thames AAR chronology showing 2 std. dev. Confidence Limits derived by ANOVA.

Figure 7.7a: Belhus Park, M25 (MIS 9)

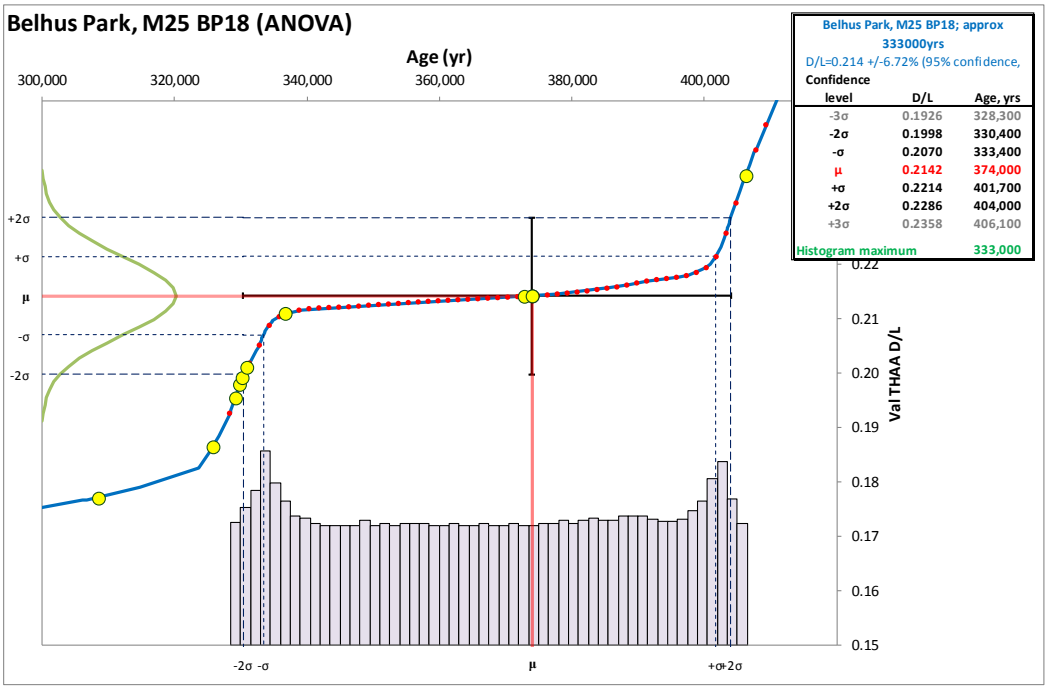
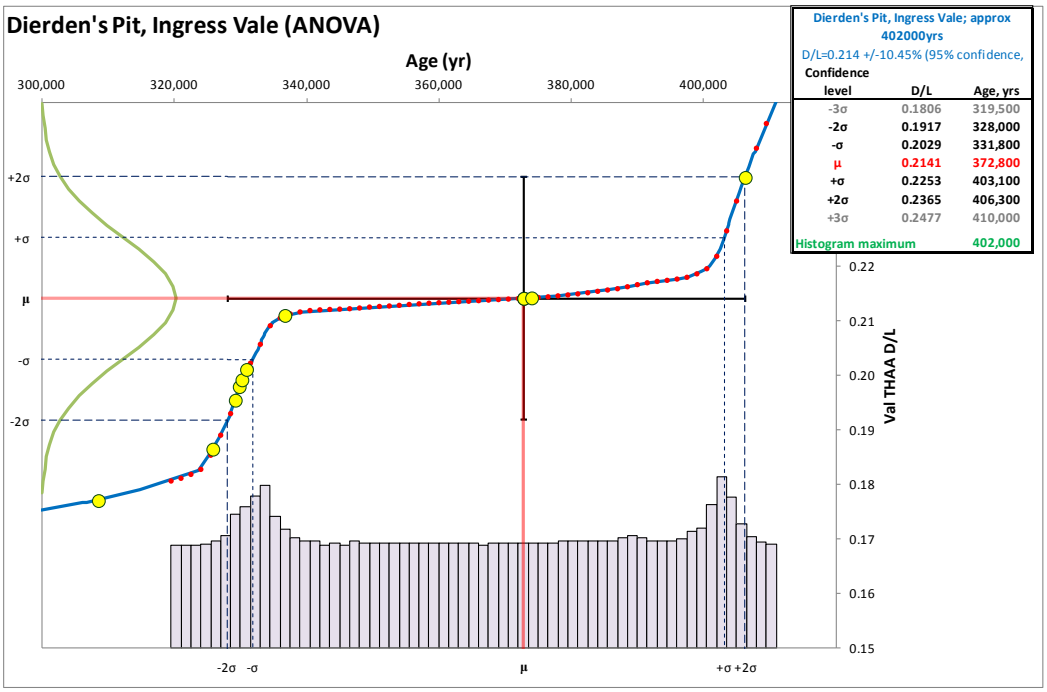


Figure 7.7b: Dierden's Pit (MIS 11)



Re-evaluating the data using the Bayesian method described in Chapter 6, revised D/L values are obtained indicating larger differences between the two sites (and giving an uncertainty best case scenario).

Thus, revised D/L values are for Belhus Park, M25 D/L = 0.211 and for Dierden's Pit D/L = 0.227. Figure 7.8 shows the improved resolution of the two sites and the unidirectional uncertainty (2 std dev) that results. Figure 7.9a & b show how these sites now resolve into their expected respective stages determined from independent stratigraphic data. Belhus Park, M25 (Figure 7.9a) has the highest probability of being a MIS 9 site, whilst Dierden's Pit (Figure 7.9b) has a higher probability of belonging to MIS 11, with neither likely to occur during the intervening glacial MIS 10

The effect of applying the Bayes method on uncertainty can be further seen with the Swanscombe data. From the original sequence (Table 6.5) based on ANOVA, Swanscombe can be seen to have one of the largest uncertainty estimates; Val D/L = 0.236 ± 12.6% (1 std dev). Using Bayes this reduces to give a D/L = 0.218 ± 1.43%. The effect on age uncertainty estimates (2 std devs) is also shown in Figure 7.10 a and b where the same axis scales have been retained.

Figure 7.8: Thames Terrace Sequence (uncertainty derived by Bayes ±2 std dev).

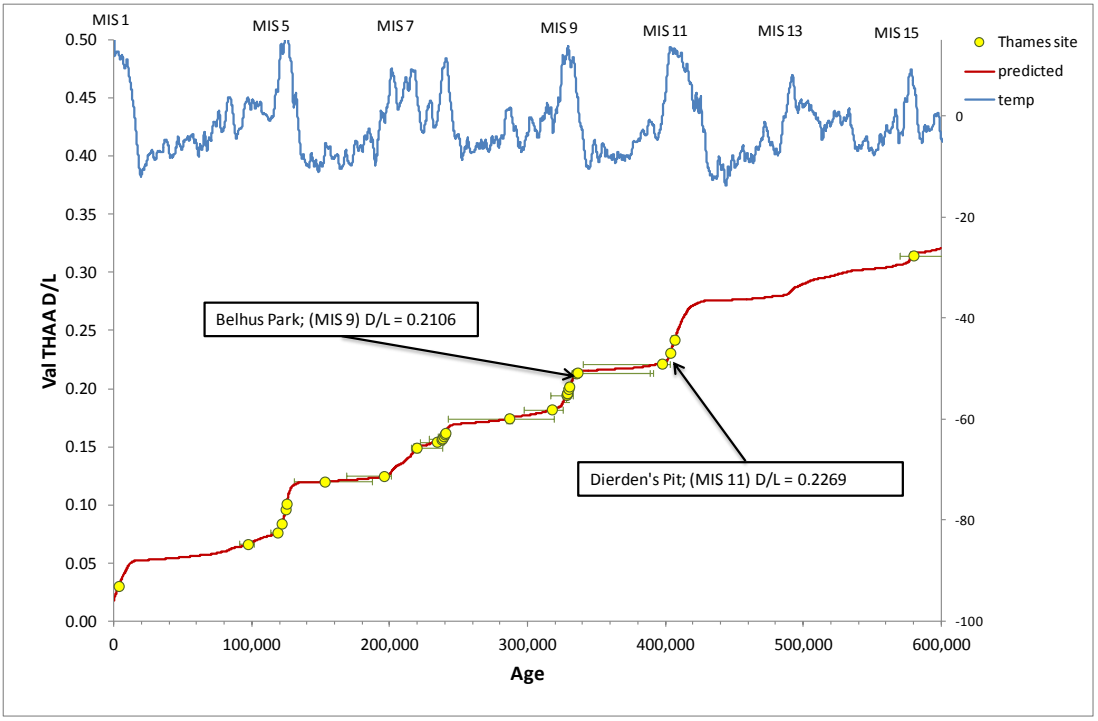


Figure 7.9 Part of a Revised Thames AAR chronology showing 2 std. dev. Confidence Limits derived using Bayes.

Figure 7.9a: Belhus Park, M25 (MIS 9)

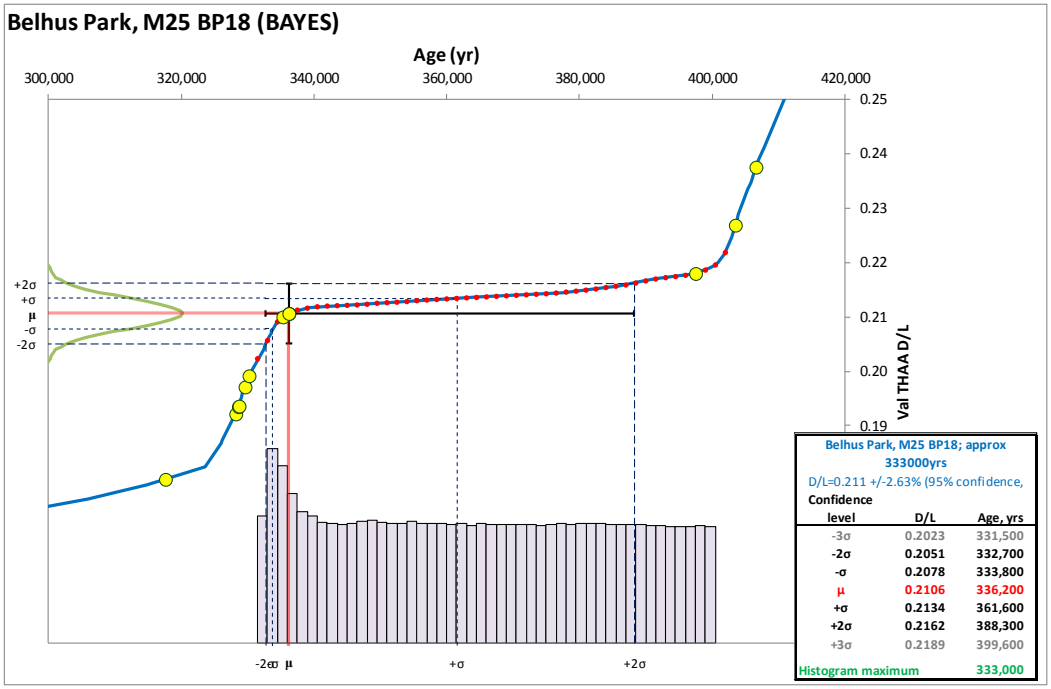


Figure 7.9b: Dierden's Pit (MIS 11)

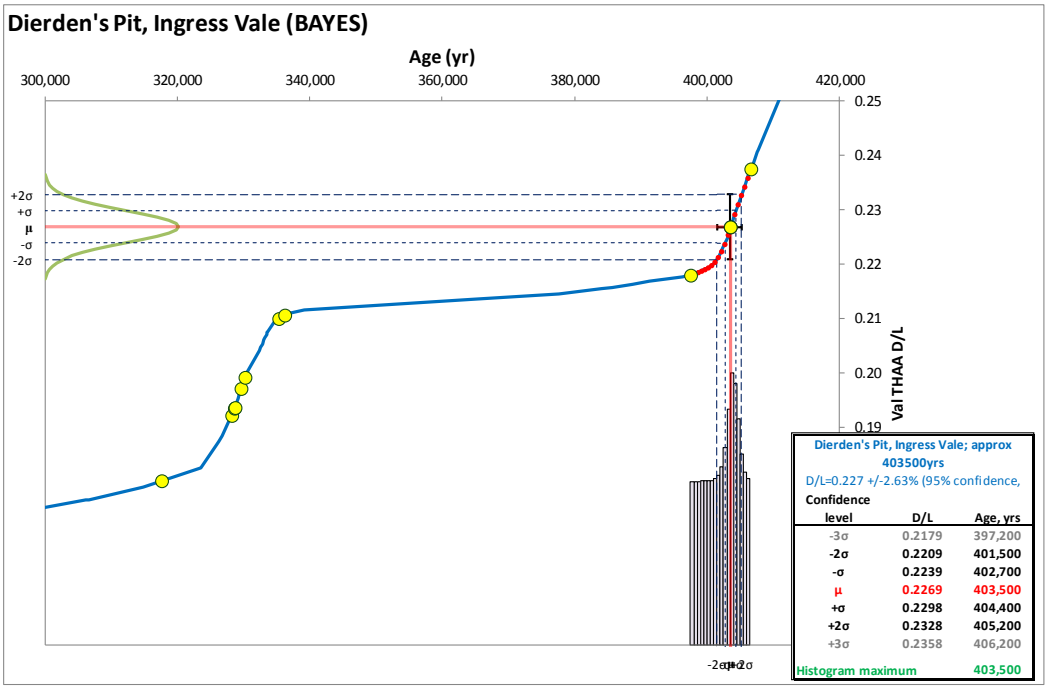


Figure 7.10: Part of a Revised Thames AAR chronology showing 2 std. dev. Confidence Limits for Swanscombe (MIS 11) (MIS derived from independent stratigraphic data)

Figure 7.10a: Swanscombe; uncertainty derived by ANOVA

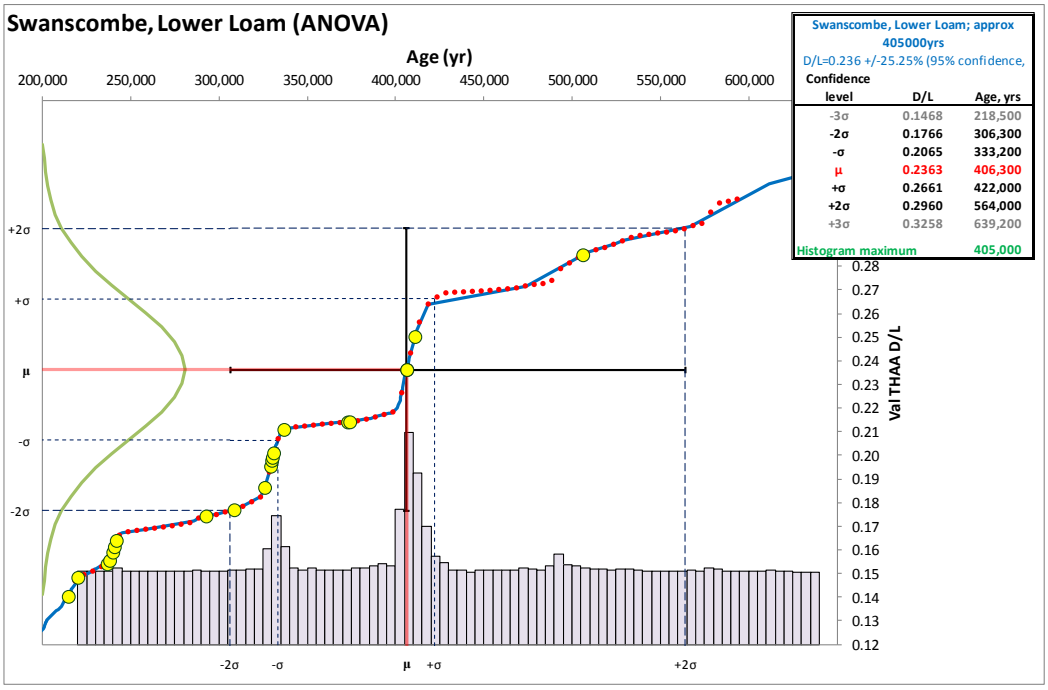
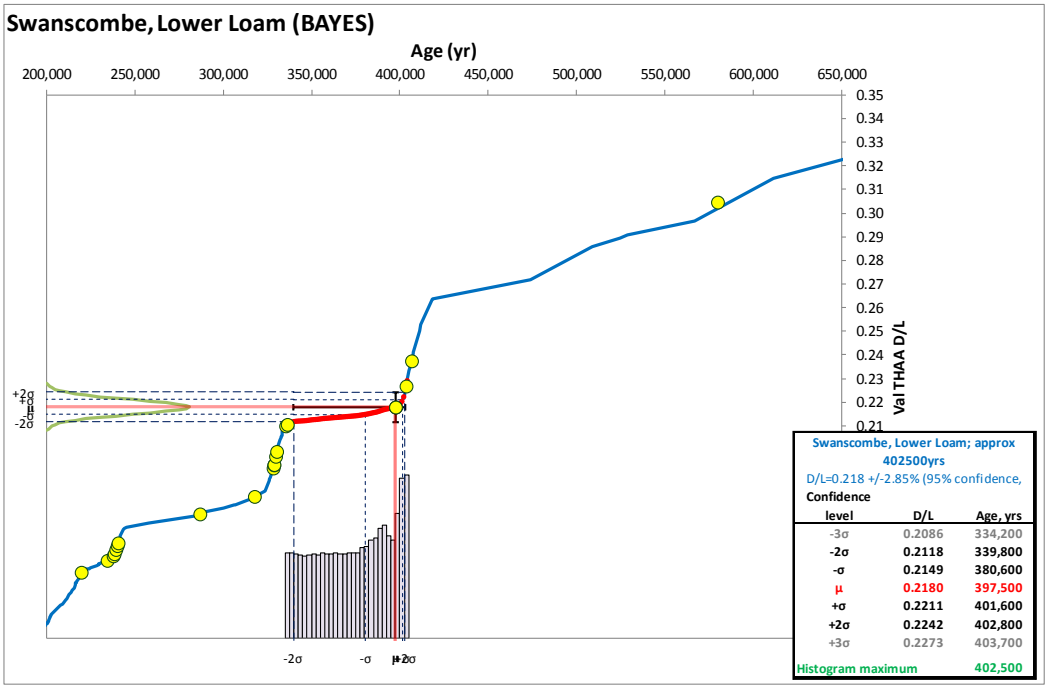


Figure 7.10b: Swanscombe; uncertainty derived by Bayes



7.7 Conclusion

A prototype model has been developed which enables quantitative AAR dating based on reaction kinetics and a palaeoclimate reconstruction. Probability based age uncertainty estimates can also be determined and plotted as histograms indicating regions of greatest or least confidence in the true age.

Using the Swanscombe data as an example, the Bayesian method described in Chapter 7 (to provide an integrated valine D/L value based on protein decomposition), demonstrates the impact that using a weighted approach to uncertainty estimation can have. Using a D/L uncertainty of 2.86% (2 std dev) (see Figure 7.10b) an age uncertainty range of 402,800 – 339,800 yrs is obtained, a difference of 63,000 yrs. Whilst this is an asymmetrical confidence interval, however, for comparative purposes, if it were to be interpreted as a normal distribution, then this would give an equivalent standard deviation of about 4% [63,000 / 2 = ± 31,500 yrs (2 std dev); 31,500 yrs / 2 = ± 15,750 yrs (1 std dev); 15,750 / 397,500(μ) x 100 = 4%].

The model is able to resolve differences between isotope stages but further correlation with independently dated sites is required, particularly at glacial/interglacial boundaries. There is the suggestion that wider temperature differences have been observed than are reflected by the temperature record used. This may indicate more significant continental influence on the palaeoclimate of the UK that previously considered. Correlation with more dated sites and stratigraphic information could help to reveal further detail in the temperature record.

However, whilst initial results are very encouraging, a considerable amount of further work is required to provide a robust and justifiable calibration, with perhaps a little more integrity than the empirical approach used here. Nonetheless it provides a compelling argument for further research.

Chapter 8. Further work

8.1 Intra-laboratory

8.1.1 Method Validation

There is an urgent need for a formal method validation to be carried out. Measurement uncertainty is currently determined as the standard deviation between replicate injections but this does not reflect the uncertainty of the method applied to samples. Uncertainty estimates for repeatability are determined by analyses of several independent samples within the same run (each having been prepared separately). Reproducibility uncertainty also needs to be determined over an extended period of time, by different operators and instruments. Materials analysed need to represent those typically analysed routinely by the laboratory and separate samples need to be worked up through the entire method, including preparation and extraction stages, independently of each other. This way the intermediate or intra-laboratory reproducibility thus derived will represent the worst case scenario. This is necessary so that there can be confidence that the measurement result of any sample being submitted for analysis will be at least as good as, and probably better than their maximum quoted uncertainty. If the best (smallest) uncertainty estimate was used then that says the measurement result will be no better than and possibly worse....with no outer boundary!

A protocol has been compiled (Barwick and Ellison, 2000a) in order to guide laboratories through uncertainty determination from method validation data. It is recognised that bias estimation probably cannot be currently determined due to the absence of reference materials although spiking experiments based on standard addition may be informative and provide recovery information.

Left over test materials from the proficiency study may be particularly beneficial in these capacities as they provide a number of homogeneous individual samples that now have consensus values associated with them. However, until reference materials are available

that can be used for bias correction, any consensus values are probably best viewed with caution and the measurement procedure considered empirical at the laboratory level.

Intra-laboratory reproducibility precision estimates can then be used as the uncertainty accompanying each measurement result. Estimates need only then be checked from time to time, perhaps when an instrument is moved or a change to the routine method occurs. It is also advisable that different uncertainty estimates are determined for different matrices.

Validation should also incorporate carefully designed ruggedness testing to evaluate the stages of the method that potentially may have the biggest influence. This includes timings, temperatures and acid concentrations. The influence of each critical stage or component needs to be carefully considered and experiments designed to vary factors and the effect of the measurement result observed. Data need then to be evaluated using statistical guidelines on the evaluation of ruggedness test data.

Guidance on this and other requirements for validation can be found at (EURACHEM, 1998; Barwick and Ellison, 2000b; Barwick *et al.*, 2000; EURACHEM / CITAC, 2001; Thompson *et al.*, 2002; ISO / IEC 17025, 2005; ISO 21748, 2010).

8.1.2 Quality control

Standard solutions are currently run routinely but are probably not being fully utilised. The only available reference material is in-house standard solutions and ILC materials. Neither are formally defined. However, information regarding the preparation of the standard solutions provides original weights and in the absence of stability data, these values could be used to correctly adjust the internal standard and determine recovery factors within each individual run. It is also observed that current practice that applies a correction to both the L and D together, is the same as no correction at all since they cancel,

This would not be an onerous task as it would simply require a couple of extra columns to be added to the current spreadsheet. It is recommended that for perhaps an extended period of time, both sets of data (corrected and uncorrected) are derived or else a series of analyses carried out that cover the full range of D/L values. This will enable direct correlation with corrected and uncorrected data which can then be used to correct historical values as required. It is noted that whilst response factors are derived using standard solutions, they will at least correct for instrumental effects, even if they can't correct for losses during extraction.

It is important NOT to use the same reference material for two different control activities. Therefore, since the standard solutions have quantifiable analytes, these should be used for calibration using the internal standards and response factors. The solid matric ILC material is therefore best suited for use as a QC material, that is run inline with test samples and control charts used to assess compliance with precision expectations (determined during validation).

Defined values for QC purposes are not required as long as long term trends can be observed. Results will be laboratory specific but changes in instrument settings, drift and instability will still be detectable. The frequency of use is dependent on a number of things but as a minimum it is suggested that perhaps one every 10 samples may be reasonable. A centralised system to automate plotting or results would be ideal and enable sharing of data. Ideally, individual samples would be worked up and prepared with each set of samples, this will identify problems with preparation reagents etc too. However a more practical solution may be to prepare a large number of individual hydrolysates that can then be rehydrated as required. This will only monitor the analytical aspects of the method for consistency but may be an acceptable compromise considering the time and effort required in including an extra sample each time?

8.2 Inter-laboratory

Because of the issues with the lack of reference materials, this probably shares equal priority to the need for validation and precision estimates.

The European Commission has recognised the difficulties laboratories face when there are not suitable commercial reference materials available (Quevauviller, 1998). To address this problem, a co-ordinated inter-laboratory collaborative trial can help to bridge the gap. This process requires sufficient homogeneous material to be available of sufficient stability and sufficient in quantity to service the user community for a significant time into the future. A collaborative trail requires all participants to apply the same analytical method, as prescribed by the organisers. This avoids additional uncertainty influences arising due to method differences. However the result provides the user community with an inter-laboratory validated method (often with previously validated intra-laboratory evaluation data being available), precision estimates, (repeatability and reproducibility), and consensus values with known uncertainty estimates. Subsequent use of this material then provides a

method for individual laboratories to evaluate, control and correct their own laboratory or method bias.

It is recommended that a set of calibration solutions is first developed with known L or D amino acid concentrations, using a primary method of analysis such as qNMR (quantitative Nuclear Magnetic Resonance). Two series of calibration standards at five concentrations traceable back to SI or a primary method should be developed using a commercially-available L- amino acid standard reference solution (NIST SRM 2389a) and a specially made Primary D-amino acid solution using commercially-available D-amino acid powders with known purities (Sigma-Aldrich). Initial D-amino acid concentrations can then be determined using qNMR. qNMR is considered a primary reference method as the signal intensity is directly related to the numbers of protons present, so does not need a standard reference of the same material. This technique is already employed by Sigma-Aldrich for the preparation of organic standard reference materials (SRMs). Serial dilutions of these two primary solutions would provide two series of calibration reference solutions (CalSols 1-5 and 6-10), for the L and D amino acids respectively. These should be dispensed and stored frozen at -20^{o}C. Once developed, these can be used to derive response factors to correct concentration data and/or correction factors to correct D/L values. However, calibration using the internal standard should be at the concentration level, not D/L.

Once in place, a set of biomineral reference materials should be developed, with parameters defined through collaborative trial (using the calibration solutions for correction). These matrix specific materials can then be used for QC purposes but even for external calibration (with internal standard calibration still being used) and corrected for losses during preparation and extraction.

8.3 Quantitative AAR

8.3.1 D/L Uncertainty

An integrated method based on protein decomposition, has been used to derived a revised valine THAA D/L value with associated uncertainty estimate. The Bayesian approach has been compared to uncertainty estimation based on ANOVA and found to make substantial reductions as a result of the increased confidence arising from the use of multiple estimates. The next stage would be to extend this approach to other amino acids in a similar fashion, and compare the results with those obtained by Monte Carlo simulation.

8.3.2 Calibration Curve

The model developed in the previous chapter is a simple prototype, exploring the possibility and potential application of quantitative AAR dating without the need for external calibration. However, the calibration curve presented, requires substantiating using independently derived dates and stratigraphy. The model uses only a single kinetic model, the RFOK model, and other functions should be explored, assessing the applicability to different D/L regions and potentially other amino acids and perhaps mollusc species. Similarly the use of different palaeoclimate models will likely result in a different calibration and these will need careful evaluation and accommodation into uncertainty estimates or else a drop down selection so as to be able to select the preferred calibration.

However, it would seem reasonable to utilise the majority of these sources of information and produce a single calibration curve that accommodates the variability reflected in the various datasets, or enables selection from a range of palaeoclimate records.

Similarly, calibration curves could be determined for different amino acids. The benefit of using a faster racemising amino acid such as aspartic acid would be that it could provide better resolution for the younger samples. Alanine could be used in the same way for the mid ranging values. A Monte Carlo approach or Gaussian process could be applied to derive a single expression for age and an uncertainty range based on these combined probability distributions. Ultimately, the results of several calibration curves could potentially be combined to give a single calibration curve with uncertainty regions similar to IntCal09 used by the radiocarbon community (Heaton *et al.*, 2009), itself derived from a number of sources.

Having derived a suitable calibration curve, one further avenue to explore would be to utilise existing Bayesian expertise used in the development of chronology building software such as Oxcal, by applying it to AAR data. In principle, it should be possible to simply replace any existing radiocarbon based calibration curve with an AAR temperature dependent one for a simplified calculation of age and uncertainty limits.

At the moment age estimates are not to be inferred from the charts presented in the last chapter. However, initial evaluation indicates that the modelling approach to quantitative AAR is potentially very powerful and can be used to reflect realistic uncertainty ranges. Much more work is required and evaluations performed before it can be put to its intended purpose, this project has just been the beginning.

Glossary of Abbreviations, Symbols, Terms & Definitions

Abbreviations

ANOVA	Analysis of Variance
CRM	Certified Reference Material
CV	Coefficient of Variation
EQC	External Quality Control
IQC	Internal Quality Control
MU	Uncertainty of Measurement / Measurement Uncertainty
PT	Proficiency test
QA	Quality Assurance
QC	Quality Control

Symbols

k	Coverage Factor
RMS_{bias}	Bias Root Mean Square
$RSD_L\%$	Relative Between Sample Standard Deviation (expressed as a percentage)
$RSU\%$	Relative Standard Uncertainty (expressed as a percentage)
$RSD\%$	Relative standard deviation (expressed as a percentage)
$RSD_r\%$	Relative Repeatability standard deviation (expressed as a percentage)
$RSD_R\%$	Relative Reproducibility standard deviation (expressed as a percentage)
s_{an}	(Homogeneity) Analytical Precision
s_{an}^2	(Homogeneity) Analytical Variance
s_{sam}	(Homogeneity) Sampling Precision
s_{sam}^2	(Homogeneity) Sampling Variance
s_{all}^2	(Homogeneity) Total Permissible Sampling Variance

s, SD or σ	Standard Deviation
s_L	Between-sample standard deviation
s_r	Repeatability Standard Deviation
s_R	Reproducibility Standard Deviation (Inter-Laboratory)
s_{RW}	Reproducibility Standard Deviation (Intra-Laboratory) or Intermediate Precision
σ_p	Target Standard Deviation
σ_h	Homogeneity Target standard deviation
$\hat{\sigma}$	Assigned Value standard deviation
$u(x)$	Standard Uncertainty
$u(\hat{X})$	Standard Uncertainty of the Assigned Value
$u(bias)$	Standard Uncertainty due to Bias
$u(\bar{x})$	Standard Uncertainty of Participant's Results
u_c	Combined (standard) Uncertainty
U	Expanded Uncertainty
x *or* x_i	Submitted Result or Value
$\bar{x}$	Measurement Result / Mean submitted result
$\hat{X}$	Assigned Value

Terms and Definitions

Specific references for terms that can be found in International Standards or guidance documents have been given in brackets at the end of each definition. Here, **VIM** refers to '*International vocabulary of metrology*' (JCGM 200, 2008), **GUM** refers to the '*Guide to the expression of uncertainty in Measurement'* (JCGM 100, 2008) and **ISO (1)**,refers to (ISO 5725-1, 1994) on the '*Accuracy (trueness and precision) of measurement methods and results'*. Terms shown in bold indicate further definitions that may be found in this section.

Readers are recommended to consult these documents for additional notes and comments not included here.

Accuracy

closeness of agreement between a measured result and the true value (if it could be known), or a reference value. (VIM 2.13)

NOTE 1; Accuracy is a concept that cannot be directly quantified. It does not possess a numerical value.

NOTE 2; Accuracy describes **random** and **systematic error** effects and as such is composed of both **precision** and **bias** components.

Analysis of Variance (ANOVA)

A group of statistical techniques that enable the different contributions from various sources of the observed variance in experimental data to be separated and estimated. (Currell and Dowman, 2005; Miller and Miller, 2005).

factor plus different precision

NOTE 1; A one-way ANOVA uses the F-test to compare the effect of one the experimental precision, eg; the effect of the measurement process on samples, (between-sample variance) against the inherent experimental (within-sample variance).

NOTE 2; Whilst it is possible to carry out the analysis by hand more commonly statsistical software packages are more convenient such as the Excel Data Analysis tools as this also carries out the F-test evaluation at the same time.

Assigned Value $\hat{X}$

The best estimate of the true value of the measurand.

NOTE; This may be the certified reference value of a CRM, a reference value from a reference laboratory or the consensus value from participants' results calculated as the robust mean, median or mode.

Assigned Value standard deviation ($\hat{\sigma}$)

Standard deviation of the assigned value.

NOTE; This may be the robust standard deviation, sMAD (median absolute deviation) or SEM (standard error of the mode)

Between-sample standard deviation (S_L);

The precision or dispersion between independent measurements carried out on different samples of the same material under **reproducibility conditions.**

NOTE: it includes the between-operator, between-day, between-instruments, and between-laboratory variability's, etc. and is a component of **reproducibility standard deviation**. It is determined using **ANOVA**, such that;

$$s_L = \sqrt{\frac{between\ group\ mean\ square - within\ group\ mean\ square}{n}}$$

Bias

estimate of a systematic measurement error (VIM 2.18)

$$bias = (\bar{x} - \hat{X})$$

Bias Root Mean Square (RMS_{bias})

A component of the bias standard uncertainty taking into account both the bias and bias variation. See **Standard uncertainty due to bias ($u(bias)$).**

Certified Reference Material (CRM);

a reference material accompanied by certified traceable measurement and uncertainty values determined using validated procedures (VIM 5.14)

Cochran's Test

A statistical test that detects extreme variances between observations by calculating the Cochran's (C) value as the ratio between the largest squared difference (D_{max}^2) to the sum of all the squared differences ($\sum D_i^2$) and comparing this against tabulated critical values. (ISO 5752-2: 1994)

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

Coefficient of Variation ($CV\%$) (expressed as a percentage).

See **Relative standard deviation ($RSD\%$)**

Combined (standard) Uncertainty (u_c)

The combined standard uncertainty of a measurement result taking into account various contributions from different standard uncertainty sources. (GUM 2.3.4)

uncertainty NOTE 1; There are two common rules for the combination of **standard** values which depend on the model used for deriving the measurement value;

Eg; a). If the model involves the addition or subtraction of values, i.e.; $y = (a + b + c \ldots)$then the combined standard uncertainty, $u_c(y)$ is given by;

$$u_c(y(a, b, c \ldots)) = \sqrt{u(a)^2 + u(b)^2 + u(c)^2 + \ldots}$$

Eg; b). If the model involves the product or quotient of values, i.e.; $y = (a \times b \times c \ldots)$ or $y = a/(b \times c \ldots)$ then the combined standard uncertainty, $u_c(y)$ is given by;

$$u_c(y(a, b, c \ldots)) = y\sqrt{\left(\frac{u(a)}{a}\right)^2 + \left(\frac{u(b)}{b}\right)^2 + \left(\frac{u(c)}{c}\right)^2 + \ldots}$$

NOTE 2; For proficiency testing the format given in the first example has been used, thus;

$$u_C = \sqrt{S_{Rw}^2 + u(\bar{x})^2 + u(\hat{X})^2 + (bias)^2}$$

Where; $\sqrt{S_{Rw}^2}$ = uncertainty due to precision, and

$\sqrt{u(\bar{x})^2 + u(\hat{X})^2 + (bias)^2}$ = $u(bias)$ i.e.; the **uncertainty due to bias.**

Coverage Factor (k)

Factor used to multiply the combined uncertainty by in order to derive the Expanded uncertainty value.

NOTE; For large data sets where the distribution approximates to normality the value of *k* to use is taken from the level of confidence required in the measurement result. Most often a 95% or 2 standard deviation level of confidence is required for the reporting of measurement results, thus $k=2$.

For smaller data sets where the distribution of measurement results is better described by a t-distribution, the equivalent t-value is used as the multiplier, thus $k=t_{(0.5,df)}$.

Error

measured quantity value minus a reference value or true value (VIM 2.16)

NOTE 1; To some extent the concept of error is a theoretical one as it is not possible to be sure of a measurand's true value, only a best estimation of it from measurement determinations. If a reference value is to be used then it

is more accurate to determine the precision and bias as estimates of random and systematic error contributions which can be quantified.

Expanded Uncertainty (U)

A quantity defined by a specified interval (i.e.; 2 standard deviations) or confidence level (i.e.; 95% confidence) about the measurement result and describes the dispersion where a large number of repeated **measurement results** would be expected to lie.

$U = u_c \times k$ where k = the **coverage factor**, and

u_c= the **combined uncertainty**

Experimental standard deviation of the mean.

See **Standard Uncertainty ($u(x)$)**

External Quality Control (EQC)

See **Quality Control (QC).**

F_1 and F_2

Are constants used to test the hypothesis that there is no significant evidence that the sampling standard deviation exceeds the allowable fraction of the target standard deviation and that the test for sufficient homogeneity has been passed (Fearn, T. and Thompson, M., 2001).

$$s_{sam}^2 = F_1 s_{all}^2 + F_2 s_{an}^2$$

Values for F_1 and F_2 may be derived from statistical tables;

$F_1 = \frac{\chi^2_{(m-1,0.95)}}{m-1}$ where m = the number of samples measured in duplicate

$$F_2 = \frac{F_{(m-1,m,0.95)} - 1}{2}$$

NOTE; The (Fisher) F-Test is a test for significant differences between the variances of two data sets and compares random error effects. The F-test may also be used within other tests such as ANOVA, (Currell, G., & Dowman, A.,2005, Miller, J.N, & Miller, J.C., 2005)

Thus; F-statistic $F = {s_a^2}/{s_b^2}$ or $= {MS_{between}}/{MS_{within}}$

(Homogeneity) Analytical Precision (s_{an})

The homogeneity within-sample standard deviation for the replicate values (i.e.; a and b) used in the test for sufficient homogeneity of the test materials. Calculated from the ANOVA within group mean square;

$$s_{an} = \sqrt{MS_w}$$

(Homogeneity) Analytical Variance (s_{an}^2)

The square of the analytical precision. . Calculated from the ANOVA within group mean square;

$$s_{an}^2 = MS_w$$

(Homogeneity) Sampling Precision (s_{sam})

The homogeneity between-sample standard deviation for the samples (i.e.; 1, 2…10) used in the test for sufficient homogeneity of the test materials. Calculated from the ANOVA between and within group mean square values;

$$s_{sam} = \sqrt{\frac{MS_b - MS_w}{2}}$$

(Homogeneity) Sampling Variance (s_{sam}^2)

The square of the sampling precision. Calculated from the ANOVA between and within group mean square values;

$$s_{sam}^2 = \frac{MS_b - MS_w}{2}$$

Homogeneity Target standard deviation (σ_h).

In the absence of an external value for target standard deviation (σ_p), a target value sufficient homogeneity (σ_h)can be determined using fitness-for-purpose criteria.

(Homogeneity) Total Permissible Sampling Variance (s_{all}^2)

The total allowable between-sample variance that must not be exceeded by the sampling variance in order for the test materials to be considered homogeneous. $\boldsymbol{s_{all}^2}$ is derived from the homogeneity target standard deviation (either σ_p *or* σ_h).

$$s_{all}^2 = (0.3 \times \sigma_p)^2$$

Intermediate conditions

Independent measurement results obtained for identical test items using the same measurement procedure under a specified set of conditions within the same laboratory that include, different operators, different operating conditions, different locations over any given

period of time, (VIM 2.22). See **Reproducibility Standard Deviation (Intra-Laboratory) or Intermediate Precision (S_{RW})**

Internal Quality Control (IQC)

See **Quality Control (QC)**

Measurement Result / Mean submitted result ($\bar{x}$)

The average of an individual participant's replicate measurement results for the same analyte in the proficiency test.

Measurement Uncertainty

See **Uncertainty of Measurement**

Precision

closeness of agreement between repeated measurement results on the same material under specified conditions (VIM 2.15)

> NOTE 1; Precision can be quantified and usually expressed as a measure of imprecision such as standard deviation, variance, relative std dev or CV and is a measure of random error.
>
> NOTE 2; Specific measurement conditions can be repeatability, intermediate or reproducibility conditions.

Proficiency test (PT);

An **external quality control (EQC)** procedure through which the **accuracy** of a laboratory's measurement result can be objectively evaluated. Performance is assessed by providing a comparison of **trueness** with other participating laboratories

> NOTE: **Trueness** is determined through the evaluation of laboratory **bias** against a reference value. This may be presented as **z-scores** or other assessment of **bias**.

Quality Assurance (QA);

Documented procedures that describe a quality management system designed to control activities and maintain a quality output.

Quality Control (QC);

Specific activities that are carried out in order to implement the procedures documented under the **Quality Assurance** programme.

NOTE; This may be in the form of **Internal Quality control (IQC)** that are carried out internally by the organization such as method validation, calibration, control charts, etc, or **External Quality Control (EQC)** coordinated by an external organization such as interlaboratory comparisons eg; proficiency tests or collaborative trails.

Random error

component of measurement error that in replicate measurements varies unpredictably (VIM 2.19)

NOTE 1; A random error value is determined as the precision that would result from a number of replicate measurements of the same measurand, expressed as a distribution.

Relative Bias % (expressed as a percentage)

Bias divided by the assigned value (x 100)

$$relative\ bias\ \% = \frac{(\bar{x} - \hat{X})}{\hat{X}} \text{ x } 100$$

Relative Between Sample Standard Deviation ($RSD_L\%$), (expressed as a percentage)

The **between-sample standard deviation** divided by the (average) measurement result (x 100)

$$RSD_L\% = \left({}^{S_L}/_{\bar{x}}\right) \times 100$$

Relative Standard Uncertainty ($RSU\%$), (expressed as a percentage)

The **standard uncertainty** divided by the (average) measurement result (x 100)

$$RSU\% = \left({}^{u(\bar{x})}/_{\bar{x}} \right) \times 100$$

Relative standard deviation ($RSD\%$) or **Coefficient of Variation ($CV\%$)** (expressed as a percentage)

The **standard deviation** divided by the (average) measurement result (x 100)

$$RSD\%\ or\ CV\% = \left({}^{S}/_{\bar{x}}\right) \times 100$$

Relative Repeatability standard deviation ($RSD_r\%$), (expressed as a percentage)

The **repeatability standard deviation** divided by the (average) measurement result (x 100)

$$RSD_r\% = \left({}^{S_r}/_{\bar{x}}\right) \times 100$$

Relative Reproducibility standard deviation ($RSD_R\%$), expressed as a percentage

The **Reproducibility standard deviation** divided by the (average) measurement result (x 100)

$$RSD_R\% = \left({}^{S_R}/_{\bar{x}}\right) \times 100$$

Repeatability conditions ;

Independent measurement results are obtained for identical test items under a specified set of conditions that include the same measurement procedure, same measurement system or laboratory, same operators, same operating conditions, same location and in as short a time as period as possible, (VIM 2.20, ISO (1) 3.14). See **Repeatability Standard Deviation (S_r)**

Repeatability Standard Deviation (S_r)

The dispersion or precision of replicate measurement values carried out under repeatability conditions (ISO (1) 3.15)

NOTE; Often calculated using **ANOVA** from the within group mean square (MS), such that;

$$s_r = \sqrt{within\ group\ mean\ square}$$

Eg; a).Within-sample (or instrumental/analytical) repeatability standard deviation is the dispersion of replicate instrumental measurements carried out on the same sample in the same analytical run, eg; an individual laboratory's replicate PT results.

b). Intra-laboratory (or method + analytical) repeatability standard deviation is the dispersion of independent measurements carried out by a single laboratory on different samples of the same material, under repeatability conditions, eg. From Intra-laboratory method validation data or homogeneity analytical precision data (s_{an}).

c). Inter-laboratory repeatability (laboratory+method+analytical) standard deviation is the dispersion of independent measurements carried out by more than one laboratory on different samples of the same material, under repeatability conditions, eg, collaborative trial precision data.

Reproducibility Conditions;

Independent measurement results obtained for identical test items using the same measurement procedure under a specified set of conditions that include, different measurement systems and laboratories, different operators, different operating conditions, different locations over any given period of time, (VIM 2.24, ISO (1) 3.18). See **Reproducibility Standard Deviation (Inter-Laboratory) (S_R)**

Reproducibility Standard Deviation (Inter-Laboratory) (S_R)

The overall dispersion or precision of independent measurement values carried out on different samples of the same material by different laboratories, under **reproducibility conditions** and incorporates both within (repeatability) and between-sample precision estimates (ISO (1) 3.19)

Thus; $s_R = \sqrt{s_r^2 + s_L^2}$

Eg; a). The Inter-laboratory reproducibility standard deviation (S_R) obtained from a collaborative trial represents the maximum dispersion for the measurement procedure carried out across laboratories and provides an estimate of best practice for the measurement procedure for a specified matrix / analyte/ concentration. Providing a laboratory's own repeatability is in agreement with the inter-laboratory repeatability precision estimate, then the laboratory can claim the Reproducibility standard deviation from a collaborative trial as their own **standard uncertainty** estimate.

Reproducibility Standard Deviation (Intra-Laboratory) or Intermediate Precision (S_{RW})

The overall dispersion or precision of independent measurement values carried out on different samples of the same material by the same laboratory, under **reproducibility conditions** and incorporates both within (repeatability) and between-sample precision estimates (VIM 2.23)

Thus; $s_{RW} = \sqrt{s_r^2 + s_L^2}$

Eg; Intra-laboratory reproducibility standard deviation (S_{RW}) represents the maximum dispersion for the measurement procedure carried out by an individual laboratory and is often used in method validation as the method precision for a particular matrix / analyte /concentration and used as the **standard uncertainty**.

Standard Deviation (s, sd or σ)

A term used to describe the dispersion or spread of measurement values and has the same units as the measurement value.

NOTE; by convention the symbol used for standard deviation depends on whether it is describing sample statistics or population parameters.

Thus;

Sample statistics; $s = \sigma_{n-1} = \sqrt{\frac{\sum_1^n (x_i - \bar{x})^2}{n-1}}$

Population parameters; $\sigma = \sqrt{\frac{\sum_1^n (x_i - \mu)^2}{n}}$

Where x_i = individual measurement values

$\bar{x}$ = average measurement value for the sample

μ = population mean

n = number of measurement values or population size

Standard Error of the Mean.

See **Standard Uncertainty ($u(x)$)**

Standard Uncertainty ($u(x)$)

The uncertainty of a measurement result expressed as a standard deviation, (GUM 2.3.1)

NOTE; When determined from a series of repeated measurements this can also be found referred to in texts as the experimental standard deviation or standard error of the mean.

Thus; $u(x) = {}^{s}/_{\sqrt{n}}$

Standard Uncertainty of the Assigned Value ($u(\hat{X})$)

The uncertainty of the **Assigned Value,** expressed as a standard deviation, (GUM 2.3.1).

$u(\hat{X}) = {}^{\hat{\sigma}}/_{\sqrt{m}}$ where $\hat{\sigma}$ = the **assigned value** std dev

and m = the number of participants' measurement results

NOTE; $u(\hat{X})$ is also a component of the **standard uncertainty due to bias** $\boldsymbol{u(bias)}$.

Standard Uncertainty due to Bias ($u(bias)$).

The uncertainty of the bias component of a participant's measurement result, expressed as a standard deviation, (GUM 2.3.1).

NOTE 1; An individual laboratory's standard uncertainty due to bias for a single proficiency test, is given as;

$$u(bias) = \sqrt{(bias)^2 + u(\bar{x})^2 + u(\hat{X})^2}$$

NOTE 2; An individual laboratory's standard uncertainty due to bias over multiple proficiency tests, is given as;

$$u(bias) = \sqrt{RMS_{bias}{}^2 + u(\hat{X})^2}$$

where; RMS_{bias} = the **bias root mean square** and given as;

$$RMS_{bias} = \sqrt{\frac{\sum(bias_i)^2}{m}}$$

and $u(\hat{X})$= the average standard uncertainty of the assigned value;

$$u(\hat{X}) = {\sum \hat{\sigma}_i}\Big/{\sqrt{\sum n_i}}$$

m = the number of proficiency tests or number of bias values, and

n = the number of participants' measurement results in each PT.

NOTE 3; It often helps to carry out these calculations as the relative percentage values.

Standard Uncertainty of Participant's Results ($u(\bar{x})$)

The uncertainty of a participant's submitted replicate results, expressed as a standard deviation, (GUM 2.3.1).

$u(\bar{x}) = {s_{\bar{x}}}/{\sqrt{n}}$ where $s_{\bar{x}}$ = the std dev of replicate values

and n = the number of replicate values submitted

NOTE; $u(\bar{x})$ is also a component of the **standard uncertainty due to bias** $\boldsymbol{u(bias)}$.

Submitted Result or Value (x or x_i)

An individual participant's submitted measurement result for the proficiency test.

Systematic Error

component of measurement error that in replicate measurements remains constant or varies predictably (VIM 2.17)

NOTE 1; A systematic error value is determined as the bias, i.e.; the difference between a measured result and the true or reference value. Measurement results should always be corrected where significant bias is detected.

Target Standard Deviation (σ_p)

The target value for standard deviation for the proficiency test used to calculate z-scores and assess homogeneity data.

NOTE; often determined independently from data external to the proficiency test, such as the reproducibility standard deviation (RSD_R%) from a collaborative trail or using a predictive model such as the Horwitz function when appropriate of fitness-for purpose criteria. The target std dev is usually matrix / analyte specific.

Eg; a) From a collaborative trial; $\sigma_p = \frac{RSD_R}{100} \times c$

where RSD_R = Relative Standard Deviation of Reproducibility from collaborative trial data, expressed as %

and *c* = concentration, i.e. the assigned value, $\hat{X}$, expressed in relevant units.

eg; b) Using the Horwitz equation; $\sigma_p = 0.02c^{0.8495}$

Or modified form; for concentrations less than 120ppb (1.2×10^{-7}); $\sigma_p = 0.22c$

and for concentrations greater than 13.8% (0.138); $\sigma_p = 0.01c^{0.5}$

Where the concentration (*c)* is expressed as a mass fraction as shown in () above.

Trueness

closeness of agreement between the average of a large number of replicate measurement results and the true value (if it could be known) or a reference value (VIM 2.14)

NOTE 1; Trueness is a concept that cannot be directly quantified. It does not possess a numerical value.

NOTE 2; Trueness is usually expressed as bias and a measure of systematic error.

t-value

2-tailed t-value is used as a correction factor in the determination of confidence intervals for small values of n. Derived from the t-distribution for sample data sets and described using $t(\bar{x}, s)$, compared to the normal distribution for populations described as $N(\mu, \sigma)$. Values for t may be obtained from statistical tables. (Currell and Dowman, 2005; Miller and Miller, 2005).

Such that, for a 95% confidence interval; $CI = \bar{x} \pm \left[t_{(2,o.o5,df)} \times \frac{\sigma}{\sqrt{n}}\right]$

NOTE; The (student's) t-Test is a test for significant differences between the mean of two data sets and compares systematic error effects.

Thus; t-statistic $t = {(x-\mu)}/{s/\sqrt{n}}$

Uncertainty of Measurement / Measurement Uncertainty (MU)

A parameter associated with a measurement result (taken as the best estimate of the true value) and characterizes the dispersion of values that could be attributed to the measurement result, taking into account both random and systematic error contributions from all possible sources and represents the degree of doubt associated with the measurement result (GUM 2.2).

Welch-Satterthwaite formula

Formula used for deriving the effective degrees of freedom for the calculation of Expanded uncertainty, when various standard uncertainties are combined with differing degrees of freedom.

$$v_{eff} = u_c^4(y) / \sum \frac{u_i^4(y)}{v_i}$$

Where v_{eff} = the effective degrees of freedom,

v_i = degrees of freedom of individual uncertainty components,

u_c = combined standard uncertainty

u_i = individual uncertainty components.

z-Score

A standardized measure of laboratory bias derived from the assigned value and target standard deviation, enabling a comparison of performance between laboratories. Satisfactory performance is considered if a $|z| \leq 2$.

$$z = \frac{(x - \hat{X})}{\sigma_p}$$

References

ABELSON, P. H. 1954. Amino acids in fossils. *Science,* 119**,** 576.

ABELSON, P. H. 1955. Organic constituents of fossils. *Carnegie Institute of Washington Year Book,* 54**,** 107-109.

AGUIRRE, E. & PASINI, G. 1985. The Pliocene - Pleistocene boundary. *Episodes,* 8**,** 118-20.

AIELLO, G., BARRA, D. & BONADUCE, G. 1996. The genus Cytheropteron Sars, 1866 (Crustaccea: Ostracoda) in the Pliocene-Early Pleistocene of the Mount San Nicola section (Gela, Sicily). *Micropalaeontology,* 42**,** 167-178.

AITKEN, M. J. 1990. *Science-based Dating in Archaeology,* London & New York, Longman Archaeology series.

AITKEN, M. J. & STOKES, S. 1997. Climatostratigraphy. *In:* TAYLOR, R. E. & AITKIN, M. J. (eds.) *Chronometric Dating in Archaeology.* New York & London: Plenum Press.

ALVAREZ-PRIETO, M., JIMÉNEZ-CHACÓN, J. & MONTERO-CURBELO, Á. 2009. Do we need to consider metrological meanings of different measurement uncertainty estimations? *Accreditation and Quality Assurance,* 14**,** 623-634.

ANDREWS, D. F., BICKEL, P. J., HAMPEL, F. R., HUBER, P. J., ROGERS, W. H. & TURKEY, J. W. 1972. *Robust estimates of location*, Princeton University Press.

ARIAS, C., AZZAROLI, A., BIGAZZI, G. & BONADONNA, F. 1980. Magneto-stratigraphy and Pliocene-Pleistocene boundary in Italy *Quaternary Research,* 13**,** 65-74.

BADA, J. L. 1970. Marine sediments: dating by the racemization of amino acids. *Science,* 170**,** 730-732.

BADA, J. L. 1972. The dating of fossil bones using the racemization of isoleucine. *Earth & Planetary Science Letters,* 15**,** 223.

BADA, J. L., *ET AL*., 1984. Accelerator mass spectrometry radiuocarbon ages of amino acid extracts from Californian palaeoindian skeletons. *Nature,* 312**,** 442-444.

BADA, J. L. & MAN, E. H. 1973. Racemization of isoleucine in cores from leg 15, site 148. *In:* HEEZAN, B. C. & ET AL. (eds.) *Initial reports of the deep sea drilling project.* US Government Pronting Office, Washington D.C.

BADA, J. L. & SCHROEDER, R. A. 1972. Racemization of isoleucine in calcareous marine sediments: Kinetics and mechanism. *Earth and Planetary Science Letters,* 15**,** 1-11.

BADA, J. L., SCHROEDER, R. A. & CARTER, G. F. 1974. New evidence for the antiquity of man inNorth America deduced from aspartic acid racemization. *Science* 184**,** 791-793.

BADA, J. L., SHOU, M.-Y., MAN, E. H. & SCHROEDER, R. A. 1978. Decomposition of hydroxy amino acids in foraminiferal tests; kinematics, mechanism and geochronological implications. *Earth and Planetary Science Letters,* 41**,** 67-76.

BAKEMAN, V. R. 2006. *Pacific and Atlantic coast mollusk shells: Chromatographic amino acid racemization kinetics and interlaboratory comparisons. MS Thesis.* MS Thesis, Newark.

BARWICK, V. J. & ELLISON, S. L. R. 2000a. Development and Harmonisation of Measurement Uncertainty Principles Part (d): Protocol for uncertainty evaluation from validation data. *VAM Technical Report,* LGC/VAM/1998/088.

BARWICK, V. J. & ELLISON, S. L. R. 2000b. The evaluation of measurement uncertainty from method validation studies. Part 1; Description of a laboratory protocol. *Accreditation and Quality Assurance: Journal for Quality, Comparability and Reliability in Chemical Measurement,* 5**,** 47-53.

BARWICK, V. J., ELLISON, S. L. R., RAFFERTY, M. J. Q. & GILL, R. S. 2000. The evaluation of measurement uncertainty from method validation studies. Part 2; The practical application of a laboratory protocol. *Accreditation and Quality Assurance: Journal for Quality, Comparability and Reliability in Chemical Measurement,* 5**,** 104-113.

BASSETT, M. 1985. Towards a "common language" in stratigraphy. *Episodes,* 8**,** 87-92.

BINTANJA, R., VAN DE WAL, R. S. W. & OERLEMANS, J. 2005. Modelled atmospheric temperatures and globel sea levels over the past million years. *Nature,* 437/1**,** 125-128.

BLAAUW, M. 2012. Out of tune: the dangers of aligning proxy archives. *Quaternary Science Reviews,* 36**,** 38-49.

BLAAUW, M. & CHRISTEN, J. A. 2010. Random walk simulations of fossil proxy data. *The Holocene,* 20**,** 645-649.

BLAAUW, M., *ET AL*., 2010. Were the last glacial climate events simultaneous between Greenland and France?A quantitative comparison using non-tuned chronologies. *Journal of Quaternary Science,* 25**,** 387-394.

BLOCKLEY, S. P. E., *ET AL*., 2012. Synchronisation o palaeoenvironmental records over the last 60,000 years and an extended INTIMATE event stratigraphy to 48,000 b2k. *Quaternary Science Reviews,* 36**,** 2-10.

BOWEN, D. Q. 1999. Chapter 1, On the correlation and classification of Quatrnary deposits andland-sea correlations. *In:* BOWEN, D. Q. (ed.) *A Revised Correlation of Quaternary Deposits in the British Isles.* Bath, UK: The Geological Society.

BOWEN, D. Q. 2000. Revised amino stratigraphy for land-sea correlations from the northeastern North Atlantic margin. *In:* GOODFRIEND, G. A. & ET AL. (eds.) *Perspectives in Amino Acid and Protein Geochemistry.* New York: Oxford University Press.

BOWEN, D. Q., HUGES, S., SUKES, G. A. & MILLER, G. H. 1989. Land-sea correlations in the Pleistocene based on isoleucine epimerization in non-marine molluscs. *Nature,* 340**,** 49-51.

BRIANT, R. M., KILFEATHER, A. A., PARFITT, S., PENKMAN, K. E. H., PREECE, R. C., ROE, H. M., SCHWENNINGER, J. L., WENBAN-SMITH, F. F. & WHITTAKER, J. E. 2012. Integrated chronological control on an archaeologically significant Pleistocene river terrace sequence: the Thames-Medway, eastern Essex, England. *Proceedings of the Geologists' Association,* 123**,** 87-108.

BRIDGLAND, D., MADDY, D. & BATES, M. 2004a. River terrace sequences: templates for Quaternary geochronology and marine -terrestrial correlation. *Journal of Quaternary Science,* 19**,** 203-218.

BRIDGLAND, D. E. 1994. *Quaternary of the Thames,* London, Chapman & Hall.

BRIDGLAND, D. R. 2006. The Middle and Upper Pleistocene sequence in the Lower Thames: a record of Milankovitch climatic fluctuation and early human occupation of southern Britain: Henry Stopes Memorial Lecture 2004. *Proceedings of the Geologists' Association,* 117**,** 281-305.

BRIDGLAND, D. R., SCHREVE, D. C., KEEN, D. H., MEYRICK, R. & WESTAWAY, R. 2004b. Biostratigraphical correlation between the late Quaternary sequence of the Thames and key fluvial localities in central Germany. *Proceedings of the Geologists' Association,* 115**,** 125-140.

BRIGHT, J. & KAUFMAN, D. S. 2011. Amino acid racemization in lacustrine ostracodes, part I: effect of oxidizing pre-treatments on amino acid composition. *Quaternary Geochronology,* 6**,** 154-173.

BRIGHT, J., KAUFMAN, D. S., FORMAN, S. L., MCINTOSH, W. C., MEAD, J. I. & BAEZ, A. 2010. Comparative dating of a Bison-bearing late-Pleistocene deposit, Térapa, Sonora, Mexico. *Quaternary Geochronology,* 5**,** 631-643.

BRONK RAMSEY, C., HIGHAM, T. & LEACH, P. 2004. Towards High-precision AMS: progress and limitations. *Radiocarbon,* 46**,** 17-24.

BROOKMAN, B. 1998. Guidelines for the in-house production of reference materials. LGC/VAM/1998/040.

BROOKS, A. S., HARE, P. E., KOKIS, J. E., MILLER, G. H., ERNST, R. D. & WENDOF, F. 1990. Dating Pleistocene archaeological sites by protein diagenesis in ostrich egg shell. *Science,* 248**,** 60-64.

BUCK, C. E. 2004. Bayesian Chronological Data Interpretation: Where now? *In:* BUCK, C. E. & MILLARD, A. R. (eds.) *Tools for constructing Chronologies.* USA: Springer-Verlag London Ltd.

BUCK, C. E., CAVANAGH, W. G. & LITTON, C. D. 1996. *Bayesian Approach to Interpreting Archaeology,* England John Wiley & Sons Ltd.

CATT, J. A., GIBBARD, P. L., LOWE, J. J., MCCARROLL, D., SCOURSE, J. D., WALKER, M. J. C. & WYMER, J. J. 2006. Chapter 17 Quaternary: ice sheets and their legacy. *In:* BENCHLEY, P. J. & RAWSON, P. F. (eds.) *The Geology of England and Wales.* 2nd ed. London: The Geological Society.

CAYRE, O., *ET AL*., 1999. Paleoceanographic reconstrustions from planktonic foraminifera off the Iberian margin: temperature , salinity and Heinrich events. *Paleogeography,* 14**,** 384-396.

CHARMAN, D., *ET AL*., 2009. Climate drivers for peatland palaeoclimate records. *Quaternary Science Reviews,* 28**,** 1811-1819.

CLARKE, S. J., MILLER, G. H., MURRAY-WALLACE, C. V., DAVID, B. & PASVEER, J. M. 2007. The geochronological potential of isoleucine epimerisation in cassowary and megapode eggshells from archaeological sites. *Journal of Archaeological Science,* 34**,** 1051-1063.

CLARKE, S. J. & MURRAY-WALLACE, C. V. 2006. Mathematical expressions used in amino acid racemisation geochronology - A review. *Quaternary Geochronology,* 1**,** 261-278.

COHEN, K. M. & GIBBARD, P. 2011. Global chronostratigraphical correlation table for the last 2.7 million years. *Subcommission on Quaternary Stratigraphy (International Commission on Stratigraphy).* Cambridge, England.

COLLINS, M. J. & RILEY, M. S. 2000. Chpt 11, Amino acid racemization in biominerals: the impact of protein degradation. *In:* GOODFRIEND, G. A., *ET AL*. (ed.) *Perspectives in Amino Acid and Protein Geochemistry.*

CONWAY, D. & LIBBY, W. F. 1958. The measurement o very slow reaction rates; decarboxylation of alanine. *Journal of Americal Chemical Society,* 80**,** 1077-1084.

CURRELL, G. & DOWMAN, A. 2005. *Essential Mathematics and Statistics for Science,* Chichester, John Wiley & Sons Ltd.

DA SILVA, R., SANTOS, J. & CAMÕES, M. 2006. A new terminology for the approaches to the quantification of the measurement uncertainty. *Accreditation and Quality Assurance: Journal for Quality, Comparability and Reliability in Chemical Measurement,* 10**,** 664-671.

DANSGAARD, W., *ET AL.,* 1993. Evidence for general instability of past climatefrim 250-kyr ice core record. *Nature,* 364**,** 218-220.

DE BIÈVRE, P. 2006. Accuracy versus uncertainty. *Accreditation and Quality Assurance: Journal for Quality, Comparability and Reliability in Chemical Measurement,* 10**,** 645-656.

DE BIÈVRE, P. 2008. “Errors” continue to lead a persisting life. *Accreditation and Quality Assurance: Journal for Quality, Comparability and Reliability in Chemical Measurement,* 13**,** 429-430.

DEAGLE, B. E., EVESON, J. P. & JARMAN, S. N. 2006. Quantification of damage in DNA recovered from highly degraded samples - a case study on DNA in faeces. *Frontiers in Zoology,* 3:11**,** np.

DEMARCHI, B., COLLINS, M. J., TOMIAK, P. J., DAVIES, B. J. & PENKMAN, K. E. H. 2013. Intra-crystalline protein diagenesis (IcPD) in Patella vulgata. Part II: Breakdown and temperature sensitivity. *Quaternary Geochronology. **In Press***.

DEMARCHI, B., WILLIAMS, M. G., MILNER, N., RUSSELL, N., BAILEY, G. & PENKMAN, K. 2011. Amino acid racemization dating of marine shells: A mound of possibilities. *Quaternary International,* 239**,** 114-124.

DÉSENFANT, M. & PRIEL, M. 2006. Road map for measurement uncertainty evaluation. *Measurement,* 39**,** 841-848.

DOBBERSTEIN, R. C., HUPPERTZ, J., VON WURMB-SCHWARK, N. & RITZ-TIMME, S. 2008. Degradation of biomolecules in artificially and naturally aged teeth: Implications for age estimation based on aspartic acid racemization and DNA analysis. *Forensic Science International,* 179**,** 181-191.

ELIAS, S. A. 2007. Introduction: History of Quaternary Science. *In:* ELIAS, S. A. (ed.) *Encyclopedia of Quaternary Science.* Elsevier.

ELLISON, S. 2002a. Kernal.xla, version1.0e. Kernal density estimation based on RSC AMC Technical Brief No 4. *Available to download from; http://www.rsc.org/Membership/Networking/InterestGroups/Analytical/AMC/Software/index.asp*. RSC.

ELLISON, S. 2002b. Robstat.xla version 1.0. Robust Statistics Tool Kit based on RSC AMC Technical Brief No 6. *Available to download from: http://www.rsc.org/Membership/Networking/InterestGroups/Analytical/AMC/Software/index.asp*.

ELLISON, S. L. R. & WILLIAMS, A. 1998. Measurement uncertainty and its implications for collaborative study method validation and method performance parameters. *Accreditation and Quality Assurance: Journal for Quality, Comparability and Reliability in Chemical Measurement,* 3**,** 6-10.

EMILIANI, C. 1955. Pleistocene temperatures. *Journal of Geology,* 63**,** 539-578.

EMILIANI, C. 1966a. Isotopic Paleotemperatures. *Science,* 154**,** 851-857.

EMILIANI, C. 1966b. Paleaotemperature analysis of Caribbean cores P6304-8 and P6304-9 and a generalised temperature curve for the past 425,000 years. *Journal of Geology,* 74**,** 109-126.

EMONS, H. 2006. The 'RM family' - Identification of all its members. *Accreditation and Quality Assurance: Journal for Quality, Comparability and Reliability in Chemical Measurement,* 10**,** 690-691.

EPICA COMMUNITY MEMBERS 2004. Eight Glacial cycles from an Antarctic ice core. *Nature,* 429**,** 623-628.

EURACHEM 1995. *Quantifying uncertainty in analytical measurement,* London, UK (ISBN 0-948926-08-2), LGC.

EURACHEM 1998. The fitness of purpose of Analytical methods - A Laboratory Guide to method Validation and related topics. Available from; www.eurachem.org [online]: LGC, UK.

EURACHEM / CITAC 2000. Guide CG 4: Quantifying Uncertainty in Analytical Measurements. 2 ed.: Available from; http://www.citac.cc/QUAM2000-1.pdf.

EURACHEM / CITAC 2001. Guide to Quality in Analytical Chemistry, An aid to Accreditation. 2 ed.: Available from; http://www.citac.cc/.

EURACHEM/EUROLAB/CITAC/NORDTEST/AMC 2007. EURACHEM / CITAC Guide; Measurement uncertainty arising from sampling. A guide to methods and approaches. RAMSEY, M. H. and ELLISON, S. L. R. (eds). 1st ed. Available online from the Eurachem secretariat.

EUROLAB 2006. Technical Report No. 1/2006. Guide to the evaluation of measurement uncertainty for Quantitative test results. *Available from; http://www.eurolab.org/docs/technical%20report/EL_11_01_06_387%20Technical%20report%20-%20Guide_Measurement_uncertainty.pdf*.

EUROLAB 2007. Technical Report No. 1/2007. Measurement uncertainty revisited: Alternative approaches to uncertainty evaluation. *Available from; http://www.eurolab.org/pub/i_pub.html*.

FINNEY, S. C. 2010. Formal definition of the Quaternary System/Period and rededinition of the Pleistocene Series/Epoch. *Episodes,* 33**,** 159-163.

GELMAN, A., CARLIN, J. B., STERN, H. S. & RUBIN, D. B. 2004. *Texts in Statistical Science. Bayesian Data Analysis. 2nd Ed.*, Chapman & Hall/CRC.

GIBBARD, P. & KOLFSCHOTEN, V. 2006. The Pleistocene and Holocene Epochs. *In:* GRADSTEIN, F. M., OGG, J. G. & SMITH, A. G. (eds.) *A Geological Time Scale 2004.* Cambridge, UK: Cambridge University Press.

GIBBARD, P. L. 1994. *Pleistocene History of the Lower Thames Valley,* Cambridge, UK, Cambridge University Press.

GIBBARD, P. L. 2007. Climatostratigraphy. *In:* ELIAS, S. A. (ed.) *Encyclopedia of Quaternary Science.* Oxford Uk: Elsevier B V.

GIBBARD, P. L., BOREHAM, S., ANDREWS, J. E. & MAHER, B. A. 2010. Sedimentation, geochemistry and palaeomagnetism of the West Runton Freshwater Bed, Norfolk, England. *Quaternary International,* 228**,** 8-20.

GIBBARD, P. L. & HEAD, M. J. 2009a. The definition of the Quaternary system/Era and Pleistocene Series/Epoch. *Quaternaire,* 20**,** 125-133.

GIBBARD, P. L. & HEAD, M. J. 2009b. IUGS ratification of the Quaternary System/Period and redininition of the Pleistocene Series/Epoch with a base at 2.58 Ma. *Quaternaire,* 20**,** 411-412.

GIBBARD, P. L. & HEAD, M. J. 2010. The newly ratified definition of the Quaternary System/Period and redefinition of the Pleistocene Series/Epoch, and comparison of proposals advanced prior to formal ratification. *Episodes,* 33**,** 152-158.

GIBBARD, P. L., HEAD, M. J., WALKER, M. J. C. & SUBCOMMISSION ON QUATERNARY STRATIGRAPHY 2009. Formal ratification of the Quaternary System/Period and the Pleistocene Series/Epoch with a base at 2.58 Ma. *Journal of Quaternary Science,* 25**,** 96-102.

GLASS, G. V., PECKHAM, P. D. & SANDERS, J. R. 1972. Consequences of failure to meet assumptions underlying fixed effects analyses of variance and covariance. *Rev. Educ. Res.,* 42**,** 237-288.

GOODFRIEND, G. A., KASHGARIAN, M. & HARASEWYCH, M. G. 1995. Use of aspartic acid racemization and post-bomb 14C to reconstruct growth rate and longevity of the deep-water slit shell Entemnotrochus adansonianus. *Geochimica et Cosmochimica Acta,* 59**,** 1125-1129.

GOODFRIEND, G. A. & MEYER, V. R. 1991. A comparative study of the kinetics of amino acid racemization/epimerization in fossil and modern mollusk shells. *Geochimica et Cosmochimica Acta,* 55**,** 3355-3367.

GRIFFIN, R. C., PENKMAN, K. E. H., MOODY, H. & COLLINS, M. J. 2010. The impact of random natural variability on aspartic acid racemization ratios in enamel from different types of human teeth. *Forensic Science International,* 200**,** 148-152.

GRIP PROJECT MEMBERS 1993. Climate instability during the last interglacial period recorded in the GRIP ice core. *Nature,* 364**,** 203-207.

GRØN, C., HANSEN, J. B., MAGNUSSON, B., NORDBOTTEN, A., KRYSELL, M., ANDERSEN, K. J. & LUND, U. 2006. Uncertainty from Sampling - A Nordtest Handbook for sampling planners on sampling quality assurance and uncertainty estimation. Nordtest Report 604, 1st Ed. *Available from, www.nordicinnovation.net/nordtest.cfm [online].* Nordic Innovation Centre.

HAMMER, C. U., *ET AL*., 1997. Greenland Summit Ice Cores. Greenland Ice sheet Project 2/Greenland Ice Core Project. *Journal of Geophysical Research,* 102**,** 26,315-26,886.

HARE, P. E. 1969. Geochemistry of proteins, peptides and amino acids. *In:* EGLINTON, G. & MURPHY, M. T. J. (eds.) *Organic Geochemistry.* Berlin: Springer.

HARE, P. E. & ABELSON, P. H. 1968. Racemization of amino acids in fossil shells. *Carnegie Institute of Washington Year Book***,** 205-208.

HARE, P. E. & MITTERER, R. M. 1967. Nonprotein amino acids in fossil shells. *Carnegie Institute of Washington Year Book,* 65**,** 362-364.

HARE, P. E. & MITTERER, R. M. 1969. Laboratory simulation of amino acid diagenesis in fossils. *Carnegie Institute of Washington Year Book***,** 205-208.

HARWELL, M. R., RUBINSTEIN, E. N., HAYES, W. S. & OLDS, C. C. 1992. Summarizing Monte Carlo results in methodological research: the one- and two-factor fixed effects ANOVA cases. *J. Educ. Stat.,* 17**,** 315-339.

HAUGEN, J.-E. & SEJRUP, H.-P. 1992. Isoleucine epimerization kineticsin the shell of *Arctica islandica*. *Norsk Geologisk Tidsskrift,* 72**,** 171-180.

HAYES, J. D., IMBRIE, J. & SHACKLETON, N. J. 1976. Variations in the earth's orbit: Pacemaker of the Ice Ages. *Science,* 1941121-1132.

HAYES, J. D., IMBRIE, J., AND SHACKLETON, N.J., 1976. Variation in the Earth's Orbit: Pacemaker of the ice ages. *Science,* 194**,** 1121-1132.

HEARTY, P. J. & KAUFMAN, D. S. 2009. A Cerion-based chronostratigraphy and age model from the central Bahama Islands: Amino acid racemization and 14C in land snails and sediments. *Quaternary Geochronology,* 4**,** 148-159.

HEARTY, P. J., MILLER, G. H., STEARNS, C. E. & SZABO, B. J. 1986. Aminostratigraphy of of Quaternary shorelines in the Mediteranean basin. *Geological Society of America Bulletin,* 97**,** 850-858.

HEATON, T. J., BLACKWELL, P. G. & BUCK, C. E. 2009. A Bayesian approach to the estimation of radiocarbon calibration curves: the IntCal09 methodology. *Radiocarbon,* 51**,** 1151-64.

HEDBERG, H. D. 1976. *Internaional Stratigraphic Guide. A Guide to stratigraphic Classification, Terminology and Procedure,* New York, Wiley.

HENDY, E. J., TOMIAK, P. J., COLLINS, M. J., HELLSTROM, J., TUDHOPE, A. W., LOUGH, J. M. & PENKMAN, K. E. H. 2012. Assessing amino acid racemization variability in coral intra-crystalline protein for geochronological applications. *Geochimica et Cosmochimica Acta,* 86**,** 338-353.

HILL, A. R. C. & VON HOLST, C. 2001a. A comparison of simple statistical methods for estimating analtical uncertainty, talking into account predicted frequency distributions. *Analyst,* 126**,** 2044-2052.

HILL, A. R. C. & VON HOLST, C. 2001b. Factor transformation to produce statistics describing the uncertainty of analytical data. *Analyst,* 126**,** 2053-2060.

HORWITZ, W. 1995. IUPAC Protocol for the design, conduct and interpretation of method-performance studies.

HORWITZ, W. 1998. Uncertainty--a chemist's view. *J. AOAC Int.,* 81 (4)**,** 785-794.

HORWITZ, W. 2003. The certainty of Uncertainty. *J. AOAC Int.,* 86 (1)**,** 109-111.

HOVIND, H., MAGNUSSON, B., KRYSELL, M., LUND, U. & IRMA MÄKINEN 2007. Internal Quality Control, Handbook for Chemical Laboratories (Trollbboken - Troll book), Nordtest Report TR 569, 3rd Edition,. *Available from, www.nordicinnovation.net/nordtest.cfm [online].* Nordic Innovation Centre.

HUND, E., MASSART, D. L. & SMEYERS-VERBEKE, J. 2001. Operational definitions of uncertainty. *Trends in Analytical Chemistry,* 20 (8)**,** 394-406.

HUND, E., MASSART, D. L. & SMEYERS-VERBEKE, J. 2003. Comparison of different approaches to estimate the uncertainty of a liquid chromatographic assay. *Analytica Chimica Acta,* 480**,** 39-52.

HUNTLEY, J. W., KAUFMAN, D. S., KOWALEWSKI, M., ROMANEK, C. S. & NEVES, R. J. 2012. Sub-centennial resolution amino acid geochronology for the freshwater mussel Lampsilis for the last 2000 years. *Quaternary Geochronology,* 9**,** 75-85.

IMBRIE, J., *ET AL*., 1984. The orbital theory of Pleistocene climate: support form a revised chronolgy of the marine δ10Ο ρεχορδ. *In:* BERGER, A. & ET AL. (eds.) *Milankovitch and climate.* Reidel, Dordrecht.

IMBRIE, J. & IMBRIE, K. P. 1979. *Ice Ages: Solving the Mystery.,* London, MacMillan.

ISO 5725-1 1994. Accuracy (trueness and precision) of measurement methods and results. Part 1: General principls and definitions.: International Standards Organisation.

ISO 5725-2 1994. Accuracy (trueness and precision) of measurement methods and results. Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method.: International Standards Organisation.

ISO 5725-3 1994. Accuracy (trueness and precision) of measurement methods and results. Part 3: Intermediate measures of the precision of a standard measurement method.: International Standards Organisation.

ISO 5725 1994. Accuracy (trueness and presicion) of measurement methods and results - Parts 1-6.: International Standards Organisation.

ISO 21748 2010. Guidance for the use of repeatability, reproducibility and trueness estimates in measurement uncertainty estimation. International Standards Organisation.

ISO / IEC 17025 2005. General requirements for the competence of testing and calibration laboratories. International Standards Organisation.

JCGM 100 2008. Evaluation of measurement data - Guide to the expression of uncertainty in measurement (GUM). 1 ed.: Available from; http://www.bipm.org/utils/common/documents/jcgm/JCGM_100_2008_E.pdf.

JCGM 101 2008. Evaluation of measurement data - Supplement 1 to the "Guide to the expression of uncertainty in measurement" - Propagation of distributions using a Monte Carlo method.

JCGM 200 2008. International Vocabulary of Metrology - Basic and general concepts and associated terms (VIM). Available from; http://www.bipm.org/en/publications/guides/vim.html

JOHNSEN, S., *ET AL*., 2001. Oxygen isotope and palaeotemperature records from six Greenland ice-core stations: Camp Century, Dye-3, GRIP, GISP2, Renland and northGRIP. *Journal of Quaternary Science,* 16**,** 299-308.

JULICHER, B., GOWIK, P. & UHLIG, S. 1999. A top-down in-house validation based approach for the investigation of the measurement uncertainty using fractional factorial experiments. *Analyst,* 124**,** 537-545.

KAUFMAN, D. S. 2000. Amino acid racemization in ostracodes. *In:* GOODFRIEND, G. A., COLLINS, M. J., FOGEL, M. L., MACKO, S. A. & WEHMILLER, J. F. (eds.) *Perspectives in Amino Acid and protein Geochemistry.* Oxford, UK: Oxford University Press Inc.

KAUFMAN, D. S. 2003. Amino acid paleothermometry of Quaternary ostracodes from the Bonneville Basin, Utah. *Quaternary Science Reviews,* 22**,** 899-914.

KAUFMAN, D. S. 2006. Temperature sensitivity of aspartic and glutamic acid racemization in the foraminifera Pulleniatina. *Quaternary Geochronology,* 1**,** 188-207.
KAUFMAN, D. S. & MANLEY, W. F. 1998. A New Procedure for determining DL amino acid ratios in fossils using reverse phase liquid chromatography. *Quaternary Geochronology,* 17**,** 987-1000.
KAUFMAN, D. S. & MILLER, G. H. 1992. Overview of amino acid geochronology. *Comparative Biochemistry and Physiology Part B: Comparative Biochemistry,* 102**,** 199-204.
KING JR., K. & HARE, P. E. 1972. Amino acid composition as a test of taxonomic character for living and fossil foraminifera. *Micropalaeontology,* 18**,** 285-293.
KIRKUP, L. & FRENKEL, R. B. 2006. *An Introduction to Uncertainty in Measurement*, Cambridge University Press.
KOSNIK, M. A. & KAUFMAN, D. S. 2008. Identifying outliers and assessing the accuracy of amino acid racemization measurements for geochronology: II. Data screening. *Quaternary Geochronology,* 3**,** 328-341.
KOSNIK, M. A., KAUFMAN, D. S. & HUA, Q. 2008. Identifying outliers and assessing the accuracy of amino acid racemization measurements for geochronology: I. Age calibration curves. *Quaternary Geochronology,* 3**,** 308-327.
KOSNIK, M. A., KAUFMAN, D. S. & HUA, Q. 2013. Radiocarbon-calibrated multiple amino acid geochronology of Holocene molluscs from Bramble and Rib Reefs (Great Barrier Reef, Australia). *Quaternary Geochronology,* 16**,** 73-86.
KRAGTEN, J. 1994. Calculating standard deviations and confidence intervalswith a universally applicable spreadsheet technique. *Analyst,* 119**,** 2161-2166.
KRIAUSAKUL, N. & MITTERER, R. M. 1978. Isoleucine epimerization in peptides and proteins: kinetic factors and application to fossil proteins. *Science,* 201**,** 1011-1014.
KRUSCHKE, J. K. 2011. *Doing Bayesian Data Analysis, A tutorial with R and Bugs,* USA, Elsevier Inc.
KUKLA, G. J. 1977. Pleistocene land-sea correlations, I, Europe. *Earth Science Reviews,* 13**,** 307-74.
KVENVOLDEN, K. A. 1980. Interlaboratory Comparison of Amino Acid Racemization in Pleistocene Mollusk, *Saxidomus giganteus*. *In:* HARE, P. E., HOERING, T.C., AND KING, K. (ed.) *Biogeochemistry of Amino Acids.* USA: John Wiley & Sons Inc.
KVENVOLDEN, K. A., PETERSON, E., WEHMILLER, J. F. & HARE, P. E. 1973. Racemization of amino acids in marine sediments by Gas Chromatography. *Geochimica et Cosmochimica Acta,* 37**,** 2215-2225.
LIRA, I. 2002. *Evalutaing the Measurement Uncertainty; Fundamentals and practical guidance,* Bristol and Philadelphia, Institute of Physics Publishing Ltd.
LISIECKI, L. E. & RAYMO, M. E. 2005. A Pliocene-Pleistocene stack of 57 globally distributed benthic $\delta^{18}O$ records. *Paleoceanography,* 20**,** 1003.
LIX, L. M., KESELMAN, J. C. & KESELMAN, H. J. 1996. Consequences of assumption violations revisited: A quantitative review of alternatives to the one-way analysis of variance F test. *Rev. Educ. Res.,* 66**,** 579-619.
LOURENS, L., HILGEN, F., SHACKLETON, N. J., LASKAR, J. & WILSON, D. 2005. The Neogene Period. *In:* GRADSTEIN, F. M., OGG, J. G. & SMITH, A. G. (eds.) *A Geological Time Scale 2004.* Cambridge UK: Cambridge University Press.
LOURENS, L. J. 2008. On the Neogene-Quaternary debate. *Episodes,* 31**,** 239-242.
LOWE, J. J. & WALKER, M. J. C. 1997. *Reconstructing Quaternary Environments,* Essex, Longman Group Ltd.
MACLEOD, N. 2004a. Prospectus & Regression 1. *Palaeo-math 101. Available from; http://www.palass.org/modules.php?name=palaeo_math&page=1:* The Palaeontological Association.

MACLEOD, N. 2004b. Regression 2. *Palaeo-math 101. Available from; http://www.palass.org/modules.php?name=palaeo_math&page=1:* The Palaeontological Association.

MAGEE, J. W., MILLER, G. H., SPOONER, N. A., QUESTIAUX, D. G., MCCULLOCH, M. T. & CLARK, P. A. 2009. Evaluating Quaternary dating methods: Radiocarbon, U-series, luminescence, and amino acid racemization dates of a late Pleistocene emu egg. *Quaternary Geochronology,* 4**,** 84-92.

MAGNUSSON, B., NAYKKI, T., HOVIND, H. & KRYSELL, M. 2004. NORDTEST Report TR 537. Handbook for calculation of measurement uncertainty in Environmental Laboratories. *Available from; http://www.nordicinnovation.net/nordtestfiler/tec537.pdf*. 2 ed.

MANLEY, W. F. & MILLER, G. H. 2000. Kinetics of aspartic acid racemization in *Mya* and *Hiatella*: modelling age and palaeotemperature of high-latitude Quaternary molluscs. *In:* GOODFRIEND, G. A., COLLINS, M. J., FOGEL, M. L., MACKO, S. A. & WEHMILLER, J. F. (eds.) *Perspectives in Amino Acid and protein Geochemistry.* Oxford, UK: Oxford University Press Inc.

MARKOVIĆ, S. B., HAMBACH, U., STEVENS, T., KUKLA, G. J., HELLER, F., MCCOY, W. D., OCHES, E. A., BUGGLE, B. & ZÖLLER, L. 2011. The last million years recorded at the Stari Slankamen (Northern Serbia) loess-palaeosol sequence: revised chronostratigraphy and long-term environmental trends. *Quaternary Science Reviews,* 30**,** 1142-1154.

MAROTO, A., RIU, J., BOQUÉ, R. & XAVIER RIUS, F. 1999. Estimating uncertainties of analytical results using information from the validation process. *Analytica Chimica Acta,* 391**,** 173-185.

MARSCHAL, A. 2004. Measurement uncertainties and specified limits: what is logical or common sense in chemical measurement? *Accreditation and Quality Assurance,* 9**,** 642-643.

MARTINSON, D. G., *ET AL*., 1987. Age dating and the orbital theory of the ice ages: development of a high-resolution 0-300,000 year chronstratigraphy. *Quaternary Research,* 27**,** 1-29.

MAUL, L. C. & PARFITT, S. A. 2010. Micromammals from the 1995 Mammoth Excavation at West Runton, Norfolk, UK: Morphometric data, biostratigraphy and taxonomic reappraisal. *Quaternary International,* 228**,** 91-115.

MCCOY, W. D. 1987. The precision of amino acid geochronology and paleothermometry. *Quaternary Science Reviews,* 6**,** 43-54.

MCDONALD, J. H. 2009. Handbook of Biological Statistics. *Available from; http://www.lulu.com/product/5507346*. 2nd ed. Baltimore, Maryland: Sparky House Publishing.

MCMILLAN, A. A., HAMBLIN, R. J. O. & MERRITT, J. W. 2005. An overview of the lithostratigraphical framework for the Quaternary and Neogene deposits of Great Britain. Nottingham, UK: British Geological Survey.

MEESE, D. A., *ET AL*., 1997. The Greenland Ice sheet Project 2 depth-age scale:methods and results. *Journal of Geophysical Research,* 102**,** 26,411-26,423.

MESLEY, R. J., POCKLINGTON, W. D. & WALKER, R. F. 1991. Analytical quality assurance - a review. *Analyst,* 116**,** 975-990.

MILLER, G. H. 1985. Aminostratigraphy of Baffin Island shell-bearing deposits *In:* ANDREWS, J. T. (ed.) *Quaternary Environments, the Eastern Canadian Arctic,, Baffin Island and west Greenland.* Winchester Massachusetts: Allen & Unwin.

MILLER, G. H., BEAUMONT, P. B., JULL, A. J. T. & JOHNSON, B. 1992. Pleistocene geochronology and palaeothermometry from protein diagenesis in ostrich eggshells: implications for the evloution of modern humans. *Philosphical Transactions of the Royal Society of London,* B 337**,** 149-157.

MILLER, G. H. & CLARKE, S. J. 2007. AMINO-ACID DATING. *In:* EDITOR-IN-CHIEF: SCOTT, A. E. (ed.) *Encyclopedia of Quaternary Science.* Oxford: Elsevier.

MILLER, G. H., *ET AL*., 2000. Isoleucine epimerization in eggshells of the flightless Australian birds *Genyornis* and *Dromaius*. *In:* GOODFRIEND, G. A., COLLINS, M. J., FOGEL, M. L., MACKO, S. A. & WEHMILLER, J. F. (eds.) *Perspectives in Amino Acid and protein Geochemistry.* Oxford, UK: Oxford University Press Inc.

MILLER, G. H. & HARE, P. E. 1980. Amino acid geochronology: integrity of carbonate matrix and potential of molluscan fossils. *In:* HARE, P. E., HOERING, T. C. & K KING JR. (eds.) *Biogeochemistry of amino acids.* New York: John Wiley & Sons.

MILLER, G. H., HOLLIN, J. T. & ANDREWS, J. T. 1979. Aminostratigraphy of Pleistocene Deposits. *Nature,* 281**,** 539-543.

MILLER, G. H., MAGEE, J. W., JOHNSON, B. J., FOGEL, M. L., SPOONER, N. A., MCCULLOCH, M. T. & AYCLIFFE, L. K. 1999. Pleistocene extinction of *Genyornis newtoni*: human impact on Australian megafauna. *Science,* 283**,** 205-208.

MILLER, J. N. & MILLER, J. C. 2005. *Statistics and Chemometrics for Analytical Chemistry,* Harlow, England., Pearson Education Ltd.

MITCHELL, G. F., PENNY, L. F., SHOTTEN, F. W. & WEST, R. G. 1973. A correlation of Quaternary deposits in the Britsh Isles. *Special Report of the Geological Society of London* Edinburgh.

MITTERER, R. M. & KRIAUSAKUL, N. 1989. Calculation of amino acid racemization ages based on apparent parabolic kinetics. *Quaternary Science Reviews,* 8**,** 353-357.

MONEGATTI, M. & RAFFI, S. 2001. Taxonomic diversity and stratigraphic distribution of Mediterranean Pliocene bivalves. *Palaeogreography, Palaeoclimatology,Palaeoecology,* 165**,** 171-193.

MONTGOMERY, D. C. 2001. Discovering dispersion effects. *In:* MONTGOMERY, D. C. (ed.) *Design and Analysis of Experiments.* 5th ed. New york: John Wiley.

MORIGI, A., SCHREVE, D. & WHITE, M. 2011. Part 1 - The Ice Ages: Palaeogeography, Palaeolithic Archaeology and Pleistocene Environments. *In:* DODD, A. (ed.) *Thames Through Time, The Archaeology of the gravel terraces of the upper and middle Thames.* Oxford, UK: Oxford Archaeology.

MURRAY-WALLACE, C. V. 1995. Aminostratigraphy of Quaternary coastal sequences in southern Australia — An overview. *Quaternary International,* 26**,** 69-86.

MURRAY-WALLACE, C. V., BOURMAN, R. P., PRESCOTT, J. R., WILLIAMS, F., PRICE, D. M. & BELPERIO, A. P. 2010. Aminostratigraphy and thermoluminescence dating of coastal aeolianites and the later Quaternary history of a failed delta: The River Murray mouth region, South Australia. *Quaternary Geochronology,* 5**,** 28-49.

NAYLOR, J. C. & SMITH, A. F. M. 1988. An archaeological inference problem. *Journal of the America Statistical Association,* 83**,** 588-595.

NEAVE, H. R. 1978. *Statistics Tables: For mathematicians, Engineers, Economists and the Behavioural and Management Sciences*, Routledge.

NELSON, A. R. 1978. *Quaternary glacial and marine stratigraphy of the Qivitu Peninsula, northern Cumberland Peninsula, Baffin Island, Canada. .* Ph.D. Dissertation, Univesity of Colorado, Boulder, CO.

NMS. accessed 2009a. What is measurement uncertainty and why is it important? *National Measurement System Chemical and Biological Metrology Website.* http://www.nmschembio.org.uk/GenericArticle.aspx?m=399&amid=5715 [Online].

NMS. accessed 2009b. Introduction to quality assurance training resource. *National Measurement System Chemical and Biological Metrology Website.* http://www.nmschembio.org.uk/GenericArticle.aspx?m=456&amid=4650 [Online].

NMS. accessed 2009c. Key Topics: The VAM Principles. *National Measurement System Chemical and Biological Metrology Website.* http://www.nmschembio.org.uk/GenericListing.aspx?m=108 [Online].

NORTH GREENLAND ICE CORE PROJECT MEMBERS 2004. High-resolution record of northern hemisphere climate extending into the last interglacial period. *Nature,* 431**,** 147-151.

OCHES, E. A. & MCCOY, W. D. 2001. Historical developments and recent advances in amino acid geochronology applied to loess research: examples from North America, Europe, and China. *Earth-Science Reviews,* 54**,** 173-192.

ORTIZ, J. E., TORRES, T., JULIÀ, R., DELGADO, A., JUAN LLAMAS, F., SOLER, V. & DELGADO, J. 2004. Numerical dating algorithms of amino acid racemization ratios from continental ostracodes. Application to the Guadix-Baza Basin (southern Spain). *Quaternary Science Reviews,* 23**,** 717-730.

OWEN, L. A., BRIGHT, J., FINKEL, R. C., JAISWAL, M. K., KAUFMAN, D. S., MAHAN, S., RADTKE, U., SCHNEIDER, J. S., SHARP, W., SINGHVI, A. K. & WARREN, C. N. 2007. Numerical dating of a Late Quaternary spit-shoreline complex at the northern end of Silver Lake playa, Mojave Desert, California: A comparison of the applicability of radiocarbon, luminescence, terrestrial cosmogenic nuclide, electron spin resonance, U-series and amino acid racemization methods. *Quaternary International,* 166**,** 87-110.

PARFITT, S. A., BARENDREGT, R. W., BREDA, M., CANDY, I., COLLINS, M. J., COOPE, G. R., DURBRIDGE, P., FIELD, M. H., LEE, J. R., LISTER, A. M., MUTCH, R., PENKMAN, K. E. H., PREECE, R. C., ROSE, J., STRINGER, C. B., SYMMONS, R., WHITTAKER, J. E., WYMER, J. J. & STUART, A. J. 2005. The earliest record of human activity in northern Europe. *Nature,* 438**,** 1008-1012.

PARNELL, A. C., *ET AL*., 2008. A flexibl approach to assessing synchroneity of past events using Bayesian reconstructions of sedimentation history. *Quaternary Science Reviews,* 27**,** 1872-1885.

PENKMAN, K. E. H. 2005. *Amino acid geochronology: a closed system approach to test and refine the UK model.*, Unpublished PhD thesis, University of Newcastle.

PENKMAN, K. E. H., KAUFMAN, D. S., MADDY, D. & COLLINS, M. J. 2008. Closed-system behaviour of the intra-crystalline fraction of amino acids in mollusc shells. *Quaternary Geochronology,* 3**,** 2-25.

PENKMAN, K. E. H., PREECE, R. C., BRIDGLAND, D. R., KEEN, D. H., MEIJER, T., PARFITT, S. A., WHITE, T. S. & COLLINS, M. J. 2011. A chronological framework for the British Quaternary based on *Bithynia* opercula. *Nature,* 476**,** 446-449.

PENKMAN, K. E. H., PREECE, R. C., KEEN, D. H. & COLLINS, M. J. 2010. Amino acid geochronology of the type Cromerian of West Runton, Norfolk, UK. *Quaternary International,* 228**,** 25-37.

PENKMAN, K. E. H., PREECE, R. C., KEEN, D. H., MADDY, D., SCHREVE, D. C. & COLLINS, M. J. 2007. Testing the aminostratigraphy of fluvial archives: the evidence from intra-crystalline proteins within freshwater shells. *Quaternary Science Reviews,* 26**,** 2958-2969.

PETIT, J. R., *ET AL*., 1999. 420,000 years of climate and atmospheric history revealed by the Vostock dep Antarctic Ice Core. *Nature,* 399**,** 429-436.

POINAR, H. N., HOSS, M., BADA, J. L. & PAABO, S. 1996. Amino Acid Racemization and the preservation of Ancient DNA. *Science,* 272**,** 864.

POWELL, J., COLLINS, M. J., CUSSENS, J., MACLEOD, N. & PENKMAN, K. E. H. 2013. Results from an amino acid racemization inter-laboratory proficiency study; design and performance evaluation. *Quaternary Geochronology,* 16**,** 183-197.

POWELL, J. & OWEN, L. 2002. Reliability of Food Measurements: The Application of Proficiency Testing to GMO Analysis. *Accreditation and Quality Assurance: Journal for Quality, Comparability and Reliability in Chemical Measurement,* 7**,** 392-402.

QUEVAUVILLER, P. 1998. Requirements for production and use of Certified Reference Materials for speciation analysis: A European Commisssion perspective. *Spectrochimica Acta Part B:,* 53**,** 1261-1279.

RASMUSSEN, C. E. & WILLIAMS, K. I. 2006. *Gaussian Processes for Machine Learning, Available from; www.GaussianProcess.org/gpml [online]*, the MIT Press.

REICHERT, K. L., LICCIARDI, J. M. & KAUFMAN, D. S. 2011. Amino acid racemization in lacustrine ostracodes, part II: Paleothermometry in Pleistocene sediments at Summer Lake, Oregon. *Quaternary Geochronology,* 6**,** 174-185.

REIMER, P. J., BAILLIE, M. G. L., BARD, E., BAYLISS, A., BECK, J. W., BLACKWELL, P. G., BRONK RAMSEY, C., BUCK, C. E., BURR, G. S., EDWARDS, R. L., FRIEDRICH, M., GROOTES, P. M., GUILDERSON, T. P., HAJDAS, I., HEATON, T. J., HOGG, A. G., HUGHEN, K. A., KAISER, K. F., KROMER, B., MCCORMAC, F. G., MANNING, S. W., REIMER, R. W., RICHARDS, D. A., SOUTHON, J. R., TALAMO, S., TURNEY, C. S. M., VAN DER PLICHT, J. & WEYHENMEYER, C. E. 2009. IntCal09 and Marine09 radiocarbon age calibration curves, 0-50,000 years cal BP. *Radiocarbon,* 51**,** 1111-50.

RENFREW, C. & BAHN, P. 2012. *Archaeology; Theories, Methods and Practice,* London.

RINK, W. J., SCHWARCZ, H. P., STUART, A. J., LISTER, A. M., MARSEGLIA, E. & BRENNAN, B. J. 1996. ESR dating of the type Cromerian freshwater bed at West Runton, U.K. *Quaternary Science Reviews,* 15**,** 727-738.

RIO, D., SPROVIERI, R., CASTRADORI, D. & DI STEFANO, E. 1998. The Gelasian Stage (Upper Pliocene): A new unit of the global standard chronostratigraphic scale. *Episodes,* 21**,** 82-87.

ROVERI, M. & TAVIANI, M. 2003. Calcarenite and sapropel deposition in the Mediterranean Pliocene: shallow- and deep-water record of astronomical driven climate events. *Terra Nova,* 15**,** 279-286.

RSC ANALYTICAL METHODS COMMITTEE 1989. AMC Technical Briefs; Robust Statistics-how not to reject outliers: Part 1, Basic concepts. *The Analyst,* 114**,** 1693-1697.

RSC ANALYTICAL METHODS COMMITTEE 1995. Uncertainty of Measurement: Implications of its use in Analytical Science. *The Analyst,* 120**,** 2303-2308.

RSC ANALYTICAL METHODS COMMITTEE 2001. AMC Technical Briefs No 6; Robust Statistics: a method of coping with outliers. Available from; http://www.rsc.org/Membership/Networking/InterestGroups/Analytical/AMC/TechnicalBriefs.asp.

RSC ANALYTICAL METHODS COMMITTEE 2003a. AMC Technical Briefs No 13 Sept 2003. Terminology - the key to understanding analytical science. Part 1: Accuracy, precision and uncertainty. . Available from; http://www.rsc.org/Membership/Networking/InterestGroups/Analytical/AMC/TechnicalBriefs.asp.

RSC ANALYTICAL METHODS COMMITTEE 2003b. AMC Technical Briefs No 14; A glimpse into Bayesian Statistics. Available from; http://www.rsc.org/Membership/Networking/InterestGroups/Analytical/AMC/TechnicalBriefs.asp.

RSC ANALYTICAL METHODS COMMITTEE 2006. AMC Technical Briefs No 4; Representing data distributions with kernal density estimates. Available from; http://www.rsc.org/Membership/Networking/InterestGroups/Analytical/AMC/TechnicalBriefs.asp.

RSC ANALYTICAL METHODS COMMITTEE 2008a. AMC Technical Briefs No 30; The standard deviation of the sum of several variables. Available from; http://www.rsc.org/Membership/Networking/InterestGroups/Analytical/AMC/TechnicalBriefs.asp.

RSC ANALYTICAL METHODS COMMITTEE 2008b. AMC Technical Briefs No 32: Optimising your uncertainty - a case study. Available from; http://www.rsc.org/Membership/Networking/InterestGroups/Analytical/AMC/TechnicalBriefs.asp.

RSC ANALYTICAL METHODS COMMITTEE 2009. AMC Technical Briefs No 40; The duplicate method for the estimation of measurement uncertainty from sampling. *Available from; http://www.rsc.org/Membership/Networking/InterestGroups/Analytical/AMC/TechnicalBriefs.asp*.

RSC ANALYTICAL METHODS COMMITTEE 2010. AMC Technical Brief No 46: Internal Quality Control in routine analysis. *http://www.rsc.org/images/internal-quality-control-routine-analysis-technical-brief-46_tcm18-214818.pdf*.

SCHWANDER, J., *ET AL.*, 2001. A tentative chronology for the EPIVA Dome Concordia. *Geophysical Research Letters,* 28**,** 4243-4246.

SHACKLETON, N. J., BERGER, A. & PELTIER, W. A. 1990. An alternative astronomical calibration of the lower Pleistocene time scale based on ODP site 677. *Transactions of the Royal Societ of Edinburgh, Earth Sciences,* 81**,** 251-261.

SHACKLETON, N. J., *ET AL.*, 2004. Absolute calibration of the Greenland timescale:implications for Antarctic timescales for the $\Delta^{14}C$. *Quaternary Science Reviews,* 23**,** 1513-1522.

SHACKLETON, N. J. & HALL, M. A. 1989. Stable isotope history of the Pleistocene at ODP site 677. *Proceedings of the Ocean Drilling Program, Vol III. College Station, TX.*

SHACKLETON, N. J., HALL, M. A. & PATE, D. 1995. Pliocene stable isotope stratigraphy pf Site 846. *In:* PISIAS, N. G. & ET AL. (eds.) *Proceedings of the Ocean Drilling Program, Scientific Results, Vol 138 College Station TX.*

SHACKLETON, N. J. & OPDYKE, N. D. 1973. Oxygen isotope and paleomagnetic stratigraphy of equatorial Pacific Core V28-238: oxygen isotope temperatures and ive volumes on a 10^5 and 10^6 year scale. *Journal of Quaternary Research,* 3**,** 39-55.

SMITH, C. I., CHAMBERLAIN, A. T., RILEY, M. S., STRINGER, C. & COLLINS, M. J. 2003. The thermal history of human fossils and the likelihood of successful DNA amplification. *Journal of Human Evolution,* 45**,** 203-217.

STRINGER, C. 2006. *Homo Britannicus,* London, Penguin Books Ltd.

STRINGER, C. 2011. Chapter 1 - The Changing Landscapes of the Earliest Human Occupation of Britain and Europe. *In:* NICK ASHTON, S. G. L. & CHRIS, S. (eds.) *Developments in Quaternary Sciences.* Elsevier.

STUIVER, M. & GROOTES, P. M. 2000. GISP2 oxygen isotope ratios. *Quaternary Research,* 53**,** 266-284.

SVENSSON, A., ANDERSEN, K. K., BIGLER, M., CLAUSEN, H. B., DAHL-JENSEN, D., DAVIES, S. M., JOHNSEN, S. J., MUSCHELER, R., RASMUSSEN, S. O., RÖTHLISBERGER, R., PEDER STEFFENSEN, J. & VINTHER, B. M. 2006. The Greenland Ice Core Chronology 2005, 15-42 ka. Part 2: comparison to other records. *Quaternary Science Reviews,* 25**,** 3258-3267.

SYKES, G. A., COLLINS, M. J. & WALTON, D. I. 1995. The significance of a geochemically isolated intracrystalline organic fraction within biominerals. *Organic Geochemistry,* 23**,** 1059-1065.

TAYLOR, B., N. & KUYATT, C., E., 1994. Guidelines for Evaluating and Expressing the uncertainty of NIST Measurement Results. NIST Technical Note 1297. Washington, DC. Also available online from the NIST website: US Government Printing Office.

THOMPSON, M. 1995. Editorial. Uncertainty in an uncertain world. *Analyst,* 120**,** 117N-118N.

THOMPSON, M., ELLISON, S. L. R. & WOOD, R. 2002. IUPAC Harmonized Guidelines for single-laboratory validation of methods of analysis, (Technical Report). *Pure and Applied Chemistry,* 74**,** 835-855.

THOMPSON, M. & FEARNE, T. 1996. What exactly is Fitness for Purpose in Analytical Chemistry? *Analyst,* 121**,** 275-278.

THOMPSON, M. & WOOD, R. 1995. IUPAC Harmonized Guidelines for Internal Quality Control in Analytical Chemistry Laboratories , (Technical Report). *Pure and Applied Chemistry,* 67 (4)**,** 649-666.

THOMPSON, W. G. & GOLDSTEIN, S. L. 2006. A radiometric calibration of the SPECMAP timescale. *Quaternary Science Reviews,* 25**,** 3207-3215.

TOMIAK, P. J., PENKMAN, K. E. H., HENDY, E. J., DEMARCHI, B., MURRELLS, S., DAVIS, S. A., MCCULLAGH, P. & COLLINS, M. J. 2013. Testing the limitations of artificial protein degradation kinetics using known-age massive Porites coral skeletons. *Quaternary Geochronology. **In Press***.

TUREKIAN, K. K. & BADA, J. L. 1972. Calibration of Hominid evolution: recent advances in isotopic and other dating methods applicable to the origin of man. *Proceedings of the symposium held at Burg, Wartenstein, Austria, 3rs-12th July 1971.* Edinburgh: Scottish Academic Press.

VERSTEEGH, V. J. M. 1997. The onset of major northern hemisphere glaciations and their impact on dinoflagellate cysts and acritachs from the singa section, Calabria (southern Italy) and DSDP Holes 607/607A (North Atlantic). *Marine Micropaleobtology,* 30**,** 319-43.

VISSER, R. 2004. Measurement uncertainty: practical problems encountered by accredited testing laboratories. *Accreditation and Quality Assurance: Journal for Quality, Comparability and Reliability in Chemical Measurement,* 9**,** 717-723.

VISSER, R. G. 2002. Measurement uncertainty: Opinions of the Government, the Accreditation Council and the Candidate accredited laboratory. *Accreditation and Quality Assurance: Journal for Quality, Comparability and Reliability in Chemical Measurement,* 7**,** 124-125.

WAGNER, G. A. 1998. *Age Determination of Young Rocks and Artifacts,* Germany Springer.

WALKER, M. 2005. *Quaternary Dating Methods,* Chichester, England, John Wiley & Sons ltd.

WALKER, M. 2008. *Quaternary Dating Methods,* West Sussex, UK, John Wiley & Sons Ltd.

WALKER, M., *ET AL*., 2009. Formal definition and dating of the GSSP (Global Stratotype Section and Point) for the base of the Holocene using the Greenland NGRIP ice core, and selected auxillary reords. *Journal of Quaternary Science,* 24**,** 3-17.

WALKER, M. J., LÓPEZ-MARTÍNEZ, M., CARRIÓN-GARCÍA, J. S., RODRÍGUEZ-ESTRELLA, T., SAN-NICOLÁS DEL-TORO, M., SCHWENNINGER, J. L., LÓPEZ-JIMÉNEZ, A., ORTEGA-RODRIGÁÑEZ, J., HABER-URIARTE, M., POLO-CAMACHO, J. L., GARCÍA-TORRES, J., CAMPILLO-BOJ, M., AVILÉS-FERNÁNDEZ, A. & ZACK, W. 2012. Cueva Negra del Estrecho del Río Quípar (Murcia, Spain): A late Early Pleistocene hominin site with an "Acheulo-Levalloiso-Mousteroid" Palaeolithic assemblage. *Quaternary International*.

WALKER, M. J. C., BJORK, S. & LOWE, J. J. 2001. Integration of ice core, marine and terrestrial records (INTIMATE) from around the North Atlantic regios: an introduction. *Quaternary Science Reviews,* 20.

WALKER, M. J. C., *ET AL*., 1999. Isotopic "events" in the GRIP ice core: a stratotype for the late Pleistocene. *Quaternary Science Reviews,* 18**,** 1143-1150.

WEHMILLER, J. F. 1981. Kinetic model options for interpretation of amino acid enantiomeric ratios in Quaternary mollusks: Comments on a paper by Kvenvolden et al. (1979). *Geochimica et Cosmochimica Acta,* 45**,** 261-264.

WEHMILLER, J. F. 1984. Interlaboratory comparison of amino acid enantiomeric ratios in fossil Pleistocene mollusks. *Quaternary Research,* 22**,** 109-120.

WEHMILLER, J. F. 1992. Aminostratigraphy of Southern California Quaternary Marine Terraces. *In:* FLETCHER, C. H. I. & WEHMILLER, J. F. (eds.) *Quaternary Coats of the United States: Marine and Lacustrine Systems.* Tusla, SEPM (Society for Sedimentary Geology). Special Edition No 48, p 317-321.

WEHMILLER, J. F. & HARE, P. E. 1971. Racemization of amino acids in marine sediments. *Science,* 173**,** 907-911.

WEHMILLER, J. F., HARRIS, W. B., BOUTIN, B. S. & FARRELL, K. M. 2012. Calibration of amino acid racemization (AAR) kinetics in United States mid-Atlantic Coastal Plain Quaternary mollusks using 87Sr/86Sr analyses: Evaluation of kinetic models and estimation of regional Late Pleistocene temperature history. *Quaternary Geochronology,* 7**,** 21-36.

WEHMILLER, J. F. & MILLER, G. H. 2000. Aminostratigraphic Dating Methods in Quaternary Geology. *In:* NOLLER, J. S., SOWERS, J. M., COLMAN, S. M. & PIERCE, K. L. (eds.) *Quaternary Geochronology: methods and Applications.* Washington DC: American Geophysical Union, Reference Shelf Series 4.

WEHMILLER, J. F., THIELER, E. R., MILLER, D., PELLERITO, V., BAKEMAN KEENEY, V., RIGGS, S. R., CULVER, S., MALLINSON, D., FARRELL, K. M., YORK, L. L., PIERSON, J. & PARHAM, P. R. 2010. Aminostratigraphy of surface and subsurface Quaternary sediments, North Carolina coastal plain, USA. *Quaternary Geochronology,* 5**,** 459-492.

WEHMILLER, J. F., YORK, L. L. & BART, M. L. 1995. Amino acid racemization geochronology of reworked Quaternary mollusks on U.S. Atlantic coast beaches: implications for chronostratigraphy, taphonomy, and coastal sediment transport. *Marine Geology,* 124**,** 303-337.

WERNIMONT, G. T. 1985. *Use of statistics to develop and evaluate analytical methods,* Virginia, USA, Association of Analytical Chemists.

WESTAWAY, R. 2009. Calibration of decomposition of serine to alanine in Bithynia opercula as a quantitative dating technique for Middle and Late Pleistocene sites in Britain. *Quaternary Geochronology,* 4**,** 241-259.

WESTAWAY, R., MADDY, D. & BRIDGLAND, D. 2002. Flow in the lower continental crust as a mechanism for the Quaternary uplift of south-east England: constraints from the Thames terrace record. *Quaternary Science Reviews,* 21**,** 559-603.