

Low impact surface hardness testing (Equotip) on porous surfaces – Advances in methodology with implications for rock weathering and stone deterioration research

Katrin Wilhelm^{1*}, Heather Viles¹, and Órlaith Burke²

¹ Oxford Rock Breakdown Laboratory (OxRBL), School of Geography and the
Environment, University of Oxford, Oxford OX1 3QY, UK

² Nuffield Department of Population Health, University of Oxford, Oxon OX3 7LF, UK

* katrin.wilhelm@ouce.ox.ac.uk

Abstract

The Equotip surface hardness tester is becoming a popular method for rock and stone weathering research. In order to improve the reliability of Equotip for on-site application this study tested four porous limestones under laboratory conditions. The range of stone porosity was chosen to represent likely porosities found in weathered limestones in the field. We consider several key issues: (i) its suitability for soft and porous stones; (ii) the type of probe required for specific on-site applications; (iii) appropriate (non-parametrical) statistical methods for Equotip data; (iv) sufficient sampling size. This study shows that the Equotip is suitable for soft and porous rock and stone. From the two tested probes the DL probe has some advantages over the D probe as it correlates slightly better with open porosity and allows for more controlled sampling in recessed areas and rough or curved areas. We show that appropriate sampling sizes and robust non-parametric methods for subsequent data evaluation can produce meaningful measures of rock surface hardness derived from the Equotip. The novel Hybrid dynamic hardness, a combination of two measuring procedures (single impact method (SIM) and repeated impact method (RIM)), has been adapted and is based on median values to provide a more robust data evaluation. For the tested stones in this study we propose a sample size of 45 readings (for a confidence level of 95%). This approach can certainly be transferred to stone and rock with similar porosities and hardness. Our approach also allows for consistent comparisons to be made across a wide variety of studies in the fields of rock weathering and stone deterioration research.

KEYWORDS: rock and stone surface hardness testing; Equotip; limestone; non-parametric statistics; outliers

1. Introduction

Weathering manifests itself in the near surface zone as changes in stone properties such as porosity and intergranular bonds (McCabe et al., 2015). Quantifying these changes is important for rock weathering and stone deterioration research to understand spatio-temporal weathering behaviour and establishing decay rates (e.g. Meierding, 1993; Inkpen et al., 2012). Results may further inform decision making on heritage conservation strategies and provide hard evidence of stone response to impacts such as climate-change and air-pollution (e.g. Ross and Butlin, 1989; Smith et al., 2011; Viles and Cutler, 2012).

Surface induration or weakening are common property changes induced through environmental impacts (e.g. Inkpen et al., 2012; Moses et al., 2014). A common method to investigate such surface changes on-site is surface hardness testing. As a portable, non-destructive method it avoids the need to take samples as required to perform other common destructive tests like unconfined compressive strength. Although a long-established proxy method for relative dating of surface exposure in geomorphology (e.g. Aydin and Basu, 2005; Goudie, 2006; Fort et al., 2013; Stahl et al., 2013), only a few studies in built heritage science have used this method to quantify the state of preservation or deterioration of monuments (e.g. Török 2003, 2007, 2008; Cutler et al. 2013; Fort et al., 2013).

The most popular device for geomorphological applications is the Schmidt Hammer (e.g. Aydin and Basu, 2005; Goudie, 2006; Fort et al., 2013; Stahl et al., 2013). However, due to its high impact energy (Type L = 0.735 N m and type N = 2.207 N m its application on soft and porous or easily damaged stone is limited (Pope, 2000; Viles et al., 2011). In contrast, the impact energy of the Equotip with probe D is 0.0115 N m which is only a fraction of that of the Schmidt Hammer (probe versions with similarly low impact energy are Type C = 0.003 N m and Type G = 0.090 N m (Proceq© SA, 2010). Therefore, the Equotip is suitable for measuring a wide range of stone and rock surfaces (e.g. gypsum, tuff, limestone, granite) at different stages of weathering, as well as detecting subtle changes in surface hardness (e.g. Hack et al., 1993; Verwaal and Mulder, 1993; Aoki and Matsukura, 2007; Viles et al., 2011; Alberti et al., 2013; Coombes et al., 2013; Hansen et al., 2013). Low rebound values indicate soft, porous and/or weathered stone surfaces, higher values less weathered or case hardened surfaces.

The overall aim of this study is to develop a reliable methodology for using the Equotip for rock weathering and stone deterioration research. This paper answers the following questions: How do the Equotip D and DL probes compare? What are the most appropriate

statistical methods to handle Equotip data? How should outliers be treated? And what is an adequate sample size to collect on porous stone?

2. The Equotip family of devices and probes

The Equotip devices relevant for this paper are Equotip 3 and Equotip Piccolo 2, which come with a range of different probes (Table 1). They measure the difference between impact and rebound velocity of a (small) hard metal impact body traveling in a probe and propelled by spring force against the surface (Proceq© SA, 2010). The D probe is the most commonly used in stone weathering research to date with a small impact body (27 mm) measuring 3 mm in diameter (Figure 1a). In contrast, the DL probe has a slim long (82 mm) front section and slightly smaller diameter end (2.78 mm) (Figure 1b), and is suitable for confined spaces and recessed surfaces (Proceq© SA, 2010). To our knowledge the DL probe has not been trialled for rock weathering or stone deterioration research. It may provide a useful addition to weathering studies for collecting data on rough and / or porous surfaces.

The main differences between the Equotip devices are the range of impact bodies that can be attached to them and the evaluation software. The Equotip 3 is more versatile and comes with a separate recording unit. Obtained data is directly comparable for both devices, when using the same probe type, whereas the probes are not comparable among themselves (i.e. the DL probe gives higher readings than the D). Hardness data is expressed on the 'Leeb hardness' scale (1 - 999) and can be converted directly to all common hardness scales (e.g. Vickers, Rockwell etc.(Proceq©)). Furthermore, data is stored automatically and Equotip 3 and Equotip Piccolo 2 calculate and record basic descriptive statistics such as mean values and standard deviation (SD) during the measurement process.

3. Challenges for using Equotip devices on rock and stone surfaces

While Equotip devices offer a useful non-destructive means of testing the hardness of stone and rock surfaces, there are several challenges associated with the use of this equipment.

3.1. Effect of natural variability of rock and stone and weathered surfaces on Equotip data

Feal-Pérez and Blanco-Chao (2012) find surface roughness of weathered clasts affects Equotip measurements on-site. Similarly Aoki and Matsukura (2008) report data scatter

obtained from unweathered limestone and andesite surfaces due to subtle roughness and large pores of the particular stone types. Thus, natural property variations of fresh stone have an effect on Equotip data and such variations are likely to increase as weathering proceeds. Nevertheless, McCarroll (1991), who observed a similar effect for Schmidt Hammer measurements, states that surface roughness and weathering are intimately related. Therefore, instead of defining it as a limitation he suggested it could be utilized for comparison in cases where "surfaces have displayed similar surface textures prior to the influence of weathering" (McCarroll 1991, p. 479).

Previous research has found good correlations between Equotip measurements and unconfined compressive strength (Alvarez Grima and Babuška, 1999; Aoki and Matsukura, 2008; Yilmaz, 2013). Yilmaz (2013) tested a range of unweathered carbonate rocks (dolomite, limestone, travertine and marble) with densities between 2.24 up to 2.80 g/cm³ and open porosities between 0.14 and 7.00 %. Aoki and Matsukura (2007) tested weathered sandstone on-site, which originally had a density of 2.69 g/cm³ and open porosity of 6.9%. Both Yilmaz (2013) and Aoki and Matsukura (2007) utilized two Equotip application methods. For the single impact method (SIM) individual measurements are randomly distributed over the stone surface. Obtained values reflect on the elastic and plastic properties of the stone surface. In contrast, with the repeated impact method (RIM) repeated measurements on one point are taken, which reflects the elastic and plastic properties of the surface and subsurface of the stone. Yilmaz (2013) and Aoki and Matsukura (2007) combined both methods to gain deeper insight in stone surface and subsurface characteristics. Aoki and Matsukura (2007) introduced the *k*-value, whereas Yilmaz (2013) calculated the hybrid dynamic hardness (HDH) measure. For both porosity characteristics of the stone are taken into account and thus, natural stone variations are better reflected.

3.2. Methodology gaps

At present there is no consensus on methodology for the use of Equotip devices in the field or laboratory, nor in the evaluation of the data obtained (Viles et al., 2011; Yilmaz, 2013). This is a major limitation if reliable and comparable data are to be collected by different studies. Table 2 summarises the approaches taken by a range of researchers using the Equotip within the geomorphology and heritage science fields, and illustrates the need for further investigations into the most efficient sample size, and the best approach to statistical analysis given variable and often non-normal data, with outliers. As explained earlier several different methods can be used to quantify surface hardness with Equotip devices, including the SIM, RIM, and combinations of the two using *k*-value (Aoki and Matsukura, 2008) or hybrid dynamic hardness (HDH) (Yilmaz, 2013). These offer

solutions to address problems like surface roughness and porous stone and have been utilized and adapted for porous limestone using alternative statistical approaches in this study.

3.3. What sample size is needed to get reliable data from rock and stone surfaces?

One key issue that needs to be addressed when applying Equotip devices to stone and rock surfaces is the number of readings that should be taken, and how this affects the reliability of statistical tests applied to the data collected. For example, studies with the Schmidt hammer have shown that the number of readings taken has bearing on the meaningfulness of subsequent statistical tests (Niedzielski et al., 2009). The implication is that only a sufficiently big sample size will reflect the true surface hardness of a material, and how big is sufficient depends on the material being tested and its weathering-stress history. Table 2 shows sample sizes used in a selection of previous studies that have applied the Equotip 3 and Piccolo 2 devices. The number of readings taken ranges from 10 (Aoki and Matsukura, 2007) up to 80 (Coombes et al., 2013). It is not clear, however, how well any of these sample sizes used reflect the true surface hardness and Viles et al. (2011) suggest that a sample size of > 50 is needed in some circumstances. No consistent approach has been taken in previous research, and no justification has been given for the choice of sample sizes in most of these studies. How can researchers cope with natural variability of stone and the need for large sample sizes?

3.4. What is the best statistical methodology to handle Equotip data?

As well as being more variable than data from the Schmidt Hammer (Viles et al., 2011), it is likely that Equotip data obtained from porous and/or weathered rock and stone surfaces will be non-normally (asymmetrically) distributed. Accordingly Hansen et al. (2013) and Alberti et al. (2013) find that Equotip data derived from on-site measurements on weathered stone are affected. Nevertheless, standard parametric statistical methods were employed (e.g. t-test, analysis of variance (ANOVA), Fisher's least significant difference (LSD)), whereas for the evaluation of Equotip (and Schmidt Hammer) data robust methods may have been more beneficial (i.e. Mann-Whitney U, Kruskal-Wallis test and Spearman correlation)(Niedzielski et al., 2009). In cases of non-normal data, the reliability of statistical estimates based on the assumption of normality may be affected and parametric tests are largely inappropriate (Tukey, 1977; Fowler et al., 1998; Filzmoser and Todorov, 2013).

3.4.1 Data transformation

Semi-parametric tests (a hybrid of parametric and non-parametric (Powell, 1996) are one solution to treat non-normal data, and have been applied to Equotip data by Alberti et al. (2013). However, semi-parametric tests often require data modification. This involves decision making (i.e. normalization, defining thresholds, trimming, outlier removal etc.) before analysis, using appropriate outlier-detecting methods (Reimann, 2008; Good and Hardin, 2009). Depending on the statistical program used to define outliers, different procedures can be applied, and these are not always obvious or consistent between different studies. Furthermore, transformation and modification of data does not always lead to an evaluable dataset. For example, Alberti et al. (2013) modified 24 Equotip datasets using two methods (in one instance using only the 50% highest values and in another removing the eight extreme values from datasets), and yet some datasets remained nonnormally distributed.

3.4.2 Outliers

One factor associated with non-normal data is the occurrence of outliers, which may be present in a dataset as a result of human and / or instrument error, or due to natural deviations in the sample population (Hodge and Austin, 2004). Outliers are frequent in Equotip datasets (Viles et al., 2011). A common approach to outliers in classical statistics is to remove them entirely from the sample, as they place restrictions on subsequent data evaluation (Rosner, 1983). However, outliers should only be removed when it is clear that their occurrence is not related to the population characteristics but have resulted from errors in the data gathering process (Field, 2009). Identifying outliers is an important part of any statistical evaluation (Lipfert, 1989; Banerjee and Iglewicz, 2007), including Equotip data, as they can provide useful information about the sample in their own right (Iglewicz and Hoaglin, 1993). However, where outliers are to be retained, such as when they are deemed to reflect inherent, true variability in the hardness of a deteriorating stone for example, a new approach to statistical evaluation is required.

3.4.3 Statistical analysis using robust measures and bootstrap

Data transformation is not necessary when robust statistical measures are used. Robust summary statistics like median and median absolute deviation (MAD) are less affected by deviations from normality (Filzmoser and Todorov, 2013). When combined with non-parametric tests like Kruskal-Wallis and Mann-Whitney U test, they may provide an appropriate solution to some of the challenges associated with the analysis of Equotip datasets. Furthermore, the bootstrap technique as robust statistical techniques offers a solution to both the natural variability of stone affecting generated data and determining

sufficient sample sizes reflecting on specific characteristics of any investigated stone type. Bootstrapping generates a predefined (large) number of new datasets from the original dataset to derive an empirical estimate of the distribution of a statistic like mean, median or confidence intervals (Mooney and Duval, 1993; Kelley, 2005). Mooney and Duval (1993) state that bootstrapping has advantages over traditional parametric statistical approaches. The latter derive probability based inferences from a sample by distributional assumptions (usually normal distribution assumed) and analytic formulas (Mooney and Duval, 1993). In contrast bootstrapping replaces those theoretical formulations by resampling with replacement from the original dataset (Erceg-Hurn and Miroseovich, 2008; Uraibi et al., 2009). Thus, rather than drawing conclusions from potentially unrealistic assumptions (using traditional approaches) bootstrapped empirical estimates of statistical quantities of interest (mean, median or confidence intervals) can further improve statistical analyses such as parameter estimation, regression, prediction models, estimation of unknown variability and any analysis of a small representative sample (Erceg-Hurn and Miroseovich, 2008; Uraibi et al., 2009). Bootstrapping is unaffected by non-normality in the original dataset to which it is applied, as is common for surface hardness data obtained from porous and weathered stone on-site. Therefore, robust bootstrapping may be used to reduce bias in statistical estimations derived from porous stone.

4. Materials and methods

4.1. Stone samples

4.1.1 Stone types tested

The tests were conducted on four porous (oolitic) limestones that have been widely used in built heritage in the City of Oxford, including the Radcliffe Camera and the University Church of St. Mary the Virgin. Table 3 summarizes the limestone properties of the following types Portland (Jordans Base Bed), Bath (Hartham Park), Clipsham and Guiting. Stone samples were obtained fresh from quarries and cut to 300 mm x 80 mm x 50 mm dimensions. Porosity has been found to influence surface hardness testing (Aoki and Matsukura, 2008) and thus limestones with a wide range of porosity values were used in this study (13.5 – 22.2%). Unconfined compressive strength (UCS), open porosity and water absorption under atmospheric pressure were determined following BS-EN standards 1926:2006, BS-EN 1936:2006 and BS-EN 13755:2008, respectively (British Standards Institute, 2006q, 2006b, 2006c). UCS was determined with 10 cubes per stone type in order to determine the correlation with surface hardness values as regression can vary for different rock types (Dinçer et al., 2004). Open porosity was determined using six cubes (50 mm x 50 mm x 50 mm dimensions) for each limestone type.

4.1.2 Sample dimensions and preparation

Three replicate blocks for each limestone type were tested with the Equotip. The measurement surface (top face, 300 mm x 80 mm in dimensions) of each specimen was finished with P120 sandpaper prior to measurement, in order to minimise measurement error and to make sure that all values obtained were 'true' values and any outliers were due to the inherent, true variability in the hardness of the limestone (i.e. porosity rather than roughness). The device was applied perpendicular to the bedding of the blocks, which were placed on a solid limestone base to prevent interference from vibration.

4.2. Equotip Piccolo 2 with D- and DL-probe

Most previous geomorphological studies have employed the Equotip 3 in combination with the D probe (e.g. Hack et al., 1993; Coombes et al., 2013). This study used the Equotip Piccolo 2 (referred to as Equotip in this paper), which in terms of impact energy and measurement scale is comparable to the Equotip 3 but more portable. The principles tested in this paper for the Equotip Piccolo 2 are equally applicable to the Equotip 3 device. The DL probe was used as well as the D probe given its advantage of being able to obtain readings in confined spaces. The Equotip was frequently checked for calibration and all measurements (except for the assessment of operator variance) were conducted by the same operator (first author) under laboratory conditions.

4.3. Surface hardness test procedure

In this study SIM and RIM were applied and HDH calculated. For SIM the Equotip randomly applied 120 times distributed over an area covering about 720 cm² (total of surface area of three blocks per group). For RIM this study followed the approach of Aoki and Matsukura (2008) and collected 20 RIM values. For further data analysis the median of the highest values in each of the three RIM testing dataset per limestone type was calculated.

4.4. Operator variance

Within the scope of this study a pilot study was conducted to assess operator variance. Three operators with varying experiences towards the Equotip device (experienced and inexperienced) applied the Equotip with the D probe 20 times to a metal test block provided by Proceq (type: calibration block for D probe for high hardness range ~55.2HRC). Two different standards to assess Leeb hardness tester accuracy can be applied, DIN 50156 and ASTM A956 (Pollok and Mennicke, 2010). Depending on the standard the Equotip with D probe is considered to be calibrated when the mean value of

>three readings on the test block are HLD 765 with a tolerance of ± 6 (ASTM A956) or ± 15 (DIN 50156). For this study the latter tolerance was used.

4.5. Statistical data analysis and sample size determination

The statistical data analysis was two-fold. In a first step SIM mean and median with SD and MAD (respectively) were determined for the two probes (D and DL). Based on these values the HDH was calculated. The hardness data collected and calculated in this study are shown in Table 4. In view of potential on-site Equotip application to porous and weathered stone (which might display increased porosity) regression analysis (Pearson's R^2 and Spearman's rank correlation coefficient (ρ or r_s) as a non-parametric version of the Pearson correlation coefficient) were used to evaluate which calculated hardness would best reflect on the porous character of the tested limestone.

In a second step, the appropriate sample sizes for Equotip data collection on limestone was determined using the bootstrap technique to calculate confidence intervals for surface hardness median values. For statistical analysis RStudio (version 0.97.551) was used. Adapting Yilmaz' (2013) approach for porous limestone this study combined SIM and RIM based on median hardness to calculate the deformation ratio (DR) and HDH (see Equations 1 and 2).

$$DR_{robust} = HLDL_{S.med} / HLDL_{R.med} \quad (\text{Equation 1})$$

The robust hybrid dynamic hardness (HDH_{robust}) is calculated as follows:

$$HDH_{robust} = DR_{robust} \times HLDL_{S.med} = (HLDL_{S.med})^2 / HLDL_{R.med} \quad (\text{Equation 2})$$

4.5.1 Normality – parametric and non-parametric statistics

4.5.2 Outliers

Following the approach of Aydin (2009) all measured values were used in the evaluation and outliers were not removed from the datasets. Nevertheless, outliers were identified in order to determine their number and gain potentially interesting information about individual stone properties (i.e. porosity). To detect outliers the MAD was used and (x_i) the boundary for extreme values (outliers) was specified using (moderately conservative) $2.5 \times MAD$ following the recommendation of Leys et al. (2013) and shown in Equation 3:

$$Median - 2.5 * MAD < x_i < Median + 2.5 * MAD \quad (\text{Equation 3})$$

4.5.3 Kruskal-Wallis and Mann-Whitney U

The Kruskal-Wallis test was used as a robust alternative to one-way ANOVA to evaluate significant differences between the tested limestone types and the two probes (D and DL) (Hodges and Lehmann, 1963). This was followed by further specifying the differences between the individual stone types using the Mann-Whitney U test (two-tailed test with a significance level of p-value 0.05, unpaired) as an alternative to the t-test (Hodges and Lehmann, 1963). The data were visualised using boxplots and density plots in order to determine skewness and detect outliers.

4.6. Sample size determination

In addition to evaluating data using robust statistical measures, the second aim of the study was to determine an appropriate sample size for the Equotip that would sufficiently reflect the true stone surface hardness, but that was also practical for on-site application. For this, the 120 readings obtained for each stone type were taken to represent the true stone surface hardness ('population'). A range of smaller sample sizes (5, 10, 20, 45 and 60 readings) were then modelled by resampling the original dataset without replacement (for each sample size this process was repeated a 100 times to simulate variation) using bootstrap in RStudio. Finally, confidence intervals for the medians of the individual modelled sample size datasets were obtained through bootstrapping.

Our assumption was that the width of the confidence intervals would vary for the different sample sizes (i.e. a small sample size would result in wider confidence interval), taken to reflect the degree of variation of the median. These intervals were calculated with 95% confidence level using the bias corrected and accelerated (bca) bootstrap for confidence intervals in R (10,000 times), the most robust version for analysing non-normal data (Efron, 1987). The bootstrapped confidence intervals for the medians of the modelled sample size datasets were then compared to the original sample confidence intervals (using the original 120 readings) by calculating the differences of confidence interval widths in percentages. Based on the results an appropriate sample size was determined.

5. Results and discussion

This section firstly evaluates the performance of two Equotip probes (D and DL) on porous limestone under laboratory conditions on four porous limestone types. It is shown that for general data analysis for data obtained with the Equotip on porous limestone it is more beneficial to use robust measures and methods in order to account for natural variability of

porous stone. Furthermore, an appropriate sample size for Equotip readings to be collected on porous limestone are determined to gain meaningful results.

Despite controlled laboratory conditions, fresh and smooth stone surfaces and a large sample size of 120 readings per stone type, the majority of the Equotip data sets show non-normal distribution (Shapiro-Wilk test, Table 5), caused by outliers and skewness (Figures 2a, 2b, 3a, 3b and Tables 5 and 6).

5.1. Probes

Figure 2a and b and Tables 5 and 6 show the data collected using the two probes (D and DL). As expected, they are not directly comparable. In every case, the DL probe produced higher hardness values, which was confirmed by Proceq® as being usual (Personal communication 28/11/2013). It would have been useful to be able to convert HLD (hardness values obtained with the D probe) values into HLDL (hardness values obtained with the DL probe) and vice versa, but due to differing variances (probably caused by limestone characteristics) in the individual probe datasets this is not possible. The coefficient for the HLD and HLDL values ranged between 1.12 and 1.32. The DL probe produced a wider data spread than the D probe (except for Portland limestone, where D obtained a wider data spread) (Figures 2a and 2b).

The Kruskal-Wallis test revealed significant differences in Equotip data for the stone types for both probes, D (df=3, chi-squared=305.904, p-value < 0.001) and DL (df=3, chi-squared=282.881, p-value < 0.000) probes. The following Mann-Whitney U tests showed significantly different hardness values (p < 0.001) for both probes on all four limestone types (Tables 7 and 8). This shows that Equotip can be used to distinguish the stone types used in this study using either probe.

5.2. Surface hardness data – Stone variance – Operator variance

Figure 4 shows no significant variance for HLD_s values (obtained on a metal test block) between the two experienced operators. The data range is well within the Equotip calibration requirements ($HLD\ 765 \pm 15$). In contrast, the HLD_s values generated by the inexperienced operator show three outliers and thus, a noticeable shift of the mean. Nevertheless, it can also be seen that the median is not affected by the three outliers and within the calibration requirements. Therefore, using the median improves the reliability of Equotip data even if an inexperienced person is using the device. Since all further measurements in this study were conducted by the same operator (first author) and robust measures are used, operator variance is not considered to be an issue. As a consequence, the variance of surface hardness data observed in this study is attributed to

the natural variability of the tested limestone as reported by Palmer (2008) and findings from Siedel and Siegesmund (2010) especially for limestone with low density and high porosity.

It was found that median values ($HLD_{S.med}$ and $HLDL_{S.med}$) showed lower variance compared to mean values ($HLD_{S.mean}$ and $HLDL_{S.mean}$) (Tables 5 and 6). The strength of the correlation for Pearson's R^2 and Spearman is categorised following Dancey and Reidy (2004), where the association with 1 = perfect, 0.7 - 0.9 = strong, 0.4 - 0.6 = moderate, 0.1 - 0.3 = weak, 0 = zero. Both the Pearson's R^2 and Spearman correlation coefficient show strong association of UCS median values with the median surface hardness values of both probes (Pearson's R^2 : D probe $R^2 = 0.99$ and DL probe $R^2 = 0.95$; Spearman: D probe ($r_s(2) = 1$, $p = 0.0833$) and DL probe ($r_s(2) = 1$, $p = 0.0833$). The correlation of all surface hardness data (see Table 4) with the median open porosity is shown in Table 9. All hardness data for both tested probes show a strong correlation, therefore reflect sufficiently on the respective porosity. Given the range of tested high porosities (13.5 – 22.2%) the implications for on-site studies on weathered limestones are, that a) high porosity can be determined using Equotip and further b) porosity changes (increase or decrease) over time through weathering could be investigated. This has implications for the potential application of the Equotip in weathering rate studies.

Although the DL probe showed higher data spread, it correlates slightly better with open porosity values of limestone in this study compared to D probe shown by the R^2 values in Table 9. The best fit is gained with the $HDH_{DL,robust}$. Furthermore, the DL probe might be more advantageous in the field, because it offers a more controlled way of sampling in recessed, rough or curved areas (typical for weathered stone and architectural geometry of built heritage). Also, it offers protection from dust for the Equotip device itself due to the long slim front section, which prevents the impact body from transporting particles into the body of the device.

5.3. Outliers

Almost every dataset contained more than one outlier (Table 10). In the case of porous limestone outliers may occur due to the heterogeneity (e.g. porosity, shells) of the stone as discussed earlier. Thus, outliers are likely to be part of the natural deviation in the population and should not be removed. For this study it is particularly noticeable that most of the outliers are higher hardness values, as might be found when the Equotip impact body strikes a hard fossil for example. Figures 2a and 2b show the effect of outliers and skewness on the mean values, which are different from the medians in the majority of cases. The difference between mean and median values is most notable for Clipsham and

Bath limestone and might be due to their particular pore size distribution and inherent material variability.

For weathered rock and stone surfaces variability in Equotip data is likely to be even higher and thus Equotip data are rarely likely to be normal. In cases of non-normal data, statistical estimates based on common statistical descriptors may be affected (Tukey, 1977) so that parametric tests are largely inappropriate (Fowler et al., 1998). Consequently, in order to account for inherent variability in surface hardness measurements caused by natural stone properties (on-site), and to avoid the need for data transformation, the robust statistical methods used in this paper are preferable to classic statistical measures and methods previously used for Equotip data evaluation.

5.4. Appropriate sample size

How many readings should be taken when applying Equotip devices to stone and rock surfaces? This study aimed to determine a sample size big enough to portray reliably the median surface hardness of the four tested stone types, but small enough to also be practical for on-site applications. Based on the original 120 readings ('population') collected per stone type several smaller sample sizes were modelled (5, 10, 20, 45 and 60) and for each the width of the confidence interval for the median was calculated to reflect the degree of variation of the median. A small degree of variation would result in narrower confidence interval widths and thus show high accuracy in prediction of the median. For all stone types a bigger sample size resulted in narrower confidence interval widths (Figures 5a and 5b, Tables 5, 6), confirming Niedzielski et al.'s (2009) statement that accuracy increases with increased sample size.

In this study the majority of confidence intervals for the median are wider for the DL probe compared to those of the D probe (exceptions are confidence intervals for a sample size of 5). The confidence intervals obtained for different stone types are noticeably distinct from each other, with wider intervals for stone with complex porosities like Clipsham and narrower intervals for stones with higher compressive strength like Portland. This indicates that an appropriate sample size is heavily dependent on the stone type and consequently its state of preservation, where changes in those properties indicate ongoing weathering processes. Therefore, either a bigger sample size is necessary for porous stone with complex pore size distributions, or a wider confidence interval needs to be tolerated for the median to fall into. Alternatively a lower confidence level could be accepted (i.e. 90%) with narrower confidence intervals.

For this study the appropriate sample size was determined using a confidence level of 95% accepting that different stone types would therefore have wider confidence intervals.

Comparing the predicted confidence intervals for the different modelled sample sizes of this study (Figures 5a and 5b) with sample sizes used by previous studies (e.g. 10 and 20 as reported by Aoki and Matsukura (2007) and Yilmaz (2013) respectively), it becomes clear that here the predicted confidence intervals within which a median is expected to appear are rather wide. For example, given 20 readings for Clipsham and Bath the confidence intervals are not substantially different and therefore the median surface hardness could not be sufficiently distinguished. In contrast, the confidence intervals of the original datasets (120 readings) are very narrow, reflect the median stone surface hardness well, and all four stone types are clearly distinguishable.

However, taking 120 readings is often not practical in the field because it is (a) time consuming and (b) would require a larger measurement area, which may limit potential subsequent investigations. Consequently, to find a good compromise between accuracy and practicality, it was aimed to define smaller appropriate sample sizes and accepting potentially wider confidence intervals, whilst ensuring that the confidence intervals of one single stone type tested in this study should not overlap with one of another stone type.

It is therefore necessary to define a general sample size that would be appropriate for all stone types tested, and that would be transferable to on-site application on stone with unknown history. As stated earlier the modelled sample size of 20 for the Clipsham limestone and Bath limestone did not show clearly separated confidence intervals, which makes it impossible to distinguish the stone types using surface hardness. Therefore, an appropriate sample size was estimated by evaluating further the width difference (as a percentage) of modelled 20, 45 and 60 sample size datasets relative to the original datasets of 120 readings.

Table 11 and 12 show that the confidence interval differences between 120 and 60 and 45 readings are smaller than between 120 and 20 readings. Although 60 readings better reflect the original data set (i.e. a smaller % difference), it would be sufficient to take 45 readings in order to obtain a narrow enough confidence interval with all confidence intervals for the four different stone types clearly distinguishable. Furthermore, Tables 5 and 6 show the lower and higher boundaries for the calculated confidence intervals for the surface median hardness. Thus, for this study every median surface hardness obtained will fall into the respective confidence interval and can clearly be attributed to a stone type. Using the range of a confidence interval rather than a single value like the median to represent stone and rock surface hardness is more applicable (i.e. versatile) for in situ applications where natural stone and rock variance is expected.

6. Conclusions and recommendations

On the basis of the results in this study we propose a number of considerations when using Equotip testing:

- 1) Scope of application of the Equotip: This study shows that the Equotip is suitable for soft and porous rock and stone. It is however, beneficial to calibrate the Equotip on fresh stone before using it on-site on weathered stone surfaces in order to establish weathering rates. Nevertheless, the Equotip application works as well as relative measure e.g. for the comparison of surfaces exposed to different aspects and/or degree of orientation and height or for quality assessment before and after stone consolidation treatment.
- 2) D and DL probe: Although the DL probe showed higher data spread, it correlates slightly better with open porosity values of limestone in this study. Further advantages are more controlled sampling in recessed areas, rough or curved areas (typical for weathered stone and architectural geometry of built heritage). The long slim front section of the probe, which prevents the impact body from transporting particles into the body of the device, offers further protection from dust for the Equotip device itself.
- 3) Non-normal data: In this study, data obtained from four different limestone under controlled conditions yielded non-normal data in the majority of cases, as a result of inherent variability in material properties such as porosity. This paper argues that Equotip data from weathered stone and rock surfaces are rarely normal and thus parametric tests are largely inappropriate and would either require data transformation to gain meaningful results or the application of robust statistical measures and methods.
- 4) Robust (non-parametric) statistical measures and methods, and outliers: Outliers and skewness were the main cause for the unsymmetrical distributed data in this study. The paper proposes to include outliers in the data analysis as their occurrence is linked to natural stone characteristics – in the case of the limestones tested they indicate the presence of fossils and other harder elements. However, including outliers in data analysis necessitates the new approaches to statistical analyses addressed in this study. The presented alternative, robust statistical approach requires no data transformation (e.g. removing outliers and more) and is more reliable for non-normally distributed data as well as being adequate for normal data. We recommend to apply robust statistical methods unaffected by non-normal data (e.g. median and MAD as alternative measures of central

tendency and variance as well as Kruskal-Wallis and Mann-Whitney test as alternatives to ANOVA and t-test.).

5) HDH: The combination of two measuring procedures (SIM and RIM) based on median values accounts for potential effect of pores/weathering especially when used with DL probe values and thus complements SIM and RIM.

6) Sufficient sample size: A big enough sample size needs to be collected and is highly dependent on the respective porosity of the tested stone. Thus, the more porous (heterogeneous) and weathered the stone the higher the sample size should be. Nevertheless, for practical reasons for on-site applications on stone with unknown history the aim was to determine a general sample size that would be a) appropriate for all stone types tested in this study, and thus include a variety of high porosities, while b) differentiate the respective stone surface median hardness and c) allow to distinguish the tested stone types. Therefore, for the tested stones in this study we propose a sample size of 45 readings (for a confidence level of 95%). It is worth mentioning that calculating sample sizes using a 95% confidence level is a conservative approach. In view of the expected variances for in situ measurements and unknown weathering-stress histories of heritage stone, it might be justified to reduce the confidence level to 90%. This would still provide reliable data output when robust measures are used, but allow for a smaller sample size to be collected. This approach can certainly be transferred to stone and rock with similar porosities and hardness.

While the study was conducted in the laboratory and took variation of natural stone into account, it used fresh, smooth stone samples and thus, research on-site is desirable to link back to results obtained in the laboratory. This study has shown that the Equotip provides valuable measures of surface hardness of porous stone which can be related to other measures such as unconfined compressive strength as found by Hack et al. (1993) and Verwaal and Mulder (1993), but also demonstrates that data evaluation can be improved by using robust measures, applying robust statistical methods and increasing sample size. The proposed methodology requires no data modification (e.g. removing outliers), is more accurate for non-normally distributed data and adequate for normal data, and thus provides a timesaving general approach to data evaluation including on-site measurements. This methodology allows for consistent comparability between different on-site research projects across the fields of rock weathering and stone deterioration research.

543 ***Acknowledgments***

544 K. Wilhelm is in receipt of an EPSRC studentship (funded via grants EP/P504287/1,
545 EP/P505216/1 and EP/P505666/1) with additional support from Proceq.

546 Dr Dan Lunn is thanked for invaluable help with the program R and introduction to
547 bootstrapping techniques. The authors want to thank Dr Helen Reeves and Marcus Dobbs
548 from the British Geological Survey (BGS), who provided technical assistance with
549 compressive strength testing. Dr Martin Coombes and Jerome Mayaud provided very
550 helpful comments on an earlier draft of this paper.

551

References

- Alberti AP, Gomes A, Trenhaile A, Oliveira M, Horacio J, 2013. Correlating river terrace remnants using an Equotip hardness tester: An example from the Miño River, northwestern Iberian Peninsula. *Geomorphology* **192**: 59–70. DOI: 10.1016/j.geomorph.2013.03.017.
- Alvarez Grima M, Babuška R, 1999. Fuzzy model for the prediction of unconfined compressive strength of rock samples. *International Journal of Rock Mechanics and Mining Sciences* **36**: 339–349. DOI: 10.1016/S0148-9062(99)00007-8.
- Aoki H, Matsukura Y, 2007. A new technique for non-destructive field measurement of rock-surface strength: an application of the Equotip hardness tester to weathering studies. *Earth Surface Processes and Landforms* **32**: 1759–1769. DOI: 10.1002/esp.1492.
- Aoki H, Matsukura Y, 2008. Estimating the unconfined compressive strength of intact rocks from Equotip hardness. *Bulletin of Engineering Geology and the Environment* **67**: 23–29. DOI: 10.1007/s10064-007-0116-z.
- Aydin A, 2009. ISRM Suggested method for determination of the Schmidt hammer rebound hardness: Revised version. *International Journal of Rock Mechanics and Mining Sciences* **46**: 627–634. DOI: 10.1016/j.ijrmms.2008.01.020.
- Aydin A, Basu A, 2005. The Schmidt hammer in rock material characterization. *Engineering Geology* **81**: 1–14. DOI: 10.1016/j.enggeo.2005.06.006.
- Banerjee S, Iglewicz B, 2007. A Simple Univariate Outlier Identification Procedure Designed for Large Samples. *Communications in Statistics: Simulation & Computation* **36**: 249–263. DOI: 10.1080/03610910601161264.
- British Standards Institute (BSI), 2006a. *Natural stone test methods – Determination of uniaxial compressive strength*. BSI: London. (BS EN 1926:2006).
- British Standards Institute (BSI), 2006b. *Natural stone test methods—Determination of real density and apparent density, and of total and open porosity*. BSI: London. (BS EN 1936:2006).
- British Standards Institute (BSI), 2006c. *Natural stone test methods—Determination of water absorption at atmospheric pressure*. BSI: London. (BS EN 13755:2008).
- Coombes MA, Feal-Pérez A, Naylor LA, Wilhelm K, 2013. A non-destructive tool for detecting changes in the hardness of engineering materials: Application of the Equotip durometer in the coastal zone. *Engineering Geology* **167**: 14–19. DOI: 10.1016/j.enggeo.2013.10.003.
- Cutler NA, Viles HA, Ahmad S, McCabe S, Smith BJ, 2013. Algal ‘greening’ and the conservation of stone heritage structures. *Science of The Total Environment* **442**: 152–164. DOI: 10.1016/j.scitotenv.2012.10.050.
- Dancey CP, Reidy J, 2004. *Statistics without maths for psychology: Using SPSS for Windows*. Prentice Hall: New York.
- Dinçer I, Acar A, Çobanoğlu I, Uras Y, 2004. Correlation between Schmidt hardness, uniaxial compressive strength and Young’s modulus for andesites, basalts and tuffs. *Bulletin of Engineering Geology and the Environment* **63**: 141–148. DOI: 10.1007/s10064-004-0230-0.
- Efron B, 1987. Better Bootstrap Confidence Intervals. *Journal of the American Statistical Association* **82**: 171–185. DOI: 10.2307/2289144.
- Erceg-Hurn DM, Mirosevich VM, 2008. Modern robust statistical methods: an easy way to maximize the accuracy and power of your research. *The American psychologist* **63**: 591–601. DOI: 10.1037/0003-066X.63.7.591.
- Feal-Pérez A, Blanco-Chao R, 2012. Characterization of abrasion surfaces in rock shore environments of NW Spain. *Geo-Marine Letters* **33**: 1–9. DOI: 10.1007/s00367-012-0300-4.
- Field AP, 2009. *Discovering statistics using SPSS*. SAGE: Los Angeles, [Calif.], London.
- Filzmoser P, Todorov V, 2013. Robust tools for the imperfect world. *Statistics with Imperfect Data* **245**: 4–20. DOI: 10.1016/j.ins.2012.10.017.

- Fort R, Alvarez de Buergo M, Perez-Monserrat EM, 2013. Non-destructive testing for the assessment of granite decay in heritage structures compared to quarry stone. *International Journal of Rock Mechanics and Mining Sciences* **61**: 296–305. DOI:10.1016/j.ijrmms.2012.12.048
- Fowler J, Cohen L, Jarvis P, 1998. Practical statistics for field biology. John Wiley: Chichester.
- Good PI, Hardin JW, 2009. Common errors in statistics (and how to avoid them). Wiley: Hoboken, N.J.
- Goudie AS, 2006. The Schmidt Hammer in geomorphological research. *Progress in Physical Geography* **30**: 703–718. DOI: 10.1177/0309133306071954.
- Hack H, Hingira J, Verwaal W, 1993. Determination of discontinuity wall strength by Equotip and ball rebound tests. *International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts* **30**: 151–155. DOI: 10.1016/0148-9062(93)90707-K.
- Hansen CD, Meiklejohn, KI, Nel W, Loubser MJ, Van Der Merwe BJ., 2013. Aspect-controlled Weathering Observed on a Blockfield in Dronning Maud Land, Antarctica. *Geografiska Annaler: Series A, Physical Geography* **95**: 305–313. DOI: 10.1111/geoa.12025.
- Hodge VJ, Austin J, 2004. A Survey of Outlier Detection Methodologies. *Artificial Intelligence Review* **22**: 85–126. DOI: 10.1007/s10462-004-4304-y.
- Hodges, JL. Jr., Lehmann, EL., 1963. Estimates of Location Based on Rank Tests. *The Annals of Mathematical Statistics* **34**: 598–611. DOI: 10.2307/2238406.
- Iglewicz B, Hoaglin DC, 1993. How to detect and handle outliers. ASQC Quality Press: Milwaukee, Wis.
- Inkpen RJ, Viles H, Moses C, Baily B, Collier P, Trudgill ST, Cooke RU, 2012. Thirty years of erosion and declining atmospheric pollution at St Paul's Cathedral, London. *Atmospheric Environment* **62**: 521–529. DOI: 10.1016/j.atmosenv.2012.08.055.
- Kelley K, 2005. The effects of nonnormal distributions on confidence intervals around the standardized mean difference. *Educational and Psychological Measurement* **65**: 51–69. DOI: 10.1177/0013164404264850.
- Leys C, Ley C, Klein O, Bernard P, Licata L, 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology* **49**: 764–766. DOI: 10.1016/j.jesp.2013.03.013.
- Lipfert FW, 1989. Atmospheric damage to calcareous stones: Comparison and reconciliation of recent experimental findings. *Atmospheric Environment (1967)* **23**: 415–429. DOI: 10.1016/0004-6981(89)90587-8.
- McCabe S, McAllister D, Warke PA, Gomez-Heras M, 2015. Building sandstone surface modification by biofilm and iron precipitation: emerging block-scale heterogeneity and system response. *Earth Surface Processes and Landforms* **40**: 112–122. DOI: 10.1002/esp.3665.
- McCarroll D, 1991. The Schmidt Hammer, weathering and rock surface roughness. *Earth Surface Processes and Landforms* **16**: 477–480. DOI: 10.1002/esp.3290160510.
- Meierding TC, 1993. Marble Tombstone Weathering and Air Pollution in North America. *Annals of the Association of American Geographers* **83**: 568–588. DOI: 10.1111/j.1467-8306.1993.tb01954.x.
- Mol L, Viles HA. 2012. The role of rock surface hardness and internal moisture in tafoni development in sandstone. *Earth Surface Processes and Landforms* **37**: 301–314. DOI:10.1002/esp.2252.
- Mooney CZ, Duval RD, 1993. Bootstrapping: A nonparametric approach to statistical inference. Sage Publications: Newbury Park, Calif.
- Moses C, Robinson D, Barlow J, 2014. Methods for measuring rock surface weathering and erosion: A critical review. *Earth-Science Reviews* **135**: 141–161. DOI: 10.1016/j.earscirev.2014.04.006.
- Niedzielski T, Migoń P, Placek A, 2009. A minimum sample size required from Schmidt hammer measurements. *Earth Surface Processes and Landforms* **34**: 1713–1725. DOI: 10.1002/esp.1851.

- Palmer T, 2008. Limestone petrography and durability in English Jurassic Freestones. In *England's heritage in stone: Proceedings of a conference, Tempest Anderson Hall, York, 15-17 March, 2005*, Doyle P (ed). English Stone Forum: Folkestone, Kent; 66–78.
- Pollok H, Mennicke RT, 2010. *Using Equotip hardness test blocks*. www.proceq.com. Accessed 09/2015.
- Pope GA, 2000. Weathering of petroglyphs: direct assessment and implication for dating methods. *Antiquity* **74**: 833–843.
- Powell J, 1996. Estimation of semiparametric models. In *Handbook of econometrics*. North-Holland: Amsterdam; 2444–2514.
- Proceq© SA, 2010. Operating instructions portable metal hardness tester. www.proceq.com. Accessed 1. September 2012.
- Reimann C, 2008. Statistical data analysis explained: Applied environmental statistics with R. John Wiley & Sons: Chichester, England, Hoboken, NJ.
- Rosner B, 1983. Percentage Points for a Generalized ESD Many-Outlier Procedure. *Technometrics* **25**: 165–172. DOI: 10.2307/1268549.
- Ross KD, Butlin RN, 1989. Durability tests for building stone. Building Research Establishment: Garston.
- Siedel H, Siegesmund S, 2010. Characterisation of stone deterioration on buildings. In *Stone in Architecture: Properties, Durability*, Siegesmund S, Snethlage R, Winkler E (eds). Springer: Berlin; 347–410.
- Smith BJ, McCabe S, McAllister D, Adamson C, Viles HA, Curran JM, 2011. A commentary on climate change, stone decay dynamics and the 'greening' of natural stone buildings: new perspectives on 'deep wetting'. *Environmental Earth Sciences* **63**: 1691–1700. DOI: 10.1007/s12665-010-0766-1.
- Stahl T, Winkler S, Quigley M, Bebbington M, Duffy B, Duke D, 2013. Schmidt hammer exposure-age dating (SHD) of late Quaternary fluvial terraces in New Zealand. *Earth Surface Processes and Landforms* **38**: 1838–1850. DOI: 10.1002/esp.3427.
- Török Á, 2003. Surface strength and mineralogy of weathering crusts on limestone buildings in Budapest. *Building Stone Decay: Observations, Experiments and Modeling* **38**: 1185–1192. DOI: 10.1016/S0360-1323(03)00072-6.
- Török Á, 2007. Characteristics and morphology of weathering crusts on porous limestone, the role of climate and air pollution. In *Preservation of Natural Stone and Rock Weathering*, Olalla C, Estaire J, Sola P (eds). Taylor & Francis; 61–66.
- Török Á, 2008. Black crusts on travertine: factors controlling development and stability. *Environmental Geology* **56**: 583–594. DOI: 10.1007/s00254-008-1297-x.
- Tukey JW, 1977. Exploratory data analysis. Addison-Wesley Pub. Co.: Reading (Mass.).
- Uraibi HS, Midi H, Talib BA, Yousif JH, 2009. Linear Regression Model Selection Based on Robust Bootstrapping Technique. *American Journal of Applied Sciences* **6**: 1191–1198. DOI: 10.3844/ajassp.2009.1191.1198.
- Verwaal W, Mulder A, 1993. Estimating rock strength with the Equotip hardness tester. *International Journal of Rock Mechanics and Mining Sciences & Geomechanics Abstracts* **30**: 659–662. DOI: 10.1016/0148-9062(93)91226-9.
- Viles H, Cutler N, 2012. Global environmental change and the biology heritage structures. *Global Change Biology* **18**: 2406–2418.
- Viles H, Goudie A, Grab S, Lalley J, 2011. The use of the Schmidt Hammer and Equotip for rock hardness assessment in geomorphology and heritage science: a comparative analysis. *Earth Surface Processes and Landforms* **36**: 320–333. DOI: 10.1002/esp.2040.
- Yilmaz NG, 2013. The Influence of Testing Procedures on Uniaxial Compressive Strength Prediction of Carbonate Rocks from Equotip Hardness Tester (EHT) and Proposal of a New Testing Methodology: Hybrid Dynamic Hardness (HDH). *Rock Mechanics and Rock Engineering* **46**: 95–106. DOI: 10.1007/s00603-012-0261-y.

717 **Figures**



718
719 Figure 1a. Equotip Piccolo 2 with impact body D on-site at Radcliffe Camera, Oxford.



720
721 Figure 1b. Equotip Piccolo 2 with impact body DL in the Oxford Rock Breakdown Laboratory
722 (OxRBL).
723

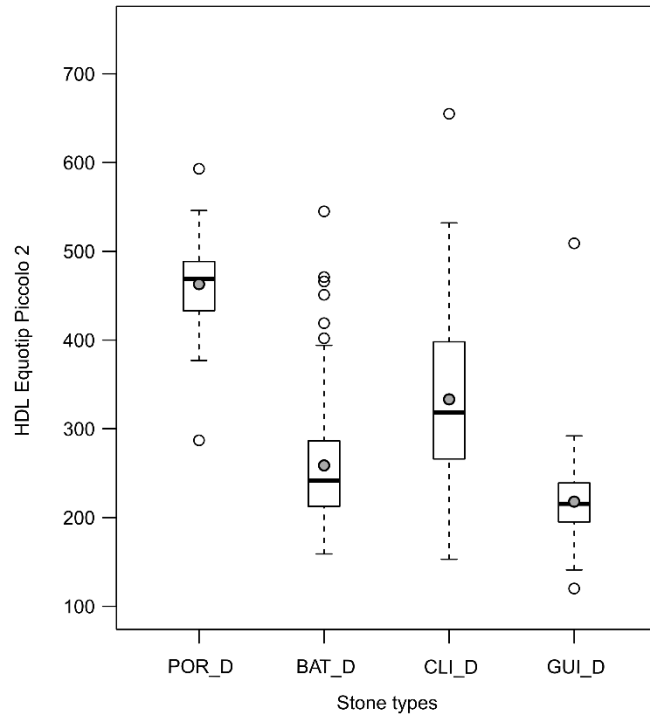


Figure 2a. Boxplot of surface hardness values, with median (black line) and mean (grey dot) and outliers (white dots), four different stone types (Portland=POR, Bath=BAT, Clipsham=CLI, Guiting=GUI) with smooth surfaces (ground with sandpaper P.120), Equotip Piccolo 2 probe D, n=120.

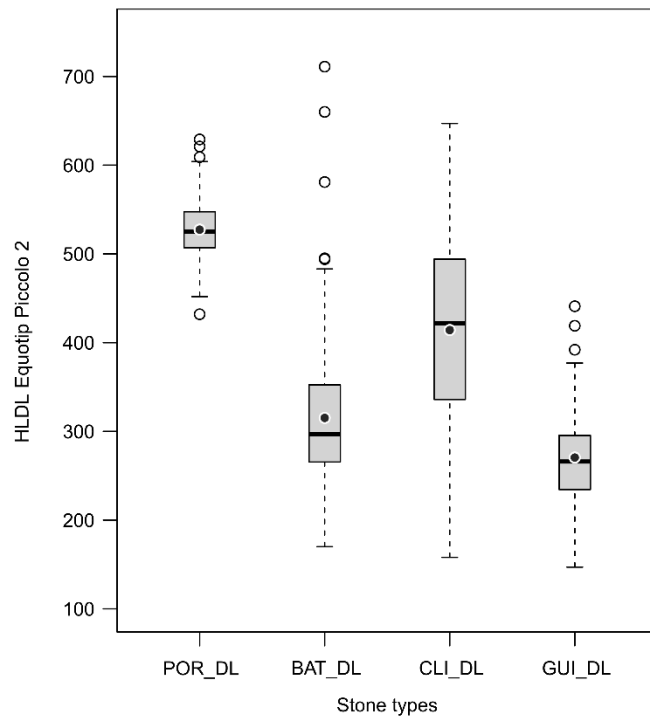


Figure 2b. Boxplot of surface hardness values, with median (line) and mean (black dot) four different stone types (Portland=POR, Bath=BAT, Clipsham=CLI, Guiting=GUI) with smooth surfaces (ground with sandpaper P.120), Equotip Piccolo 2 probe DL, n=120.

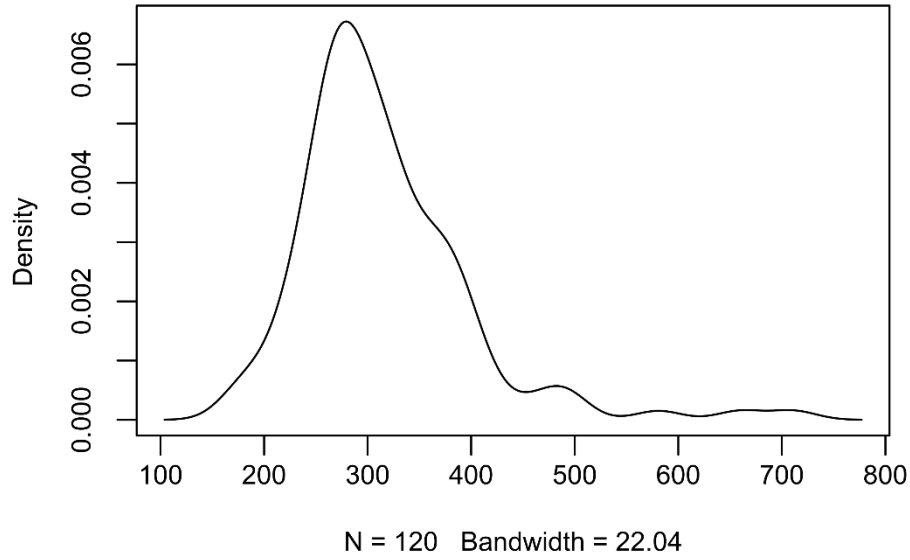


Figure 3a. Example for skewed data in this study, density plot for distribution of surface hardness values (HLDL) for Bath limestone showing positive skew, Equotip Piccolo 2 with impact body DL, 120 readings.

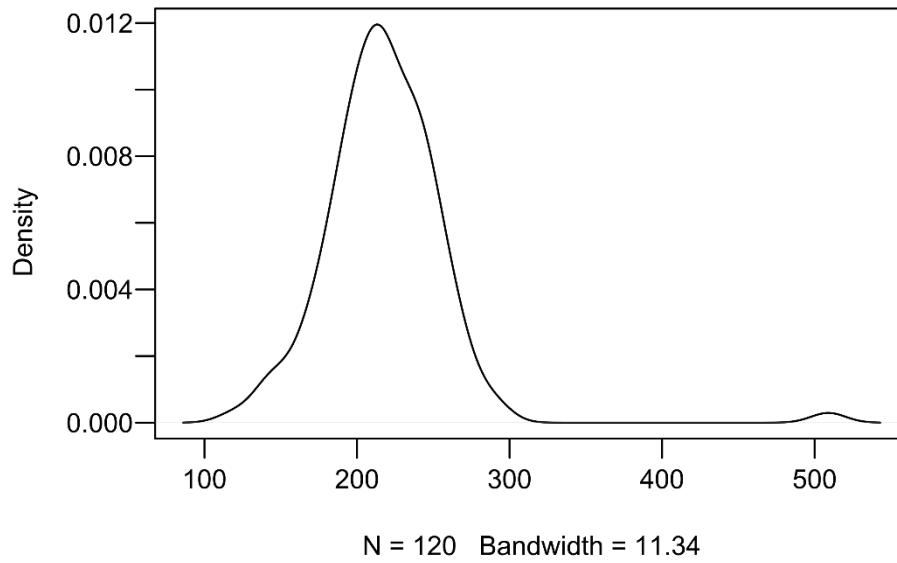


Figure 3b. Example for skewed data in this study, density plot for distribution of hardness values (HLD) for Guiting limestone showing positive skew, Equotip Piccolo 2 with impact body D, 120 readings.

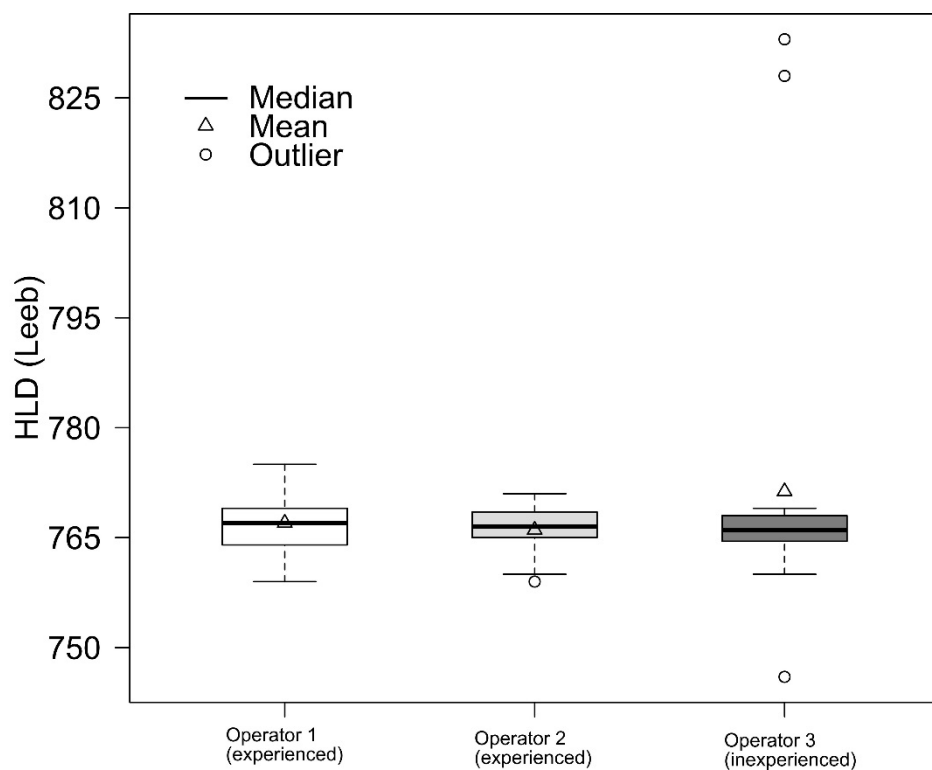


Figure 4. Boxplot showing 3 surface hardness datasets (20 single impact readings on a metal test block) generated by three different operators. Operator 1 and 2 had experience with the Equotip and operator 3 was using the device for the first time.

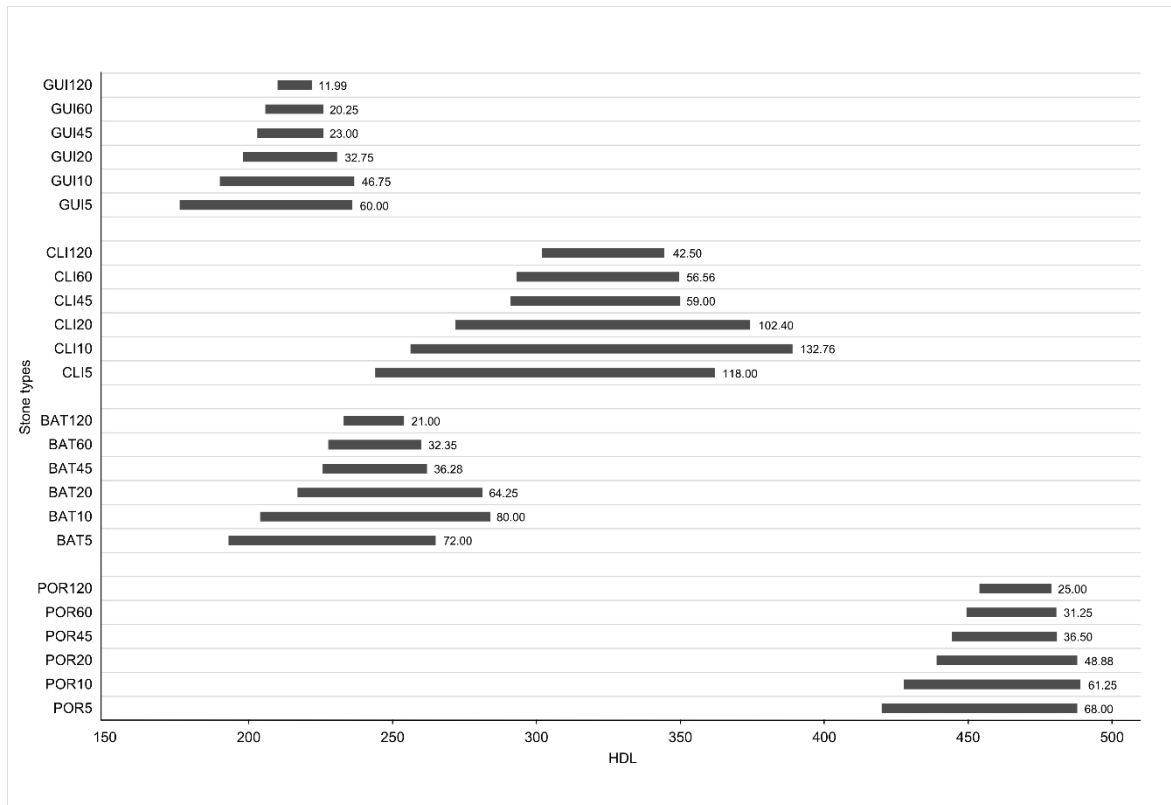


Figure 5a. Predicted confidence intervals for medians of Equotip Piccolo 2 D probe data for different modelled sample sizes (numbers on the y-axis) on four different limestone in this study (Portland=POR, Bath=BAT, Clipsham=CLI, Guiting=GUI). Confidence intervals are obtained applying bias-corrected accelerated bootstrap to datasets of 120 readings. Modelled samples sizes are 5, 10, 20, 45, and 60 readings, resampled from the original dataset (120). Bars show confidence interval width (numeric value indicated) for median to occur within at 95% confidence level (See also table 5).

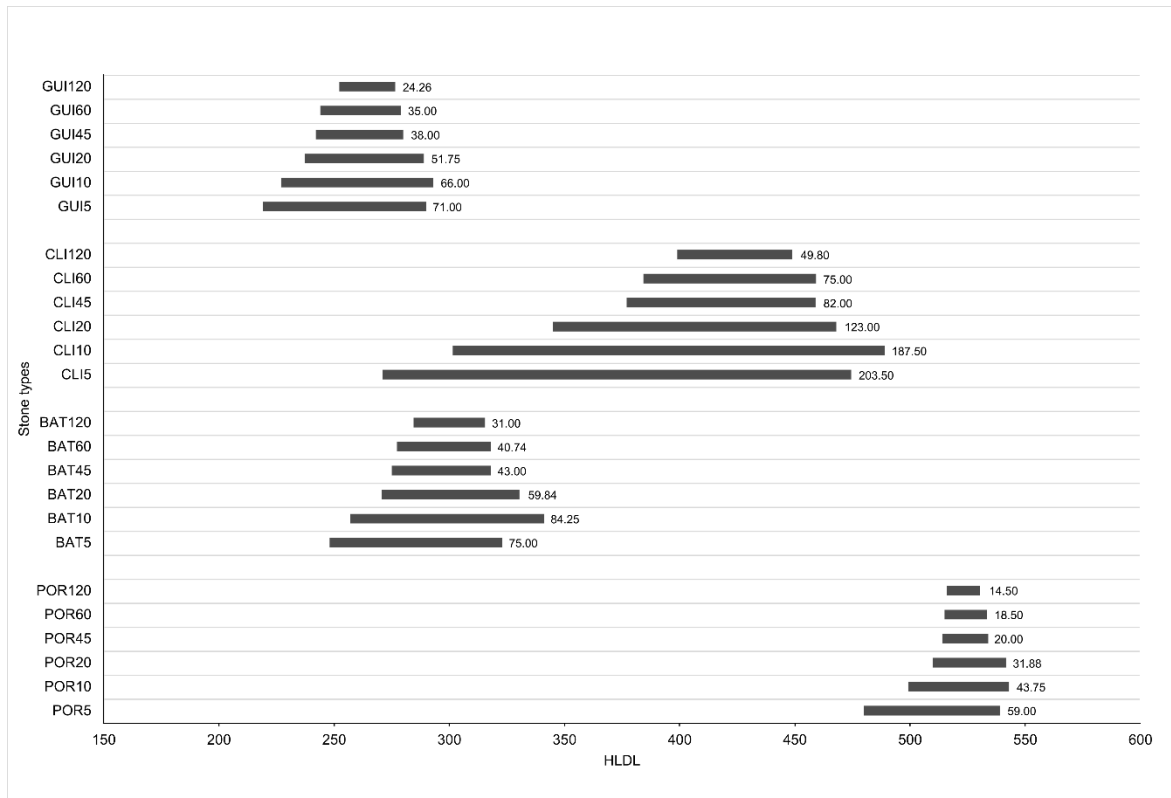


Figure 5b. Predicted confidence intervals for medians of Equotip Piccolo 2 DL probe data for different modelled sample sizes (numbers on the y-axis) on four different limestone in this study (Portland=POR, Bath=BAT, Clipsham=CLI, Guiting=GUI). Confidence intervals are obtained applying bias-corrected accelerated bootstrap to datasets of 120 readings. Modelled samples sizes are 5, 10, 20, 45, and 60 readings, resampled from the original dataset (120). Bars show confidence interval width (numeric value indicated) for median to occur within at 95% confidence level (See also table 6).

Tables

Table 1 Key characteristics of the Equotip 3 and Equotip Piccolo 2 mobile hardness testing devices

Key characteristics of probes			Equotip 3	<i>Equotip Piccolo 2</i>
Impact body	Ball intender type and diameter	Impact energy (Nmm)		
C	tungsten carbide, 3.0 mm	3.0	x	
<i>D</i>	<i>tungsten carbide, 3.0 mm</i>	<i>11.5</i>	<i>x</i>	<i>x</i>
DC	tungsten carbide, 3.0 mm	11.5	x	
<i>DL</i>	<i>tungsten carbide, 2.78 mm</i>	<i>11.1</i>	<i>x</i>	<i>x</i>
E	polycrystalline diamond	11.5	x	
G	tungsten carbide, 5.0 mm	90.0	x	
S	ceramics, 3.0 mm	11.5	x	
Measuring range			1 – 999 HL	<i>150 – 950 HLD / 250 – 970 HLDL</i>
Measuring accuracy			± 4 HL (0.5% at 800 HL)	<i>± 4 HL (0.5% at 800 HLD / HLDL)</i>
Impact direction automatic compensation			Yes (except DL probe)	Yes
Software			x	x
Weight			780 g	<i>142 g</i>

Note: Devices and probes tested in this study are shown in italic typeface.

768 Table 2 Existing research on Equotip in the field of rock and stone testing

Study	Device - Probe	Tested stone types / location	Sample size SIM/(RIM)	Application method/ Surface preparation (Y/N)	Data evaluation / Test for normality (Y/N) / Outliers (Y/N) / Modification (Y/N)
Aoki and Matsukura (2007)	Equotip 3 - D	Sandstone /On-site	10 (20)	SIM, RIM, k-value / N	Parametric / n.a. / n.a. / n.a.
Viles et al. (2011)	Equotip 3 - D Piccolo 2 - D	Limestone, sandstone, dolerite, basalt /On-site	50	SIM / Y & N	Parametric / n.a. / n.a. / n.a.
Mol and Viles (2012)	Equotip 3 - D	Sandstone /On-site	10	SIM / N	Parametric / n.a. / n.a. / n.a.
Yilmaz (2013)	Equotip 3 - D	Limestone, dolomite, marble, travertine /Laboratory	20 (10-20)	SIM, RIM, HDH / Y	Parametric / n.a. / n.a. / n.a.
Coombes et al. (2013)	Equotip 3 - D	Limestone, granite, concrete /On-site	80	SIM / N	Parametric / Y / n.a. / n.a.
Alberti et al. (2013)	Equotip 3 - D	Quartzite /On-site	600 total on 25 clasts at each of 24 outcrops	n.a./ Y	Parametric and non-parametric/ Y / Y / Y
Hansen et al. (2013)	Equotip 3 - D	Dolerite /On-site	15 per aspect, per clast (210 values in total)	SIM / N	Parametric / n.a./ n.a./ Y

¹ SIM = single impact method, RIM = repeated impact method with combinations of the two (SIM and RIM) being k-value and HDH = Hybrid Dynamic Hardness, Y=Yes, N=No (table modified after Yilmaz (2013)).

771 Table 3 Physico-mechanical properties of the tested stone (derived using standard procedures)
772 and surface hardness results D probe (HLD) and DL probe (HLDL).

Stone type	UCS [MPa] (min-max of n=10)	Open porosity [%] (min-max of n=6)	WAAP [Mass %] (min- max of n=6)	Apparent density [kg/m ³] (min-max of n=6)
Portland Base Bed	μ 55.98 med 52.65 (43.20-75.73)	μ 13.5 med 13.63 (13.12 – 13.82)	μ 6.71 (6.49 – 6.87)	μ 2205.99 (2177.65- 2223.31)
Bath Hartham Park	μ 16.04 med 15.76 (14.32-20.09)	μ 22.2 med 22.11 (21.11-23.51)	μ 11.84 (11.07-12.68)	μ 1984.45 (1954.6-2017.51)
Clipsham	μ 26.71 med 26.19 (17.37-50.65)	μ 15.63 med 16.33 (12.48-17.97)	μ 7.89 (6.23-9.53)	μ 2123.27 (1975.66- 2284.59)
Guiting	μ 11.15 med 11.82 (6.15 – 17.15)	μ 21.3 med 21.94 (16.1 – 24.96)	μ 11.55 (7.94 – 14.23)	μ 2004.54 (1796.41- 2376.79)

773 ¹ Water absorption under atmospheric pressure (WAAP) was tested using BS EN 13755. Unconfined
774 compressive strength (UCS) was tested using BS EN 1926:2006 and for porosity BS EN 1936:2006; μ=mean,
775 med=median.

776 Table 4 Overview of surface hardness data collected and calculated in this study

Hardness unit	Definition
$HLD_{S,mean}$	D-probe, single impact method, mean
$HLD_{S,SD}$	D-probe, single impact method, standard deviation
$HLD_{S,med}$	D-probe, single impact method, median
$HLD_{S,MAD}$	D-probe, single impact method, median absolute deviation
$HLDL_{S,mean}$	DL-probe, single impact method, mean
$HLDL_{S,SD}$	DL-probe, single impact method, standard deviation
$HLDL_{S,med}$	DL-probe, single impact method, median
$HLDL_{S,MAD}$	DL-probe, single impact method, median absolute deviation
$HLD_{R,med}$	D-probe, median of the 3 highest values in each of the 3 repeated impact method (RIM) datasets of 20 readings
$HLDL_{R,med}$	DL-probe, median of the 3 highest values in each of the 3 repeated impact method (RIM) datasets of 20 readings
$HDH_{D,robust}$	D-probe, robust hybrid dynamic hardness (combination of SIM and RIM)
$HDH_{DL,robust}$	DL-probe, robust hybrid dynamic hardness (combination of SIM and RIM)

777

Table 5 Surface hardness results for this study (120 readings per stone type)

Stone type	$HLD_{S,med}$ ($HLD_{S,MAD}$)	Conf.int. $HLD_{S,med}$ low	Conf.int. $HLD_{S,med}$ high	$HLD_{R,med}$	$HDH_{D,robust}$	$HLD_{S,mean}$ ($HLD_{S,SD}$)	Shapiro- Wilk test (p-value)	Skewness	Kurtosis
P	469.00 (27)	454.00	479.00	681.00	318.32	462.99 (42.29)	0.012	-0.423	1.848
B	241.50 (38)	233.00	254.00	583.00	99.28	258.72 (68.90)	<0.000	0.272	2.909
C	318.50 (62)	302.00	344.50	623.00	161.63	333.18 (90.46)	0.029	1.506	0.321
G	215.50 (22.5)	210.01	222.00	622.00	78.68	217.89 (41.84)	0.000	1.873	18.798

¹ HLD=values obtained with D-probe. P=Portland Jordans Base Bed, B=Bath Hartham Park, C=Clipsham, G=Guiting; Conf.int=confidence interval, low=lower boundary, high=upper boundary. Subscript key: Med=median. MAD=median absolute deviation, S=SIM (single impact method) and R=RIM (repeated impact method). (See also Figure 5a)

Table 6 Surface hardness results for this study (120 readings each stone type)

Stone type	$HLDL_{S,med}$ ($HLDL_{S,MAD}$)	Conf.int. $HLDL_{S,med}$ low	Conf.int. $HLDL_{S,med}$ high	$HLDL_R$ med	$HDH_{DL,robust}$	$HLDL_{S,mean}$ ($HLDL_{S,SD}$)	Shapiro -Wilk test (p- value)	Skewnes s	Kurtosis
P	525.00 (20)	516.00	500.00	766.00	363.94	527.35 (36.78)	0.294	-0.423	0.300
B	297.00 (40)	284.50	315.50	637.00	138.98	315.15 (83.88)	<0.001	0.272	5.893
C	422.00 (76.5)	399.00	448.80	758.00	267.93	414.15 (112.93)	0.117	1.506	-0.588
G	266.00 (30.5)	252.24	276.50	631.00	106.40	270.53 (51.86)	0.019	1.873	0.801

[†] HLDL=values obtained with DL-probe. P=Portland Jordans Base Bed, B=Bath Hartham Park, C=Clipsham, G=Guiting; Conf.int=confidence interval, low=lower boundary, high=upper boundary. Subscript key: Med=median. MAD=median absolute deviation, S=SIM (single impact method) and R=RIM (repeated impact method). (See also Figure 5b).

790 Table 7 Results of the Mann-Whitney U test for the **D** probe and the single limestone types
 791 (Portland=POR, Bath=BAT, Clipsham=CLI, Guiting=GUI). All stones of this study can significantly
 792 be distinguished from each other

Groups D probe		U	p-value
POR	BAT	335	<0.001
POR	CLI	353.5	<0.001
POR	GUI	32	<0.001
BAT	CLI	3415.5	<0.001
BAT	GUI	4440.5	<0.001
CLI	GUI	1439	<0.001

793

794 Table 8 Results of the Mann-Whitney U test for the **DL** probe and the single limestone types
795 (Portland=POR, Bath=BAT, Clipsham=CLI, Guiting=GUI). All stones of this study can significantly
796 be distinguished from each other

Groups DL probe		U	p-value
POR	BAT	411.5	<0.001
POR	CLI	2556	<0.001
POR	GUI	1	<0.001
BAT	CLI	3371	<0.001
BAT	GUI	4607	<0.001
CLI	GUI	2062.5	<0.001

797

798 Table 9 Pearson's R^2 and Spearman correlations for varying surface hardness data calculations
799 and median open porosity of the limestones tested in this study

	R^2	Spearman
<i>HLD</i>_{S.mean}	0.9082	-0.8
<i>HLD</i>_{S.med}	0.8971	-0.8
<i>HDH</i>_{D.robust}	0.8626	-0.8
<i>HLDL</i>_{S.mean}	0.9452	-0.8
<i>HLDL</i>_{S.med}	0.9757	-0.8
<i>HDH</i>_{DL.robust}	0.9785	-0.8

800

801 Table 10 Results of outlier detection using equation 3 (section 5.5.2). Notice the majority of outliers
802 is beyond the upper bound (i.e. extreme high hardness values) indicating the presence of fossils
803 and other harder elements

Stone	Probe	Total outliers	Beyond lower bound	Beyond upper bound
Portland	D	2	1	1
Portland	DL	6	1	5
Bath	D	8	0	8
Bath	DL	7	0	7
Clipsham	D	1	0	1
Clipsham	DL	0	0	0
Guiting	D	2	1	1
Guiting	DL	4	1	3

804

805 Table 11 **D** probe with Equotip Piccolo 2, percentage (%) differences of confidence interval widths
 806 for sampling sizes of 20, 45 and 60 readings (resampled) in comparison to a sample size of 120
 807 (original 'population')

D probe	% difference in confidence interval width			
Readings	POR [%]	BAT [%]	CLI [%]	GUI [%]
120 to 60	25.00	54.06	33.09	68.86
120 to 45	46.00	72.75	38.82	91.79
120 to 20	95.54	205.95	140.94	173.09

808

809 Table 12 **DL** probe with Equotip Piccolo 2, percentage (%) differences of confidence interval widths
 810 for sampling sizes of 20, 45 and 60 readings (resampled) in comparison to a sample size of 120
 811 (original 'population')

DL probe	% difference in confidence interval width			
Sample sizes compared	POR [%]	BAT [%]	CLI [%]	GUI [%]
60 and 120	27.59	31.40	50.62	44.25
45 and 120	37.93	38.71	64.67	56.62
20 and 120	119.85	93.02	147.01	113.29

812