1. Introduction

A fundamental concern of lithic analysts today is the exploration of the extent of reduction, and especially of mass removed from retouched flakes. Many different approaches to measuring extent of reduction have been devised and the imperative of assessing the relative reliability and analytical power of these approaches is now recognized. To this end we undertook a series of experiments designed to evaluate the robustness of key archaeological measures of flake retouching intensity (Hiscock and Clarkson, 2005a,b; Clarkson and Hiscock, 2008). We concluded that the most powerful measurement, under many circumstances, was a variant of the Geometric Index of Unifacial Retouch (GIUR) proposed by Kuhn (1990).

Our confidence in the GIUR was based on its strong, positive correlations with amount of mass lost through retouching, measured as log(%)weight loss, and our finding that blank cross-section has minimal impact on that capacity of the index to predict mass loss as some archaeologists had feared (Dibble, 1995). The strong correlation we found between GIUR and mass loss was obtained from experiments in which we retouched very different flake blanks, on different raw materials, and with different retouching strategies including soft and hard hammer as well as uni-marginal and bi-marginal (see details in Hiscock and Clarkson, 2005a,b; Clarkson and Hiscock, 2008). We argued that this clear, strong experimental correlation provides a sound basis on which to predict mass loss from GIUR recorded on archaeological specimens and hence GIUR can be a reliable index of extent of retouching in ancient tools. Furthermore, in our experiments we calculated alternative indices of retouching proposed in the literature and found that all were relatively weakly correlated with mass loss. Consequently we recommended that in appropriate circumstances a GIUR should be preferred to other measures, except where assemblage specific contraindications exist.

Eren and Sampson (2009) carried out an experiment, based on but not duplicating our procedures, and they obtained a high correlation ($r = 0.91, n = 147$) between GIUR and percentage of original flake weight lost. Clearly this confirms our experimental results and should be interpreted as supporting our conclusions about the value of the GIUR. Eren and Sampson, however, argue that our experiments were flawed and our conclusions incorrect. Their critique is founded on two arguments: 1) the validity of our experiments and hence our capacity to apply experimentally derived principles to archaeological specimens and hence GIUR can be a reliable index of extent of retouching in ancient tools. Furthermore, in our experiments we calculated alternative indices of retouching proposed in the literature and found that all were relatively weakly correlated with mass loss. Consequently we recommended that in appropriate circumstances a GIUR should be preferred to other measures, except where assemblage specific contraindications exist.

Eren and Sampson (2009) carried out an experiment, based on but not duplicating our procedures, and they obtained a high correlation ($r = 0.91, n = 147$) between GIUR and percentage of original flake weight lost. Clearly this confirms our experimental results and should be interpreted as supporting our conclusions about the value of the GIUR. Eren and Sampson, however, argue that our experiments were flawed and our conclusions incorrect. Their critique is founded on two arguments: 1) the validity of our experiments and hence our capacity to apply experimentally derived principles to archaeological specimens, and 2) the estimates of mass loss that result from the GIUR approach are imprecise. Neither point is correct and their evidence actually supports our conclusion that some GIUR versions will be reliable measures of the extent of retouching in appropriate circumstances. Eren and Sampson (2009) amplify the critique rehearsed recently by Eren and Prendergast (2008). In this paper we will explain why those comments have no value for lithic analysts. In particular we show there are no substantial flaws in the experimental design Hiscock

* Corresponding author.
E-mail address: peter.hiscock@anu.edu.au (P. Hiscock).
and Clarkson (2005a,b) published. We reiterate that our experiments reveal our variants of the GIUR to be robust measures of flake reduction, in appropriate circumstances, and that this approach is currently better than other procedures for estimating the extent of dorsal retouch on flakes.

2. Procedures in reduction experiments

Eren (Eren et al., 2005; Eren and Prendergast, 2008) was previously unable to obtain results comparable to those we published in 2005, and argued that this discrepancy should be investigated. In their paper Eren and Sampson (2009) conclude that the discrepancy mainly reflects the peculiarities of our experimental design and that the GIUR is therefore not experimentally verified. We disagree and begin our rejoinder by addressing their misunderstandings about the 2005 publications that described our first experiments.

In our first experimental series (Hiscock and Clarkson, 2005a), we held the shape and length of retouched edges constant, and delivered 10 flake-detaching blows to the ventral surface, spread evenly along a single margin (Fig. 1). The GIUR was calculated from paired flake thickness (T) and scar height (t) measurements taken at three places at equal intervals along the edge. Weight was also recorded. The procedure was then repeated, removing a further 10 flakes followed by measuring the resulting retouched flake. This process continued until each specimen broke. Thirty flakes were retouched in this way, yielding 348 data points. Correlations we presented, between GIUR and mass lost, were calculated using all data points, with multiple points from each retouched flake representing the relationship of these variables at different times in the retouching process. Details of experimental design are available in Hiscock and Clarkson (2005b).

Eren and Sampson (2009) argue that our choice of these procedures determined the experimental outcome. This is untrue and we can show why by addressing a number of their key concerns. Firstly, they assert our regression of multiple data points for each retouched flake is problematic because this did not replicate the ‘static’ condition of tools encountered by archaeologists in the field and artificially inflated the correlation coefficient we obtained by adding more points to an inevitable trend. These concerns are unfounded. Our experiments were designed to understand the relationship of GIUR and mass loss, and describing the changing relationship throughout the reduction of individual flakes was as relevant to this goal as describing the relationship expressed on different specimens with different reduction histories. In establishing principles by which archaeologists can infer mass loss from the GIUR it is not necessary to experiment using ‘static’ archaeological conditions, indeed as ethnoarchaeological and actualistic theory recognizes the purpose of experiments is to provide a connection between the ‘static’ archaeological observations and processes that create them but which cannot be directly observed in the archaeological record. The central question in these retouching experiments is whether the relationship of GIUR and mass loss is broadly similar for every retouched flake. By measuring only one point in the retouching history of each specimen it would be difficult to evaluate the similarity in reduction trends between flakes; our experimental design facilitated detailed assessments of the relationship between the GIUR and amount of material removed through retouching.

The proposition that our use of multiple data points from each specimen inflated the correlation coefficient is easily refuted by further analyses of our original experimental data. To this end we undertook 30 separate simulations in which a single data point was randomly selected from each retouched specimen, a procedure that allows us to determine what would happen to the correlation if we measured each specimen only once (such as at the point of abandonmen). The result was a series of extremely strong correlations between the percentage of initial weight lost (log transformed) and the GIUR, coefficients having the following statistics: \( r = 0.935 \pm 0.019 \), with a range of r values from 0.886 to 0.964. This average coefficient calculated with only one data point per specimen is statistically indistinguishable from the one we calculated with all 348 data points (\( r = 0.933 \)). Additionally, when we repeated the simulations excluding all GIUR values of 1, to approximate the limits of Eren and Sampson’s experiment, we still obtained similar correlation coefficients: \( r = 0.926 \pm 0.019 (N = 30) \), range from 0.884 to 0.955. All of these correlations are comparable to the one we published, demonstrating that our analytical procedure did not inflate the strength of the measured relationship. Hence the lower GIUR–mass loss coefficients produced experimentally by Eren et al. (2005) and Eren and Prendergast (2008) are not explained as a result of whether or not the experiment recorded multiple data points on each retouched flake; some other factor must be involved.

Another criticism of our experimental design is that our interval between measurements (10 flake-removing blows) was ‘unrealistic’. We discussed this choice of measurement unit elsewhere (Hiscock and Clarkson, 2005b); one of our concerns with Kuhn’s (1990) initial experiments was that his use of an ill-defined amount of retouching in each ‘retouch event’ created events of unequal magnitude. Our arbitrary but standardized unit of retouching provided sensitivity to identify the non-linear relationship between extent of retouch and GIUR (Hiscock and Clarkson, 2005b:12), and the increased sensitivity this procedure provides justifies its incorporation in our experiments. Eren and Sampson (2009) retained a version of Kuhn’s procedure, where they use their judgment as to how many flakes should be removed between measurements, and argue that our experiments did not ‘properly resharp’ edges. However we pose the questions ‘when is an edge properly retouched’ and ‘how do Eren and Sampson know it to be proper?’ The rules used to specify when an edge was ‘properly’ retouched are not provided. Normative and potentially ethnocentric judgements about what are correctly finished tools, such as...
Eren and Sampson have employed, cannot represent the diversity of decisions made by prehistoric knappers around the world. The view that a replicated flake must be ‘properly’ retouched fails to account for the numerous archaeological specimens which demonstrate that ancient knappers were prepared to make notched tools, denticulated tools, tools with only small amounts of retouch and tools massively retouched. These conceptual straight-jackets about how tools should be retouched threaten to make Eren and Sampson’s experiments less realistic. For example, their presumptions about acceptable tool designs led them to remove only small amounts of retouch from many of their experimental flakes (see below). Yet we know of no evidence that suggests prehistoric knappers always applied minimal retouch to flakes. The level of retouch, as well as the nature of retouch, was variable and contingent on many factors (for examples of these complexities see Hiscock, 2004, 2009; Clarkson, 2002, 2007). It was specifically to avoid subjective and unwarranted judgments on functionality that our experiments continued until specimens broke (Hiscock and Clarkson, 2005b).

We expand this point to emphasise that Eren and Sampson’s (2005) experiments did not duplicate our own. For instance, they claim that previous experiments by Eren had reduced blanks to ‘exhaustion’, yet the amounts of mass removed were typically small. Eren et al. (2005) removed no more than 23% of the original flake mass and most specimens were far less reduced ($\bar{x} = 8.4 \pm 4.7\%$). Eren and Sampson (2009) reduced one specimen to almost half its original weight, but approximately 60% of their data points represent reduction of less than 15%. In contrast data points in our experiments were more widely and evenly spread. In our first experiments (Hiscock and Clarkson, 2005a,b) the average mass removed was substantially higher ($\bar{x} = 22.8 \pm 20.3\%$), some specimens were retouched until more than 70% of original weight had been removed and approximately 15% of data points represented reduction more than 50% of initial flake mass. Extended reduction, like the diverse flake blanks we employed, was a way to determine the range of conditions in which GIUR reliably indicates the amount of mass loss. By extending reduction as far as mechanically possible we explored the operation of the GIUR/mass loss relationship in extreme situations without imposing our notions of appropriate tool form on the experiment. By comparison to our experiments those of Eren and Sampson (2009) as well as Eren et al. (2005) and Eren and Prendergast (2008) were limited explorations of the relationship between reduction and measure GIUR and mass loss.

There are other differences between our experimental procedures and the ones used by Eren and Sampson. For instance, it seems Eren and Sampson measured $T$ only once, making it possible for their calculated GIUR values to decrease on extensively retouched specimens, where retouching had removed the thickest part of the flake, whereas we always took paired $t/T$ measurements making it impossible for the GIUR value to decrease as retouch continued.

We make these points to explain that Eren and Sampson’s (2009) carried out an experiment which was quite different from our own. This is a critical point. Since Eren and Sampson (2009) obtained a strong correlation between GIUR and mass loss with a different experimental design they have actually demonstrated that the coefficient we reported was not merely a product of our experimental design but reflects the robust relationship between this index and the extent of mass removed through retouching! This robust relationship can be identified in experiments of different designs as long as they are sensitive to the non-linear co-variation that exists.

However, nothing we said in our 2005 papers suggested the GIUR we used in experiments would be the most appropriate measure of GIUR in all circumstances. It is important to understand that there are many variants of the GIUR concept, and that they have different qualities and strengths. For instance, in our first experimental series we retouched a single margin and calculated average GIUR from measurements at three equally spaced intervals. This approach would be less suited to flakes which had been retouched on two or three margins, and could result in a retouched margin being characterized by only one measurement. In situations of extended marginal retouch a more elaborate variant would be desirable. For this reason we employed a zonal system in our work on Middle Palaeolithic retouched flakes in France (Hiscock and Clarkson, 2007, 2008). In this system each specimen was divided into eight zones, as shown in Fig. 2, and one or more GIUR measurements was taken for each zone containing retouch. This approach allows average GIUR to be calculated on specimens with more than one retouched edge, and hence the experiment that Eren and Sampson (2009) carried out is not a test of the applicability of the GIUR measurements we recently used on Middle Palaeolithic tools. In fact we have already experimentally verified the applicability of this zonal approach.

We have published two series of experiments, each with its own production rules. Only our first experiments are the subject of critique by Eren and Sampson, and we now explore the implications of our second experiment (Clarkson and Hiscock, 2008). In this experiment we retouched both lateral margins as well as the distal end. As shown in Fig. 3, we notionally divided each flake into eight segments and then applied 10 flake-removing blows to each segment, expanding one zone at a time along both lateral margins until the entire perimeter was retouched. Retouch began at the distal end and when retouch had been applied to all zones we again retouched the distal segment and then adjoining zones. This pattern was repeated until the flake broke. After each zone was retouched we measured the $t/T$ values and re-calculated the average GIUR based on all $t/T$ pairs currently present on the specimen. Details are provided in Clarkson and Hiscock (2008). Even with multiple margins being retouched, and retouch expanding around the flake perimeter, there was an extremely high correlation between GIUR and percentage of mass lost (log), with a coefficient of 0.883 ($N = 316$). This experiment demonstrates the applicability of a GIUR in estimating mass loss in appropriate circumstances, even on specimens with two retouched margins. Furthermore, this was another, quite different, experimental design which also revealed the strong correlation between mass removed and GIUR. We reiterate that the diversity of experiments showing very strong correlations refute Eren and Sampson’s claim that such correlations are an inevitable outcome of our 2005 experimental
between 0.004 and 0.26, virtually the entire range of values 0.12–0.13 represented an 8% mass loss for one specimen and 17% for specimen but for another specimen a 24% mass loss, while an ERP of an ERP value of 0.25–0.26 represented an 8% mass loss for one percentage' (ERP). In experiments published by Eren et al. (2005) of variation are documented for Eren's own 'estimated reduction indices and is not specific to the GIUR. Comparable levels a particular reduction index. This pattern is shared by many rather than a problem with a particular experimental design or they reflect the massive variability inherent in reducing flakes lost in reduction. These observations are entirely correct, although

3. Predicting amount of mass lost in retouching

Eren and Sampson seek an ideal measure of mass lost that would have only a single mass loss value, or a very small range of values, for any particular reduction index value. This would be a desirable situation, but the real world is complex and archaeological data rarely display simple one-to-one relationships that allow us to predict with certainty a non-material value from preserved physical materials. For instance radiocarbon age-estimates may measure a value which could be one of several ages. Flake reduction is another example because an amount of mass can be removed in many ways from a flake, sometimes creating different morphologies for the same amount of reduction.

Does this mean the GIUR cannot measure the amount of mass removed during retouching? In assessing this question Eren and Sampson (2009), like Eren and Prendergast (2008), focus on the range of mass loss values for specific GIUR. Using their experimental results Eren and Sampson (2009) point out that for GIUR values of 0.4–0.59 estimated mass loss might vary by 8% between specimens, and differences in mass loss between specimens were greater for specimens with higher GIUR. Eren and Prendergast (2008) previously pointed out that our experiments (Hiscock and Clarkson, 2005a) produced results that also contained similar variation in the relationship of GIUR and the percentage of weight lost in reduction. These observations are entirely correct, although they reflect the massive variability inherent in reducing flakes rather than a problem with a particular experimental design or a particular reduction index. This pattern is shared by many reduction indices and is not specific to the GIUR. Comparable levels of variation are documented for Eren's own 'estimated reduction percentage' (ERP). In experiments published by Eren et al. (2005) an ERP value of 0.25–0.26 represented an 8% mass loss for one specimen but for another specimen a 24% mass loss, while an ERP of 0.12–0.13 represented an 8% mass loss for one specimen and 17% for another. In fact an 8–9% mass loss was represented by ERP values between 0.004 and 0.26, virtually the entire range of values produced by Eren et al. (2005) in their experiments. The logic of Eren and Sampson (2009) therefore demands their view, that experiments do not verify the GIUR as a reliable measure of retouch extent, must be equally applied to Eren's own ERP index. The ERP, like the GIUR that Eren and Sampson want to dismiss, has a complex relationship with mass lost through retouching. Should both measures consequently be abandoned as useless?

Imperfect as it may be the GIUR remains valuable, robust measure of the extent of retouching, and our experiments demonstrated this conclusion in several ways. The first is that the diversity of experimentally defined mass loss/GIUR values should be assessed in light of the variability in the sizes and shapes of flakes which were retouched. As discussed above our experiments purposefully employed flakes of very different sizes and cross-sectional shapes. This was done explicitly to determine how variable the relationship between mass loss and GIUR would be with extreme differences in blanks and the experimental design had the effect of exaggerating variability in that relationship. Since in many assemblages flakes selected for retouching will be less varied than ones in our experiments, our correlations, and conclusions we developed from them, are conservative. This enhances confidence in the analytical power of a GIUR measurement.

More importantly Eren and Sampson (2009) are statistically naive in their critique of our experimental results. Their use of the range of values to understand the relationship between the two variables of interest is by definition an expression of only extreme observations and does not depict the distribution or central tendency of observations. Their focus on range guarantees an imperfect as it may be the GIUR remains valuable, robust measure of the extent of retouching, and our experiments demonstrated this conclusion in several ways. The first is that the diversity of experimentally defined mass loss/GIUR values should be assessed in light of the variability in the sizes and shapes of flakes which were retouched. As discussed above our experiments purposefully employed flakes of very different sizes and cross-sectional shapes. This was done explicitly to determine how variable the relationship between mass loss and GIUR would be with extreme differences in blanks and the experimental design had the effect of exaggerating variability in that relationship. Since in many assemblages flakes selected for retouching will be less varied than ones in our experiments, our correlations, and conclusions we developed from them, are conservative. This enhances confidence in the analytical power of a GIUR measurement.

More importantly Eren and Sampson (2009) are statistically naive in their critique of our experimental results. Their use of the range of values to understand the relationship between the two variables of interest is by definition an expression of only extreme observations and does not depict the distribution or central tendency of observations. Their focus on range guarantees an imperfect as it may be the GIUR remains valuable, robust measure of the extent of retouching, and our experiments demonstrated this conclusion in several ways. The first is that the diversity of experimentally defined mass loss/GIUR values should be assessed in light of the variability in the sizes and shapes of flakes which were retouched. As discussed above our experiments purposefully employed flakes of very different sizes and cross-sectional shapes. This was done explicitly to determine how variable the relationship between mass loss and GIUR would be with extreme differences in blanks and the experimental design had the effect of exaggerating variability in that relationship. Since in many assemblages flakes selected for retouching will be less varied than ones in our experiments, our correlations, and conclusions we developed from them, are conservative. This enhances confidence in the analytical power of a GIUR measurement.
interval (CI) for each GIUR category rarely overlaps with the CI of neighbouring categories and there is a consistent increase in CI as GIUR increases, so that very different GIUR values clearly indicate very different mass loss values. For instance GIUR values of 0.2, 0.6 or 0.9 all predict different mean mass loss values within a relatively small CI. While the size of the CI increases as GIUR increases it is nevertheless still reasonable to estimate mass loss from the GIUR. Of course this procedure does not give correct mass loss statements for every specimen, outliers will exist, but it provides a sound guide to the likely mass removed for the vast majority of specimens.

The confidence lithic analysts can have in predicting likely mass loss from the GIUR is also expressed through the effect size: an objective measure of the magnitude of effect that change in the x axis (GIUR) has on the value plotted on the y axis (mass loss). The ability of a reduction index (x) to accurately estimate the likely mass removed (y) can be expressed by the coefficient of determination (r²) for the regression line describing the relationship between the two variables. It is a standard statistical proposition, reflecting smaller residuals for higher r² values. This coefficient is often converted into a percentage (multiply r² by 100) to give a figure that expresses the percentage of y variation that is described or explained by the x value. Of course we are not asserting that the GIUR literally causes the mass loss value, we are simply using the strong relationship to predict mass loss from the measured GIUR.

In our published experiments coefficient of determinations are high. For retouch on a single margin 87% (r² = 0.871) of variation in mass loss was expressed by reference to the GIUR value (Hiscock and Clarkson, 2005a:1020). For retouch on both margins 78% (r² = 0.779) of variation in mass loss was expressed by GIUR value. When we combined both experiments 78% (r² = 0.778) of variation in mass loss was still explicable in terms of the GIUR value, irrespective of where or how the specimens were retouched (Clarkson and Hiscock, 2008:12). Such coefficients are remarkably strong and create the opportunity to predict mass loss from GIUR. We remind readers that this predictive power exists in spite of our retouching blanks of widely varying shapes.

The inferential power of the GIUR in our experiments is dramatically greater than that offered by other proposed indices for unifacial reduction, revealed by a coefficient of determination of only 45% (r² = 0.454) for Eren's ERP (Eren et al., 2005), or a mere 7% (r² = 0.067) for Dibble’s surface-area/platform-area ratio (Hiscock and Clarkson, 2005a:1021). In light of the lower predictive capacity of those indices it is curious that Eren and Sampson (2009) prefer them to a version of the GIUR.

4. Conclusions

There is an inherent contradiction in Eren and Sampson’s statement that GIUR will inevitably increase as mass is lost and their conclusion that GIUR cannot predict mass loss. Given that their experiment duplicated the extraordinarily strong correlation between GIUR and mass loss that we had found there can be little doubt that a GIUR offers the opportunity to predict likely mass loss on individual specimens, in appropriate circumstances. Eren's inability to find this strength of correlation in earlier experiments indicates they were not sensitive enough, whereas the experimental design employed by Eren and Sampson (2009) has largely overcome those deficiencies and hence documents the strong relationship of GIUR and mass lost.

There is no ambiguity in the existing empirical support for GIUR as a more powerful predictor of mass loss than touted competitors for estimating the extent of unifacial retouch. For this reason researchers should use the GIUR in preference to other methods. We have previously acknowledged that it is also clear that there are conditions in which a GIUR, as we have measured it, may be less than optimal: extremely flat flakes with no ventral curvature, retouch scars of extremely different heights on opposing margins, and so on. As we have explained, in such sub-optimal circumstances analysts have the choice to filter out inappropriate specimens (on which the GIUR would be unreliable) or adopt alternative or additional indices. Furthermore, we encourage the kind of conservative approach to interpreting the GIUR which we employ. For example, our analyses of Middle Palaeolithic assemblages used the experimentally derived GIUR confidence intervals as an indicator of likely mass lost through retouching, but we cross-checked those inferences against other characteristics that also indicate the extent of retouching, such as the preserved angle of unretouched dorsal/ventral surfaces on retouched edges and the marginal extent of retouch scars (Hiscock and Clarkson, 2007, 2008). Interpreting the extent of retouch on individual flakes is a complex task and the GIUR is currently the most powerful single index available to lithic analysts. We have demonstrated that the evidence presented by Eren and Sampson (2009) supports our conclusion that some GIUR versions are reliable reduction indices in appropriate circumstances, and that at the moment this approach is better than and
should be preferred to other measures of estimating the extent of dorsal retouch on flakes.

References