Supplementary Material 1

This document contains a more in-depth explanation of the performed statistical tests.
The interested reader is also advised to read the accompanying HTML report (FkergStats.html).

- Statistical power

We used this test to answer the question whether 40 specimens are enough to separate populations at a given power/for a given difference between average and measured value.
We evaluated statistical power for the valve area data from the two samples, with the highest/lowest standard deviation among the Termination I dataset (PS2498, 40cm and PS1654, 1202cm, respectively). The former sample (Avg. area = 109.66 μm^2; St.dev. = 35.88) comes from an interglacial interval, the latter (Avg. area = 189.36 μm^2; St.dev. = 90.68) from a glacial interval.
We also tested an alternative formulation of the statistical power test (i.e., what is the separable value, via a t-test and at 0.90 power, if the sample size is 40) for the whole of the valve area measurements for MIS 1 and MIS 2.

Results

A sample size of 40 is enough to separate, via a t-test and at 0.90 power, the estimated average value for sample PS2498, 40cm (109.66 μm^2) from another measurement, provided the latter is smaller than 100 or larger than 120 μm^2.
A sample size of 40 is enough to separate, via a t-test and at 0.90 power, the estimated average value for sample PS1654, 1202cm (189.36 μm^2) from another measurement, provided the latter is smaller than 150 or larger than 230 μm^2.
The separable valve area size (via a t-test, at 0.90 power, and sample size of 40) for all interglacial and all glacial valve area measurements are 158.50 and 160.17 μm^2, respectively.

Conclusions

The average values (separable size in brackets) for all interglacial and all glacial valve area measurements are 126.18 (158.50) and 198.39 (160.17) μm^2, respectively. A sample size of 40 measured specimens is therefore adequate to be able to separate, at power 0.90, a sample coming from a glacial interval from a sample coming from an interglacial interval, and vice versa.
This is a first indication of what later confirmed by both the Two-paired Kolmogorov-Smirnov test and the Wilcoxon rank sum test (i.e., the glacial and interglacial valve area averages/medians are significantly different).

- Chi-square test

We checked whether the valve area measurements were normally distributed, as this as a pre-requisite for further testing (e.g., t-test to evaluate whether populations are statistically different at a certain confidence level).
We tested the two samples used for the statistical power test above (PS2498, 40cm and PS1654, 1202cm), and the whole of Termination I MIS 1 and MIS 2 valve area measurements.
The Chi-square test performs a chi-square goodness-of-fit test of the default null hypothesis that the data in a vector are a random sample from a normal distribution with mean and variance estimated from the data themselves, against the alternative that the data are not normally distributed with the estimated mean and variance.
The test is performed by grouping the data into bins, calculating the observed and expected counts for those bins, and computing the chi-square test statistic.
We ran two versions of the test:

a) one with automatic selection of bins, in order to comply with the minimum requisite of five observations per bin;

b) the other still versus the normal distribution, but with a fixed number of bins: ten

Results
In the version of the test using a fixed number of bins, the null hypothesis was rejected at the 5% significance level (i.e., distributions are not normal) for all tested data (sample PS2498, 40cm, sample PS1654, 1202cm , all valve area measurements for MIS 1, all valve area measurements MIS 2).
In the version of the test using an automatic selection for the number of bins (pooling data from adjacent bins), the null hypothesis was rejected at the 5% significance level (i.e., distributions are not normal) for sample PS1654, 1202cm (chi square statistic = 9.26, $p<0.01$), all valve area measurements for MIS 1 (chi square statistic = 86.68, $p<0.01$), and all valve area measurements MIS 2 (chi square statistic = 171.74, $p<0.01$).
The null hypothesis could not be rejected (albeit at a high level of p) at the 5% significance level for sample PS2498, 40cm (chi square statistic = 1.81, $p=0.40$).

Conclusions
As most of the tested populations had a non-normal distribution, the further tests we used to evaluate our datasets were non-parametric.

- Two different non-parametric tests
We applied both the Two-sample Kolmogorov-Smirnov test and the Wilcoxon rank sum test/Mann-Whitney U-test to evaluate whether the glacial and interglacial populations, as characterized by their valve area size, are statistically different.

The Two-sample Kolmogorov-Smirnov test compares the distributions of the values in two data vectors, with the null hypothesis they are from the same continuous distribution, against the alternative hypothesis that they are from different continuous distributions.

Results
The test rejected the null hypothesis, i.e. all tested populations differed significantly at the 5% significance level (see details below).
Tested populations: Termination I MIS 1 vs. MIS 2 valve area measurements from each core (PS2498, PS2499, PS1654), and all three cores grouped together.

All Termination 1 MIS1 vs. MIS2 valve area measurements
The test rejected the null hypothesis at the 5% significance level, i.e. the distributions in the two groups differed significantly (Two-sample K-S statistic = 0.48, $p<0.01$).

PS2498 - Termination 1 MIS1 vs. MIS2 valve area measurements
The test rejected the null hypothesis at the 5% significance level, i.e. the distributions in the two groups differed significantly (Two-sample K-S statistic = 0.62, $p<0.01$).

PS2499 - Termination 1 MIS1 vs. MIS2 valve area measurements
The test rejected the null hypothesis at the 5% significance level, i.e. the distributions in the two groups differed significantly (Two-sample K-S statistic = 0.56, $p<0.01$).

PS1654 - Termination 1 MIS1 vs. MIS2 valve area measurements
The test rejected the null hypothesis at the 5% significance level, i.e. the distributions in the two groups differed significantly (Two-sample K-S statistic = 0.28, $p<0.01$).

The Wilcoxon rank sum test (equivalent to the Mann-Whitney U-test) performs a two-sided rank sum test of the null hypothesis that the values in two data vectors are independent samples from identical continuous distributions with equal medians, against the alternative that they do not have equal medians.

Results

The test rejected the null hypothesis, i.e. all tested populations and their medians differed significantly at the 5% significance level (see details below).
Tested populations: Termination I MIS 1 vs. MIS 2 valve area measurements from each core (PS2498, PS2499, PS1654), and all three cores grouped together.

All Termination 1 MIS1 vs. MIS2 valve area measurements

The medians for the MIS 1 and MIS 2 size groups were 112.31 and 188.85 μm^2, respectively.
The test rejected the null hypothesis at the 5% significance level, i.e. the distributions in the two groups differed significantly (rank sum value = $3.69*e^5$ = 547.64, z-statistic = -21.4, n_1 (sample size MIS 1) = 555, n_2 (sample size MIS2) = 1914, $P<0.01$ two-tailed).

PS2498 - Termination 1 MIS1 vs. MIS2 valve area measurements

At this location, the medians for the MIS 1 and MIS 2 size groups were 100.22 and 185.41 μm^2, respectively.
The test rejected the null hypothesis at the 5% significance level, i.e. the distributions in the two groups differed significantly (rank sum value = 44376, z-statistic = -16.5, n_1 (sample size MIS 1) = 245, n_2 (sample size MIS2) = 467, $P<0.01$ two-tailed).

PS2499 - Termination 1 MIS1 vs. MIS2 valve area measurements

At this location, the medians for the MIS 1 and MIS 2 size groups were 100.78 and 190.91 μm^2, respectively.
The test rejected the null hypothesis at the 5% significance level, i.e. the distributions in the two groups differed significantly (rank sum value = 22863, z-statistic = -10.4, n_1 (sample size MIS 1) = 88, n_2 (sample size MIS2) = 1269, $P<0.01$ two-tailed).

PS1654 - Termination 1 MIS1 vs. MIS2 valve area measurements

At this location, the medians for the MIS 1 and MIS 2 size groups were 135.02 and 177.95 μm^2, respectively.
The test rejected the null hypothesis at the 5% significance level, i.e. the distributions in the two groups differed significantly (rank sum value = 42168, z-statistic = 5.6, n_1 (sample size MIS 1) = 222, n_2 (sample size MIS2) = 178, $P<0.01$ two-tailed).

Conclusions

Both tests confirmed how the valve area values measured in samples coming from glacial and interglacial intervals are significantly different, both within the same core (tested for PS2498, PS2499, PS1654) and for the whole Termination I dataset.