AGU PUBLICATIONS

Geochemistry, Geophysics, Geosystems

RESEARCH ARTICLE
10.1002/2013GC005135

Special Section:
Magnetism From Atomic to Planetary Scales: Physical Principles and Interdisciplinary Applications in Geo- and Planetary Sciences

Key Points:
- Standard definitions for paleointensity statistics are proposed
- A large paleointensity meta-analysis is conducted to investigate data selection
- Modifications based on SD predictions improve the effectiveness of selection

Supporting Information:
- SPD v1.0
- ReadMe

Correspondence to:
G. A. Paterson,
greig.paterson@mail.iggcas.ac.cn

Citation:
Paterson, G. A., L. Tauxe, A. J. Biggin, R. Shaar, and L. C. Jonestrask (2014), On improving the selection of Thellier-type paleointensity data, *Geochem. Geophys. Geosyst.*, *15*, 1180–1192, doi:10.1002/2013GC005135.

Received 6 NOV 2013
Accepted 1 MAR 2014
Accepted article online 5 MAR 2014
Published online 22 APR 2014

On improving the selection of Thellier-type paleointensity data

Greig A. Paterson[1], Lisa Tauxe[2], Andrew J. Biggin[3], Ron Shaar[2], and Lori C. Jonestrask[2]

[1]Key Laboratory of Earth's Deep Interior, Institute of Geology and Geophysics, Chinese Academy of Sciences, Beijing, China, [2]Geosciences Research Division, Scripps Institution of Oceanography, University of California San Diego, La Jolla, California, USA, [3]Geomagnetism Laboratory, Department of Geology and Geophysics, School of Environmental Sciences, University of Liverpool, Oliver Lodge Laboratories, Liverpool, UK

Abstract The selection of paleointensity data is a challenging, but essential step for establishing data reliability. There is, however, no consensus as to how best to quantify paleointensity data and which data selection processes are most effective. To address these issues, we begin to lay the foundations for a more unified and theoretically justified approach to the selection of paleointensity data. We present a new compilation of standard definitions for paleointensity statistics to help remove ambiguities in their calculation. We also compile the largest-to-date data set of raw paleointensity data from historical locations and laboratory control experiments with which to test the effectiveness of commonly used sets of selection criteria. Although most currently used criteria are capable of increasing the proportion of accurate results accepted, criteria that are better at excluding inaccurate results tend to perform poorly at including accurate results and vice versa. In the extreme case, one widely used set of criteria, which is used by default in the Thellier-Tool software (v4.22), excludes so many accurate results that it is often statistically indistinguishable from randomly selecting data. We demonstrate that, when modified according to recent single domain paleointensity predictions, criteria sets that are no better than a random selector can produce statistically significant increases in the acceptance of accurate results and represent effective selection criteria. The use of such theoretically derived modifications places the selection of paleointensity data on a more justifiable theoretical foundation and we encourage the use of the modified criteria over their original forms.

1. Introduction

Reconstructing the evolution of Earth's geodynamo requires detailed and accurate records of long-term geomagnetic field variations. Obtaining reliable estimates of the ancient geomagnetic field strength (paleointensity) can be a challenge and numerous factors, such as chemical alteration or multidomain (MD) grains, result in high data rejection rates, or the acceptance of inaccurate results. The use of appropriate selection criteria to discriminate against such factors is essential if we want to understand long-term paleointensity behavior. At present paleointensity data selection is a notoriously arbitrary process that lacks a solid theoretical foundation. The modern approach to data selection was established 35 years ago by *Coe et al*. [1978], but despite numerous investigations of paleointensity statistics [e.g., *Tauxe and Staudigel*, 2004; *Chauvin et al*., 2005; *Biggin et al*., 2007; *Paterson et al*., 2012], no consensus exists as to how to select paleointensity data appropriately. In this study, we lay the foundations for an approach to paleointensity data selection that removes ambiguity and transforms the selection process into a more theoretically justifiable endeavor.

To date, more than 40 paleointensity statistics have been proposed and are used regularly in modern studies. Differences in the details of the calculations between different laboratories or software packages often mean that data, and statistical characterization thereof, are inconsistent. To overcome this, we introduce the Standardized Paleointensity Definitions (section 2), which is a new reference document that outlines the definitions and calculations of paleointensity data. In section 3, we describe the largest-to-date compilation of raw paleointensity data from experiments where the true paleointensity is known. This is now available for download from the MagIC database (earthref.org/MAGIC/). In this section, we also outline current data selection criteria sets and new analyses to quantify and assess their effectiveness. Using these analyses and the compiled data set, we assess the effectiveness of commonly used sets of selection criteria in section 4. We then modify these criteria sets according to theoretically predicted ideal single domain (SD) behavior [*Paterson et al*., 2012; *Paterson*, 2013] and demonstrate that these modifications improve the overall success of the selection process.

2. Standardizing Paleointensity Statistics

Through the authors' experiences, discussions with the paleomagnetic community, examination of open source code, and through reanalysis of published data, it has become clear that there are inconsistencies in the quantification of paleointensity statistics. Many of these inconsistencies are small and may be attributed to numerical rounding. Others, however, are more substantial and may influence the outcome of data selection and the comparison of studies. One example is the fraction (f) of natural remanent magnetization (NRM) used for the best-fit on the Arai plot [as defined by *Coe et al.*, 1978], which is one of the most widely used paleointensity statistics. Through our work, we have encountered three different methods of calculating f, which can be substantially different.

To ensure that paleointensity statistics are consistently calculated we have written the Standard Paleointensity Definitions (SPD). SPD outlines both the textual and mathematical definitions for the calculation of paleointensity estimates and for over 40 statistics used to select data. SPD also describes the theory and mathematics of applying corrections to paleointensity data affected by anisotropic thermoremanent magnetization (TRM) [*Veitch et al.*, 1984; *Chauvin et al.*, 2000] and nonlinear TRM acquisition [*Selkin et al.*, 2007].

The SPD is a reference document to allow paleointensity analysts to consistently quantify their data. To facilitate this, the SPD includes numerical and programming advice that will help to ensure that paleointensity data are accurately and efficiently determined across all platforms of analysis. In addition, we provide a reference data set, which contains the raw paleointensity data and the statistics for 20 specimens. This data set can be used by developers of paleointensity software to ensure that their analyses are consistent with SPD.

Version 1.0 of the SPD is attached as supporting information. This and future versions of SPD along with examples code are available from http://www.paleomag.net/SPD. We welcome all comments and suggestions to help further improve SPD and the consistency of paleointensity analysis.

3. Data and Methods

3.1. Historical Data

To investigate the effectiveness of paleointensity data selection, we require an extensive data set from specimens where the expected paleointensity is known. To that end, we have compiled the raw data from 395 specimens obtained from historical volcanoes or laboratory experiments. The data set is summarized in Table 1 and consists of data from 13 studies, which represent 15 localities or laboratory experiments (18 unique heating events). Full details of the experimental protocols and measurements are given in the respective references.

All studies used variants of the Coe or IZZI protocols [*Coe*, 1967; *Yu et al.*, 2004] using either conventional thermal [e.g., *Yamamoto and Hoshi*, 2008] or microwave techniques [*Biggin et al.*, 2007]. Thirty-six specimens ($\sim$9%) are from the microwave technique [*Biggin et al.*, 2007] and 54 ($\sim$14%) from the IZZI protocol [*Paterson et al.*, 2010b; *Shaar et al.*, 2010]. The largest data set from a single event is that of *Bowles et al.* [2006], which contains 53 specimens and constitutes $\sim$13% of the data set. Our compilation is composed of a range of different materials and includes geological as well as archeological materials. Basalts and andesites are the most abundant material in the data set contributing 128 ($\sim$33%) and 117 ($\sim$30%) specimens, respectively. Synthetic magnetite specimens, which are typically MD in nature [*Muxworthy*, 1998; *Krása et al.*, 2003], constitute $<$3% of the entire data set. The diversity of the data set means that our results should not be systematically biased by data from any single study.

The availability of the original raw data allows the calculation of paleointensity statistics that were not used in the original studies. All statistics are calculated according to SPD v1.0. Where required, anisotropy and nonlinear TRM corrections have been made, but cooling rate corrections are not necessary or are negligible ($\ll$10%) for these specimens, due to either grain size considerations [*Biggin et al.*, 2013] or given that NRM and laboratory TRM cooling rates are known to be identical. To allow comparison between different studies, paleointensity estimates are normalized by the expected values. All of the data are available for download from the MagIC database or from http://www.paleomag.net/SPD.

3.2. Paleointensity Selection Criteria

Most studies use a unique combination of selection criteria, but a number of criteria sets are frequently used in the literature and their effectiveness will be tested here. These criteria sets include PICRIT03 [*Kissel and Laj*,

Table 1. The Paleointensity Data Sets Used in this Study

Reference	Location(s)	N	pTRM Checks	pTRM Tail Checks	Method	Material[a]	B_{Exp}[b] (μT)	Comment
Pick and Tauxe [1993]	East Pacific Rise: 1990	12	Yes	No	Coe	SBG	37.0	
Muxworthy [1998]	N/A	4	No	Yes	Coe	Synthetic magnetite	100.0	Average grain size 7.5–27.5 μm
Selkin et al. [2000]	Stillwater complex	8	Yes	No	Coe	Anorthosite	25.0	Laboratory induced remanence. Corrected for anisotropy
Krása et al. [2003]	N/A	7	Yes	Yes	Coe	Synthetic magnetite	25.0, 60.0	Average grain size 0.023–12.1 μm
Yamamoto et al. [2003]	Hawaii: 1960	22	Yes	No	Coe	Basaltic lava		
Bowles et al. [2006]	East Pacific Rise: 1991/92	53	Yes	Yes	Coe	SBG	35.8	pTRM tail checks on 31 specimens only
Biggin et al. [2007]	Mt. Etna: 1950, 1979, 1983	36	Yes	Yes	Coe	Basaltic lava	43.3, 44.1, 44.2	Microwave. Eighteen specimens partially AF cleaned
Donadini et al. [2007]	Helsinki: 1906	8	Yes	Yes	Coe	Brick	49.6	
Yamamoto and Hoshi [2008]	Sakurajima: 1914, 1946	72	Yes	Yes	Coe	Andesitic lava	45.7, 46.0	
Paterson et al. [2010b]	Mt. St. Helens: 1980; Láscar: 1993	86	Yes	Yes	Coe (52) + IZZI (34)	Andesite, basalt (Mt. St. Helens only), and dacite	55.6; 24.0	Lithic clasts within pyroclastic deposits
Shaar et al. [2010]	N/A	20	Yes	Yes	IZZI	Remelted copper slag	30.0, 60.0, 90.0	All specimens anisotropy corrected. Ten specimens corrected for nonlinear TRM
Muxworthy et al. [2011]	Parícutin: 1943; Vesuvius: 1944	64	Yes	Yes	Coe	Basaltic lava	45.0; 44.0	
Tanaka et al. [2012]	Krafla: 1984	3	Yes	No	Coe	Basaltic lava	52.1	

[a] SBG, submarine basaltic glass.
[b] With the exception of the data sets that are laboratory based, all expected field values are based on DGRF models.

2004], SELCRIT2 [*Biggin et al.*, 2007], and the ThellierTool A and B criteria sets [*Leonhardt et al.*, 2004] (herein referred to as TTA and TTB, respectively). The definitions of the statistics used in these sets are given in the SPD and the threshold values used for selection are given in Table 2. The TTA and TTB criteria are the default values from v4.22 of the ThellierTool. We note that PICRIT03 uses the criterion $\alpha' \leq 15°$, where α' is the angular difference between the anchored best fit direction from the paleointensity experiment and an independent measure of the paleomagnetic direction (e.g., from a separate demagnetization experiment or a known direction). Only about half of the specimens are oriented such as to allow a comparison with an expected direction. We therefore apply PICRIT03 without the $\alpha' \leq 15°$ criterion. Not all of the data include partial TRM (pTRM) or pTRM tail checks; however, in these cases only four specimens do not have pTRM checks and 65 (~16%) do not have tail checks. In these minority cases, absence of a check is regarded as failing to pass the check criteria, but the excluded data contribute to the assessment of the effectiveness of selection.

Recently, *Paterson et al.* [2012] developed a stochastic paleointensity model of ideal SD specimens that experience expected levels of experimental noise. They used this model to investigate the behavior of the

Table 2. The Sets of Selection Criteria Investigated[a]

Criterion	PICRIT03	PICRIT03 (Modified)	SELCRIT2	SELCRIT2 (Modified)	TTA	TTA (Modified)	TTB	TTB (Modified)
n	≥4	≥4	≥4	≥4	≥5	≥5	≥5	≥5
f	≥0.35	≥0.35	≥0.15	≥**0.35**	≥0.5	≥**0.35**	≥0.3	≥**0.35**
β	≤0.1	≤0.1	≤0.1	≤0.1	≤0.1	≤0.1	≤0.15	≤0.15
q	≥2	≥2	≥1	≥1	≥5	≥5	≥0	≥0
MAD_{Anc}	≤7	≤7	≤15	≤15	≤6	≤6	≤15	≤15
α	–	–	≤15	≤15	≤15	≤15	≤15	≤15
n_{pTRM}	≥3	≥3	–	–	–	–	–	–
$DRAT$	≤7	≤**10**	≤10	≤10	–	–	–	–
$CDRAT$	≤10	≤**11**	–	–	–	–	–	–
$DRAT_{Tail}$	–	–	≤10	≤10	–	–	–	–
δCK	–	–	–	–	≤5	≤**7**	≤7	≤**9**
δpal	–	–	–	–	≤5	≤**10**	≤10	≤**18**
δTR	–	–	–	–	≤10	≤10	≤20	≤20
$\delta t*$	–	–	–	–	≤3	≤**9**	≤99	≤99

[a] Definitions of the various statistics are given in SPD v1.0 (supporting information). Thresholds in bold are modified from their original values.

paleointensity results and various selection statistics. Paterson et al. took the 95th percentiles of the selection statistic distributions to define limits of how ideal SD specimens behave. The 95th percentiles represent the upper limit of values that can be produced by ideal SD specimens in the presence of experimental noise. For example, when the laboratory and ancient fields are of equal strength (i.e., $B_{Lab} = B_{Anc}$) and an NRM fraction of $f \geq 0.15$ is used, 95% of ideal SD specimens will have *DRAT* values of ≤ 16.6 for both the Coe and IZZI protocols. On the basis of minimizing the influence of noise and maximizing sensitivity to nonideal factors Paterson et al. suggested a minimum acceptable NRM fraction of $f \geq 0.35$.

In comparison with typical selection criteria, Paterson et al. noted that the behavior of ideal SD specimens frequently exceeds these arbitrarily defined thresholds: Some common selection criteria are too strict. Based on this theoretically predicted SD behavior we suggest modifying commonly used criteria sets to prevent the overly strict rejection of ideal SD specimens that are subject to ever present experimental noise. The modified criteria sets are given in Table 2. Although our data set has a range of B_{Lab}/B_{Anc} ratios (from $\sim$0.18 to $\sim$1.33), for simplicity, we modify the criteria according to the $B_{Lab} = B_{Anc}$ results of *Paterson et al*. [2012], which also follow SPD v1.0. For all criteria sets, the minimum fraction is set to be $\geq$0.35. For the PICRIT03, SELCRIT2, and TTA criteria, the thresholds values are modified to the 95th percentiles suggested by *Paterson et al*. [2012]. The TTB criteria, however, are designed to be more relaxed than the TTA criteria. We have therefore relaxed the threshold values for the modified TTB criteria to the 99th percentiles.

Paterson [2013] extended the stochastic model to simulate the effects of anisotropic and nonlinear TRM on paleointensity data. He demonstrated that paleointensity selection statistics are unaffected by these nonideal factors and that, after correction, the results were nearly identical to those from ideal SD specimens. Therefore, the modifications that we propose here are also valid for anisotropic and nonlinear TRM corrected data.

Several studies have demonstrated that the Coe and IZZI protocols can differ in how the data respond to some nonideal factors (e.g., MD behavior) [*Yu et al*., 2004; *Biggin*, 2006]. Both *Paterson et al*. [2012] and *Paterson* [2013], however, found that for ideal SD behavior the results for the Coe and IZZI protocols were identical. Because of this, we can combine the Coe and IZZI data outlined in section 3.1 to assess the effectiveness of the suggested modifications, which are based on theoretical SD predictions.

The above described selection criteria are widely used in paleointensity studies, but it is also common to specify additional, but unquantified selection criteria. One such example is Arai plot curvature, which is often visually assessed [e.g., *Spassov et al*., 2010; *Calvo-Rathert et al*., 2011; *Neukirch et al*., 2012]. Recently, *Paterson* [2011] proposed a statistic to quantify Arai plot curvature ($|\vec{k}|$) as produced by MD grains. $|\vec{k}|$ is defined as the inverse of the radius of the best fit circle to the Arai plot data. Based on an analysis of paleointensity results from laboratory experiments on 38 specimens with known grain sizes, *Paterson* [2011] proposed a strict selection threshold of $|\vec{k}| \leq 0.164$ and a more relaxed threshold of $|\vec{k}| \leq 0.270$. We therefore test a further modification of the criteria sets with the addition of the relaxed curvature criterion to demonstrate the usefulness of assessing curvature in a quantified fashion as well as the potential improvement in the effectiveness of selection criteria when considerations beyond SD effects are taken into account. We note that although the Coe and IZZI protocols can behave differently for MD dominated specimens, *Shaar et al*. [2011] demonstrated that MD Arai plots from the IZZI protocol can exhibit similar curvature to that seen from Coe protocol data.

3.3. Measuring Effectiveness

To undertake a detailed analysis of the efficacy of paleointensity selection criteria, it is necessary to define what the criteria must be effective in achieving. Ultimately, the goal of all paleointensity studies is to accurately estimate the strength of the paleomagnetic field. Therefore, irrespective of specific causes (e.g., NRM that is not a primary TRM, magnetomineralogical alteration, pseudosingle domain (PSD) or MD grains, nonlinear TRM acquisition, or anisotropic TRM), any results that are inaccurate can be classed as "nonideal" and the purpose of data selection is to bias against inaccurate results. To test the effectiveness of data selection with this approach requires a definition of accuracy. The definition that we propose is as follows. For a known field strength (B_{Exp}), a paleointensity estimate (B_{Anc}) is classified as accurate if $\frac{1}{1.1} \leq \frac{B_{Anc}}{B_{Exp}} \leq 1.1$ (i.e., the estimate is within a factor of 1.1, $\sim$10%, of the expected value). *Paterson et al*. [2012] demonstrated that, in the presence of expected levels of experimental noise, ideal SD specimens can results up to a factor of $\sim$1.06–1.07 from the expected value ($\sim$6–7%). The factor of 1.1 that we adopt is more relaxed than this lower limit and although chosen arbitrarily, a ~10% limit is widely used in paleointensity studies [e.g., *Chauvin et al*., 2005; *Bowles et al*., 2006; *Biggin et al*., 2007; *Yamamoto and Hoshi*, 2008; *Herrero-Bervera and Valet*, 2009; *Paterson et al*., 2010b; *Valet et al*., 2010; *Shaar et al*., 2011]. Therefore, we take this definition to reflect the level of accuracy that is desired by the paleointensity community.

Throughout this paper, we describe paleointensity accuracy as the deviation from the expected value. The deviation is the logarithm of the estimate normalized by the expected value, $\ln\left(\frac{B_{Anc}}{B_{Exp}}\right)$. Zero deviation is exactly the expected value; positive and negative values are over and underestimates, respectively, and are symmetric about zero. Data scatter is quantified as the standard deviation as a percentage of the mean result $\left(\delta B(\%)=\frac{s}{m}\times 100\right)$, but modified to account for differences in the number of results accepted ($\delta B_{\mathrm{N}}(\%)$) [*Paterson et al*., 2010a; *Paterson*, 2011]:

$$\delta B_N(\%)=\left|\frac{\sqrt{N}}{t_{nc}\left(1-\alpha;(N-1);\frac{m\sqrt{N}}{s}\right)}\right|\times 100,$$

where N is the number of accepted results, m and s are the estimated mean and standard deviation, respectively, and t_{nc} is the noncentral t critical value for the $(1-\alpha)$ confidence level for $(N-1)$ degrees of freedom and with noncentrality parameter $\frac{m\sqrt{N}}{s}$. This modification calculates the upper 95% confidence interval on the estimate of scatter, which allows us to say at the 95% confidence level the true scatter of the data is $\leq\delta B_N(\%)$. As N becomes larger, the difference between $\delta B(\%)$ and $\delta B_N(\%)$ decreases.

It can be noted that the question of whether a result is accurate has a binary answer (i.e., yes or no). Therefore, the likelihood of obtaining N_s successful (accurate) results from N_t trials (accepted results) is the result of a Bernoulli trial process and the proportion of accurate results accepted follows a binomial distribution. For a population distribution where the probability of randomly selecting an accurate result is P, the probability (p_r) of obtaining N_s or more accurate results by randomly selecting N_t results can be determined from the binomial cumulative distribution function, $F(N_s, N_t, P)$:

$$p_r=1-F(N_s,N_t,P)=1-\sum_{i=0}^{N_s}\binom{N_t}{i}P^i(1-P)^{N_t-i}, \quad (1)$$

where i is an integer count from 0 to N_s, $\binom{N_t}{i}$ is the binomial coefficient, and $(1-P)$ is the probability of failure (i.e., the probability of randomly selecting an inaccurate result). In the context of paleointensity selection, for a given population distribution (i.e., a data set prior to selection) with P concentration of accurate results, after applying a set of selection criteria and obtaining N_s accurate results from N_t accepted results we can calculate the probability (p_r) of randomly obtaining N_s accurate results. We can use p_r to test the null hypothesis that a concentration of accurate results greater than or equal to N_s/N_t can be obtained by a process of random selection. That is to say, p_r represents the probability that our realized paleointensity success rate occurred by chance. If $p_r\leq 0.05$, we can reject the null hypothesis at the 5% significance level in favor of the alternative hypothesis that there is a biasing factor increasing the likelihood of selecting accurate results. In the context of paleointensity selection, this biasing factor is the criteria used to select the data. If, however, $p_r>0.05$, we cannot distinguish the effects of data selection from a random selection process at the 5% significance level and the selection criteria are ineffective at isolating accurate results.

In the situation where two sets of criteria can significantly increase the likelihood of obtaining accurate results (i.e., $p_r\leq 0.05$), an additional quantification is useful to assess their effectiveness. We score each set of criteria based on their efficiency at accepting accurate results (E_A) and their efficiency at rejecting inaccurate results (E_I):

$$E_A=\frac{\text{Number of accurate results accepted}}{\text{Total number of accurate results}}, \quad (2)$$

and

Table 3. Descriptive Statistic of the Reanalyzed Published Fits Before and After the Application of Commonly Used Selection Criteria Sets[a]

			PICRIT03		SELCRIT2		TTA		TTB
	Unselected	PICRIT03	(Modified)	SELCRIT2	(Modified)	TTA	(Modified)	TTB	(Modified)
Mean	1.02	1.03	1.03	1.06	1.04	1.03	1.04	1.03	1.04
N accepted	383	221	247	293	282	93	205	229	271
δB (%)	29.4	21.8	21.6	24.5	22.6	16.8	21.3	22.4	22.7
δB_N (%)	30.7	22.8	22.6	24.9	23.2	18.2	22.3	23.5	23.5
% Accurate	41.8	51.6	50.6	47.1	47.9	45.2	47.3	44.5	47.2
N_{Accurate}	160	114	125	138	135	42	97	102	128
$N_{\text{Inaccurate}}$	223	107	122	155	147	51	108	127	143
p_r	–	0.001	0.002	0.029	0.017	**0.221**	0.047	**0.180**	0.030
E_A	1	0.713	0.781	0.863	0.844	0.263	0.606	0.638	0.800
E_I	0	0.520	0.453	0.305	0.341	0.771	0.516	0.430	0.359
S	0	0.371	0.354	0.263	0.288	0.202	0.313	0.274	0.287

[a] % Accurate refers to the proportion of accurate in the selected data set as a percentage of the number of results within the selected data set. p_r values in bold (>0.05) indicate that the criteria sets do not significantly increase the likelihood of accepting accurate results.

$$E_I = \frac{\text{Number of inaccurate results rejected}}{\text{Total number of inaccurate results}}, \tag{3}$$

respectively. E_A and E_I are identical to the statistical concepts of sensitivity and specificity, respectively. The score (S) of a set of criteria is simply $E_A \times E_I$ and lies in the interval [0, 1]. When all accurate results are accepted and all inaccurate results are rejected $S = 1$. If, however, no inaccurate results are rejected, or no accurate results are accepted, $S = 0$.

4. Results

4.1. Published Results

We reanalyze the paleointensity estimates that were published in the original studies. This analysis includes results that both passed and failed any subsequent data selection and should not be biased by any selection criteria. Given that the original authors did not publish best-fits for the data from *Muxworthy* [1998] and *Krása et al*. [2003], and for one specimen from *Yamamoto and Hoshi* [2008], these data are excluded from this reanalysis. The descriptive statistics of the data set before and after the application of the four criteria sets are given in Table 3. Before selection criteria are applied, the mean result is accurate (within a factor 1.1 of the expected value), the scatter (δB_N(%)) is ~29% of the mean and ~42% of all results are accurate. After applying the four unmodified criteria sets, all mean results are accurate and the scatters are reduced by ~6–12% with respect to the unselected data. Although all criteria increase the proportion of accurate results accepted, with the exception of PICRIT03, <50% of the accepted results are accurate. The unmodified PICRIT03 and SELCRIT2 are effective at significantly increasing the acceptance of accurate results ($p_r < 0.05$; Table 3). For both ThellierTool sets, however, $p_r \geq 0.180$ and we cannot reject the null hypothesis that the proportions of accurate results accepted occurred by chance.

The modifications to PICRIT03 and SELCRIT2 make only a small difference to the final results (Table 3). The modified TTA criteria, however, represent a large improvement over the unmodified version. Despite a decrease in the efficiency of rejecting inaccurate results (E_I decreases), the overall score of the TTA criteria increases from 0.202 to 0.313 (due to the efficiency of accepting accurate results more than doubling). The modified criteria increase the proportion of accurate results accepted (~2%) and are more effective than a random selection process ($p_r < 0.05$). For TTB, the modifications increase the proportion of accurate results accepted by ~2.5% over the original criteria and p_r decreases from 0.180 to 0.030, which indicates that the modified criteria are a significant improvement over random selection and the original criteria set. The overall score of the TTB criteria increases from 0.274 to 0.287 due to the improved acceptance of accurate results.

4.2. A Bootstrap Approach

The above results rely on reanalyzing previously published fits, which may suffer from user bias and may not represent the general behavior of a specimen and hence the true effectiveness of the selection criteria. To overcome this we adopt a bootstrap approach. For each specimen, a best fit segment is randomly fitted to the Arai plot data. To ensure a physically reasonably fit, the segment must contain at least three consecutive Arai plot points that define a segment with a negative slope, it must have a gap factor (g) > 0, and $f \geq 0.05$. The random fitting process for each specimen is repeated until a best fit segment passing these conditions is found. This fitting procedure is repeated for all specimens to obtain a pseudo-data set of 395 paleointensity estimates and statistics. The various selection criteria sets are then applied to this pseudo-data set and the descriptive statistics recorded. This whole process is repeated for 10^4 bootstraps to build up distributions for the descriptive statistics of the pseudo-data sets before and after selection. All 395 specimens in Table 1 are included in this analysis.

The probability densities for the descriptive statistics before and after selection are shown in Figure 1. Before selection the mean values from all the pseudo-data sets are accurate (|mean deviation| $\leq$ln(1.1); Figure 1a). The percentage of accurate results, however, is generally low (median ~32%) and the scatter high (median ~44%). In all cases, after selection there is a shift to higher proportions of accurate results in the accepted pseudo-data sets and a reduction in scatter. For PICRIT03 and SELCRIT2, the original and modified criteria behave similarly. The score for the modified PICRIT03 criteria is slightly higher than for the original (Figure 1g), but tends to be slightly lower for the modified SELCRIT2 when compared with the original criteria (Figure 1k).

Although more peaked around the median values, the mean deviations, the percentage of accurate results and the scatter of the modified TTA criteria behave similarly to the original (Figure 1l–1n). There is, however, a large increase in the scores of the selected data sets when the modified criteria are compared with the original criteria (median scores of 0.260 and 0.107, respectively; Figure 1o). As is the case for the analysis of the published fits, this large increase is due to a large increase in E_A (the E_A and E_I distributions are discussed in section 5). For the TTB criteria (Figure 1p–1s), both the original and modified criteria consistently yield accurate mean results and have similar proportions of accurate results (Figure 1q). The scatter remains unaffected by the modifications, but there is an increase in the scores of the modified TTB criteria with respect to the original set (Figure 1s).

The results of adding a quantitative curvature criterion to the modified criteria are represented by the gray dotted lines in Figure 1. In all cases, the distribution of the deviation of the mean values from the expected values becomes more peaked close to zero. For the proportion of accurate results in the accepted data sets, the addition of $|\vec{k}|$ increases this percentage above all of the original criteria and improves on the modified criteria. The most notable difference after including $|\vec{k}|$ is the reduction of the scatter: The median values are reduced by ~2–5% with respect to both the original and modified criteria and is most pronounced for the SELCRIT2 and TTB criteria. Curved Arai plots can produce a wide range of values that both over and underestimate the expected paleointensity, whereas near-linear Arai plots yield only a narrow range of values. It is the exclusion of this wide range of possible values from curved Arai plots that results in lower scatters.

The percentage of bootstrap results that yield p_r value >0.05 (i.e., bootstrap results where the selection process cannot be distinguished from random selection) are shown in Table 4. For the original and modified PICRIT03 and SELCRIT2 criteria, these percentages are low ($\leq$0.49%), which indicates that these criteria are effective at improving the acceptance of accurate results. The TTB criteria are moderately effective in their original form, with ~5.6% of the bootstraps yielding $p_r > 0.05$. This is further reduced ($\leq$0.95%) by the modifications suggested here. The TTA criteria are much less effective, with one in four bootstraps (26.02%) yielding results that are no better than random chance. The modified criteria greatly reduce this to 1.93% of bootstraps and the inclusion of a curvature criterion reduces this further (1.37%).

5. Discussion

The definition of accuracy used here (within a factor of 1.1) may be viewed as strict and a more relaxed value, say, a factor 1.2 (~20%), may be more appropriate. For comparison with the median values shown in Figure 1, the median values of the descriptive statistics from the bootstrapped results using a factor of 1.2 are given in Table 5. As would be expected, the definition of accuracy has no influence on the mean deviation or the scatter of the results, but the proportion of accurate results increases by 23–33%. Given that in the accurate/inaccurate binary system, regardless of definition, the number of accurate results is correlated with the number of inaccurate results, the relaxed definition has little influence on the median values of the

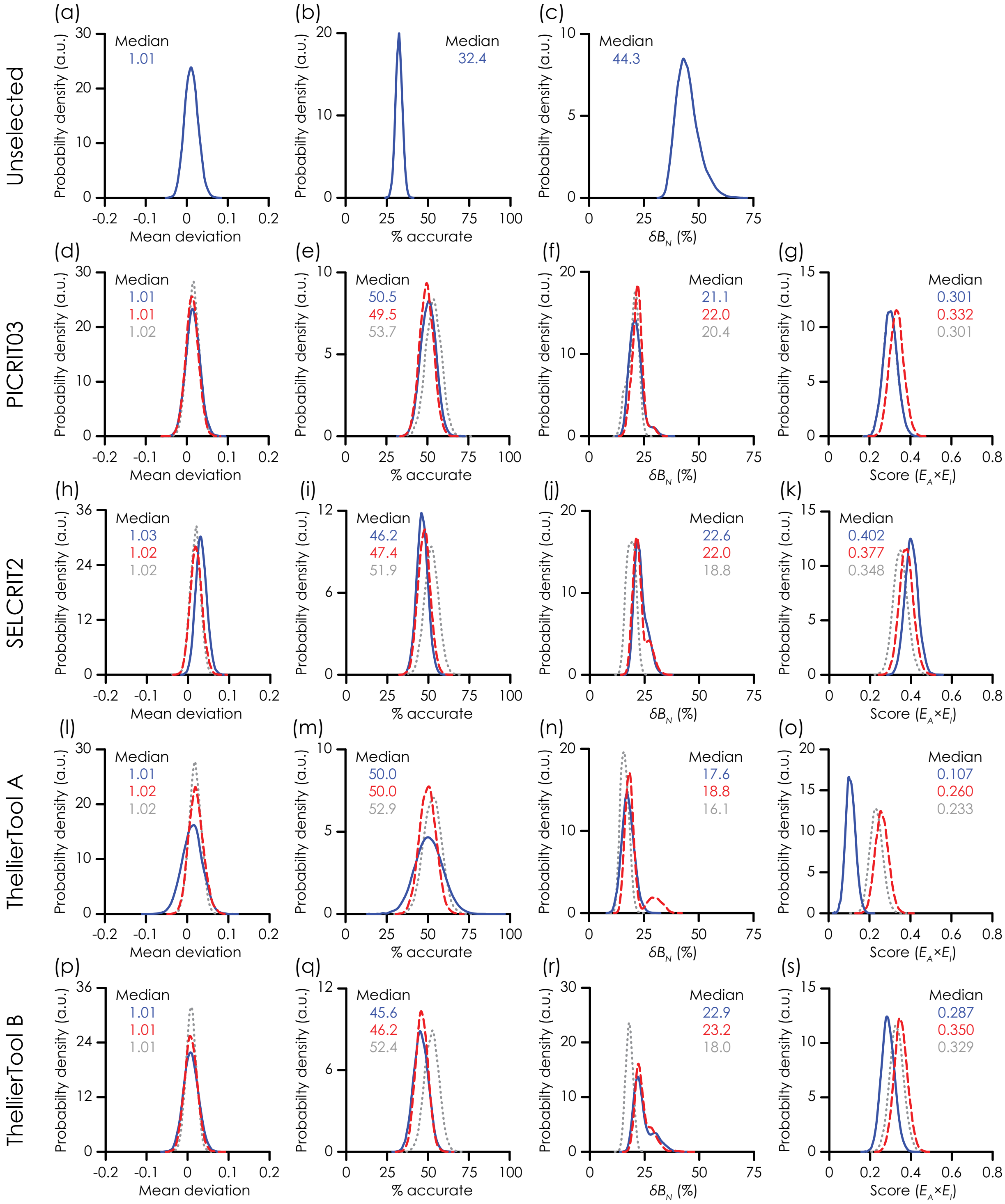

Figure 1. Probability density functions of the bootstrapped descriptive statistics from (a–c) the unselected data, selected following (d–g) PICRIT03, (g–k) SELCRIT2, (l–o) TTA, and (p–s) TTB. Densities are in arbitrary units. In all plots, the blue solid curves represent the unselected data or the data selected using the original criteria. The red dash curves represent the modified criteria and the gray dotted curves represent the modified criteria with the inclusion of the Arai plot curvature criterion.

score. The main difference when using a relaxed definition of accuracy is in the percentage of bootstraps where the criteria are statistically indistinguishable from randomly selecting data. For PICRIT03, SELCRIT2, and TTB, before and after modification the percentages remain low using a factor of 1.2 (Table 5). The

Table 4. The Percentage of Bootstrapped Results With $p_r > 0.05$

	Original	Modified	Modified Incl. $\lvert\vec{k}\rvert$
PICRT03	0.49	0.28	0.10
SELCRIT2	0.06	0.22	0.07
TTA	26.02	1.93	1.37
TTB	5.64	0.95	0.02

original TTA criteria are not effective at isolating estimates that are within a factor of 1.1 of the expected result (Table 4), but they are effective at isolating results within a factor of 1.2 and only 1.56% of bootstraps are no better than random chance. Despite this improvement for the original TTA criteria, we still prefer to use the factor 1.1 in our definition of accuracy. First, this definition is used in at least five studies that do not involve the authors of this current work [*Chauvin et al.*, 2005; *Bowles et al.*, 2006; *Yamamoto and Hoshi*, 2008; *Herrero-Bervera and Valet*, 2009; *Valet et al.*, 2010] and reflects the general view of the paleointensity community. Second, we have demonstrated that, for the data set compiled here, various sets of selection criteria can be effective in isolating results that are within a factor 1.1 of the expected values. In cases where criteria are ineffective (e.g., TTA), theoretically justified modifications can make these criteria effective at isolating accurate results.

Throughout paleointensity literature, sets of selection criteria are variably described as strict or relaxed. The strictness and effectiveness of a set of selection criteria are a balance between E_A and E_I. Relaxed criteria tend to accept larger numbers of data and produce high E_A values, but yield low E_I values. Conversely, strict criteria reject large numbers of data and tend to sacrifice E_A to maximize E_I. By quantifying E_A and E_I, we define relaxed selection criteria, where $E_A > E_I$; strict selection criteria, where $E_A < E_I$; and ideal criteria, where $E_A = E_I = 1$. The distributions of E_A and E_I from the bootstrapped results are shown in Figure 2. In general, the sets of criteria tested here would be classified as strict (i.e., $E_A < E_I$). This is most extreme for the original TTA criteria, where the median E_I is 0.947 (i.e., 94.7% of all inaccurate results are rejected), but the median E_A is only 0.113 (11.3% of accurate results are accepted; Figures 2e and 2f). Although the TTA criteria are effective at rejecting inaccurate results they are ineffective are accepting accurate results, which is why in many cases they are no better than making a random selection of data. In this sense, the original TTA criteria are too strict. The modifications suggested here reduce the effectiveness of rejecting inaccurate results by a small amount, but allow for a relatively large increase in the acceptance of accurate results (Figure 2e). The other criteria sets are more balanced, in that they achieve high E_I values while maintaining moderate E_A values. For PICRIT03 and TTB, the modified criteria sets maintain or increase both E_A and E_I (Figures 2a, 2b, 2g, and 2h). For SELCRIT2, the improved rejection of inaccurate results is offset by a reduction in the acceptance of accurate results (lower E_A), which results in the small decrease in the overall score (Figure 1k). Despite this slight decrease, the theoretical basis for these modifications is a stronger justification for using the modified criteria.

While the modified criteria can improve the proportions of accurate results and increase the general effectiveness of accepting accurate results and rejecting inaccurate results (given by S), the scatters remain high (typically >20%). The addition of an experimentally constrained and quantitative measure of Arai plot curvature, however, can improve this. The addition of $\lvert\vec{k}\rvert$ reduces the overall scatter of the results through an improved rejection of inaccurate results (Figure 2). This is offset by a reduction in E_A, but results in an overall improvement of the results when compared to the original criteria (Figure 1). We note that, although $\lvert\vec{k}\rvert$ is one of only a few selection statistics that has been quantitatively constrained from independent control experiments, further work is needed to achieve the same level of theoretical understanding as has been achieved for the modification based on SD predictions.

Of the original criteria sets investigated, the TTA criteria are the most ineffective at successfully isolating accurate results. This criteria set is widely used in the modern literature and, although the mean values of

Table 5. Median Values of the Descriptive Statistics From the Bootstrapped Data Set Using a Factor of 1.2 to Define Accurate Results

			PICRIT03		SELCRIT2		TTA		TTB
	Unselected	PICRIT03	(Modified)	SELCRIT2	(Modified)	TTA	(Modified)	TTB	(Modified)
Mean	1.01	1.01	1.01	1.03	1.02	1.01	1.02	1.01	1.01
δB_N (%)	44.3	21.1	22.0	22.6	22.0	17.6	18.8	22.9	23.2
% Accurate	54.9	78.1	76.9	74.4	75.3	83.3	79.7	73.5	73.3
S		0.294	0.329	0.438	0.394	0.107	0.261	0.292	0.358
% with $p_r > 0.05$		0	0	0	0	1.56	0	0.11	0.02

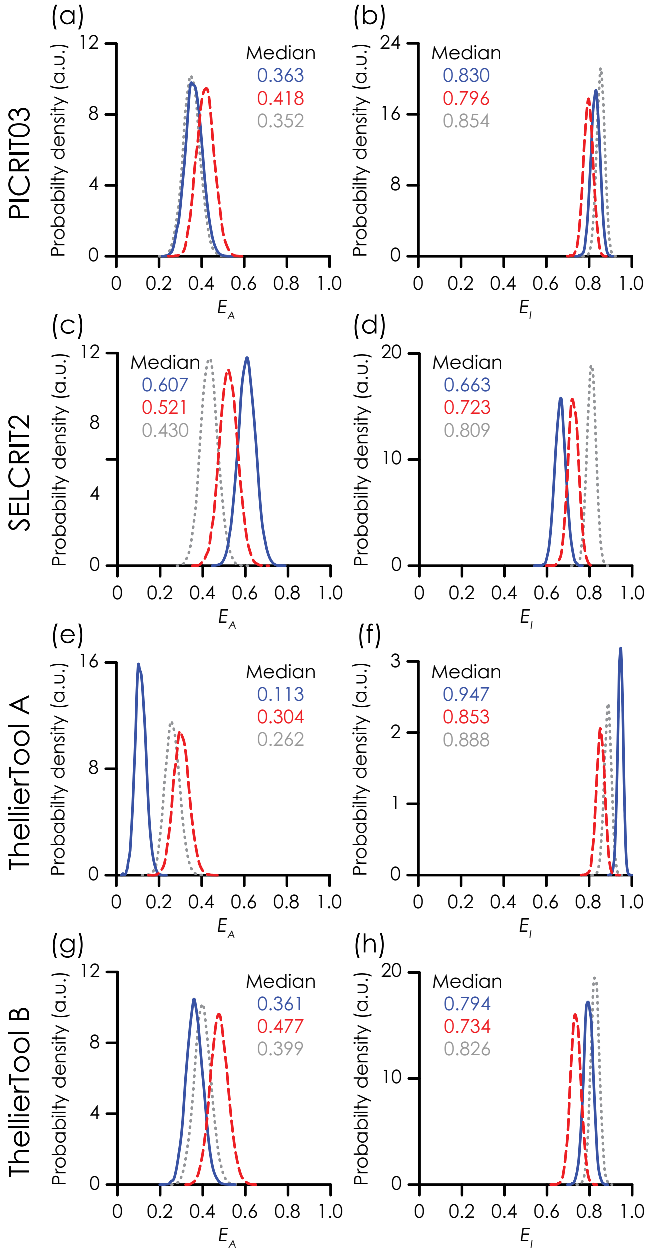

Figure 2. Probability density functions of E_A and E_I values from the bootstrapped data after selection following (a–b) PICRIT03, (c–d) SELCRIT2, (e–f) TTA, and (g–h) TTB. Densities are in arbitrary units. The line colors are the same as in Figure 1.

the bootstrapped data set are accurate (Figure 1l), they have a worryingly high rate of being no more effective than randomly selecting data (Table 4). We recommend that alternative, but demonstrably effective selection criteria, such as the modified TTA criteria be used instead of the original TTA criteria. Our suggested modifications make changes to threshold values for four statistics (f, δCK, δpal, and δt^*), which results in the median score increasing from 0.107 to 0.260 (Figure 1o). To identify which statistic accounts for the poor performance of the original TTA criteria, we repeat the bootstrap process described in section 4.2, but modifying one statistic at a time. When we modify only f, the median score is 0.132, for δCK the median score is 0.126, modifying only δpal yields a median score of 0.148, and δt^* produces a median score of 0.124. For the four individual modifications (f, δCK, δpal, and δt^*), the percentage of bootstraps where $p_r > 0.05$, are 20.40%, 14.70%, 18.96%, and 24.19%, respectively (cf. 26.02% for the original and 1.93% for the fully modified criteria). The modifications to δpal and δCK produce the largest improvements to the median score and the percentage of bootstraps with $p_r > 0.05$. No single modification, however, produces the large improvement seen from the fully modified criteria. The modifications are most effective when combined as a set, which indicates that multiple factors are influencing the data set compiled here. It also suggests that the interplay of different selection statistics is an important consideration when developing a better understanding of how the data selection process works.

The poor performance of the original TTA criteria may not be isolated: Many studies use unique sets of selection criteria, the effectiveness of which remains unknown. Studies that use unique sets of selection criteria should define selection thresholds that fit with our latest quantitative understanding of how the selection statistics behave, which may be derived from theoretical or empirical constraints. In addition, we recommend that the effectiveness of selection criteria sets be assessed using a data set that is independent of how the criteria were defined and that provides a means to quantify the accuracy of the paleointensity results.

In addition to the Coe and IZZI protocols, the original Thellier protocol [*Thellier and Thellier*, 1959] and the Aitken protocol [*Aitken et al*., 1988] are in current use. *Paterson et al*. [2012] and *Paterson* [2013] both demonstrated that when $B_{Lab} = B_{Anc}$ the Aitken protocol behaves similarly to the Coe and IZZI protocols, but when $B_{Lab} > 2 \times B_{Anc}$ some statistics deviate slightly from the Coe and IZZI trends. Nevertheless, the modifications suggested here should be valid for the Aitken protocol, although without any data currently available, we cannot test this. Similarly, modifications could be suggested for the criteria sets when applied to Thellier protocol data, however, in the absence of control data with which to assess the criteria sets we cannot test the validity of possible modifications.

6. Conclusions

Our ability to make reliable inferences about the ancient geomagnetic field strength relies heavily on consistently and reliably screening paleointensity data for nonideal behavior, but until now the efficacy of the data selection process has been unquantified. We have outlined the largest-to-date analysis of selection statistics and presented a series of tools that allow the quantitative assessment of paleointensity data selection. These tools have been used to assess the effectiveness of widely used criteria sets and the effectiveness of modifications based on theoretical SD paleointensity behavior.

We have proposed a new standard document to define paleointensity data and help to remove ambiguities and inconsistencies that exist in data quantification. The Standard Paleointensity Definitions is intended to be a useful reference document for the paleomagnetic community. If readers have comments, suggestions, corrections, or criticisms, we warmly invite them to contact G.A.P. as all input that can help to further improve our ability to consistently select reliable paleointensity data is appreciated.

We have demonstrated that, when theoretical SD paleointensity behavior is considered, paleointensity data selection can be improved. This improvement comes in two forms. First, the modifications we propose can be used to improve the results of the data selection process and increase the likelihood of obtaining accurate and low scatter paleointensity estimates. Second, and importantly, these modifications are based on theoretical predictions that are derived independently of the large data set that we have compiled here. By taking this approach to modifying sets of paleointensity selection criteria we are beginning to place the data selection process on a more defensible theoretical foundation, which is essential for ensuring that we can reliably isolate accurate paleointensity data.

Despite this improvement, considerable work remains before we fully understand the quantitative manifestation of nonideal factors in selection statistics and whether or not they provide an effective means to screen these undesirable effects from our data sets. Although the necessary systematic studies can be difficult and time consuming, the improvements that we see when applying a quantitative measure of Arai plot curvature demonstrate that they are a valuable undertaking.

We recommend that all sets of selection criteria used in future paleointensity studies follow some basic requirements to ensure the validity of the obtained results.

1. The threshold values used for selection must conform to the most recent quantitative understanding of selection statistic behavior, be this from theoretical considerations or from systematic control experiments that provide empirical constraints.

2. The efficacy of selection criteria must be demonstrated using an independent data set, such as the one compiled here or any other relevant independent data set that allows the accuracy of results to be quantified.

3. The criteria should not exhibit a high proportion of instances where the average results are inaccurate with respect to an expected value or where the effectiveness of selection is no better than randomly selecting data. We suggest that, in a bootstrap style test (cf., section 4.2), fewer than 5% of all bootstraps should yield results with an inaccurate mean and/or $p_r \leq 0.05$.

For future studies, we strongly recommend the use of the modified criteria sets outlined here. These criteria sets fit the above requirements and remove a large degree of arbitrariness from the selection of paleointensity data.

Our understanding of the paleointensity selection process is rapidly evolving and the latest information can raise questions over the reliability of older data sets. We strongly encourage the use of online data repositories, such as the MagIC database, which allow the raw paleointensity data to be stored and reanalyzed as paleointensity data analysis develops, the usefulness of which is demonstrated here. Without such data availability, vast amounts of legacy data will be rendered unreliable, which will only hinder our ability to understand geomagnetic field evolution.

Acknowledgments
Julie Bowles, Fabio Donadini, Roman Leonhardt, Adrian Muxworthy, Peter Selkin, Hidefumi Tanaka, and Yuhji Yamamoto are thanked for providing data. The authors are grateful for the efforts of the MagIC team (earthref.org/MAGIC/). G.A.P. acknowledges funding from the Key Laboratory of Earth's Deep Interior, IGGCAS and L.T. is supported by the National Science Foundation grant EAR1141840 to L.T. The authors thank the Associate Editor (Suzanne McEnroe) and two anonymous reviewers for their comments and suggestions.

References

Aitken, M. J., A. L. Allsop, G. D. Bussell, and M. B. Winter (1988), Determination of the intensity of the Earth's magnetic field during archaeological times: Reliability of the Thellier technique, *Rev. Geophys.*, *26*, 3–12, doi:10.1029/RG026i001p00003.

Biggin, A. J. (2006), First-order symmetry of weak-field partial thermoremanence in multi-domain (MD) ferromagnetic grains: 2. Implications for Thellier-type palaeointensity determination, *Earth Planet. Sci. Lett.*, *245*, 454–470, doi:10.1016/j.epsl.2006.02.034.

Biggin, A. J., M. Perrin, and M. J. Dekkers (2007), A reliable absolute palaeointensity determination obtained from a non-ideal recorder, *Earth Planet. Sci. Lett.*, *257*, 545–563, doi:10.1016/j.epsl.2007.03.017.

Biggin, A. J., S. Badejo, E. Hodgson, A. R. Muxworthy, J. Shaw, and M. J. Dekkers (2013), The effect of cooling rate on the intensity of thermoremanent magnetization (TRM) acquired by assemblages of pseudo-single domain, multidomain and interacting single-domain grains, *Geophys. J. Int.*, *193*, 1239–1249, doi:10.1093/gji/ggt078.

Bowles, J., J. S. Gee, D. V. Kent, M. R. Perfit, S. A. Soule, and D. J. Fornari (2006), Paleointensity applications to timing and extent of eruptive activity, 9°–10°N East Pacific Rise, *Geochem. Geophys. Geosyst.*, *7*, Q06006, doi:10.1029/2005GC001141.

Calvo-Rathert, M., A. Goguitchaichvili, M.-F. Bógalo, N. Vegas-Tubía, Á. Carrancho, and J. Sologashvili (2011), A paleomagnetic and paleointensity study on Pleistocene and Pliocene basaltic flows from the Djavakheti Highland (Southern Georgia, Caucasus), *Phys. Earth Planet. Inter.*, *187*, 212–224, doi:10.1016/j.pepi.2011.03.008.

Chauvin, A., Y. Garcia, P. Lanos, and F. Laubenheimer (2000), Paleointensity of the geomagnetic field recovered on archaeomagnetic sites from France, *Phys. Earth Planet. Inter.*, *120*, 111–136, doi:10.1016/S0031-9201(00)00148-5.

Chauvin, A., P. Roperch, and S. Levi (2005), Reliability of geomagnetic paleointensity data: The effects of the NRM fraction and concave-up behavior on paleointensity determinations by the Thellier method, *Phys. Earth Planet. Inter.*, *150*, 265–286, doi:10.1016/j.pepi.2004.11.008.

Coe, R. S. (1967), Paleo-intensities of the Earth's magnetic field determined from Tertiary and Quaternary rocks, *J. Geophys. Res.*, *72*, 3247–3262, doi:10.1029/JZ072i012p03247.

Coe, R. S., S. Grommé, and E. A. Mankinen (1978), Geomagnetic paleointensities from radiocarbon-dated lava flows on Hawaii and the question of the Pacific nondipole low, *J. Geophys. Res.*, *83*, 1740–1756, doi:10.1029/JB083iB04p01740.

Donadini, F., M. Kovacheva, M. Kostadinova, L. Casas, and L. J. Pesonen (2007), New archaeointensity results from Scandinavia and Bulgaria: Rock-magnetic studies inference and geophysical application, *Phys. Earth Planet. Inter.*, *165*, 229–247, doi:10.1016/j.pepi.2007.10.002.

Herrero-Bervera, E., and J.-P. Valet (2009), Testing determinations of absolute paleointensity from the 1955 and 1960 Hawaiian flows, *Earth Planet. Sci. Lett.*, *287*, 420–433, doi:10.1016/j.epsl.2009.08.035.

Kissel, C., and C. Laj (2004), Improvements in procedure and paleointensity selection criteria (PICRIT-03) for Thellier and Thellier determinations: Application to Hawaiian basaltic long cores, *Phys. Earth Planet. Inter.*, *147*, 155–169, doi:10.1016/j.pepi.2004.06.010.

Krása, D., C. Heunemann, R. Leonhardt, and N. Petersen (2003), Experimental procedure to detect multidomain remanence during Thellier-Thellier experiments, *Phys. Chem. Earth*, *28*, 681–687, doi:10.1016/S1474-7065(03)00122-0.

Leonhardt, R., C. Heunemann, and D. Krása (2004), Analyzing absolute paleointensity determinations: Acceptance criteria and the software ThellierTool4.0, *Geochem. Geophys. Geosyst.*, *5*, Q12016, doi:10.1029/2004GC000807.

Muxworthy, A. R. (1998), *Stability of Magnetic Remanence in Multidomain Magnetite*, Univ. of Oxford, Oxford, U. K.

Muxworthy, A. R., D. Heslop, G. A. Paterson, and D. Michalk (2011), A Preisach method for estimating absolute paleofield intensity under the constraint of using only isothermal measurements: 2. Experimental testing, *J. Geophys. Res.*, *116*, B04103, doi:10.1029/2010JB007844.

Neukirch, L. P., J. A. Tarduno, T. N. Huffman, M. K. Watkeys, C. A. Scribner, and R. D. Cottrell (2012), An archeomagnetic analysis of burnt grain bin floors from ca. 1200 to 1250 AD Iron-Age South Africa, *Phys. Earth Planet. Inter.*, *190–191*, 71–79, doi:10.1016/j.pepi.2011.11.004.

Paterson, G. A. (2011), A simple test for the presence of multidomain behaviour during paleointensity experiments, *J. Geophys. Res.*, *116*, B10104, doi:10.1029/2011JB008369.

Paterson, G. A. (2013), The effects of anisotropic and non-linear thermoremanent magnetizations on Thellier-type paleointensity data, *Geophys. J. Int.*, *193*, 694–710, doi:10.1093/gji/ggt033.

Paterson, G. A., D. Heslop, and A. R. Muxworthy (2010a), Deriving confidence in paleointensity estimates, *Geochem. Geophys. Geosyst.*, *11*, Q07Z18, doi:10.1029/2010GC003071.

Paterson, G. A., A. R. Muxworthy, A. P. Roberts, and C. Mac Niocaill (2010b), Assessment of the usefulness of lithic clasts from pyroclastic deposits for paleointensity determination, *J. Geophys. Res.*, *115*, B03104, doi:10.1029/2009JB006475.

Paterson, G. A., A. J. Biggin, Y. Yamamoto, and Y. Pan (2012), Towards the robust selection of Thellier-type paleointensity data: The influence of experimental noise, *Geochem. Geophys. Geosyst.*, *13*, Q05Z43, doi:10.1029/2012GC004046.

Pick, T., and L. Tauxe (1993), Geomagnetic palaeointensities during the Cretaceous normal superchron measured using submarine basaltic glass, *Nature*, *366*, 238–242, doi:10.1038/366238a0.

Selkin, P. A., W. P. Meurer, A. J. Newell, J. S. Gee, and L. Tauxe (2000), The effect of remanence anisotropy on paleointensity estimates: A case study from the Archean Stillwater Complex, *Earth Planet. Sci. Lett.*, *183*, 403–416, doi:10.1016/S0012-821X(00)00292-2.

Selkin, P. A., J. S. Gee, and L. Tauxe (2007), Nonlinear thermoremanence acquisition and implications for paleointensity data, *Earth Planet. Sci. Lett.*, *256*, 81–89, doi:10.1016/j.epsl.2007.01.017.

Shaar, R., H. Ron, L. Tauxe, R. Kessel, A. Agnon, E. Ben-Yosef, and J. M. Feinberg (2010), Testing the accuracy of absolute intensity estimates of the ancient geomagnetic field using copper slag material, *Earth Planet. Sci. Lett.*, *290*, 201–213, doi:10.1016/j.epsl.2009.12.022.

Shaar, R., H. Ron, L. Tauxe, R. Kessel, and A. Agnon (2011), Paleomagnetic field intensity derived from non-SD: Testing the Thellier IZZI technique on MD slag and a new bootstrap procedure, *Earth Planet. Sci. Lett.*, *310*, 213–224, doi:10.1016/j.epsl.2011.08.024.

Spassov, S., J.-P. Valet, D. Kondopoulou, I. Zananiri, L. Casas, and M. Le Goff (2010), Rock magnetic property and paleointensity determination on historical Santorini lava flows, *Geochem. Geophys. Geosyst.*, *11*, Q07006, doi:10.1029/2009GC003006.

Tanaka, H., Y. Hashimoto, and N. Morita (2012), Palaeointensity determinations from historical and Holocene basalt lavas in Iceland, *Geophys. J. Int.*, *189*, 833–845, doi:10.1111/j.1365-246X.2012.05412.x.

Tauxe, L., and H. Staudigel (2004), Strength of the geomagnetic field in the Cretaceous Normal Superchron: New data from submarine basaltic glass of the Troodos Ophiolite, *Geochem. Geophys. Geosyst.*, *5*, Q02H06, doi:10.1029/2003GC000635.

Thellier, E., and O. Thellier (1959), Sur l'intensité du champ magnétique terrestre dans le passé historique et géologique, *Ann. Geophys.*, *15*, 285–376.

Valet, J.-P., E. Herrero-Bervera, J. Carlut, and D. Kondopoulou (2010), A selective procedure for absolute paleointensity in lava flows, *Geophys. Res. Lett.*, *37*, L16308, doi:10.1029/2010GL044100.

Veitch, R. J., I. G. Hedley, and J.-J. Wagner (1984), An investigation of the intensity of the geomagnetic-field during Roman times using magnetically anisotropic bricks and tiles, *Arch. Sci.*, *37*, 359–373.

Yamamoto, Y., and H. Hoshi (2008), Paleomagnetic and rock magnetic studies of the Sakurajima 1914 and 1946 andesitic lavas from Japan: A comparison of the LTD-DHT Shaw and Thellier paleointensity methods, *Phys. Earth Planet. Inter.*, *167*, 118–143, doi:10.1016/j.pepi.2008.03.006.

Yamamoto, Y., H. Tsunakawa, and H. Shibuya (2003), Palaeointensity study of the Hawaiian 1960 lava: Implications for possible causes of erroneously high intensities, *Geophys. J. Int.*, *153*, 263–276, doi:10.1046/j.1365-246X.2003.01909.x.

Yu, Y. J., L. Tauxe, and A. Genevey (2004), Toward an optimal geomagnetic field intensity determination technique, *Geochem. Geophys. Geosyst.*, *5*, Q02H07, doi:10.1029/2003GC000630.