Radiocarbon, Vol 58, Nr 1, 2016, p 115–130 DOI:10.1017/RDC.2015.10

# AMS DATES FROM TWO ARCHAEOLOGICAL SITES OF KOREA: BLIND TESTS

Jangsuk Kim[1*] • David K Wright[1] • Youngseon Lee[2] • Jaeyong Lee[2] • Seonho Choi[3] • Junkyu Kim[1] • Sung-Mo Ahn[4] • Jongtaik Choi[5] • Chuntaek Seong[6] • Chang Ho Hyun[7] • Jaehoon Hwang[1] • Hyemin Yang[6] • Jiwon Yang[1]

[1]Department of Archaeology and Art History, Seoul National University, Seoul 151-742, South Korea.
[2]Department of Statistics, Seoul National University, Seoul 151-742, South Korea.
[3]Department of Physics, Seoul National University, Seoul 151-742, South Korea.
[4]Department of Archaeology and Art History, Wonkwang University, Iksan 570-749, South Korea.
[5]Department of Archaeology and Art History, Korea University, Sejong 339-700, South Korea.
[6]Department of History, Kyung Hee University, Seoul 130-791, South Korea.
[7]Division of Science Education, Daegu University, Kyeongsan 712-714, South Korea.

**ABSTRACT.** In interpreting radiocarbon dating results, it is important that archaeologists distinguish uncertainties derived from random errors and those from systematic errors, because the two must be dealt with in different ways. One of the problems that archaeologists face in practice, however, is that when receiving dating results from laboratories, they are rarely able to critically assess whether differences between multiple $^{14}$C dates of materials are caused by random or systematic errors. In this study, blind tests were carried out to check four possible sources of errors in dating results: repeatability of results generated under identical field and laboratory conditions, differences in results generated from the same sample given to the same laboratory submitted at different times, interlaboratory differences of results generated from the same sample, and differences in the results generated between inner and outer rings of wood. Five charred wood samples, collected from the Namgye settlement and Hongreyonbong fortress, South Korea, were divided into 80 subsamples and submitted to five internationally recognized $^{14}$C laboratories on a blind basis twice within a 2-month interval. The results are generally in good statistical accordance and present acceptable errors at an archaeological scale. However, one laboratory showed a statistically significant variance in ages between batches for all samples and sites. Calculation of the Bayesian partial posterior predictive *p* value and chi-squared tests rejected the null hypothesis that the errors randomly occurred, although the source of the error is not specifically known. Our experiment suggests that it is necessary for users of $^{14}$C dating to establish an organized strategy for dating sites before submitting samples to laboratories in order to avoid possible systematic errors.

**KEYWORDS:** blind test, South Korea, errors, interbatch difference, Bayesian *p* value.

## INTRODUCTION

Despite radiocarbon laboratories' continuous efforts to increase the accuracy and precision of measurements, uncertainty regarding the general reliability of $^{14}$C dates to correctly date past human activity has long been one of the primary concerns of archaeologists (e.g. van der Plicht and Bruins 2001; Pettitt et al. 2003; Mellars 2006; Buck et al. 2007; Faught 2008; Graf 2009). Although archaeologists are well aware that uncertainty of $^{14}$C dates is inevitable and dating results should be understood probabilistically, there remains a strong desire to obtain exact dates for target events.

Sources of uncertainty in $^{14}$C dating can be divided into two components: those derived from random errors and those from systematic errors (Ward and Wilson 1978; Scott et al. 2007). The distinction between the two components is important to archaeologists, because they must be dealt with in different ways. Random errors are to be treated statistically, and by increasing sample size, uncertainty can be decreased. When using an accelerator mass spectrometer (AMS), precision is enhanced when extra time is taken to count the numbers of $^{13}$C and $^{14}$C isotopes. On the other hand, systematic errors should be controlled before results are interpreted. Systematic errors can further be subdivided into archaeological and nonarchaeological systematic errors. The former includes erroneous stratigraphic interpretations during fieldwork, failure to detect later inclusion of materials, contamination of samples during sampling, and the so-called old-wood effect (Schiffer 1986), all of which should be avoided or carefully controlled

*Corresponding author. Email: jangsuk@snu.ac.kr.

https://doi.org/10.1017/RDC.2015.10 Published online by Cambridge University Press

by archaeologists. Nonarchaeological systematic errors are caused by physical and chemical factors affecting concentration of $^{14}$C in dated materials, and many of them such as the Suess effect (Stuiver and Suess 1966), marine reservoir effect (Keith and Anderson 1963; Stuiver et al. 1986), hardwater effect (Shotton 1972), contamination through exposure to volcanic ash (Pichler and Friedrich 1976), and bone apatite diagenesis (Price et al. 1992; Nielsen-Marsh and Hedges 2000a, 2000b) have been reported. From the perspective of archaeologists, non-taphonomic systematic errors are harder to determine and may include contamination of materials during pretreatment or erroneous measurement of calibration standards. These errors may systematically lead to anomalous results even if the reported dates were statistically treated and taphonomic contamination effects were controlled.

One of the problems that archaeologists face in practice is that when receiving dating results from laboratories, they are rarely able to critically assess whether differences between multiple $^{14}$C dates of materials that are expected to be the same age are caused by random or systematic errors, or whether the error is in their expectations of the temporal accumulation of archaeological deposits of the site. In such cases, archaeologists are often at a loss as to whether the results should be statistically treated or controlled in different ways or merely discarded. Although many statistical methods have been developed to deal with random errors (e.g. Ward and Wilson 1978; Christen 1994; Christen and Buck 1998; Buck and Millard 2004; Scott et al. 2007; Bronk Ramsey 2009; Scott 2011), unless archaeologists are able to distinguish systematic errors affecting the amount of $^{14}$C during measurement from random errors, statistical treatments of conflicting $^{14}$C dates (e.g. statistically combining multiple dates) are not meaningful.

To a certain degree, archaeologists' practical problems can be mitigated if they can distinguish purely random errors and possible systematic errors. One way to distinguish the two types of errors is to measure a sample believed to be taphonomically consistent under various conditions to test whether systematic errors occur by comparing the results. If the comparison of results indicates that certain conditions repeatedly and consistently produce different results, they may be viewed as possible causes of systematic errors, which should be considered before undertaking a statistical analysis of the ages.

This paper reports the results of an experiment designed to check possible causes of errors of $^{14}$C dating of charcoal, by dating samples from single archaeological contexts under a variety of conditions. The experiment attempted to check four possible sources of variability:

(1) *Repeatability under identical conditions*: When one object is dated under presumably identical conditions in the same laboratory at the same point of time, how different are the results? What is the range of random errors of different aliquots?

(2) *Interbatch differences in a laboratory*: When multiple subsamples from the same bulk sample are submitted to a laboratory at different points in time, how much does the difference in timing of the analysis affect the outcomes? Do possible differences in measurement background and laboratory settings significantly affect the results?

(3) *Interlaboratory difference*: The International Radiocarbon Intercomparison (IRI) has been carried out five times thus far (Rozanski et al. 1992; Scott et al. 2003, 2010), but all of the laboratories participating in the experiments were aware that their dating results would be compared with other laboratories. There could be a temptation to treat IRI samples differently than commercial samples if a source of systematic error is suspected. Laboratories may repeat dating the IRI sample, choose some dates considered to be close to the "consensus date" and report them. What if laboratories measure the same sample

https://doi.org/10.1017/RDC.2015.10 Published online by Cambridge University Press

following their normal protocols, without knowing that they are participating in an interlaboratory experiment (cf. Potter and Reuther 2012)?

(4) *Difference between inner and outer rings of wood*: When dating long-lived wood from archaeological sites, it is conventional wisdom to select near-surface outer rings rather than inner rings in order to get results closer to an archaeological target event (Bowman 1990). However, in many parts of the world where preservation of organic material is poor, it can be difficult to discriminate from which aspect of a tree the sampled charred wood pieces are derived. How much does this factor affect the results in a given context? Does it lie beyond or within the statistical error range? Although this difference is not related to the uncertainty of $^{14}$C dates *per se* but is context dependent, here we test how much it affects the results in the Korean context because long-lived wood pieces are common in archaeological deposits and are usually used for $^{14}$C dating due to the difficulty of identifying *in situ* seeds that have not been bioturbated.

To examine these possible sources of uncertainty, a blind test was carried out at five different AMS laboratories across the world, which had no prior knowledge of the experiment in order to statistically compare the results from a suite of 80 samples. Five bulk charcoal samples from two archaeological sites in the central Korean peninsula were divided into multiple subsamples, and submitted blindly to the laboratories. This article reports the results of the experiment as they pertains to the four potential sources of systematic errors in $^{14}$C dating described above.

## SITES AND SAMPLES

Samples for experiment were collected from two archaeological sites in the central Korean Peninsula (Figure 1).

### Namgye, Yeoncheon (37°00′22″N, 127°05′53″E)

Namgye is a settlement with four subterranean houses previously known to date to the Proto Three Kingdoms period (100 BC–AD 300) of Korea (Seoul National University Museum 2014). The site is located on a sandy river terrace in the Hantan River Valley. This site was excavated by Seoul National University Museum in 2013.

### Hongryeonbong, Seoul (37°33′07″N, 127°00′54″E)

Hongryeonbong is a fortress of the Koguryeo (37 BC–AD 668), an ancient state in northern Korea and northeast China (Choi 2014). This fortress is located on a hilltop on the north bank of the Han River. Historical documents and archaeological evidence indicate that it was constructed around and occupied by Koguryeo's southernmost frontline troops until the mid-6th century AD, and then reused by Silla, another ancient state competing with Koguryeo, between the late 6th and 7th centuries AD (Choi 2014). The site was excavated by Korea University from 2007 to 2013 (Choi et al. 2007; Korea Institute for Archaeology and Environment 2012).

The sites are located within the humid continental/subtropical climate of the central Korean Peninsula, which includes cold, dry winters and warm, humid summers (Kim et al. 2012). The bedrock is comprised primarily of Tertiary granites uplifted as the result of the formation of backarc basins that formed distally to the continental arc as the Pacific Plate subducted orthogonally under the Asian continent (Chough 2013). Therefore, the present-day landscape includes high topographic relief with strongly seasonal monsoonal rainfall. The resulting vegetation mosaic is dominated by coniferous trees that grow on the northern aspects of the mountains with deciduous hardwood species located on the southern aspects (Kim et al. 2012).

https://doi.org/10.1017/RDC.2015.10 Published online by Cambridge University Press

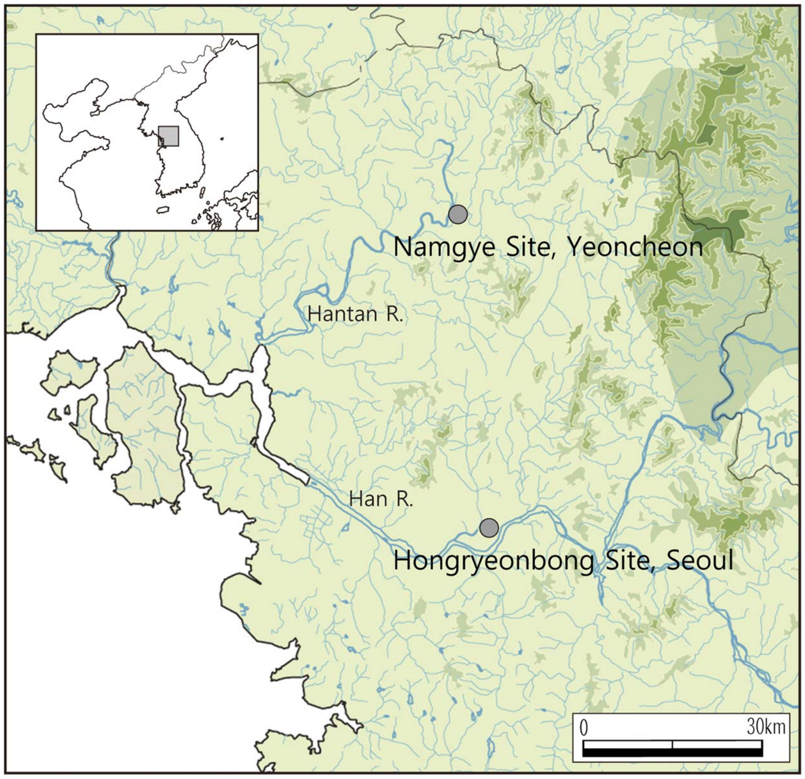

Figure 1 Locations of Namgye and Hongryeonbong

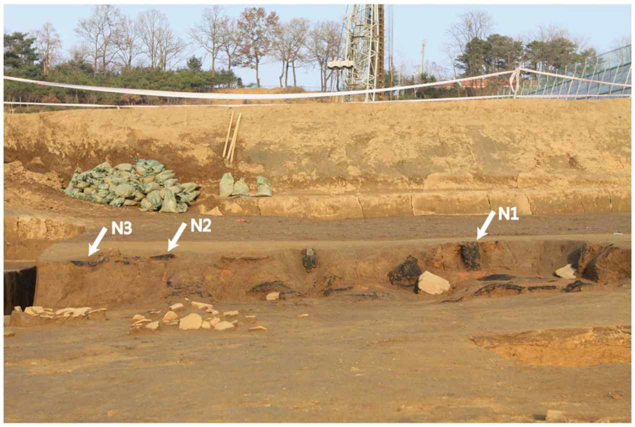

Figure 2 Sampling locations of N1, N2, and N3

Three bulk charcoal samples were collected from Namgye (Namgye 1, 2, 3; hereafter N1, N2, N3) and two from Hongryeonbong (Hongryeonbong 1 and 2; hereafter H1, H2). N1, N2, and N3 (Figure 2) were charred wood from a subterranean house feature abandoned following a fire (House No. 3). These three bulk samples are inferred to have been used as support beams for the wall installed when the house was constructed and are expected to have the same dates as one

https://doi.org/10.1017/RDC.2015.10 Published online by Cambridge University Press

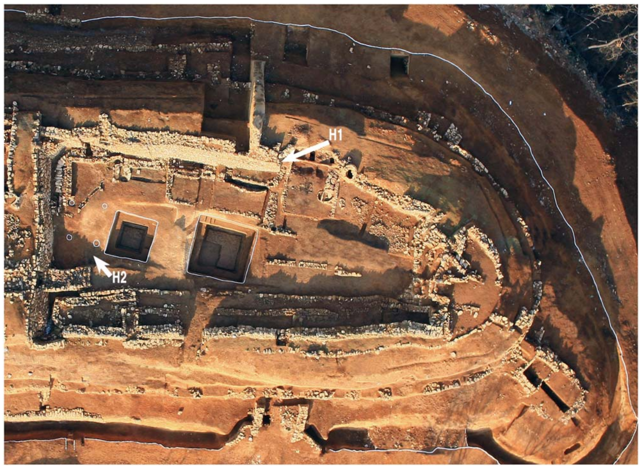

Figure 3 Sampling locations of H1 and H2.

Table 1 Species of bulk samples.

| Bulk sample | Species |
|---|---|
| N1 | *Quercus* (species unidentifiable) |
| N2 | *Quercus* (species unidentifiable) |
| N3 | *Quercus acutissima* |
| H1 | *Quercus serrata* |
| H2 | *Quercus acutissima* |

another within the standard range of error. However, we cannot eliminate the possibility that the beams might have been reused from earlier contexts. H1 and H2 were support beams for one of the inner stonewalls of the fortress (Figure 3), likely installed during reinforcement. H1 and H2 are also comprised of charred wood. H1 and H2 are also expected to be contemporaneous at an archaeological timescale. The samples were collected using rubber gloves and trowels, scooping charcoal into clean aluminum foil during excavations in collaboration between the excavation teams (Seoul National University Museum for Namgye and Korea University for Hongryeonbong) and our team in 2013.

All bulk samples from the two sites were identified as variants of oak (*Quercus* sp.), which is an abundant genus in Korea (Table 1). Because our aim was to compare the dating results measured under various conditions by dividing samples into many aliquots, we selected large pieces of wood charcoal as bulk samples, although we are aware of possible problems that may arise during pretreatment of charcoal (Gillespie 1997; Bird et al. 1999), homogeneity issues (Scott et al. 2004), age differences from archaeological target events, and the “old wood problem” (Schiffer 1986). Because of the readily available sources of standing hardwood and humid summers present in the region, “old” or recycled wood is not a taphonomic situation commonly considered in dating archaeological sites in Korea.

## METHODS

The purpose of our blind test was to check the four potential sources of uncertainty of $^{14}$C dates discussed in the previous section, by dating the same samples under multiple conditions.

https://doi.org/10.1017/RDC.2015.10 Published online by Cambridge University Press

Materially coherent bulk samples were divided into smaller aliquots using clean tweezers and knives for simultaneous and staggered submittal to the five different AMS laboratories being tested. Each subsample was assigned a subsample identification number to anonymize its relationship to the bulk sample within the suite of materials submitted to the laboratories. Depending on size, each bulk charcoal sample was divided into either 20 (N1, H1, and H2) or 10 (N2 and N3) subsamples; thus, a total of 80 subsamples were sent to the laboratories. Clear division between inner and outer rings of samples was only possible for H1, and the age difference between the inner and outer rings was considered to be approximately 10 to 15 yr, although the number of rings between the two parts was not exactly counted. For N1, N2, N3, and H2, outer parts of bulks were sampled. When dividing bulk samples into aliquots, we were careful to avoid possible contamination. As part of our sampling protocol, rings in similar ages were assayed to homogenize the aliquots from each bulk sample as much as possible (N1: 25.6–45.0 mg; N2: 42.6–59.3 mg; N3: 53.2–67.6 mg; H1 inner: 76.7–96.4 mg; H1 outer: 50.6–75.6 mg; H2: 47.3–51.4 mg) to avoid introducing systematic errors from dating different aspects of tree wood (*sensu* Scott et al. 2003, 2004). We did not pulverize samples because archaeologists rarely pulverize samples when they submit samples to laboratories for dating.

Samples were submitted to five AMS laboratories: two in the USA, one in the UK, one in Korea, and one in Japan. We do not specify names of the laboratories subjected to the test here; instead, we randomly assign the laboratory codes as A, B, C, D, and E. Each laboratory measured 16 samples. Among the five laboratories, four (Labs A to D) measured samples twice within a 2-month interval, while Lab E received its 16 samples at one time. Only site names, locations, and subsample identification numbers assigned by the research team were provided to the laboratories, and we did not inform staff at the laboratories that a test was being performed.[1]

Following receipt of the results of $^{14}$C dating from the respective laboratories, dates were analyzed using Bayesian methods. First, medians of uncalibrated BP dates were estimated for each bulk sample by inferring a posterior distribution with the Markov chain Monte Carlo technique. Then, for a subset of samples that consistently showed different dates, the Bayesian $p$ value (Bayarri and Berger 2000), which determines the probability of occurrence of data more deviant than the observed data for relevant statistics, was calculated to assess whether the difference was statistically significant. Non-Bayesian chi-squared tests (Ward and Wilson 1978; Bronk Ramsey 2009) were also carried out for these samples to supplement the Bayesian $p$ value calculation.

## RESULTS AND DISCUSSION

Among the 80 samples submitted, one sent to Lab A was determined to be undatable; therefore, we report dating results of 79 samples (Tables 2 and 3). Detailed statistical and physical analyses of the results are now in progress; thus, we briefly comment on some aspects of the experiment here.

Precisions of the dates, presented as the standard deviations of uncalibrated BP dates, vary with laboratories, ranging from 15 to 60 yr within the 1σ confidence interval, probably due to differences in isotope counting procedures among laboratories. In general, most dates from each site are in good statistical agreement with one another and the repeatability of measurement under identical conditions appears to be met. Statistically significant differences among subsamples assayed from bulk samples [i.e. what Ward and Wilson (1978) call “Case II error”]

[1] After the experiment, we informed all laboratories of our test, including the research purpose and results. The laboratories are not aware of the names of other laboratories in the experiment.

https://doi.org/10.1017/RDC.2015.10 Published online by Cambridge University Press

Table 2 AMS dates of Namgye bulk samples. [Some labs did not provide $\delta^{13}C$ or fraction modern carbon (FMC) values. Only Lab E reported that $\delta^{13}C$ values were measured by AMS.]

| Bulk sample | Subsample ID | Lab | Batch | BP | 1σ | $\delta^{13}C$ (‰) | FMC | ± |
|---|---|---|---|---|---|---|---|---|
| N 1 | MRR2013-1 | A | 1 | 1840 | 50 | –27.47 | | |
| | MRR2013-3 | A | 1 | 1840 | 50 | –29.14 | | |
| | MRR2013-2 | A | 2 | 1860 | 40 | –24.14 | | |
| | MRR2013-4 | A | 2 | 2170 | 60 | –37.10 | | |
| | MRR2013-17 | B | 1 | 1710 | 30 | –25.6 | 0.8083 | 0.0030 |
| | MRR2013-19 | B | 1 | 1780 | 30 | –25.7 | 0.8012 | 0.0030 |
| | MRR2013-18 | B | 2 | 1910 | 30 | –25.8 | 0.7884 | 0.0029 |
| | MRR2013-20 | B | 2 | 1800 | 30 | –26.0 | 0.7993 | 0.0030 |
| | MRR2013-49 | C | 1 | 1860 | 25 | –25.6 | 0.7932 | 0.0022 |
| | MRR2013-51 | C | 1 | 1910 | 25 | –25.0 | 0.7882 | 0.0021 |
| | MRR2013-50 | C | 2 | 1850 | 15 | –25.5 | 0.7945 | 0.0012 |
| | MRR2013-52 | C | 2 | 1870 | 15 | –25.8 | 0.7922 | 0.0012 |
| | MRR2013-33 | D | 1 | 1900 | 20 | –27.21 | | |
| | MRR2013-35 | D | 1 | 1875 | 20 | –24.39 | | |
| | MRR2013-34 | D | 2 | 1865 | 20 | –26.45 | | |
| | MRR2013-36 | D | 2 | 1865 | 20 | –25.88 | | |
| | MRR2013-67 | E | 1 | 1868 | 38 | | 0.7925 | 0.0037 |
| | MRR2013-69 | E | 1 | 1882 | 37 | | 0.7911 | 0.0036 |
| | MRR2013-68 | E | 1 | 1866 | 37 | | 0.7927 | 0.0037 |
| | MRR2013-70 | E | 1 | 1907 | 37 | | 0.7886 | 0.0037 |
| N 2 | MRR2013-5 | A | 1 | N/A | — | — | | |
| | MRR2013-6 | A | 2 | 1790 | 40 | –26.75 | | |
| | MRR2013-21 | B | 1 | 1790 | 30 | –27.2 | 0.8002 | 0.003 |
| | MRR2013-22 | B | 2 | 1810 | 30 | –28.6 | 0.7983 | 0.003 |
| | MRR2013-53 | C | 1 | 2275 | 25 | –27.1 | 0.7535 | 0.0021 |
| | MRR2013-54 | C | 2 | 1860 | 15 | –26.9 | 0.7932 | 0.0012 |
| | MRR2013-37 | D | 1 | 1830 | 20 | –31.2 | | |
| | MRR2013-38 | D | 2 | 1850 | 20 | –26.68 | | |
| | MRR2013-71 | E | 1 | 1905 | 34 | | 0.7889 | 0.0034 |
| | MRR2013-72 | E | 1 | 1910 | 42 | | 0.7884 | 0.0041 |
| N 3 | MRR2013-7 | A | 1 | 1850 | 40 | –27.24 | | |
| | MRR2013-8 | A | 2 | 1880 | 40 | –26.89 | | |
| | MRR2013-23 | B | 1 | 1690 | 30 | –28.0 | 0.8103 | 0.003 |
| | MRR2013-24 | B | 2 | 1870 | 30 | –28.4 | 0.7923 | 0.003 |
| | MRR2013-55 | C | 1 | 1835 | 25 | –27.6 | 0.7959 | 0.0022 |
| | MRR2013-56 | C | 2 | 1855 | 15 | –27.5 | 0.7936 | 0.0012 |
| | MRR2013-39 | D | 1 | 1870 | 20 | –28.82 | | |
| | MRR2013-40 | D | 2 | 1840 | 20 | –26.87 | | |
| | MRR2013-73 | E | 1 | 1913 | 35 | | 0.7881 | 0.0034 |
| | MRR2013-74 | E | 1 | 1830 | 33 | | 0.7963 | 0.0032 |

were not detected in this experiment, as we expected during collection, although it was not perfectly certain that all the bulk samples from each site were contemporaneous when we collected them in the field. A statistical estimation of median dates using the Markov chain Monte Carlo technique does not show significant differences among bulk samples from each site (N1 = 1866.59, N2 = 1856.51, N3 = 1855.06, H1 = 1504.87, and H2 = 1493.27).

Table 3 AMS dates of Hongryeonbong bulk samples. [Some labs did not provide $\delta^{13}$C or fraction modern carbon (FMC) values. Only Lab E reported that $\delta^{13}$C values were measured by AMS.]

| Bulk sample | Sample ID | Lab | Ring | Batch | BP | 1σ | $\delta^{13}$C (‰) | FMC | ± |
|---|---|---|---|---|---|---|---|---|---|
| H 1 | MRR2013-9 | A | in | 1 | 1550 | 40 | –24.13 | | |
| | MRR2013-11 | A | out | 1 | 1500 | 40 | –24.76 | | |
| | MRR2013-10 | A | in | 2 | 1400 | 50 | –26.39 | | |
| | MRR2013-12 | A | out | 2 | 1510 | 40 | –23.73 | | |
| | MRR2013-25 | B | in | 1 | 1350 | 30 | –25.0 | 0.8453 | 0.0032 |
| | MRR2013-27 | B | out | 1 | 1370 | 30 | –25.0 | 0.8432 | 0.0031 |
| | MRR2013-26 | B | in | 2 | 1490 | 30 | –25.0 | 0.8307 | 0.0031 |
| | MRR2013-28 | B | out | 2 | 1440 | 30 | –24.9 | 0.8359 | 0.0031 |
| | MRR2013-57 | C | in | 1 | 1445 | 25 | –25.6 | 0.8356 | 0.0023 |
| | MRR2013-59 | C | out | 1 | 1440 | 25 | –25.8 | 0.8361 | 0.0023 |
| | MRR2013-58 | C | in | 2 | 1465 | 15 | –25.4 | 0.8331 | 0.0015 |
| | MRR2013-60 | C | out | 2 | 1485 | 20 | –26.0 | 0.8314 | 0.0016 |
| | MRR2013-41 | D | in | 1 | 1515 | 20 | –26.35 | | |
| | MRR2013-43 | D | out | 1 | 1520 | 20 | –28.18 | | |
| | MRR2013-42 | D | in | 2 | 1475 | 20 | –25.27 | | |
| | MRR2013-44 | D | out | 2 | 1495 | 20 | –25.69 | | |
| | MRR2013-75 | E | in | 1 | 1533 | 35 | | 0.8263 | 0.0036 |
| | MRR2013-77 | E | out | 1 | 1572 | 32 | | 0.8222 | 0.0033 |
| | MRR2013-76 | E | in | 1 | 1527 | 32 | | 0.8269 | 0.0033 |
| | MRR2013-78 | E | out | 1 | 1584 | 32 | | 0.8211 | 0.0032 |
| H 2 | MRR2013-13 | A | — | 1 | 1480 | 40 | –26.75 | | |
| | MRR2013-15 | A | — | 1 | 1540 | 50 | –27.99 | | |
| | MRR2013-14 | A | — | 2 | 1560 | 50 | –28.15 | | |
| | MRR2013-16 | A | — | 2 | 1450 | 40 | –27.32 | | |
| | MRR2013-29 | B | — | 1 | 1390 | 30 | –27.0 | 0.8411 | 0.0031 |
| | MRR2013-31 | B | — | 1 | 1340 | 30 | –25.9 | 0.8464 | 0.0032 |
| | MRR2013-30 | B | — | 2 | 1520 | 30 | –26.0 | 0.8276 | 0.0031 |
| | MRR2013-32 | B | — | 2 | 1500 | 30 | –26.2 | 0.8297 | 0.0031 |
| | MRR2013-61 | C | — | 1 | 1450 | 25 | –26.3 | 0.8350 | 0.0023 |
| | MRR2013-63 | C | — | 1 | 1425 | 25 | –26.6 | 0.8372 | 0.0023 |
| | MRR2013-62 | C | — | 2 | 1475 | 15 | –27.7 | 0.8324 | 0.0014 |
| | MRR2013-64 | C | — | 2 | 1510 | 15 | –26.9 | 0.8288 | 0.0014 |
| | MRR2013-45 | D | — | 1 | 1505 | 20 | –26.93 | | |
| | MRR2013-47 | D | — | 1 | 1515 | 15 | –26.17 | | |
| | MRR2013-46 | D | — | 2 | 1505 | 20 | –25.74 | | |
| | MRR2013-48 | D | — | 2 | 1455 | 20 | –26.63 | | |
| | MRR2013-79 | E | — | 1 | 1505 | 33 | | 0.8292 | 0.0034 |
| | MRR2013-81 | E | — | 1 | 1452 | 33 | | 0.8347 | 0.0034 |
| | MRR2013-80 | E | — | 1 | 1438 | 31 | | 0.8361 | 0.0032 |
| | MRR2013-82 | E | — | 1 | 1507 | 25 | | 0.8290 | 0.0025 |

Therefore, multiple bulk samples from each site can be seen as statistically overlapping between analyses even between different laboratories. The estimated median of 39 subsamples from Namgye settlement (N1 to N3) is 1859 ± 14 BP, and that of the other 40 samples from Hongryeonbong (H1 and H2) is 1492 ± 15 BP (Lee et al. 2014).

Table 4 Estimated BP dates using Markov chain Monte Carlo simulation by laboratory.

| | Laboratory | | | | |
|---|---|---|---|---|---|
| Site | A | B | C | D | E |
| Namgye | 1866 | 1830 | 1863 | 1863 | 1877 |
| 95% credible set | (1832, 1901) | (1784, 1873) | (1843, 1881) | (1827, 1901) | (1852, 1899) |
| Hongryeonbong | 1499 | 1462 | 1495 | 1496 | 1510 |
| 95% credible set | (1465,1533) | (1416,1507) | (1475,1516) | (1459,1533) | (1484,1534) |

There are a few outliers in the data, which warrant discussion. One date of N1 bulk sample (MRR2013-4: 2170 ± 60 BP) measured by Lab A and one of N2 (MRR2013-53: 2275 ± 25 BP) by Lab C fall outside the 2σ confidence interval from the aggregate confidence interval generated from all samples. When the two anomalous values are manually removed, the median Markov chain Monte Carlo age of Namgye becomes 1853 ± 12 BP.

The interlaboratory variance does not seem significant in general. However, the results from Lab B tend to be younger than the other laboratories' results (Table 4). A closer look at the results suggests that this tendency results from Lab B's interbatch differences: dates measured from Batch 1 were consistently younger than those measured in Batch 2 two months later, for all samples regardless of site and bulk sample (Figures 4 and 5). Comparison of the dates with those measured by the other laboratories indicates that the dates of Lab B Batch 2 are in closer agreement with the dates generated from the other laboratories, unlike those of Batch 1 (Figures 6, 7, 8, and 9).

To assess the amount of possible bias with Lab B Batch 1, we calculated the Bayesian $p$ value. In our study, $y_{\text{obs}}$ was the observed data, $y_{\text{rep}}$ represented replicated data, $\theta$ was the parameter, and $T$ was the statistic representing deviation of the data. While the classical $p$ value is defined by $p_c = P(T(y_{\text{rep}}) \leq T(y_{\text{obs}})|\theta)$, for fixed $\theta$, the Bayesian posterior predictive $p$ value is defined by $p_b = P(T(y_{\text{rep}}) \leq T(y_{\text{obs}})|y_{\text{obs}})$, which is the probability that the replicated data deviate from the current model more than the observed data when using all the information available. Although it is convenient to use and is consequently popular, the Bayesian posterior predictive $p$ value has been criticized for double-using data to calculate both the test statistic and the posterior probability (Tsui and Weerahandi 1989; Berger and Boos 1994). To avoid this problem, we calculate the partial posterior predictive $p$ value (Bayarri and Berger 2000) defined by $P_{\text{ppp}} = P(T(y_{\text{rep}}) \leq t_{\text{obs}}|y_{\text{obs}}/t_{\text{obs}})$, where $t_{\text{obs}}$ is the observed test statistic and $y_{\text{obs}}/t_{\text{obs}}$ is the part of the data not involved in calculating $t_{\text{obs}}$. By dividing $y_{\text{obs}}$ to $t_{\text{obs}}$ and $y_{\text{obs}}/t_{\text{obs}}$, the partial posterior predictive $p$ value avoids the issue of circular validation.

In practice, often the division of $y_{\text{obs}}$ to $t_{\text{obs}}$ and the rest is not obvious. In the current analysis, the division is rather obvious, because $t_{\text{obs}}$ is a test statistic based on Lab B Batch 1. We set $y_{\text{obs}}/t_{\text{obs}}$ as all the data except Lab B Batch 1. We estimated parameters related to Namgye and Hongryeonbong dates without using eight dates from Lab B Batch 1, and eliminate the influence on parameters by using Monte Carlo integration. Then, Bayesian $p$ values of three test statistics (mean, minimum, and maximum) for uncalibrated BP dates from Namgye and Hongryeonbong were calculated. Specifically, we calculated (1) the probability that each mean of four data points replicated from Namgye and Hongryeonbong dates, respectively, are smaller than mean of the four observed data points (i.e. mean dates from Lab B Batch 1; Namgye = 1742.5 and Hongryeonbong = 1362.5); (2) the probability that minimums of four replicated data points are smaller than those of the four observed data points (Namgye = 1690;

https://doi.org/10.1017/RDC.2015.10 Published online by Cambridge University Press

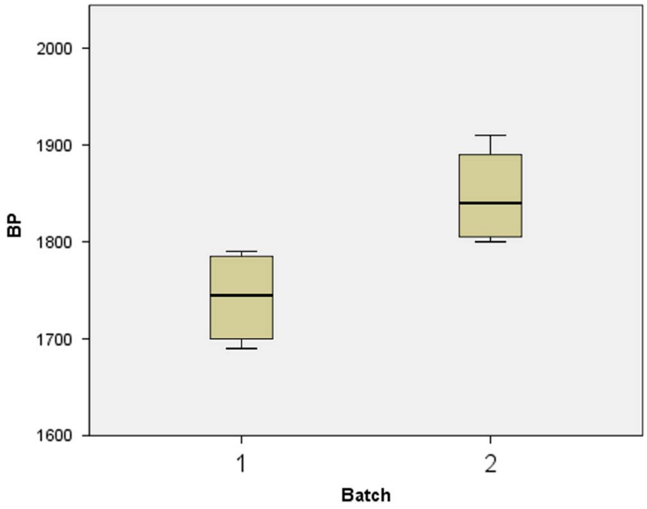

Figure 4 Distribution of Namgye BP dates measured by Lab B (Batch 1 median = 1745 BP; Batch 2 median = 1840 BP).

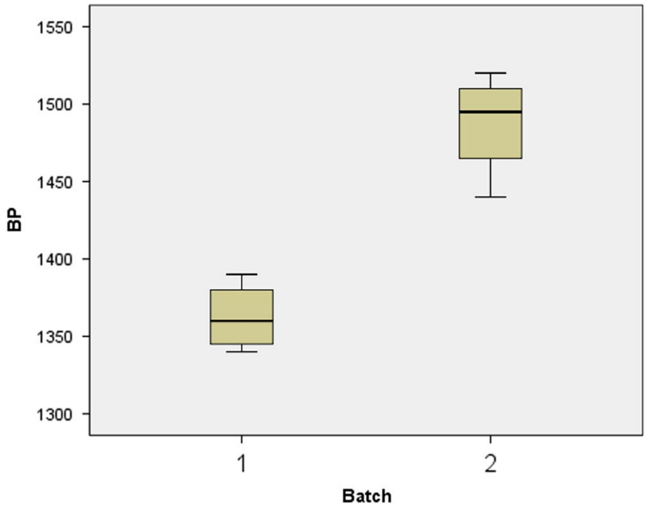

Figure 5 Distribution of Hongryeonbong BP dates measured by Lab B (Batch 1 median = 1360 BP; Batch 2 median = 1495 BP).

Hongryeonbon = 1340); and (3) the probability that a maximum of four replicated data points are smaller than those of the four observed data points (Namgye = 1790; Hongryeonbong = 1390).

Posterior distributions of all three Bayesian *p* values reject the null hypothesis that the variant statistical distribution of $^{14}$C ages generated by Lab B Batch 1 for both sites are the product of random errors (Table 5). This suggests that the eight measurements of Lab B Batch 1 are likely to have a systemic error in some aspect of the taphonomic, handling, or analytical measurement of the samples. Based on the available data, it is unknowable whether this consistent difference resulted from contamination during collection or handling of the sample, pretreatment,

https://doi.org/10.1017/RDC.2015.10 Published online by Cambridge University Press

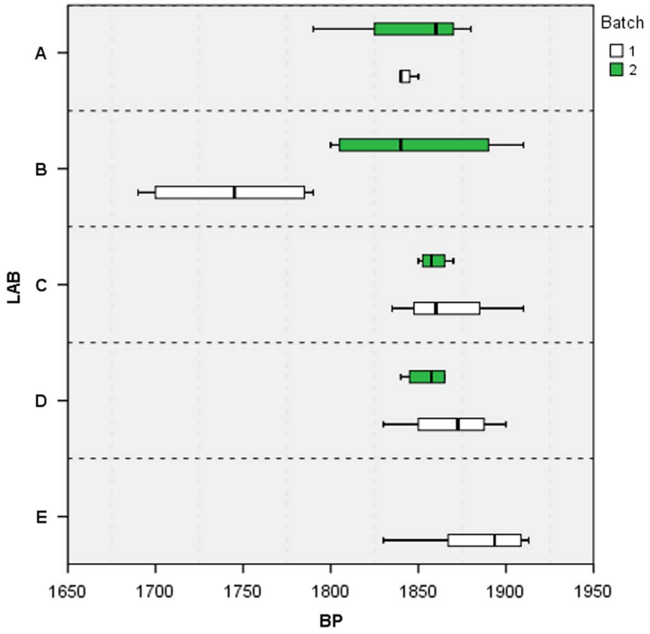

Figure 6 Comparison of Namgye BP dates by laboratory and batch

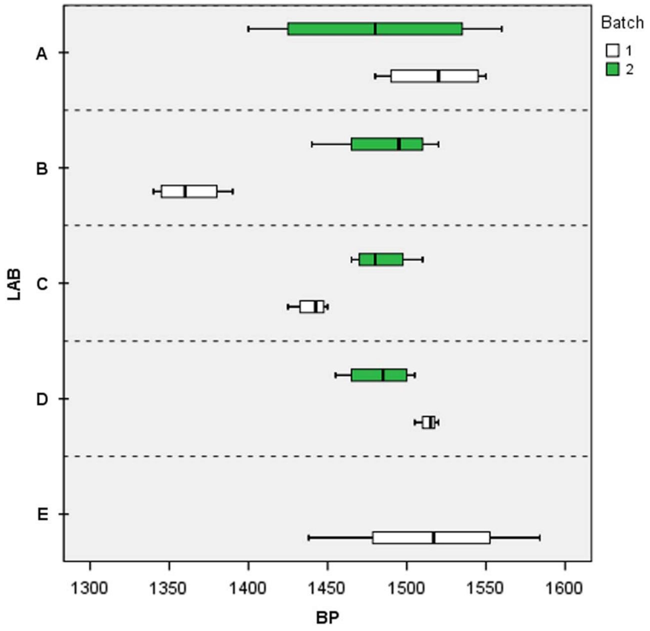

Figure 7 Comparison of Hongryeonbong BP dates by laboratory and batch.

erroneous measurement of standards, or some changes in background of measurement. Taphonomic circumstances for the nonmatching age sets are not suspected since postdepositional contamination would have likely affected the samples equally. The same is true about handling,

https://doi.org/10.1017/RDC.2015.10 Published online by Cambridge University Press

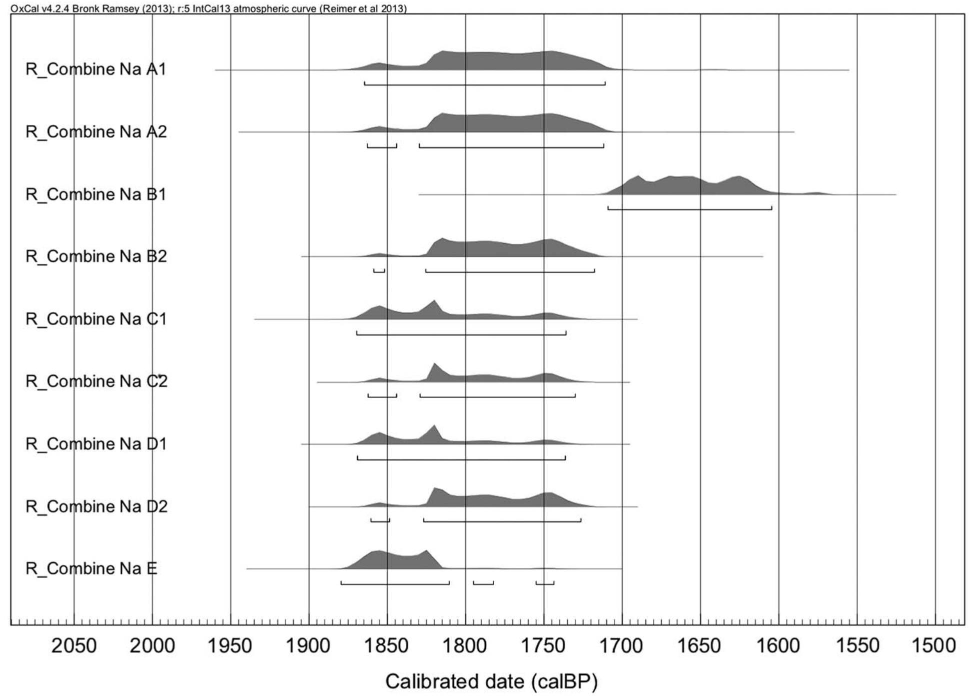

Figure 8 Distribution of combined BP dates of Namgye by laboratory and batch using OxCal v 4.2 and IntCal13 (Reimer et al. 2013).

storage, and shipping, but, given the size of the artifact, it is plausible that one portion was inadvertently mishandled despite the protocols.

Mainly due to this statistically significant difference between batches, there is a nonrandom variance in the agreement of dates measured by Lab B compared to those of other laboratories that participated in this test, and non-Bayesian chi-squared tests (Ward and Wilson 1978) using OxCal v 4.2 demonstrate similar results (Tables 6 and 7). In the case of Namgye, only Lab B's $T$ value is significant at the 0.05 level, with 77.6% agreement. Hongryeonbong dates measured by Lab B also demonstrate high $T$ value and low agreement, although some labs' results also have $T$ values significant at 0.05 level.

An experiment on the potential differences between inner and outer rings was carried out only on the H1 sample, and consistent differences in age outcomes were not detected (Figure 10). The number of rings in H1 was not rigorously counted by a botanical specialist, but our observation during aliquot division suggests the age difference between the two parts of H1 was only 10 to 15 yr. Also, the diameter of the bulk sample (oak tree) was 15 cm, suggesting that the age of the tree would not have been older than 25 yr in the typical central Korean environment (Byun et al. 2010). Thus, although outer rings should theoretically provide a younger age than inner rings (Bowman 1990), the difference appears to lie within the statistical error range in this case, probably owing to the young age of the tree at the time it was felled to use as construction material.

Overall, our blind tests demonstrate generally good concordance in the results and present acceptable errors at an archaeological scale, but interbatch differences may potentially result in uncertainty of dating results, although this was detected for only one laboratory out of five that

https://doi.org/10.1017/RDC.2015.10 Published online by Cambridge University Press

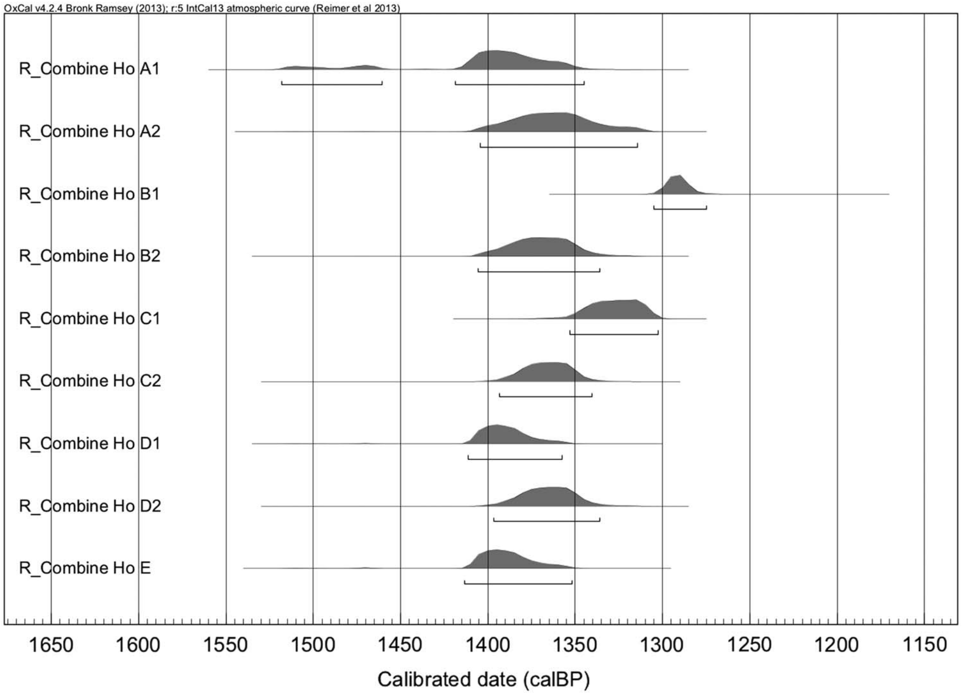

Figure 9 Distribution of combined BP dates of Hongryeonbong by laboratory and batch using OxCal v 4.2 and IntCal13 (Reimer et al. 2013).

Table 5 Bayesian *p* values of Lab B Batch 1 dates.

| | *p* values | | |
|---|---|---|---|
| | Mean | Minimum | Maximum |
| Namgye | 1.13e-05 | 0.0116 | 3.48e-05 |
| Hongryeonbong | 6.41e-06 | 0.0066 | 1.62e-06 |

were subject to the tests. Bayesian *p* values and chi-squared tests reject the null hypothesis that the errors randomly occurred.

## CONCLUSION

A total of 79 $^{14}$C samples were analyzed from five macrosamples recovered from two separate archaeological sites showing a narrow distribution of uncalibrated $^{14}$C ages. One batch of samples produced results consistently outside the 2σ distribution of the remaining 75 samples and was determined to be the result of a systematic error, but the source of the error is not specifically known. The tests performed in this experiment were not designed to highlight deficiencies or successes of individual laboratories or identify unreliable laboratories, but rather to determine potential anomalies in data generation for $^{14}$C dating, in general. Users of $^{14}$C dating should be aware of different sources of potential uncertainty resulting from the metabolic lifecycle of the organism, burial, taphonomy, recovery, handling, shipping, and laboratory treatment of the sample in order to relevantly interpret the results. Uncertainty derived from random errors can be decreased by increasing sample size, insisting on more robust isotopic

https://doi.org/10.1017/RDC.2015.10 Published online by Cambridge University Press

Table 6 Results of chi-squared tests and agreements of Namgye dates by laboratory.

| Lab | *n* | *df* | *T* | *T* at 0.05 | Agreement (%) |
|---|---|---|---|---|---|
| A | 6 | 5 | 2.8 | 11.1 | 99.1 |
| B | 8 | 7 | 41.8 | 14.1 | 77.6 |
| C | 7 | 6 | 6.0 | 12.6 | 93.0 |
| D | 8 | 7 | 8.4 | 14.1 | 93.7 |
| E | 8 | 7 | 4.9 | 14.1 | 98.3 |

Table 7 Results of chi-squared tests and agreements of Hongryeonbong dates by laboratory.

| Lab | *n* | *df* | *T* | *T* at 0.05 | Agreement (%) |
|---|---|---|---|---|---|
| A | 8 | 7 | 9.5 | 14.1 | 97.3 |
| B | 8 | 7 | 40.3 | 14.1 | 40.5 |
| C | 8 | 7 | 14.2 | 14.1 | 89.5 |
| D | 8 | 7 | 9.3 | 14.1 | 99.0 |
| E | 8 | 7 | 18.2 | 14.1 | 82.6 |

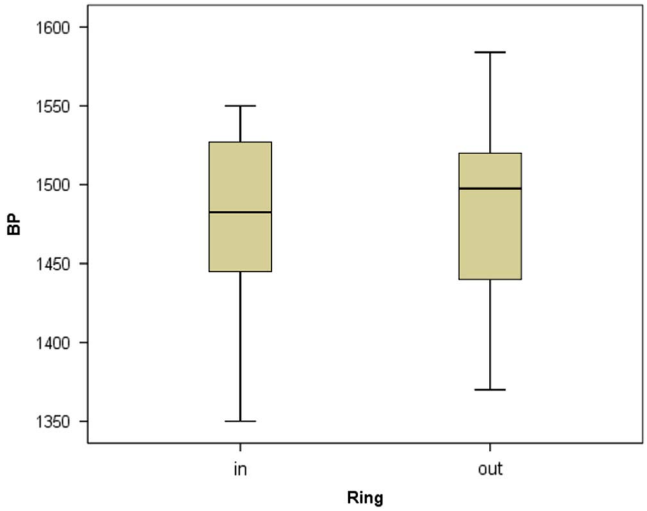

Figure 10 Comparison of inner and outer ring BP dates of H1 (in: $n = 10$, median = 1482.5, mean = 1475.0, standard deviation = 63.04; out: $n = 10$, median = 1497.5, mean = 1491.6, standard deviation = 63.58).

counting procedures and using appropriate statistical techniques. However, repeated errors could signal more significant problems in either pre-laboratory handling of samples or laboratory procedures, and it would be inappropriate to include them in the statistical sample of unaffected samples. Without determining whether differences among $^{14}$C ages are caused by random errors or possible systematic errors, it can be difficult to properly understand the uncertainty of measurements provided to consumers. However, when dating results are obtained from laboratories, most end-users will not be aware of potential errors in their data because sample sizes tend to be small, and our results suggest that results with systematic errors are erroneously included and reported in the archaeological literature.

https://doi.org/10.1017/RDC.2015.10 Published online by Cambridge University Press

Certainly, whether a data set is subject to either random errors or systematic errors is not clear-cut unless large numbers of samples are taken from one context. Even then, the division between random and systematic errors can be heuristic, depending on one's perspective. $^{14}$C laboratories may view, for example, interbatch differences detected in this experiment as an uncontrollable random error that can possibly happen as a mass spectrometer runs many times. At the same time, a systematic error may be suspected because the error repeatedly occurs outside the statistical boundaries of a truly random distribution. In such cases, archaeologists face a dilemma in interpreting the veracity of their samples. In this case, we were able to statistically determine the presence of a systematic error in the results of Lab B's Batch 1 based on a large data set of samples generated. However, few archaeological research projects can afford to generate so many ages from single macrosamples in order to identify potential sources of error. Even when identified, the source of the error is not obvious.

Large sample sizes are important for archaeologists to get accurate dates of archaeological events, but simply increasing sample size does not automatically guarantee a decrease in uncertainty unless possible systematic errors are relevantly controlled. If multiple samples are dated under the same conditions, it is possible for all results to be affected by the same systematic errors. This risk can be mitigated when samples are dated under multiple conditions and results are compared by users before ultimate age determination, although this may be costly and time consuming. Although there is no universal method for separating random and systematic errors of dating results, our experiment suggests that it is necessary for archaeologists to establish an organized strategy for dating sites before submitting samples to laboratories, which can avoid the inclusion of possible systematic errors.

## ACKNOWLEDGMENTS

This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2013S1A5B6043901). We thank the participating laboratories for their cooperation with the test and understanding of our motivation to advance the scientific potential of the discipline. We also thank Tim Jull, Mark McClure, and Alex Bayliss and one anonymous reviewer for valuable comments.

## REFERENCES

Bayarri MJ, Berger JD. 2000. *P* values for composite null models. *Journal of the American Statistical Association* 95(452):1127–42.

Berger RL, Boos DD. 1994. *P* values maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association* 89(427):1012–6.

Bird MI, Ayliffe LK, Fifield LK, Turney CSM, Cresswell RG, Barrows TT, David B. 1999. Radiocarbon dating of "old" charcoal using a wet oxidization, steeped-combustion procedure. *Radiocarbon* 41(2):127–40.

Bowman S. 1990. *Radiocarbon Dating*. Berkeley: University of California Press.

Bronk Ramsey C. 2009. Bayesian analysis of radiocarbon dates. *Radiocarbon* 51(1):337–60.

Buck CE, Millard A. 2004. *Tools for Constructing Chronologies: Crossing Disciplinary Boundaries*. London: Springer.

Buck CE, Christen JA, Kenworthy JB, Litton CD. 2007. Estimating the duration of archaeological activity using $^{14}$C determinations. *Oxford Journal of Archaeology* 13(2):229–40.

Byun JG, Lee WK, Nor DK, Kim SH, Choi JK, Lee YJ. 2010. The relationship between tree radial growth and topographic and climatic factors in red pine and oak in central regions of Korea. *Journal of Korean Forest Society* 99(6):908–13.

Choi J. 2014. *Acha Montain Fortresses and the Southward Expansion of Koguryo*. Seoul: Seokyeong Press.

Choi J, Lee SJ, Oh EJ, Cho SY. 2007. *Excavation Report of Hongryeonbong Fortress II*. Jochiwon: Korea University.

Chough SK. 2013. *Geology and Sedimentology of the Korean Peninsula*. Waltham: Elsevier.

Christen JA. 1994. Summarizing a set of radiocarbon determinations - a robust approach. *Applied Statistics-Journal of the Royal Statistical Society Series C* 43(3):489–503.

Christen JA, Buck CE. 1998. Sample selection in radiocarbon dating. *Applied Statistics-Journal*

https://doi.org/10.1017/RDC.2015.10 Published online by Cambridge University Press

*of the Royal Statistical Society Series C* 47(4): 543–57.

Faught MK. 2008. Archaeological roots of human diversity in the New World: a compilation of accurate and precise radiocarbon ages from the earliest sites. *American Antiquity* 73(4):670–98.

Gillespie R. 1997. Burnt and unburnt carbon: dating charcoal and burnt bone from the Willandra Lakes, Australia. *Radiocarbon* 39(3):225–36.

Graf KE. 2009. The good, the bad, and the ugly": evaluating the radiocarbon chronology of the middle and late Upper Paleolithic in the Enisei River valley, south-central Siberia. *Journal of Archaeological Science* 36(3):694–707.

Keith MS, Anderson GM. 1963. Radiocarbon dating: fictitious results with mollusk shells. *Science* 141(3581):634–7.

Kim J, Lee MB, Kong WS, Kim TH, Kang CS, Park K, Park BI, Park HD, Song HH, Son MW, Yang HG, Lee SH, Choi YE. 2012. *Physical Geography of Korea*. Seoul: Seoul National University Press.

Korea Institute for Archaeology and Environment. 2012. *Preliminary Excavation Report of Hongryeonbong Fortresses I and II*. Jochiwon: Korea University.

Lee Y, Lee J, Kim J. 2014. Bayesian analyses of uncertainty of radiocarbon dating. *Proceedings of the 38th Annual Meeting of the Korean Archaeological Society*. p 337–47.

Mellars P. 2006. A new radiocarbon revolution and the dispersal of modern humans in Eurasia. *Nature* 439(7079):932–5.

Nielsen-Marsh CM, Hedges RE. 2000a. Patterns of diagenesis in bone I: the effects of site environments. *Journal of Archaeological Science* 27(12):1139–50.

Nielsen-Marsh CM, Hedges RE. 2000b. Patterns of diagenesis in bone II: effects of acetic acid treatment and the removal of diagenetic $CO_3^{2-}$. *Journal of Archaeological Science* 27(12):1151–9.

Pettitt PB, Davies W, Gamble CS, Richards MB. 2003. Palaeolithic radiocarbon chronology: quantifying our confidence beyond two half-lives. *Journal of Archaeological Science* 30(12):1685–93.

Pichler H, Firedrich W. 1976. Radiocarbon dates of Santorini volcanics. *Nature* 262(5567):373–4.

Potter BA, Reuther JD. 2012. High resolution radiocarbon dating at the Gerstle River Site, central Alaska. *American Antiquity* 77(1):71–98.

Price TD, Blitz J, Burton J, Ezzo JA. 1992. Diagenesis in prehistoric bone: problems and solutions. *Journal of Archaeological Science* 19(5):513–29.

Reimer PJ, Bard E, Bayliss A, Beck JW, Blackwell PG, Bronk Ramsey C, Buck CE, Cheng H, Edwards RL, Friedrich M, Grootes PM, Guilderson TP, Haflidason H, Hajdas I, Hatté C, Heaton TJ, Hoffmann DL, Hogg AG, Hughen KA, Kaiser KF, Kromer B, Manning SW, Niu M, Reimer RW, Richards DA, Scott EM, Southon JR, Staff RA, Turney CSM, van der Plicht J. 2013. IntCal13 and Marine13 radiocarbon age calibration curves 0–50,000 years cal BP. *Radiocarbon* 55(4): 1869–87.

Rozanski K, Stichler W, Gofiantini R, Scott EM, Beukens RP, Kromer B, van der Plicht J. 1992. The IAEA $^{14}$C intercomparison exercise 1990. *Radiocarbon* 34(3):506–19.

Schiffer MB. 1986. Radiocarbon dating and the "old wood" problem: the case of the Hohokam chronology. *Journal of Archaeological Science* 13(1):13–30.

Scott EM. 2011. Models, data, statistics, and outliers: a statistical revolution in archaeology and $^{14}$C dating. *Radiocarbon* 53(4):559–62.

Scott EM, Bryant C, Carmi I, Cook GT, Gulliksen S, Harkness DD, Heinemeier J, McGee E, Naysmith P, Possnert G, van der Plicht J, Van Strydonck M. 2003. Homogeneity testing. In: The Third International Radiocarbon Intercomparison (TIRI) and the Fourth International Radiocarbon Intercomparison (FIRI) 1990–2002. Results, Analyses, and Conclusions [special issue]. *Radiocarbon* 45(2):144–9.

Scott EM, Boaretto E, Bryant C, Cook GT, Gulliksen S, Harkness DD, Heinemeier J, McGee E, Naysmith P, Posssnert G, van der Plicht H, Van Strydonck M. 2004. Future needs and requirements for AMS $^{14}$C standards and reference materials. *Nuclear Instruments and Methods in Physics Research B* 223–224:382–7.

Scott EM, Cook GT, Naysmith P. 2007. Error and uncertainty in radiocarbon measurements. *Radiocarbon* 49(2):427–40.

Scott EM, Cook GT, Naysmith P. 2010. The Fifth International Radiocarbon Intercomparison (VIRI): an assessment of laboratory performance in stage 3. *Radiocarbon* 53(2–3):859–65.

Seoul National University Museum. 2014. *Preliminary Excavation Report of Namgyeri Settlement, Yeonchon*. Unpublished report submitted to Cultural Heritage Administration of Korea.

Shotton FW. 1972. An example of hard-water error in radiocarbon dating of vegetable matter. *Nature* 240(5382):460–1.

Stuiver M, Suess HE. 1966. On the relationship between radiocarbon dates and true sample ages. *Radiocarbon* 8(1):534–40.

Stuiver M, Pearson GW, Branziunas TF. 1986. Radiocarbon age calibration of marine samples back to 9000 cal yr BP. *Radiocarbon* 28(2):980–1021.

Tsui K-W, Weerahandi S. 1989. Generalized p-values in significance testing of hypothesis in the presence of nuisance parameters. *Journal of the American Statistical Association* 84(406):602–7.

van der Plicht J, Bruins HJ. 2001. Radiocarbon dating in Near-Eastern contexts: confusion and quality control. *Radiocarbon* 43(3):1155–66.

Ward GK, Wilson SR. 1978. Procedures for comparing and combining radiocarbon age determinations: a critique. *Archaeometry* 20(1):19–31.

https://doi.org/10.1017/RDC.2015.10 Published online by Cambridge University Press