JOURNAL OF FORENSIC SCIENCES

J Forensic Sci, 2016
doi: 10.1111/1556-4029.13349
Available online at: onlinelibrary.wiley.com

TECHNICAL NOTE

PHYSICAL ANTHROPOLOGY

Jennifer F. Byrnes,[1] *Ph.D.; Michael W. Kenyhercz,*[2] *Ph.D.; and Gregory E. Berg,*[3] *Ph.D.*

Technical Note: Examining Interobserver Reliability of Metric and Morphoscopic Characteristics of the Mandible*

ABSTRACT: Mandibular metric and morphological characteristics have long been used for sex and ancestry estimation. Currently, there are no large-scale studies examining interobserver agreement, particularly examining the role of observer experience. This study examines the interobserver agreement of six morphoscopic and eleven metric mandibular variables. Four observers with varied levels of experience scored 183 mandibles from the William M. Bass Donated Skeletal Collection. Absolute agreement and consistency were evaluated with the intraclass correlation coefficient (ICC). Additionally, technical error of measurement (TEM) and relative TEM (%TEM) were calculated for each metric variable. All analyses were conducted twice—once with all observers and again with only experienced observers. Results show mean morphoscopic agreement of 0.543 among all observers and 0.615 for experienced observers, and mean metric agreement of 0.886 among all observers and 0.911 for experienced observers. Further, no TEM exceeded 2 mm. All results were significant ($p < 0.001$).

KEYWORDS: forensic science, forensic anthropology, interobserver agreement, mandibular morphology, technical error of measurement, experience bias

Forensic anthropology practitioners have routinely used metric and morphoscopic traits of the mandible as a means of estimating sex and ancestry. Berg (1) provided the forensic anthropology community with statistical tests and their associated accuracy rates for the estimation of sex and ancestry using six morphoscopic and eleven metric characteristics of the mandible from various world populations. Sex and ancestry were shown to have high classification rates using a combination of the morphoscopic and metric variables. However, some populations showed higher classification accuracies than others (e.g., U.S. Hispanics showed lower classification accuracies than U.S. Whites). Berg (1,2) reported high rates of intra-observer agreement among all variables, with metric variables showing greater repeatability than morphoscopic variables. In addition to published discriminant functions (1), his data were also made public for use in programs such as FORDISC (3), for creating custom discriminant functions. Thus, in order for practitioners to confidently employ the discriminant functions, it is necessary to test the agreement in these traits across multiple observers.

Tests of inter- and intra-observer agreement are essential given the push within forensic science for valid and repeatable methods (4,5). Additionally, no studies have truly examined the repeatability of most mandibular morphoscopic traits among observers with differing levels of experience. Therefore, the aim of this study is to evaluate interobserver agreement in mandibular morphoscopic and metric variables among observers with varied levels of experience.

Materials and Methods

Mandibles from 183 modern U.S. Black and White individuals were included in this study (Table 1). All data were collected from the William M. Bass Donated Skeletal Collection housed at the University of Tennessee, Knoxville. All individuals observed were skeletally mature and lacked any abnormal bony remodeling due to trauma or pathology. Edentulous mandibles were excluded from this study because advanced bony resorption obscures observations and hinders accurate measurements (6).

Six morphoscopic and eleven metric variables were recorded for each individual (Table 2). Three observers held PhDs in physical anthropology and the fourth was an advanced undergraduate with some osteological training. All metric data were measured with a mandibulometer and either a Martin type sliding or digital caliper set. The line drawings and definitions in Berg (1) were consulted for accurate scoring of the morphoscopic traits, as well as for caliper positioning for the new metric variables (i.e., TML23 and XDA). The side of the mandible with the greatest expression of a morphoscopic trait was recorded, though in the event of symmetrical expression, the left side was used. No collaboration among study participants occurred during the data collection period.

As a summary measure, the modal values for each morphoscopic trait were calculated along with the frequency of any deviation from the modal value, either positive or negative. No deviation from the modal value was considered 0.

[1]Division of Social Sciences, University of Hawai'i – West O'ahu, 91-1001 Farrington Highway, Kapolei, HI 96707.
[2]Department of Anatomy, University of Pretoria, Private Bag x323, 0007, Arcadia, South Africa.
[3]Defense POW/MIA Accounting Agency, 570 Moffet Street, JBPHH, HI 96853.
*Presented in part at the 68th Annual Meeting of the American Academy of Forensic Sciences, February 22–27, 2016, in Las Vegas, NV.

Received 7 June 2016; and in revised form 30 Aug. 2016; accepted 8 Sept. 2016.

TABLE 1—*Sample demographics by sex and ancestry.*

Sample	U.S. White	U.S. Black	Total
Female	66	2	68
Male	96	19	115
Total	162	21	183

The morphoscopic and metric data collected from each observer were compared with the intraclass correlation coefficient (ICC) (8) in R (9) using the package irr (10). Two different ICC models were utilized: the first examined absolute agreement and the second consistency. Both used a two-way random model with 95% tolerance level. Informally, the ICC with absolute agreement can be stated as:

ICC (absolute agreement) = trait variability/(trait variability + variability in repeated measures + measurement error),

while ICC with consistency is further simplified as:

ICC (consistency) = trait variability/(trait variability + measurement error).

Thus, absolute agreement shows the concordance of trait scores among observers, while consistency does not penalize systematic bias in each observer's scoring practice (11). For example, if one observer always rated one expression of GAF as slight, while another observer classified the same example as medium, an ICC based on consistency will provide a more optimistic correlation, as this is an observation bias inherent to the observer. In sum, absolute agreement values should always be lower than consistency values when evaluating agreement with ICC.

TABLE 2—*Variables used in this study and their definitions.*

Variable (ABR)	Type	Definition
Chin Shape (CS)*	Morphoscopic	The chin shape is viewed superiorly and scored as blunt (smoothly rounded), pointed (the chin comes to a distinct point), square (the chin has a nearly straight front), or bilobate (the chin has a distinct central sulcus).
Lower Border of the Mandible (LBM)*	Morphoscopic	The mandible is scored on a flat surface. If the majority of the lower border of the mandible is flush against the surface, it is scored as straight. If there is a deviation of the border superiorly, typically in the region of the lower second to third molars, it is scored as undulating. If the mandible inclines near the chin (and is somewhat rounded near the gonial angle), and it rocks forward when gentle pressure is applied to the anterior dentition, it is scored as partial rocker. If the mandible is sufficiently rounded on the bottom, such that pressure on the anterior teeth causes it to rock forward and back, it is scored as a rocker.
Ascending Ramus Shape (ARS)*	Morphoscopic	The trait is scored as pinched if the ascending ramus noticeably narrows near its midpoint, or wide if it is a relatively uniform width.
Gonial Angle Flare (GAF)*	Morphoscopic	This trait is scored in five stages. The first stage is inverted, wherein the gonial process slants medially toward the midline of the mandible, absent when the gonial process is in line with the ramus, slight when the gonial process flares outward a short distance (~1–2 mm), medium when the gonial process flares beyond slight to double that distance (~2–4 mm), and everted when the process is greater than ~4 mm.
Mandibular Torus (MT)*	Morphoscopic	The mandibular torus is a bony protuberance of varying size and shape on the lingual surface of the mandible below the alveolar margin, typically near the premolars. This trait is only scored as present or absent.
Posterior Ramus Edge Inversion (PREI)*	Morphoscopic	The trait is observed on the posterior one-third of the ascending ramus. If no discernible flexure toward the midline is present, the mandible is scored as absent. If a small, but discernible flexure toward the midline is present, the trait is scored as slight. A medium expression is a very noticeable inward deviation, up to twice the distance of the slight category. The mandible is scored as turned when it is greater than double the expression of the slight category.
Chin Height (GNI)†	Metric	The direct distance from infradentale to gnathion.
Height of Mandibular Body at the Mental Foramen (HML)†	Metric	The direct distance from the alveolar process to the inferior border of the mandible perpendicular to the base at the mental foramen.
Bigonial Width (GOG)†	Metric	The direct distance between the right and left gonions.
Bicondylar Width (CDL)†	Metric	The direct distance between the most laterally projecting points on the two condyles
Minimum Ramus breadth (WRB)†	Metric	The smallest breadth of the mandibular ramus measured perpendicularly to the height of the ramus
Maximum Ramus Height (XRH)†	Metric	The direct distance from the highest point on the mandibular condyle to gonion
Mandibular Length (MLT)†	Metric	The distance from the anterior margin of the chin to a center point on the projected straight line placed along the posterior border of the two mandibular angles
Mandibular Angle (MAN)†	Metric	The angle formed by the inferior border of the corpus and the posterior border of the ramus
Mandibular Body Breath at the Mental Foramen (TML)*	Metric	The maximum width of the mandibular body taken at the mental foramen. The measurement is typically taken from a superior-to-inferior direction and the caliper arm should be parallel to the flat surface on which the mandible is resting
Mandibular Body Breath at the M2/M3 Junction (TML23)*	Metric	The maximum mediolateral breath of the corpus taken at the level of the articulation between the second and third molars. The sliding caliper arm should be parallel to the surface upon which the mandible rests. The measurement location usually corresponds to a medial–lateral thickening of the mandible at that location
Dental Arcade Width at the Third Molar (XDA)*	Metric	The maximum breadth of the dental arcade at the level of the most posterior points of the third molar crypt on the lingual surface. If necessary, a line should be drawn perpendicular to the ramus body and the tooth crypt to mark the measurement locations. If the third molars are absent, the measurement could be taken at the location of the second molar position, but should be annotated appropriately

*Definitions from Berg (1,2).
†Definitions from Moore-Jansen (7).

Additionally, technical error of measurement (TEM), and relative TEM (%TEM) were calculated in accordance with Ulijaszek and Kerr (12). For more than two observers, TEM is calculated as follows:

$$\mathrm{TEM} = \sqrt{\frac{\left(\sum_1^N \sum_1^K - \left(\sum_1^K M\right)^2 / K\right)}{N(K-1)}}$$

In the above formula, N represents the number of measurements, K is the number of raters or observers, and M is the actual measurement. Given the formula, the TEM will represent discrepancies among measurements in the units originally collected (in this instance, mm). To compare the variability across measurements of different magnitudes, relative TEM can be computed as:

$$\text{Relative TEM} = \left(\frac{\mathrm{TEM}}{\bar{x}^i}\right) \times 100$$

where $\bar{x}^i$ is the sample mean for each individual measurement. Using relative TEM will allow for direct comparison of measurement variability among measures of different magnitude. All analyses were performed twice—once with all observers (pooled) and again with only the PhDs (experienced observers). Lastly, missing data were removed via listwise deletion.

Results

Morphoscopic Trait Agreement

The frequencies of variable mode deviation by pooled and experienced observers are shown in Table 3. In each instance, the experienced observers show higher frequencies of agreement (0 deviation from modal value), and typically have fewer extreme deviations (+2/−2) than the pooled observers. The average agreement is 79.3% for pooled observers and 84.4% for experienced observers.

The ICC values for the pooled and experienced observers are shown in Table 4. Absolute agreement for all observers ranges from 0.375 (PREI) to 0.734 (MT) and absolute agreement for the experienced observers ranges from 0.430 (LBM) to 0.729 (MT). Mean absolute agreement among the pooled observers (0.543) is lower than absolute agreement among the experienced observers (0.615). The same trend holds true for consistency. All morphoscopic ICCs are significant at $p < 0.001$.

Metric Trait Agreement

The ICC values for each of the metric variables for the pooled and experienced observers are listed in Table 5. For pooled observers, absolute agreement values range from 0.710 (TML23) to 0.969 (CDL) and for the experienced observers, it ranges from 0.761 (TML23) to 0.988 (CDL). The mean absolute agreement among pooled observers (0.886) is lower than that among experienced observers (0.911). The ICC values based on consistency show the same pattern as absolute agreement. All metric ICCs are significant at $p < 0.001$.

The TEM and %TEM values for the metric traits are shown in Table 6. The pooled observers' TEM ranges from 0.63 mm (WRB) to 1.84 mm (XDA) and the experienced observers' TEM ranges from 0.35 (WRB) to 0.94 (GOG). The pooled observers' %TEM ranges from 1.00% (CDL) to 10.53% (TML) and the experienced observers' %TEM are between 0.62% (CDL) and 3.86% (TML). None of the measurements exceed an overall TEM of 2 mm. The pooled observers' TEM for MAN is 2.19°, although this might be inflated because it was noted that the

TABLE 3—*Frequency of deviation from modal value (0) from all observers (Pooled) and only PhDs (Experienced Observers).*

Trait	Experience	−2	−1	0	1	2
Ascending Ramus	Pooled	–	9.3	79.2	11.5	–
	Experienced Observers	–	6.0	88.2	5.8	–
Chin Shape	Pooled	5.2	11.0	79.1	1.5	2.5
	Experienced Observers	1.5	11.5	81.3	2.0	3.0
Gonial Flare	Pooled	0.3	19.6	72.7	6.6	0.7
	Experienced Observers	0.6	9.1	78.5	11.1	0.4
Lower Border of Mandible	Pooled	1.1	7.1	84.3	5.9	1.5
	Experienced Observers	0.2	7.6	85.8	5.0	0.9
Mandibular Torus	Pooled	–	3.4	92.7	3.8	–
	Experienced Observers	–	3.5	93.8	2.2	–
Posterior Ramus Edge Inversion	Pooled	4.6	15.1	67.7	10.8	1.5
	Experienced Observers	2.9	7.0	78.7	9.9	1.0
Mean	Pooled	2.8	10.9	79.3	6.7	1.6
	Experienced Observers	1.3	7.5	84.4	6.0	1.3

Deviations range from −2 (2 trait scores below mode) to 2 (2 trait scores above mode). All values are percentages (%).

TABLE 4—*Comparison of absolute agreement and consistency ICC values by experience for each morphoscopic trait.*

Trait	ICC Absolute Agreement (Pooled)	ICC Absolute Agreement (Experienced Observers)	ICC Consistency (Pooled)	ICC Consistency (Experienced Observers)
CS	0.665	0.713	0.685	0.743
LBM	0.391	0.430	0.407	0.452
ARS	0.437	0.693	0.457	0.721
GAF	0.656	0.658	0.675	0.684
MT	0.734	0.729	0.736	0.729
PREI	0.375	0.464	0.405	0.513
Mean	0.543	0.615	0.561	0.640

TABLE 5—*Comparison of absolute agreement and consistency ICC values by experience for each metric trait.*

Measurement	ICC Absolute Agreement (Pooled)	ICC Absolute Agreement (Experienced Observers)	ICC Consistency (Pooled)	ICC Consistency (Experienced Observers)
CDL	0.969	0.988	0.972	0.990
GNI	0.907	0.953	0.910	0.957
GOG	0.960	0.965	0.961	0.967
HML	0.884	0.894	0.891	0.898
MAN	0.871	0.890	0.879	0.895
MLT	0.902	0.933	0.903	0.935
TML	0.805	0.811	0.814	0.818
TML23	0.710	0.761	0.729	0.799
WRB	0.915	0.983	0.916	0.985
XDA	0.915	0.927	0.919	0.934
XRH	0.907	0.919	0.911	0.922
Mean	0.886	0.911	0.891	0.918

TABLE 6—*TEM and %TEM values for each metric measurement ordered from lowest %TEM (Pooled) to greatest %TEM.*

Measurement	TEM (mm) Pooled	TEM (mm) Experienced	%TEM Pooled	%TEM Experienced
CDL	1.16	0.72	1.00	0.62
GOG	1.04	0.94	1.11	0.98
MAN	2.19 (degrees)	0.83 (degrees)	1.74	0.65
WRB	0.63	0.35	1.95	1.18
MLT	1.60	0.70	2.03	0.93
XRH	1.77	0.77	2.89	1.25
HML	1.15	0.65	3.70	2.16
XDA	1.84	0.71	3.91	1.46
GNI	1.47	0.63	4.55	1.94
TML23	1.53	0.40	7.96	2.08
TML	1.62	0.60	10.53	3.86
Mean	1.45	0.66	3.76	1.56

observer with the least experience (advanced undergraduate) misread the mandibulometer in several instances, with a discrepancy of 10°. Comparatively, the experienced observers' TEM for MAN is 0.83°.

Discussion

The deviations from the modal trait scores for each of the morphoscopic traits show that the experienced observers have greater agreement (i.e., no deviation) and less extreme deviations (2 scores above or below the mode) (see Table 3). The mean percentage agreement is 79.3% for all observers and somewhat higher when only experienced observers were examined (84.4%). Overall, the higher agreement for the experienced observers implies that familiarity with the range of normal human variation reduces extreme bias in trait expression scoring. To place these results in context, agreement among observers performed much better for mandibular traits (this study) than those reported by Walker (13) for cranial morphoscopic traits. In Walker's study, average trait agreement (0 difference) was 60.9% as compared to the current study's average agreement of 79.3%.

Experienced observers show greater ICCs for each morphoscopic trait except for MT (pooled = 0.734; experienced = 0.729), in which the difference is essentially negligible. Still, the experienced observers have greater agreement on modal value as compared to agreement calculated through ICC. Interestingly, Lewis and Garvin (14) report ICC for interobserver agreement on Walker's nonmetric cranial traits and found levels of agreement comparable to our results (Lewis and Garvin range 0.06–0.83; current study range for experienced observers 0.43–0.73). However, Lewis and Garvin's (14) interobserver results were much lower than those presented by Walker (13), albeit through different statistical techniques. The discrepancy between these results might be due to the experience level of the observers, or the clarity of each of the respective method's definitions to adequately encapsulate morphoscopic variability. In sum, the results suggest that experience in the range of variation in the expression of morphoscopic traits does affect agreement, although in most instances, this effect is minimal and these traits can be confidently employed in studies of sex and ancestry.

The effect of experience is also seen in the metric data, although all levels of agreement are substantially higher than the morphoscopic traits. The lowest metric agreement is seen in TML23 with pooled observer 0.710, which also happened to have the second highest %TEM (7.96%). Overall, disagreements among the metric variables did not exceed 2 mm. However, as seen with TML and TML23, the effect of even 1 mm can substantially affect %TEM. Both measurements of mandibular thickness incur the greatest %TEMs with the pooled observers (TML23 = 7.96%; TML = 10.53%). With the experienced observers, TML23 and TML also show the greatest %TEM (TML23 = 2.08%; TML = 3.86%), although they are considerably lower than the pooled observers. Again, the discrepancy between pooled and experienced observers' TEMs shows that experience plays an important role in precisely locating anatomical landmarks. While TML23 and TML show the greatest % TEM for both the pooled and experienced observers, these measurements are the smallest in terms of absolute size; thus, even small fluctuations can give rise to large discrepancies among investigators. The range of %TEM observed in the current study is consistent with %TEM reported by Stull and colleagues (15) on cranial and postcranial elements. Additionally, Adams and Byrd found that approximately 3% deviation is to be expected between observers (16), which is consistent with the results of the experienced observers in the present study. Special care should be given when taking these measurements, particularly the measurements with lower magnitudes. Given the strong agreement shown across the metric variables, these too can be confidently used when documenting skeletal features.

A final note on measurement error for the mandibular angle: In this study, one observer (advanced undergraduate) misread the mandibulometer, adding 10 degrees to the actual angle. While it can be tempting to simply denote this as a one-off problem, it occurs quite frequently. One author (GEB) has had numerous questions on how to properly read the angle and has determined that two misreads of the mandibulometer often occur.

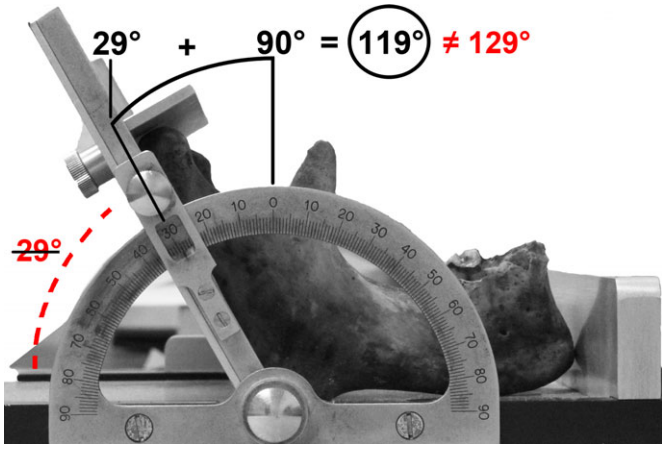

FIG. 1—*Actual reading for the mandibular angle shown above is 119° (encircled). Common mistakes in recording mandibular angle are adding 10° (129°) or reporting the complementary angle (29°, struck-through on bottom left). [Color figure can be viewed at wileyonlinelibrary.com]*

The first is to add an extra 10 degrees to the actual measurement (Fig. 1). The second error is to read the complement of the angle (see Fig. 1). In the first instance, the error is rather insidious, and not easily spotted, while the second instance should be identifiable to the experienced practitioner.

Conclusions

Overall, agreement among observers, both pooled and experienced, are comparable to other published studies of morphoscopic trait agreement and metric %TEM. The morphoscopic variables have lower ICC values than the metric variables, as the majority of the measurements are defined as point-to-point, thus limiting subjectivity. Additionally, nonmetric traits pose the inherent problem of placing the responsibility on the practitioner to decide where the variation cutoff is between scores for a trait, as compared to metric traits, which do not rely on a subjective interpretation. In nearly all of the comparisons, the experienced observers had better overall agreement. Although experience plays a role in agreement among observers, the effect is negligible so long as users familiarize themselves with the trait definitions, illustrations, and instrumentation prior to collecting data. In sum, mandibular metric and morphoscopic traits can be confidently utilized in the estimation of sex and ancestry as proposed by Berg (1,2).

Acknowledgments

The authors would like to thank Dr. Dawnie Wolfe Steadman at the University of Tennessee for access to the William M. Bass Donated Skeletal Collection. Additionally, the authors would like to thank Samantha Torres for assistance in data collection. Lastly, the authors would like to thank the Defense POW/MIA Accounting Agency for loaning extra calipers for the study.

References

1. Berg GE. Biological affinity and sex from the mandible utilizing multiple world populations. In: Berg GE, Ta'ala SC, editors. Biological affinity in forensic identification of human skeletal remains: beyond black and white. Boca Raton, FL: CRC Press, 2014;43–81.
2. Berg GE. Biological affinity and sex determination using morphometric and morphoscopic variables from the human mandible [dissertation]. Knoxville, TN: University of Tennessee, 2008.
3. Jantz RL, Ousley SD. FORDISC 3: Computerized forensic discriminant functions. 3.1 version. Knoxville, TN: University of Tennessee, 2005.
4. Committee on Identifying the Needs of the Forensic Sciences Community, National Research Council. Strengthening forensic science in the United States: a path forward. Washington, DC: The National Academic Press, 2009.
5. Daubert V. Merrell Dow Pharmaceuticals, Inc., 509 U.S. 579, 1993.
6. Parr NML, Passalacqua NV, Skorpinski K. Investigations into age-related changes in the human mandible. Proceedings of the 67th Annual Meeting of the American Academy of Forensic Sciences; 2015 Feb 16–21; Orlando, FL. Colorado Springs, CO: American Academy of Forensic Sciences, 2015.
7. Moore-Jansen PH, Ousley SD, Jantz RL. Data collection procedures for forensic skeletal material. Knoxville, TN: University of Tennessee, 1994.
8. Shrout PE, Fleiss JL. Intraclass correlations: uses in assessing rater reliability. Psychol Bull 1979;86(2):420–8.
9. R Core Team [computer program]. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing, 2014; http://www.R-project.org/ (accessed 24 October 2014).
10. Gamer M, Lemon J, Fellows I, Singh P. Various coefficients of interrater reliability and agreement. R package 0.84 version, 2012; http://CRAN.R-project.org/package=irr (accessed 22 June 2012).
11. Kim H-Y. Statistical notes for clinical researchers: evaluation of measurement error 1: using intraclass correlation coefficients. Restor Dent Endod 2013;38(2):98–102.
12. Ulijaszek SJ, Kerr DA. Anthropometric measurement error and the assessment of nutritional status. Br J Nutr 1999;82(3):165–77.
13. Walker PL. Sexing skulls using discriminant function analysis of visually assessed traits. Am J Phys Anthropol 2008;136(1):39–50.
14. Lewis CJ, Garvin HM. Reliability of the Walker cranial nonmetric method and implications for sex estimation. J Forensic Sci 2016;61(3):743–51.
15. Stull KE, Tise ML, Ali Z, Fowler DR. Accuracy and reliability of measurements obtained from computed tomography 3D volume rendered images. Forensic Sci Int 2014;238:133–40.
16. Adams BJ, Byrd JE. Interobserver variation of selected postcranial skeletal measurements. J Forensic Sci 2002;47:1193–202.

Additional information and reprint requests:
Jennifer F. Byrnes, Ph.D.
University of Hawai'i – West O'ahu
91-1001 Farrington Highway
Kapolei, HI 96707
E-mail: jfbyrnes@hawaii.edu