

Interobserver Error in a Large Scale Anthropometric Survey

CLAIRE C. GORDON¹ AND BRUCE BRADTMILLER²

¹*Anthropology Branch, Science & Advanced Technology Directorate, U.S. Army Natick Research, Development, and Engineering Center, Natick, Massachusetts 01760-5020;* ²*Anthropology Research Project, Inc., Yellow Springs, Ohio 45387*

ABSTRACT The adverse effects of interobserver error on morphometric population comparisons are well documented in the literature. While interobserver error can rarely be avoided, it can be minimized by having a single individual locate and mark relevant landmarks, by limiting the number of observers for each variable, and by reviewing repeated measures data daily to catch and correct measurer drift during data collection. In this study, two pairs of experts participated in interobserver error trials designed to pre-set observer error limits for use in the quality control of a large scale anthropometric survey. Repeatability data were also collected twice daily in the field and reviewed with the measurers. Interobserver errors obtained in the field were lower than those achieved by the experts for 27 of 30 dimensions. These results suggest that establishment of permissible interobserver error in advance of data collection and frequent review of repeated measurements during data collection can reduce the magnitude of interobserver error below that obtained by experts measuring in a laboratory setting. However, even differences of small magnitude can be serious when they are directional, and 17 of 30 dimensions exhibited statistically significant bias between measurers despite all quality control efforts. The magnitudes of interobserver error observed in this study have proven particularly useful in evaluating the biological relevance of statistically significant differences which are of relatively small magnitude.

Measurement error in anthropometry arises from numerous sources. Instrument precision and accuracy are probably the easiest sources of error to quantify and minimize. Errors in instrument assembly, instrument reading, and data recording are common, but can be minimized in a straightforward manner by utilizing computerized data entry/editing at the measuring site (Churchill et al., 1988; Healy, 1989). Inconsistent execution of the measuring protocol (commonly referred to as "observer error") is undoubtedly the most troublesome source of anthropometric error since it includes imprecision in landmark location, subject positioning, and instrument application which may be accentuated by the use of multiple observers, even when observers are trained by the same individual (Bennett and Osborne, 1986). Intraindividual fluctuations, such as diurnal variation in stature, also pose a potential source of error if their effects are not considered in the measuring protocol.

Although reliability considerations are often overlooked in problem-oriented research, the impact of measurement error on hypothesis testing can be serious, particularly when conclusions rest on univariate and multivariate statistical tests among groups. For example, when some portion of the observed variance in a body dimension is due to error, univariate tests between groups will be too conservative, i.e., the researcher loses power in detecting differences between groups (Bailey and Byrnes, 1990). Simple regression and correlation coefficients between body dimensions measured with substantial error will be biased towards zero (Healy, 1989), and multiple regression and partial correlation coefficients can be either attenuated or enhanced depending upon the

Received May 1, 1991; accepted August 31, 1991.

Send correspondence to: Dr. Claire C. Gordon, U.S. Army Natick R. D. & E. Center, ATTN: SFRNC-YBA, Natick, MA 01760-5020.

error structure present in the independent variables considered (Liu, 1988). Because covariances are also affected, observer errors are also known to compromise the interpretation of results from multivariate techniques, such as principal components and discriminant analysis (Bailey and Byrnes, 1990).

When observer error is not random, as might occur if there are consistent differences in the techniques utilized by multiple measurers, even errors of small magnitude can have devastating results. Francis and Mattlin (1986) report radical and unequal alterations in the misclassification rates of discriminant functions with biases as small as 0.2 mm. When more substantial biases between observers exist (1–7 mm), one can even get significant discrimination among measurers based upon the same subject sample (Jamison and Zegura, 1974). What is more alarming is that reliability research has clearly demonstrated that such biases between observers are not unusual, even when the observers are trained by a single professional and work closely together (Jamison and Zegura, 1974; Utermohle and Zegura, 1982; Utermohle et al., 1983; Bennett and Osborne, 1986).

Despite the serious impact that measurement errors can have on hypothesis testing, reliability studies do not appear to be standard operating procedures in biometric research (Cameron, 1984; Bailey and Byrnes, 1990). This is puzzling when one considers the many useful applications that reliability studies can serve. Test-retest data can be used in the early planning stages of a study to select the most reliable variables among a set of candidate dimensions (Utermohle et al., 1983; Clauser et al., 1986; Himes, 1989; Marks et al., 1989). Should a critical dimension prove to have low reliability, test-retest data can also be used to estimate the number of replicates needed in the protocol in order to obtain reliability of a given magnitude (Himes, 1989; Bailey and Byrnes, 1990).

Once measurement protocols are established, test-retest data can be used to guide the frequency of remeasurement during longitudinal studies (Cameron, 1984). While reliability assessment is still most often an a priori and/or post hoc event, an increasing number of methodologists are emphasizing its importance throughout the course of data collection (Cameron, 1984; Mueller and Martorell, 1988; Healy, 1989). Typically, permis-

sible error limits are established in advance of data collection using the finalized protocol, and measurer performance is monitored periodically during data collection against these a priori standards (Malina et al., 1973; Cameron, 1984; Johnston and Martorell, 1988; Gordon et al., 1989; Himes, 1989). This approach permits one to detect and correct observer drift before it adversely impacts the data.

Reliability studies conducted during data collection are also useful in data analysis. Knowledge of the error structure in a data set permits one to correct bias in regression coefficients (Healy, 1989), to transform data prior to discriminant function estimation (Jamison and Zegura, 1974; Francis and Mattlin, 1986), and/or to estimate true misclassification rates in discriminant functions (Liu, 1988). Furthermore, knowledge of the actual error structure of the data permits more informed decisions regarding the magnitudes of truly detectable differences (Cameron, 1984; Greiner and Gordon, 1990), and thus aids in the biological interpretation of statistically significant results (Utermohle and Zegura, 1982).

While interobserver error cannot be eliminated from most research designs, it can be minimized. This paper reports interobserver error for 30 International Biological Program dimensions (Table 1) that were part of a large scale anthropometric survey which deliberately incorporated a number of quality control measures. Three hypotheses are tested: 1) incorporation of computerized data entry and editing routines on site can reduce the frequency of gross errors in a large scale

TABLE 1. Dimensions reported by instrument class

Anthropometric tape	Beam caliper
Ankle circumference	Acromion-radiale length
Biceps circumference,	Biacromial breadth
flexed	Bideloid breadth
Buttock circumference	Chest breadth
Calf circumference	Chest depth
Chest circumference	Hip breadth
Head circumference	Radiale-styilion length
Neck circumference	Waist breadth
Waist circumference	Holtain caliper
Anthropometer	Bimalleolar breadth
Cervicale height	Ear breadth
Sitting height	Ear length
Stature	Heel breadth
Suprasternale height	Sliding caliper
Spreading caliper	Hand breadth
Bizygomatic breadth	Hand length
Head breadth	Menton-sellion length
Head length	

survey; 2) proper training and ongoing reliability checks can reduce the gender bias in error magnitudes reported by some researchers (Bennett and Osborne, 1986); and 3) a thoughtfully developed protocol, proper training, and ongoing reliability assessments can reduce the levels of error expected in a field situation.

This study is strictly cross-sectional. Test-retest consistency over a short period of time is of primary concern, and measurement error due to physiological variation over time is not addressed, although it is critical to the design and interpretation of longitudinal studies. Using the terminology of Mueller and Martorell (1988), only one component of overall measurement reliability is thus discussed: precision.

MATERIALS AND METHODS

Sample

The repeatability data are from the 1988 U.S. Army anthropometric survey, which included 5,692 men and 3,599 women selected randomly within age and race/ethnicity sampling strata (Gordon et al., 1989). Each subject was measured for 132 different body dimensions that had been selected for their value to research and engineering problems and defined to enhance their replicability (Clouser et al., 1986, 1988). Subjects were measured semi-nude. They visited each of eight different landmark/measurement stations and completed the full survey in approximately 90 minutes. Fifty subjects were measured daily.

Measurement protocol

The measuring team was recruited specifically for this study, had no prior anthropometric experience, and underwent a month of full-time training. Early in the training process, individuals were assigned in pairs to either the marking station, or one of seven measuring stations. At the marking station, one individual drew marks above the waist, and his/her partner drew those below the waist. At the measuring stations, both partners learned only the dimensions at their station; one measured while the other recorded, and they switched at will to alleviate fatigue and boredom. Team assignments were permanent. When a measurer became ill during the survey, his/her partner did all the measuring, and a substitute recorder was provided. Thus only one person contributed to observer error for any landmark, and

only two people contributed to observer error for any dimension.

Each measuring station was provided with a portable computer and two data editing programs (Churchill et al., 1988). The first program checked each entered value to see if it exceeded the minimum or maximum recorded for that variable to date. If so, an audible alarm sounded, and the program prompted the recorder to request remeasurement. The second program utilized multiple regression techniques to predict the value expected based on those of other variables already entered for that subject. If the entered value differed from that of the regression estimate by more than 3 standard errors, again an audible alarm sounded and the program prompted the recorder to request a remeasurement. This software permitted on site correction of many "blunders" (Healy, 1989), such as incorrectly assembled equipment, digit transpositions, or dimensions measured out of order.

The survey protocol also incorporated pre-set observer error limits that were compared to repeatability data collected twice daily at each measuring station. Repeatability data were summarized weekly in the field, and special practice sessions were held whenever errors exceeded the pre-set limits. This paper reports data from the expert error trials used to derive these pre-set error limits, and from the daily error checks that were conducted throughout the field survey at each measuring station.

Setting observer error limits

Four experts, each with 15 or more years experience in anthropometric data collection, participated in two repeatability trials of 10 subjects each (5 males and 5 females). Whereas Pair 1 had measured together often over several decades, Pair 2 had never measured together before this study. The trials were preceded by a week of practice sessions utilizing the finalized protocol for the survey.

The mean absolute difference (MAD) was chosen as the preferred statistic for establishing measurer standards because it is known to be poorly correlated with dimensional magnitude (Utermohle et al., 1983), and because its own magnitude is easily interpreted as a standard against which measurer performance can be tested on a daily basis. The highest MAD of the four expert estimates (two expert pairs \times two trials) was chosen as the maximum permissible

interobserver error for each dimension (see Gordon et al., 1989 for details). For a few very small dimensions (e.g., hand breadth), the expert MAD's suggested maximum permissible error limits of only 1 mm. Since instrument precision in this study was 1 mm, these error limits were deemed too restrictive, and were automatically elevated to 2 mm.

Intraobserver error was also recorded in the expert error trials. However, since intraobserver errors for these experts were not very different from interobserver errors, and because intraobserver error is thought to decline with experience whereas interobserver error does not (Jamison and Zegura, 1974; Utermohle et al., 1983; Nichol and Turner, 1986), only interobserver error was chosen for quality control.

Quantifying observer error

A wide variety of statistics are available for quantifying observer error, with little consistency in the literature as to which are reported (Utermohle et al., 1983). For reasons described above, the MAD was chosen for establishing measurer standards in this study. Others have suggested that ratios of error variances such as the technical error of measurement (TEM) be used to monitor measurer performance (Cameron, 1984; Mueller and Martorell, 1988; Healy, 1989). However, power calculations indicate that variance estimates must be based upon relatively large samples (40+) before ratio tests detect even two-fold differences in error variance with 90% certainty (Healy, 1989). Thus in practice, use of a variance ratio approach to quality control during data collection may lead to unacceptable trade-offs between the number of subjects measured daily, the frequency of measurer monitoring/feedback, and the power to detect departures from acceptable error levels.

Power considerations need not compromise the usefulness of TEMs in determining when to end measurer training, nor in a post hoc description of observer error. In fact, some researchers consider the TEM to be one of only two primary statistics needed to describe reliability in a data base (Mueller and Martorell, 1988). Furthermore, as noted by Utermohle et al. (1983), although the TEM is highly correlated with the MAD and thus potentially redundant, its widespread use alone justifies its inclusion in reliability

studies for comparative purposes. Thus TEMs are also calculated and reported in this paper.

Although the MAD and TEM both describe observer error magnitude, neither indicates what proportion of measurement variance is error free. This is a particularly important issue for morphometric studies since a dimension with relatively high within-individual variability compared to between-individual variability is not taxonomically useful (Bailey and Byrnes, 1990). The reliability coefficient (R), which can be conveniently computed using a random effects analysis of variance in which measurer effects are nested within subject effects, provides a measure of the proportion of variance which is error free (Cameron, 1984; Chumlea et al., 1985; Mueller and Martorell, 1988). Furthermore, because it is dimensionless, R permits clean comparisons between variables of different magnitudes. Thus Rs are also reported in this study.

Whereas the TEM and R statistics are thought by many to be all that is needed to describe measurement reliability, neither addresses the question of potential bias in the measurements of an observer. To assess bias, a two-way analysis of variance without replication is commonly used (Utermohle et al., 1983; Bennett and Osborne, 1986; Mueller and Martorell, 1988). In this test, observer and subject effects are partitioned, and the observer mean square is tested over the error mean square. When only two observers are involved, as in this study, this test reduces to a paired comparisons t-test, and these are reported for both experts and field measurers.

In this study, the general effects of quality control measures on 30 anthropometric dimensions are studied using multiple significance tests for each hypothesis. In order to avoid elevating false positive error rates and thus possibly drawing conclusions based upon spurious results, individual significance tests under each hypothesis are adjusted using a Bonferroni inequality (Koopmans, 1987). The Bonferroni adjustment is trivial to calculate; one simply divides the chosen individual significance level (usually .05) by the number of related tests to be performed. In this case, in order to ensure an experimentwise Type I error rate of .05, significance testing for each dimension is conducted at the $.05/30 = .0017$ level.

RESULTS

Post hoc editing that utilized both range and regression methods to detect gross measuring errors (such as equipment misassembly and digit transpositions) identified 3 of 1,200 (.25%) expert values as bad and 163 of 1,187,604 (.01%) field values as bad. Had gross errors been present in the survey data base in the same proportions as the expert data base, we might have expected 2,969 bad values instead of only 163. Clearly, use of computerized data entry and editing programs in the field resulted in a substantial reduction in the number of gross errors present in the data base.

Gender differences in observer error magnitudes are reported in Table 2. Male sample sizes are larger than female sample sizes because more males were processed in the

Army's survey than females. Sample sizes also vary slightly from dimension to dimension due to subject processing adjustments made early in the survey to improve efficiency, and due to measurer illnesses during the survey. Wilcoxon rank sum tests (Rosner, 1990) are used to test for error differences between males and females because, whereas the dimensions themselves and possibly their measuring errors (Utermohle et al., 1983) are normally distributed, their absolute differences are not.

Only 7 of the 30 Wilcoxon rank sum tests had outcomes with individual probabilities smaller than .05 (Table 2). In 4 of the cases, the MAD for female subjects was larger than for males; in 3 of the cases, MADs were larger for males. Only 3 of the 30 tests are statistically significant after Bonferroni correction. The MAD is greater for males in 2 of

TABLE 2. Field team MADs by subject gender

	Male MAD (n)	Female MAD (n)	P ¹
Anthropometric tape			
Ankle circumference	1.93 (201)	1.98 (137)	.209
Biceps circumference, flexed	2.94 (231)	3.25 (154)	.350
Buttock circumference	4.73 (240)	5.18 (159)	.532
Calf circumference	2.02 (201)	2.04 (137)	.201
Chest circumference	7.38 (238)	6.64 (152)	.740
Head circumference	1.18 (233)	1.48 (156)	.439
Neck circumference	3.70 (238)	3.05 (152)	.109
Waist circumference	4.87 (238)	7.00 (152)	.006
Anthropometer			
Cervicale height	2.61 (231)	2.69 (154)	.772
Sitting height	3.66 (235)	3.37 (159)	.239
Stature	3.30 (231)	3.19 (154)	.530
Suprasternale height	3.24 (231)	3.22 (154)	.625
Spreading caliper			
Bizygomatic breadth	.88 (233)	.97 (156)	.400
Head breadth	.80 (233)	.85 (156)	.685
Head length	.87 (233)	.90 (156)	.824
Beam caliper			
Acromion-radiale length	1.83 (231)	2.21 (154)	.014
Biacromial breadth	4.21 (235)	4.21 (159)	.260
Bideltoid breadth	4.40 (235)	3.56 (159)	.169
Chest breadth	3.88 (231)	3.95 (154)	.900
Chest depth	3.32 (231)	3.53 (154)	.165
Hip breadth	2.56 (231)	3.11 (154)	.037
Radiale-stylion length	3.25 (231)	2.70 (154)	.054
Waist breadth	2.48 (231)	3.58 (154)	.000*
Holtain caliper			
Bimalleolar breadth	.99 (240)	.58 (159)	.000*
Ear breadth	1.00 (233)	1.16 (156)	.124
Ear length	.86 (233)	.86 (156)	.955
Heel breadth	1.52 (240)	.76 (159)	.000*
Sliding caliper			
Hand breadth	.71 (233)	.75 (156)	.497
Hand length	1.67 (233)	1.40 (156)	.047
Menton-sellion length	1.43 (233)	1.38 (156)	.974

¹Individual Wilcoxon rank sum test probability.

*Significant at the .05 level or better after Bonferroni correction for 30 tests.

the 3: bimalleolar breadth and heel breadth. However, the magnitudes of all MADs are so close to the precision limits of the Holtain caliper itself (1 mm) that these findings may not be very interesting, even if they are statistically significant. Waist breadth, the only other dimension with a statistically significant result, provides the only substantive evidence for a gender bias in observer error. In this case, both the female MAD and median absolute difference (3.58, 3.00) are 1 mm larger than those for males (2.48, 2.00). To increase statistical power of other tests, observer error data for male and female subjects are analyzed together.

Table 3 presents expert measurer MADs, TEMs, and Rs. Expert reliabilities for this protocol were well above the 90–95% minimum commonly used as an evaluation guideline for dimension selection (Himes, 1989; Marks et al., 1989). The reliabilities of four dimensions studied, however, fell below the

90% minimum: hand length, hand breadth, ear breadth, and waist breadth. Reliabilities for the expert pair that had measured together over many years (figures in parentheses, Table 3) were much higher than for the expert pair that had measured together for only several weeks. In fact, the reliabilities for expert Pair 1 exceeded the 90% minimum for all dimensions except ear breadth.

Bias within the two expert pairs was measured using paired comparison t-tests. Of the 60 tests, 19 had individual *P*-values of .05 or better (see Table 4), but only 4 of these exhibited statistically significant observer effects after a Bonferroni correction for multiple comparisons.

Table 5 presents field measurer MADs, TEMs, and Rs. Reliabilities for these "novice" measurers performing under field conditions are exceptionally high, with the sole exception of ear breadth, a troublesome dimension for the experts also. Not only are

TABLE 3. Expert error data

	n	MAD (mm)	TEM (mm)	R (%) ¹
Anthropometric tape				
Ankle circumference	40	1.98	1.87	96.6
Biceps circumference, flexed	40	4.25	4.40	97.2
Buttock circumference	40	8.25	8.14	99.2
Calf circumference	40	2.48	2.21	98.6
Chest circumference	40	11.80	10.57	98.6
Head circumference	40	3.92	3.68	95.3
Neck circumference	40	5.20	4.78	98.0
Waist circumference	40	10.70	10.27	99.4
Anthropometer				
Cervicale height	40	4.80	5.17	99.5
Sitting height	39	5.00	4.66	98.8
Stature	40	4.15	3.90	99.5
Suprasternale height	40	3.95	3.36	99.6
Spreading caliper				
Bizygomatic breadth	40	.92	.94	98.7
Head breadth	40	.78	.92	98.1
Head length	40	1.22	1.19	97.3
Beam caliper				
Acromion-radiale length	40	2.88	2.71	99.7
Biacromial breadth	39	6.46	6.11	97.4
Bideltoid breadth	39	7.15	6.64	97.4
Chest breadth	40	5.75	5.30	99.1
Chest depth	40	3.40	3.21	98.2
Hip breadth	40	4.58	4.44	99.1
Radiale-styilion length	40	3.88	3.75	97.0
Waist breadth	40	3.80	3.51	83.4 (92.3)
Holtain caliper				
Bimalleolar breadth	40	1.02	.92	98.5
Ear breadth	40	1.78	1.68	75.7 (88.7)
Ear length	40	1.25	1.31	94.4
Heel breadth	40	1.40	1.20	97.6
Sliding caliper				
Hand breadth	40	.95	.91	75.9 (93.0)
Hand length	40	2.08	2.19	89.8 (92.6)
Menton-sellion length	40	2.12	2.07	93.0

¹Rs in parentheses are for Expert Pair 1 only.

TABLE 4. Directionality in expert errors (dimensions with P-values of .05 or better)

Dimension circumference	D ¹	(se) (mm)	t	P	Pair
Biceps circumference, flexed	3.30	(1.03)	3.20	.0047	1
Biceps circumference, flexed	3.30	(.62)	5.36	.0001*	2
Head circumference	2.50	(.82)	3.04	.0067	1
Head circumference	4.05	(.99)	4.09	.0006*	2
Neck circumference	-4.90	(1.28)	-3.82	.0011	1
Neck circumference	3.90	(1.06)	3.69	.0015	2
Sitting height	-4.60	(1.20)	-3.83	.0011	1
Cervicale height	2.70	(1.28)	2.11	.0480	2
Bizygomatic breadth	.70	(.23)	3.04	.0068	2
Head length	-1.00	(.36)	-2.81	.0111	1
Biacromial breadth	3.68	(1.59)	2.32	.0326	2
Chest breadth	-2.95	(.72)	-4.14	.0006*	2
Hip breadth	-2.80	(.66)	-4.25	.0004*	1
Waist breadth	10.05	(2.96)	3.40	.0030	2
Bimalleolar breadth	-2.00	(.70)	-2.87	.0097	1
Bimalleolar breadth	-1.90	(.59)	-3.23	.0044	2
Ear breadth	1.90	(.51)	3.71	.0015	2
Heel breadth	-.90	(.30)	-3.02	.0071	2
Hand length	-1.50	(.60)	-2.52	.0221	2

¹Mean difference (standard error); n = 20 for all dimensions.

*Significant at the .05 level after a Bonferroni adjustment for 60 tests.

TABLE 5. Field team error data

	n	MAD (mm)	TEM (mm)	R (%)
Anthropometric tape				
Ankle circumference	338	1.95	2.01	98.1
Biceps circumference, flexed	385	3.06	2.71	99.5
Buttock circumference	399	4.91	5.88	99.1
Calf circumference	338	2.03	2.02	99.4
Chest circumference	390	7.09	6.42	99.3
Head circumference	389	1.30	1.31	99.5
Neck circumference	390	3.45	3.43	99.1
Waist circumference	390	5.70	6.19	99.5
Anthropometer				
Cervicale height	385	2.64	2.38	99.9
Sitting height	394	3.54	3.34	99.5
Stature	385	3.26	2.94	99.9
Suprasternale height	385	3.23	2.92	99.9
Spreading caliper				
Bizygomatic breadth	389	.92	1.07	97.8
Head breadth	389	.82	.86	98.3
Head length	389	.88	.85	98.9
Beam caliper				
Acromion-radiale length	385	1.98	1.73	99.4
Biacromial breadth	394	4.21	4.16	97.0
Bideltoid breadth	394	4.06	3.77	99.1
Chest breadth	385	3.91	3.58	98.7
Chest depth	385	3.40	3.18	98.0
Hip breadth	385	2.78	2.78	98.6
Radiale-styilion length	385	3.03	2.71	98.3
Waist breadth	385	2.92	2.88	99.2
Holtain caliper				
Bimalleolar breadth	399	.83	.83	97.9
Ear breadth	389	1.06	1.00	86.4
Ear length	389	.86	.82	96.9
Heel breadth	399	1.22	1.22	96.0
Sliding caliper				
Hand breadth	389	.72	.70	98.9
Hand length	389	1.56	1.39	98.6
Menton-sellion length	389	1.41	1.29	97.2

field reliabilities very high, field MADs are lower than those of the experts measuring in laboratory conditions for 27 of the 30 dimensions, and equal in 2 other dimensions (Table 6). Wilcoxon rank sum tests between absolute differences observed for expert and field measurers are presented in Table 6. Of the 30 dimensions tested, 13 had individual probabilities of .05 or less, and 4 of these had experimentwise probabilities of .05 or less after a Bonferroni correction. The sole instance in which expert absolute differences were lower than those for the field measurers (head breadth) had an individual probability of .432.

Bias in field measurer errors was examined using paired comparison t-tests (Table 7). Sample sizes for some dimensions are lower than previously reported because one of the two measurers at station 4 was replaced mid-survey. Station 4 data reported

are for the second half of the survey because it had the larger sample size. Despite the fact that measurer errors were quite small as reflected in Rs and MADs, there are a striking number of significant t-tests. After Bonferroni correction, 17 of the 30 dimensions studied (57%) exhibit significant measurer bias, and these biases range in magnitude between .25 mm (head circumference) and 2.12 mm (biacromial breadth). Compared to the experts (with only $4/60 = 7\%$ bias), the field measurers were more often biased, despite the small magnitudes of observer error achieved.

DISCUSSION

Computerized data entry and editing

Post hoc data editing of the field measurer data base indicated that wild values were present at much lower frequencies than the 1.7–4% reported by Healy (1989). These re-

TABLE 6. *Field team versus expert absolute differences*

	Expert MAD (n)	Field MAD (n)	P ¹
Anthropometric tape			
Ankle circumference	1.98 (40)	1.95 (338)	.858
Biceps circumference, flexed	4.25 (40)	3.06 (385)	.445
Buttock circumference	8.25 (40)	4.91 (399)	.000*
Calf circumference	2.48 (40)	2.03 (338)	.168
Chest circumference	11.80 (40)	7.09 (390)	.003
Head circumference	3.92 (40)	1.30 (389)	.000*
Neck circumference	5.20 (40)	3.45 (390)	.017
Waist circumference	10.70 (40)	5.70 (390)	.001*
Anthropometer			
Cervicale height	4.80 (40)	2.64 (385)	.020
Sitting height	5.00 (39)	3.54 (394)	.034
Stature	4.15 (40)	3.26 (385)	.136
Suprasternale height	3.95 (40)	3.23 (385)	.085
Spreading caliper			
Bizygomatic breadth	.92 (40)	.92 (389)	.828
Head breadth	.78 (40)	.82 (389)	.432
Head length	1.22 (40)	.88 (389)	.128
Beam caliper			
Acromion-radiale length	2.88 (40)	2.21 (385)	.147
Biacromial breadth	6.46 (39)	4.21 (394)	.009
Bideloid breadth	7.15 (39)	3.56 (394)	.002
Chest breadth	5.75 (40)	3.95 (385)	.014
Chest depth	3.40 (40)	3.53 (385)	.768
Hip breadth	4.58 (40)	3.11 (385)	.001*
Radiale-styilion length	3.88 (40)	2.70 (385)	.401
Waist breadth	3.80 (40)	3.58 (385)	.076
Holtain caliper			
Bimalleolar breadth	1.02 (40)	.83 (399)	.090
Ear breadth	1.78 (40)	1.06 (389)	.005
Ear length	1.25 (40)	.86 (389)	.285
Heel breadth	1.40 (40)	1.22 (399)	.114
Sliding caliper			
Hand breadth	.95 (40)	.72 (389)	.142
Hand length	2.08 (40)	1.56 (389)	.501
Menton-sellion length	2.12 (40)	1.41 (389)	.045

¹Individual Wilcoxon rank sum test probability.

*Significant at the .05 level or better after Bonferroni correction for 30 tests.

TABLE 7. Directionality in field measurer errors

	n	D ¹	(se)	t	P
Anthropometric tape					
Ankle circumference	338	-0.70	(.150)	-4.65	.000*
Biceps circumference, flexed	385	-1.94	(.169)	-11.53	.000*
Buttock circumference	240	1.43	(.432)	3.31	.001*
Calf circumference	338	.20	(.155)	1.32	.189
Chest circumference	390	-.58	(.460)	-1.26	.208
Head circumference	227	-.25	(.104)	-2.37	.019
Neck circumference	390	-1.69	(.230)	-7.36	.000*
Waist circumference	390	1.02	(.441)	2.32	.021
Anthropometer					
Cervicale height	385	-.63	(.169)	-3.72	.000*
Sitting height	394	-.45	(.237)	-1.91	.056
Stature	385	-1.94	(.188)	-10.36	.000*
Suprasternale height	385	-.43	(.209)	-2.05	.041
Spreading caliper					
Bizygomatic breadth	227	.38	(.095)	4.00	.000*
Head breadth	227	.43	(.077)	5.54	.000*
Head length	227	-.08	(.081)	-1.04	.300
Beam caliper					
Acromion-radiale length	385	-1.27	(.106)	-11.94	.000*
Biacromial breadth	394	2.12	(.277)	7.67	.000*
Bideltoid breadth	394	-1.12	(.263)	-4.24	.000*
Chest breadth	385	-.35	(.258)	-1.36	.174
Chest depth	385	-.10	(.229)	-.42	.675
Hip breadth	385	-1.91	(.175)	-10.92	.000*
Radiale-styilion length	385	-1.41	(.182)	-7.73	.000*
Waist breadth	385	-1.82	(.186)	-9.78	.000*
Holtain caliper					
Bimalleolar breadth	240	-.55	(.080)	-6.84	.000*
Ear breadth	227	.05	(.089)	.55	.586
Ear length	227	-.25	(.068)	-3.70	.000*
Heel breadth	240	+0.63	(.126)	+5.03	.000*
Sliding caliper					
Hand breadth	227	-.14	(.061)	-2.38	.018
Hand length	227	-.36	(.128)	-2.78	.006
Menton-sellion length	227	-.22	(.119)	-1.86	.064

¹Mean difference (standard error).

*Significant at the .05 level after a Bonferroni adjustment for 30 tests.

sults indicate clearly that inclusion of on-site data entry and editing systems can reduce gross errors due to instrument misassembly, instrument misreading, omission of a measurement, and digit transposition to a minimum: in this case, approximately .01%. This results in a much shorter post-survey data editing period and avoids the difficult decisions (retain, delete, substitute a regressed value?) that must be made when a wild value is identified post-survey and the subject cannot be remeasured.

Gender differences in error

Whereas some investigators have reported that observer errors are higher for female subjects, this was clearly not the case in this study. It is possible that gender differences in reliability introduced by more extensive fat deposits on females are not so great in military subjects. It is also very

likely that extensive landmarking, measurer training, and daily test-retest aided in standardizing measurement locations and the degree of pressure applied in using the instruments.

Measurement reliability

A glance at Tables 3 and 5 immediately suggests the value of looking at both error magnitudes (MAD or TEM) and error-free proportions of variance (R) since observer differences of relatively large magnitude may actually represent relatively small proportions of total measurement variance (see waist circumference) and, on the other hand, errors of very small magnitude may have relatively large impacts on measurement variance (see hand breadth and ear breadth).

In this study, measurement errors in the field tended to be smaller in magnitude than

those of laboratory experts despite the fact that none of the measurers had had prior anthropometric experience, and despite the fact that measurers processed 50 subjects daily under field conditions. These data suggest that the magnitudes of interobserver error in field studies can be reduced considerably if pre-set limits are established and test-retest data are collected and reviewed regularly during the course of data collection. Splitting the full 132 dimension protocol into 7 measuring stations, so that a measurer need only learn 20 or so dimensions, undoubtedly also contributed to the measurers' success.

The fact that the experts' laboratory reliability was lower, and MADs were higher, than those obtained from measurers in the field also suggests that frequent measuring, with proper training and test-retest feedback, may be more important in reducing interobserver error than long experience in anthropometric techniques. Since, in practice, many studies must be undertaken without the luxury of a month's full-time measurer training, are often subject to interruptions, and sometimes of necessity involve multiple field teams, the importance of continuous reliability checks during data collection cannot be overemphasized.

Although the approaches taken in this survey resulted in very small magnitudes of interobserver error, they did not eliminate the directional biases that occur as each individual develops his/her own measuring style. Indeed, the small magnitudes of some of these biases suggest that it may not be possible to identify and eliminate such subtle stylistic differences in the use of traditional anthropometric equipment. Almost one-half (7/17) of the significantly biased dimensions have mean differences smaller than instrument precision levels (1 mm) and all except one have mean differences smaller than 2 mm. In any case, these results fully support conclusions drawn by other researchers that observer errors in anthropometry are not random, and that bias is not an unusual phenomenon (Jamison and Zegura, 1974; Bennett and Osborne, 1986).

With this in mind, it seems more important than ever that reliability studies become an integral part of anthropometric protocols. Unlike a priori or post hoc reliability trials that only estimate the error present in a data base, the reliability data in this study are based on the actual measurers and subjects in the data base, and are available

for every day in which the data were collected. This permits one to quantify rather than estimate the levels of error present in the data base, and to transform the data and/or choose analyses that minimize the consequences that interobserver errors have in morphometric research.

The magnitudes of observer error estimated from these data have proven extremely useful in the practical interpretation of statistically significant differences based on very large sample sizes. While this is an important consideration in all biometric studies, it is particularly important to the Army, because sample sizes are rarely less than 1,000 individuals for any estimates of population parameters, and because these estimates must be translated into engineering specifications with appropriate manufacturing tolerances.

ACKNOWLEDGMENTS

Thanks are especially due the four expert anthropometrists who participated in this study: Wm. Cameron Chumlea, Charles E. Clauser, Kenneth W. Kennedy, and John T. McConville. Successful execution of the 1988 Army survey is largely owed to the dedication of Stanley Holgate, Beth Ann Holloway, Robert Walker, Jeryl Neff, and their liaison and measuring teams.

LITERATURE CITED

- Bailey RC, Byrnes J (1990) A new, old method for assessing measurement error in both univariate and multivariate morphometric studies. *Syst. Zool.* 39:124-130.
- Bennett KA, Osborne RH (1986) Interobserver measurement reliability in anthropometry. *Hum. Biol.* 58:751-759.
- Cameron N (1984) *The Measurement of Human Growth*. London: Croom Helm.
- Chumlea WC, Roche AF, Mukherjee D, Steinbaugh ML (1985) Errors of measurement for methods of recumbent nutritional anthropometry in the elderly. *J. Nutr. Eld.* 5:3-11.
- Churchill TD, Bradtmiller B, Gordon CC (1988) Computer Software Used in the U.S. Army Anthropometric Survey 1987-1988. Technical Report NATICK/TR-88/045, U.S. Army Natick Research, Development, and Engineering Center, Natick, MA.
- Clauser CE, McConville JT, Gordon CC, Tebbetts IO (1986) Selection of Dimensions for an Anthropometric Data Base Volume I: Rationale, Summary, and Conclusions. Technical Report NATICK/TR-86/053, U.S. Army Natick Research, Development, and Engineering Center, Natick, MA.
- Clauser CE, Tebbetts IO, Bradtmiller B, McConville JT, Gordon, CC (1988) *Measurer's Handbook: U.S. Army Anthropometric Survey 1987-1988*. Technical Report NATICK/TR-88/043, U.S. Army Natick Research, Development, and Engineering Center, Natick, MA.
- Francis RICC, Mattlin RH (1986) A possible pitfall in the

- morphometric application of discriminant analysis: measurement bias. *Marine Biol.* 93:311-313.
- Gordon CC, Bradtmiller B, Churchill T, Clauser CE, McConville JT, Tebbetts IO, Walker RA (1989) 1988 Anthropometric Survey of U.S. Army Personnel: Methods and Summary Statistics. Technical Report NATICK/TR-89/044, U.S. Army Natick Research, Development, and Engineering Center, Natick, MA.
- Greiner TM, Gordon CC (1990) An Assessment of Long-Term Changes in Anthropometric Dimensions: Secular Trends of U.S. Army Males. Technical Report NATICK/TR-91/006, U.S. Army Natick Research, Development, and Engineering Center, Natick, MA.
- Healy MJR (1989) Measuring measuring errors. *Stat Med* 8:893-906.
- Himes JH (1989) Reliability of anthropometric methods and replicate measurements. *Am. J. Phys. Anthropol.* 79:77-80.
- Jamison PL, Zegura SL (1974) A univariate and multivariate examination of measurement error in anthropometry. *Am. J. Phys. Anthropol.* 40:197-203.
- Johnston FE, Martorell R (1988) Population Surveys. In TG Lohmann, AF Roche, and R Martorell (eds.): *Anthropometric Standardization Reference Manual*. Champaign, Illinois: Human Kinetics Books, pp. 107-110.
- Koopmans LH (1987) *Introduction to Contemporary Statistical Methods*, 2nd edition. Boston: Duxbury Press.
- Liu K (1988) Measurement error and its impact on partial correlation and multiple linear regression analyses. *Am. J. Epidemiol.* 127:864-874.
- Malina RM, Hamill PVV, Lemeshow S (1973) *Selected Body Measurements of Children 6-11 Years*. Vital and Health Statistics, Series 11, Number 123.
- Marks GC, Habicht JP, Mueller WH (1989) Reliability, dependability, and precision of anthropometric measurements. *Am. J. Epidemiol.* 130:578-587.
- Mueller WH, Martorell R (1988) Reliability and accuracy of measurement. In TG Lohmann, AF Roche, and R Martorell (eds.): *Anthropometric Standardization Reference Manual*. Champaign, Illinois: Human Kinetics Books, pp. 83-86.
- Nichol CR, Turner CG (1986) Intra- and interobserver concordance in classifying dental morphology. *Am. J. Phys. Anthropol.* 69:299-315.
- Rosner B (1990) *Fundamentals of Biostatistics*, 3rd edition. Boston: PWS-Kent Publishing Company.
- Rothman, KJ (1986) *Modern Epidemiology*. Boston: Little, Brown and Company.
- Utermohle CJ, Zegura SL (1982) Intra- and interobserver error in craniometry: a cautionary tale. *Am. J. Phys. Anthropol.* 57:303-310.
- Utermohle CJ, Zegura SL, Heathcote GM (1983) Multiple observers, humidity, and choice of precision statistics: factors influencing craniometric data quality. *Am. J. Phys. Anthropol.* 61:85-95.