Legacy Resource Management Program Project No. 11-157

The Accuracy and Adequacy of In-Field Artifact Analysis: An Experimental Test

Michael Heilen, Ph.D., Statistical Research, Inc.

Jeffrey H. Altschul, Ph.D., SRI Foundation and Statistical Research, Inc.

The Problem: Throughout the western United States, limited-collection and no-collection policies promulgated by federal agencies and State Historic Preservation Offices have led the Department of Defense (DoD) to rely on in-field analysis for analyzing most or all artifacts used to characterize and interpret archaeological sites recorded during survey. This trend counters a century-long scientific practice of collecting samples of surface artifacts to assist in dating, interpreting, and documenting sites and curating these artifacts to allow contemporary and future archaeologists access to primary scientific data. The driving force behind in-field analysis has been to minimize the DoD's curation responsibilities and expenditures. Under federal law, collections made as a result of federally funded projects must be cared for in perpetuity. Curation can be expensive, and the funding and space needed for curation are limited throughout the country.

Limited-collection and no-collection policies are based on the assumption that in-field artifact analysis is comparable in quality to artifact analysis conducted in the laboratory. Further implicit in limited-collection and no-collection policies are that contemporary and future archaeologists will accept the results of in-field analysis without question and that future research and management decisions will not need access to artifacts. The DoD and other federal agencies have expended, at a minimum, tens of millions of dollars on surveys based on these assumptions, yet no one has actually tested whether they are valid. Despite heavy reliance on in-field analysis for inventory and evaluation and considerations of performing in-field analysis during data recovery, very little is known about how accurate in-field analysis really is or how adequate it is for site interpretation. Legacy Resource Management Program (Legacy) Project No. 11-157 was designed as a preliminary experiment to test those assumptions. The project is a follow-on project to Legacy Project No. 07-353, which developed analytical tools for assessing the quality and reliability of basic data derived from archaeological survey.

Test Experiment: Legacy Project No. 11-157 assessed the adequacy and accuracy of in-field artifact analysis and digital-photograph analysis at two prehistoric archaeological sites located on military installations in the western United States. One case study was the Soldier Creek site (AZ EE:7:164 [AMS]), in southeastern Arizona, at Fort Huachuca. The other site was FB 9583, in south-central New Mexico, on the East McGregor Range of Fort Bliss Military Reservation. At each of these sites, samples of individually numbered ceramic and lithic artifacts discovered in surface contexts were analyzed by two separate field technicians. The artifacts were then collected and analyzed by trained specialists who analyzed, in a laboratory setting, either the physical artifacts or only digital photographs of the artifacts. To standardize identification methods, all analysts used the same typological system for identifying artifacts and also used manuals that described the attributes of specific artifact types identified during the project. Artifact identifications resulting from each of the analyses were entered into a relational database system, using standardized typologies and quality controls. The data from each artifact analysis were compared between and among analyses to assess the levels of agreement between analyses and the adequacy of site interpretations that could be derived from each analysis.

Several statistics were used to assess the accuracy and precision of the in-field and digital-photograph analyses: *agreement index*, *Cohen's kappa*, *McNemar's test for bias*, *false-positive rate*, and *false-*

negative rate. For most assessments, the hands-on laboratory analysis was treated as the "gold standard" for the project, and the results of the in-field and digital-photograph analyses were tested against that standard. The *agreement index* is a common method for assessing interobserver agreement that represents the proportions of observations that are in agreement with each other for a given pair of analyses. *Cohen's kappa* uses contingency tables to calculate the strength of agreement between observers based on the observed and expected levels of agreement. *McNemar's test for bias* uses the same contingency table to identify whether systematic bias may have occurred in making artifact identifications.

The levels of Type I (false positive) and Type II (false negative) errors in artifact identifications were assessed by calculating *false-negative* and *false-positive rates* for each artifact type and site. A *false negative* is a case in which something is asserted to be absent when it is, in fact, present. For example, a *false negative* would occur if the hands-on analysis identified an artifact as a specific type, such as a biface, and one or more of the other analyses failed to identify the same artifact as a biface and instead identified the artifact as something else. A *false positive* is a case in which something is asserted to be present when it is, in fact, absent. In the context of this project, a *false positive* represented a case in which an artifact was identified by one or more analyses as a specific type, and the hands-on laboratory analysis identified the artifact as a different type. For instance, if an artifact was identified by the hands-on laboratory analysis as a flake, but one of the other analyses identified it as a biface, this would represent a *false-positive* case for the artifact type of biface.

In addition to the accuracy of the in-field and digital-photograph analyses, the adequacy of results for site interpretation was also assessed. This was achieved by evaluating differences in the overall distribution of analysis results, identifying the numbers and kinds of site functions and components that could be inferred based on artifact-analysis results, and calculating, for each analysis, standard metrics used in site interpretation, such as artifact richness or numbers of lithic tools or ceramic vessels represented in a sample.

The Result: In general, results of these assessments showed that both the in-field analysis and the digital-photograph analysis were of low accuracy and were often inadequate for site interpretation. Rare and important artifact types were often misclassified, and evidence for both random error and systematic bias in artifact identification was common. Digital-photograph analysis tended to be more precise than in-field analysis, but digital-photograph analysis also tended to identify rare types incorrectly, resulting in more-precise but inaccurate inferences regarding the temporal and cultural affiliations of a site.

A striking result of the *agreement-index* assessments was that the levels of agreement between analyses were low, on average. Approximately half of the artifact identifications were in agreement between analyses.

Cohen's kappa, calculated for the more-common artifact types, showed that the strength of agreement for the most-common artifact types was often low and that none of the analyses was universally better in identifying artifact types. Typically, one analysis was better than another in identifying one or more common artifact types but not in identifying all or a majority of common artifact types. Essentially, this means that none of the in-field or digital-photograph analyses could be counted on for accurate identification of even the most-common artifact types.

McNemar's test for bias, which was calculated for the same artifact types as Cohen's kappa, demonstrated that systematic bias did occur for several analyses. Evidence for systematic bias was not discovered for the analysis of ceramic artifacts at Fort Bliss but was discovered for the analysis of ceramic artifacts at Fort Huachuca, and evidence for systematic bias was found for the analysis of common lithic-artifact types at both installations. Systematic bias was particularly pronounced at Fort Huachuca, where one of the in-field analyses frequently and consistently misidentified both ceramic and lithic artifacts as types that were actually either rare in or absent from the sample.

False-positive and *false-negative rates* showed that identifications of individual artifact types were frequently in error. In general, only the most common and recognizable of artifact types were classified

correctly a majority of the time. Rare types were commonly misidentified. In short, artifacts that would be most useful in interpreting a site are exactly the ones most likely to be misidentified.

The results of the in-field and digital-photograph analyses demonstrated the potential for such analyses to result in erroneous and incomplete interpretation of site attributes, such as cultural and temporal affiliation. Artifact richness and the numbers of ceramic vessels that could potentially be inferred based on analysis results tended to be greater for hands-on laboratory analysis and digital-photograph analysis, suggesting that the intensity and diversity of activities interpreted to have occurred at a site could be substantially lower if the results of the in-field analysis are used. In-field and digital-photograph analyses also tended to identify greater numbers of flaked stone tools than were indicated by the hands-on laboratory analysis.

All together, these results suggest that accurate site interpretation based on in-field or digital-photograph analysis is possible but not necessarily probable. Some analyses will perform well for some artifact types but not for other artifact types. Other analyses will perform poorly for many artifact types but can also perform fairly well for a few artifact types. In some cases, digital-photograph analysis performed better than in-field analysis, but in other cases, the opposite occurred. There also tended to be variation in the accuracy of artifact identifications between in-field analyses; one in-field analysis could be fairly accurate for many, but not all, artifact identifications, and the other could be inaccurate in many cases. Rare and important artifacts were missed most often by in-field analysis; digital-photograph analysis identified rare and important artifact types, but not necessarily the right ones.

An investigator might conclude, based on in-field analysis, that fewer activities occurred at a site than was the case, because of artificially low artifact richness; that ceramic-vessel use was less intense than was the case, because of erroneously low vessel counts; or that flaked-stone-tool use was either more or less intense than was the case, because of erroneously high or low flaked-stone-tool counts. Different kinds of ground stone or flaked stone technology could also be inferred based on the results of the in-field or digital-photograph analysis. For instance, one analysis might identify a substantially greater number of two-handed manos in a collection than were likely present, leading to the potential inference that intensive processing of maize occurred at a site, whereas another might identify several basin metates and milling stones that were not likely present at a site, leading to the inference that grinding activities focused on the use of less-formal tools. Fewer cultural or temporal affiliations and less-precise chronologies could potentially be inferred from the results of in-field analysis. On the other hand, digital-photograph analysis appears to have the potential for achieving a more-accurate and more-precise picture of a site's temporal components, but it can still miss rare components.

Implications: Accurate artifact-analysis data are routinely needed to evaluate significance under Section 106 of the National Historical Preservation Act of 1966, as amended (NHPA); to support the stewardship responsibilities mandated by Section 110 of the NHPA; and to consider historic properties for adverse effects. Several of the variables that many investigators have indicated as crucial to determining significance and establishing representative samples of sites for long-term preservation are derived primarily from artifact analysis: site function, assemblage diversity, temporal affiliation, and cultural affiliation. The experiment performed for Legacy Project No. 11-157 showed that all of those variables could be inaccurately assessed, sometimes grossly so, when based on in-field or digital-photograph analysis.

Consultation regarding the eligibility and treatment of historic properties is a fundamental component of the compliance process mandated by Section 106 of the NHPA. Engaging stakeholders in the identification and consideration of adverse effects is an important goal of DoD cultural resource management (CRM) programs. Establishing trust and good working relationships is crucial to ensuring that the process runs smoothly and to demonstrating that the DoD is performing its compliance responsibilities in good faith. Having accurate data that can be counted on for consulting in regard to adverse effects to historic properties is one means of establishing trust. This experiment makes clear that site interpretation based on in-field analysis can be inaccurate and that rare and important site components may be missed or misin-

terpreted because of faulty artifact data. If these possibilities become accepted truths, then other, more basic and important data, such as data on location and integrity, may also not be trusted. Lacking confidence, installations may never get buy-in from stakeholders and could either fail to successfully complete the compliance process or be required to engage in other, unwelcome and costly efforts, such as dispute resolution and termination of consultation.

Recommendations: Because in-field analysis is not likely to be abandoned and may be necessary in many cases, the following recommendations are made to better understand the quality of in-field-analysis data, to improve in-field-analysis results, and to better assess the situations in which in-field analysis should be implemented:

- **Consult with stakeholders.** Engage stakeholders in survey methods and procedures, providing the pros and cons of in-field vs. laboratory analyses and collection vs. noncollection of artifacts. Allow stakeholders the opportunity to weigh in with their concerns regarding in-field analysis and artifact collection.
- **Promote and support standardization and consistency in artifact identification.** This can be achieved by supporting the development of technical field manuals pertinent to specific regions and subregions that can be used to present standardized and explicit information to analysts about how to identify the artifact types likely to be encountered at archaeological sites in a given region. Further, support and encourage periodic training of field technicians in artifact analysis that is specific to particular subregions, survey areas, and artifact classes. Encourage the use of handheld recording devices or personal digital assistants with standardized entry forms and databases in conducting in-field analysis.
- **Promote and support testing and validation of in-field-analysis results.** Encourage CRM programs and DoD contractors to test the accuracy and adequacy of analyses performed by individuals tasked with conducting in-field artifact analysis for archaeological projects. To ensure effective management of historic properties when relying on in-field analysis data, require CRM programs to test, for a representative sample of projects, the accuracy and adequacy of the in-field artifact analyses performed during previous surveys.
- **Ensure scientific validity and preservation of CRM data.** When in-field analysis is to be conducted, require that trained specialists perform the analysis during inventory of large and important sites, during eligibility evaluations or site revisits, and during data recovery. When possible, analyze provenienced artifacts in a laboratory setting, and either curate them or return them to their discovery locations after analysis. Intensively document artifacts analyzed in the field, collect representative samples of portable artifacts from sites subjected to in-field analysis, and properly curate all data.
- **Conduct a study comparing the costs of in-field analysis to the costs of artifact collection and curation.** The assumption is that because collection and curation add lines to a budget, in-field analysis must be cheaper and more cost-effective than collection. Whether this is really true in the long-run is not known. What are the costs of revisiting sites and performing additional recording and analysis in support of evaluation efforts necessitated by inaccurate artifact data? What are the costs of investing more time in the in-field recording of artifacts and curating the related documentation? What are the social, political, and cultural-heritage costs of losing resource values by not knowing enough about a resource to interpret its significance and decide on the course of stewardship? The real costs, to the DoD and to the public, of collection, laboratory analysis, and curation vs. in-field analysis and limited collection or noncollection of artifacts need to be more thoroughly evaluated.