THE VARIABILITY OF CRATER IDENTIFICATION AMONG EXPERT AND COMMUNITY CRATER ANALYSTS. S.J. Robbins[1,*], I. Antonenko[2], M.R. Kirchoff[3], C.R. Chapman[3], C.I. Fassett[4], R.R. Herrick[5], K. Singer[6], M. Zanetti[6], C. Lehan[7], D. Huang[7], P.L. Gay[7]. [1]LASP, 3665 Discovery Dr., University of Colorado, Boulder, CO 80309; [2]Planetary Institute of Toronto, 197 Fairview Ave. Toronto, ON M6P 3A6, Canada and University of Western Ontario, 1151 Richmond St. N. London, ON N6A 5B7, Canada; [3]Southwest Research Institute; 1050 Walnut Street, Suite 300, Boulder, CO 80502; [4]Department of Astronomy, Mount Holyoke College, 50 College St., South Hadley, MA 01075; [5]Geophysical Institute, University of Alaska Fairbanks, Fairbanks, AK 99775; [6]Department of Earth and Planetary Sciences and McDonnell Center for the Space Sciences, Washington University in St. Louis, 1 Brookings Dr., Saint Louis, MO 63130; [7]The Center for STEM Research, Education, and Outreach at Southern Illinois University Edwardsville, Edwardsville, IL 62025 IL. *stuart.robbins@colorado.edu

Introduction: Statistical studies of impact crater populations have been used to model ages of planetary surfaces for several decades [1]. This assumes that crater counts are approximately invariant and a "correct" population will be identified if the analyst is skilled and diligent. However, the reality is that crater identification is somewhat subjective, so variability between analysts, or even a single analyst's variation from day-to-day, is expected [*e.g.*, 2-3]. This study was undertaken to quantify that variability within an expert analyst population and between experts and minimally trained volunteers.

Methods: Eight scientists (authors 1-8), each with at least 5 years of crater counting experience, were recruited to measure craters on two images using their preferred software. The software included ArcGIS (by ESRI) with various extensions, JMARS (by ASU), DS9 (by Smithsonian Astrophysical Obs.) with custom add-ons, and the Moon Mappers ("MM") interface (by CosmoQuest). In addition, two researchers (Antonenko and Robbins) used several interfaces to decouple differences between software packages and individuals. The first region was a 4107×2218-px segment of *Lunar Reconnaissance Orbiter* (*LRO*) Narrow-Angle Camera (NAC) M146959973L (63 cm/px) data, centered on the *Apollo 15* site. The region has ~1000 craters in the 10-400 m range and craters ≲150 m are in empirical saturation for (typical for mare [4]), representing an worst case for crater counting repeatability. Volunteers from the MM project also identified craters in this image. The second image, viewed only by the experts, was *LRO* Wide-Angle Camera (WAC) image M119455712M that contains both mare and highlands.

Individual markings were grouped for experts and volunteers using a clustering code to identify which marked features represent the same crater marked by different persons. In the expert data, craters marked in 5 or more instances (NAC) were deemed "verified" and added to a final "ensemble" crater catalog (this was reduced to ≥4 for WAC data because the number of interfaces used was less by 2). In the volunteer data, craters marked by ≥6 persons were "verified." Individuals' results were compared amongst themselves and to the ensemble catalog. Analyses were done in units of pixels so that results may be generalized. The NAC results are shown in Fig. **1**, and the crater population data are displayed as cumulative size-frequency distributions (CSFDs) and R-plots in Fig. **2**.

Results: First, experts using the MM interface were compared with volunteers to determine if there is reasonable agreement between experts and volunteers; they were also compared with experts' preferred software to determine if experts can reasonably reproduce their counts regardless of interface. Both hypotheses were validated.

Our second investigation used the CSFDs shown in Fig. **2** along with the ensemble result. These illustrate a large dispersion in the number of craters identified at any given diameter. Standard deviation from the ensemble for the NAC ranged from 21% for $D \approx 18$ px (~12 m) to 32% for $D \approx 100$ px (~70 m). This implies expert CSFD results are more consistent for smaller craters than larger craters, possibly due to fewer craters and more degradation at large sizes. This is similar to the results of [2]. WAC mare data have a minimum dispersion of 13% for $D \approx 10$ px; WAC highlands have a dispersion of 30–40% across all diameters.

Third, we studied the populations with Kolmogorov-Smirnov (K-S) tests to determine if the experts and volunteers found similar populations regardless of absolute number of craters found. The NAC data show poor agreement among experts for $D \geq 18$ px, with 54% of data pairs representing different populations (P-value <0.01). Agreement improved significantly when smaller diameters were removed, with 39% representing different populations at $D \geq 22$ px (~15 m) and only 18% being different at $D \geq 25$ px (~17 m), suggesting that aliasing effects occur at smaller diameters. Similar effects were found for the WAC data, though agreement was better at smaller diameters (researchers also identified smaller craters in WAC data). Consistency among different interfaces for individual experts was also variable. Robbins conducted NAC counts using MM and ArcGIS. His results show good agreement over the entire diameter range: the two CSFDs are within 1 standard deviation of each other's error bars over all diameters. Antonenko conducted NAC counts using MM, JMARS, and ArcGIS (with CraterHelper tools). Her results are more complicated; all three methods agree to 1 standard deviation for large craters ($D > 80$ px), ArcGIS and JMARS data differ by >1 standard deviation from the other methods

for medium ($30 < D < 80$ px), and small ($D < 25$ px) craters, respectively. For $D \geq 25$ px, K-S test P-values of <0.05 suggest that none of Antonenko's data unambiguously represent the same population. This shows that individual experts may produce varying results *via* different interfaces.

Fourth, we compared individual NAC craters between the experts and volunteers. To within the standard deviations from the weighted means of the ensemble results, all matched crater diameters agreed (*e.g.*, Fig. **1**, right). We found that volunteers generally have a 2× greater dispersion than experts in both crater diameter and location.

Fifth, we separated the craters by preservation state (Chapman and Robbins separated them into four different classes). As perhaps expected, we found volunteers have a more difficult time than experts identifying highly degraded craters, such as Fig. **1** bottom-right. We also found that the scatter in crater measurement (diameter and location) was independent of preservation for both experts and volunteers except for expert diameter measurement in NAC data, where there was better agreement for more pristine craters.

Finally, we investigated artifacts near the minimum diameter. The NAC image's was cutoff was set at $D < 18$ px, and we found that all experts were complete at those diameters with few artifacts; they accomplished this by identifying craters to at least 2 px smaller. Volunteers, however, showed significant artifacts for $18 \leq D < 21.5$ px; this was in part due to the clustering algorithm not being able to average in smaller-diameter craters. In WAC data, no cutoff was set and experts were varied in (1) where they thought their completeness was (Fig. **2**, arrows), (2) their method for determining it, and (3) their estimate's success relative to the ensemble. We also saw a disturbing feature of there being no trend in artifacts near an individual's completeness level – some showed a gradually decreasing population before a sharp decrease, others a sharp uptick, while others followed a normal population until their completeness level.

Implications: This study has significant implications for comparisons of model surface ages determined by different researchers. Results show that variability in crater counts between different experts regardless of interface is generally ~15–40% but can be as much as a factor of 2 different. When using these populations to estimate ages (despite secondary craters being included), they vary from 1.5±0.7 to 3.2±0.8 Gyr (NAC), 1.3±0.4 to 2.2±0.5 (WAC, mare), and 3.4±0.1 to 3.8±0.0 (WAC, highlands). Meanwhile, the NAC ensemble age for experts and volunteers are 2.71 and 2.72 Gyr, respectively, showing that volunteers as an ensemble can produce crater population statistics as good as experts. From this, we also conclude that it is inappropriate to quote model crater ages to three or more significant figures, and that standard Poisson uncertainties are a *minimum* because they do not factor in the single analyst's threshold of detection and the larger variations expected among other experts.

References: [1] Shoemaker & Hackman (1962), in The Moon, LPI, p. 289-300. [2] Greeley & Gault (1970), doi: 10.1007/BF00561875. [3] Hiesinger *et al.* (2012), doi: 10.1029/2011JE003935. [4] Shoemaker (1965), in Nature of Lun. Surfaces, 23-77.

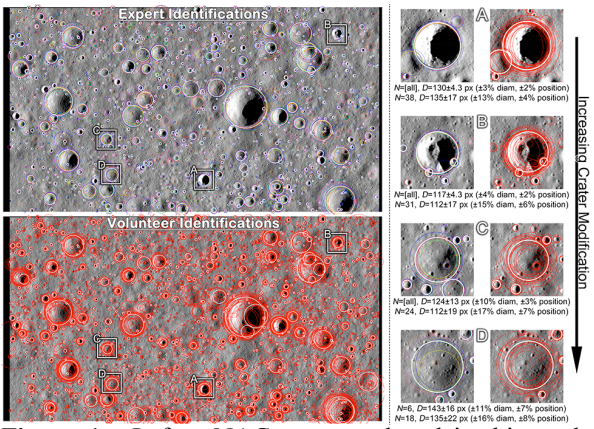

Figure 1: Left— NAC areas analyzed in this study with markings overlaid. Top are expert markings, bottom are volunteer data; both only show craters $D \geq 18$ px. Expert markings are color-coded to correspond with Fig. **2**. White circles are results from the clustering algorithm. Right— Example craters are shown with expert markings and ensemble craters (left column) and volunteer markings and ensemble craters (right column); craters are in order of increasing modification. Below each pair is the number (N) of persons who marked that crater and the mean diameter (D) with standard deviation. Values in parentheses are relative standard deviations ($\delta D/\mu_D$; $\delta(x,y)/\mu_D$).

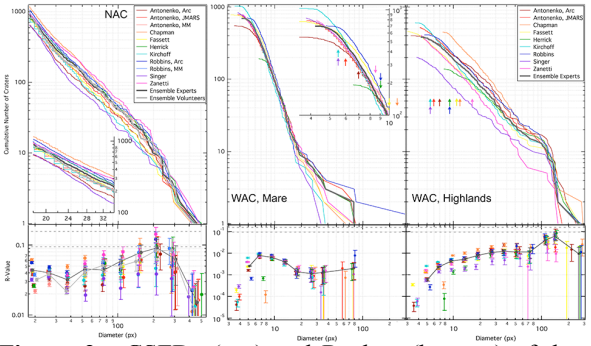

Figure 2: CSFDs (top) and R-plots (bottom) of data for craters in the NAC image (left), WAC mare (middle), and WAC highlands (right). Colors are different experts (see legend). Dark grey is the clustered expert data and light grey is the clustered volunteer data (latter is NAC only). Dashed lines on R-plots correspond to 3% and 5% of geometric saturation. Small vertical arrows (WAC) are where each expert estimated their completeness to be. Horizontal and vertical axes are different for the NAC and WAC columns because of different completeness levels. Error bars have been removed from the CSFDs for clarity; the vertical scale is the cumulative number of craters so uncertainty is $N^{1/2}$ (*e.g.*, ±10 for $N_{\text{cumulative}} = 100$).