10-2008

An Assessment of Shore-Based Counts of Gray Whales

David Rugh
National Marine Mammal Laboratory, Alaska Fisheries Science Center, National Marine Fisheries Service, NOAA

Marcia Muto
National Marine Mammal Laboratory, Alaska Fisheries Science Center, National Marine Fisheries Service, NOAA

Roderick Hobbs
National Marine Mammal Laboratory, Alaska Fisheries Science Center, National Marine Fisheries Service, NOAA

James Lerczak
National Marine Mammal Laboratory, Alaska Fisheries Science Center, National Marine Fisheries Service, NOAA

MARINE MAMMAL SCIENCE, 24(4): 864–880 (October 2008)
No claim to original US government works
DOI: 10.1111/j.1748-7692.2008.00215.x

An assessment of shore-based counts of gray whales

DAVID J. RUGH
MARCIA M. MUTO
RODERICK C. HOBBS
JAMES A. LERCZAK
National Marine Mammal Laboratory,
Alaska Fisheries Science Center,
National Marine Fisheries Service, NOAA,
7600 Sand Point Way NE,
Seattle, Washington 98115–6349, U.S.A.
E-mail: dave.rugh@noaa.gov

ABSTRACT

Counts of migrating whales depend on accurate sightings data. In this study, teams of shore-based observers independently tracked whale pods during the southbound migration of gray whales (*Eschrichtius robustus*) while a routine ("standard watch") census was underway. A comparison of sighting records showed that time and location accuracy was limited to 45 s, 3° (magnetic) horizontally, and 0.0057° (0.2 reticles) vertically. Of 242 attempts to track whale groups, 72 failed, 120 were "good tracks," and 83 qualified as "best tracks" because they had ≥8 sightings/pod, ≥16-min observation time, and unequivocal matches to sightings in the standard watch during uncompromised visibility. Between paired tracking teams, 39 attempts to conduct concurrent tracks resulted in 21 "good tracks" with complete agreement in 71% of the cases. Of 133 comparisons between trackers and the standard watch, 43% of the pod-size estimates were the same, but the standard watch overestimated 10% of the pods and underestimated 47%. Thus, according to results from tracking teams, pods recorded as size 1 by observers on the standard watch should be corrected by +0.6; pods of 2 by +0.5; pods of 3 by +0.8; and pods >3 (4–10) were overestimated and should be corrected by −0.6.

Key words: shore-based gray whale counts, *Eschrichtius robustus*, whale census, counting accuracy, sighting records.

Gray whales (*Eschrichtius robustus*) in the eastern North Pacific Ocean have made a remarkable recovery since the 19th century, when they were nearly exterminated by commercial whaling. This recovery has been documented by abundance estimates made from shore-based counts at or near Granite Canyon, in central California, since 1967 (Rugh *et al.* 2005). These counts are based on observers independently searching for and recording whale sightings during the southbound migration, generally from mid-December until mid-February. In 1986 the standardized counting procedure was evaluated during a 6-d test using paired, independent observers (Rugh *et al.* 1990), and this test was applied throughout the 2-mo census in 1987–1988 (Rugh *et al.* 1993). During each census since then, paired, independent counts have occurred during some or all of the daily watches (*e.g.*, Rugh *et al.* 2005). This has provided documentation of the degree of consistency between observers' sighting records and a characterization of each observer's performance relative to the others. Results from this study have also led to a modification of the abundance calculations by using a multiplier of approximately 1.2 to correct for whales missed within the viewing area during a watch (*e.g.*, Hobbs *et al.* 2004).

After many seasons of applying the paired, independent observer tests, it is evident that this is a valuable tool for evaluating observers' sighting records. However, there are some limitations to this technique. For instance, discrepancies in pod-size estimates and links between sightings have been treated as an undercounting error by the paired observer with fewer sightings, but there has not been a way to document overcounting, if it occurred. What has been needed is an efficient (large sample size per cost) technique to study sighting records and related variables to give a better assessment of the error range in the census data, as well as to provide improved parameterization of elements (sighting time, distance, and pod size) used in the matching algorithm.

In calculating the gray whale abundance estimate, the factors with the greatest uncertainties, and potentially the greatest unknown biases, are (1) the pod-size correction factor, (2) links made between sightings by each observer, and (3) the matching algorithm (which matches sightings between paired observers). All of these factors involve knowledge of how an observer identifies and interprets the visual cues from a pod of whales passing through the viewing area. The gray whale survey design is built on some basic assumptions: (1) each whale provides one or more visible cues when it passes through the viewing area while a watch is in effect; (2) no sighting data are recorded in the absence of whales (*i.e.*, there are no "false positives"); (3) the whales travel in fairly discrete pods that remain cohesive, at least in the area perpendicular to shore where the search is concentrated; (4) whales maintain a typical travel speed (6 km/h, Swartz *et al.* 1987), migration path (traveling parallel to the coast), and surfacing pattern (*i.e.*, average surfacing intervals of 1.3 min and long dives of 3.1 min, Swartz *et al.* 1987); and (5) the whales are traveling south (whales traveling north were treated separately in the analysis). This list of assumptions was fundamental to the way each observer linked multiple sightings of a whale pod and recognized when a new pod had been seen. These assumptions must also be met to accurately compare concurrent sighting records through the matching algorithm used in abundance estimations.

Accurate pod-size estimation is an integral component of the gray whale survey because it is more efficient, both for data recording and statistical analysis, to count pods and estimate pod size than to record individual whales. Available data and observer experience have indicated that, in the majority of circumstances, gray whale pods are sufficiently cohesive and behave in a manner predictable enough to support this approach. What remains is to determine the range of deviation from typical behavior and to quantify biases that may result from errors in linking sightings within each record, matching sightings between records, and estimating pod sizes.

Previous studies (*e.g.*, Reilly 1981, Laake *et al.* 1994) have attempted to calibrate pod-size estimates made by shore-based observers. These studies used observations from a circling aircraft to establish "true" pod size. The difficulty came in matching aerial sightings to sightings recorded from shore. In 1978, Reilly (1981) had 12 shore-based observers and an aerial team independently record pod sizes for each of 62 pods (381 matches). Radio communication between the aerial and shore-based teams was used to establish which pod was being circled. However, the shore-based observers' search effort differed slightly from their typical standard watch because of the distractions of a circling aircraft. A similar problem occurred during the aerial-calibration studies conducted in 1993 and 1994 (Laake *et al.* 1994). In many instances, the records (including 66 different pods) were based on dedicated searches by several shore-based observers, not just the one or two observers on the standard watch. This was in the interest of maximizing the effectiveness of the aerial survey, but often the standard watch had to be abandoned because of the distractions caused by the test. Although observers were asked to maintain a consistent systematic search, awareness of the test inevitably drew more attention to whale pods being circled by an aircraft.

Another study to calibrate pod sizes (DeAngelis *et al.* 1997) used video records collected from paired thermal sensors at Granite Canyon in 1995 and 1996. The sample size was large (242 pods), but the sensors were stationary and could not follow whale pods; therefore, the sample time for each pod was fairly small, especially for pods close to shore. The accuracy of pod sizes established in thermal sensor videos needs to be verified.

To address the issue of accuracy in pod-size estimates, we conducted a shore-based visual study at the standard observation site at Granite Canyon during the gray whale southbound migration in January 1997. This effort was continued in January 1998, in conjunction with the routine census of 1997–1998 (Hobbs and Rugh 1999). The intent was to examine how gray whale pods moved through the survey-viewing field and to compare these observations to data recorded by standard-watch observers for the respective pods. More specifically, the objectives were to: (1) develop and test a reliable, efficient method for tracking whale pods (*i.e.*, to concentrate on one pod at a time throughout the viewing area); (2) measure the precision of time and location data recorded during the standard watch; and (3) compare reliable sighting records made by tracking teams to pod-size estimates recorded during the standard watch. This experiment tested the assumption that teams of observers working in pairs can reliably follow, record sighting locations, and determine the pod sizes of whales migrating through the primary viewing field. These tracking records, when compared to the standard-watch records, may then be used to calibrate pod-size estimates, links between sightings, and the inter-observer matching algorithm used in abundance estimate calculations.

Methods

Standard Watch

Season and location—Systematic counts of gray whales have typically been conducted from mid-December to mid- or late-February, and sometimes as late as March (Rugh *et al.* 2005), during virtually the entire duration of the southbound migration past the Granite Canyon research station. Observations were conducted from the edge of a sea cliff, 22.5 m (eye height) above mean sea level. Sheds provided writing platforms with some protection from the elements, enabling observers to concentrate on the viewing area. Although the field of view covered $>150^\circ$, observers generally searched through an arc of only 40°–50° near the standard azimuth (a line, perpendicular to the coastline, intersecting the survey site at the magnetic bearing of 241°).

Schedule—Standard watches were 3 h each, maintaining a search through most daylight hours, from 0730 to 1630. Observers were rotated to keep a balance of effort in each of the three shifts, minimizing potential biases in observer performance as a function of possible diurnal trends in sighting rates. Each observer operated independently and hand-recorded entries onto a data form.

Data recorded—Each gray whale pod was recorded as to time, horizontal bearing, vertical angle, and pod size, generally for the sighting as close to the standard azimuth as possible. Magnetic compasses in Fujinon 7 × 50 binoculars provided a horizontal bearing, and 14 reticle marks in the binoculars provided vertical angles relative to the horizon (detailed in Rugh *et al.* 1993, Kinzey and Gerrodette 2001). A table based on average swimming speeds and sighting locations helped the observers predict the time and vertical angle at which a pod would cross the standard azimuth. In addition to sighting information, observers recorded start and end times of systematic search effort and environmental changes. These entries included visibility (VIZ: subjectively categorized from 1 to 6 for excellent to useless), wind direction, and sea state (Beaufort scale). Generally, after each watch, data were entered into a computer and quality checked before the next day's effort began.

Paired, independent counts—In addition to the primary standard watch, a second, independent watch was conducted up to three times daily most seasons since 1985 (detailed in Rugh *et al.* 1990, 1993). The field of view, altitude, and construction of the sheds used in the standard and independent watches were nearly identical. This provided paired, independent sighting records, which allowed comparisons between observers and an estimation of the number of whales missed within the viewing area.

Data Precision

Time precision—Observers recorded time to the second, using timepieces that were synchronized frequently between the different observation sites. On the standard watch, observers worked alone, so they had to quickly glance at their watches and make hasty entries, increasing the probability of making reading or recording errors. Tracking teams (described below), on the other hand, had the advantage of a dedicated recorder who could look at a watch and record the time while the primary observer concentrated on tracking a focal pod.

Bearing precision—Prior to each research season, all available binoculars were tested by one person who took bearings (magnetic) on a unique, static target, thus minimizing variables. Whale sighting data were collected from only those binoculars that were within 1° of the average of all of the binoculars. Accordingly, 1° is the best precision that can be expected from binocular compass readings. Each year, binoculars were designated to particular sheds, so the same binoculars were used by all observers at the respective shed throughout each season but not necessarily between seasons.

Reticle precision—Because of the low altitude of the research station (eye height at 22.5 m above mean sea level), distances to targets were sensitive to even small errors in vertical angle measurements. Observers were asked to make records to the nearest tenth of a reticle (0.1 reticle = 0.03°). This involved interpolations because etch marks in the binoculars were subdivided into increments of 0.1 within only the first reticle (0.0–1.0). Therefore, 0.1 was the best possible precision expected from reticles.

Static calibration target—The accuracy of each binoculars' horizontal bearing was tested prior to each watch by having observers take a reading on a calibration point (a unique rock feature, approximately 1.2 km north of the site). Initially, this was done to check for compass problems resulting from the magnetic pull of a nearby vehicle or from damage to the binoculars. In retrospect, this proved to be an excellent tool for quantifying precision in bearing readings without the challenge of aiming at ambiguous, moving targets. In many cases, observers also recorded the binocular reticles; this has served as a test of consistency of vertical measurements.

Calibrating on a ship—In January 1996, a U.S. Coast Guard ship provided location information within the viewing area while shore-based observers at Granite Canyon recorded reticle values of the waterline below the center of the ship. These reticles were converted to distances from the observers to the ship.

Calibrating on selected whale pods—During training exercises, while practicing the research protocol, pairs of observers were asked to keep records on any whale pod that was not easily confused with others. Each observer maintained independent records of sighting times and locations, although both observers openly discussed which sightings were recorded. These data provided comparisons of observers' records of time, bearing, and reticle on a moving target.

Tracking Test

The protocol used to count gray whales over the past three decades was tested by having pairs of observers (a "tracking team") work together while the standard watch continued independently. In the tracking team, one observer focused on following a whale pod while the second observer served as a recorder but could also provide sighting information. Selected pods were followed (tracked) through the viewing area for as long as practical, generally for 30 min or more. In contrast to the standard-watch observers, who had to maintain a search of the entire viewing area (40°–50°), the tracking team could focus on a single pod or pods, depending on how many were in the field of view of the binoculars (5.4°). Also, tracking teams had the advantage of open communication with each other during the tracking session; and, since there was a dedicated (primary) observer and a recorder, the primary observer could maintain a constant field of reference, which helped to keep track of the whales. The tracking teams' training sessions were not included in the analysis.

Pod selection for the tracking test—While standard counts were underway, trackers selected whale pods north of the primary search area in a zone perpendicular to the coast. Pod selection was randomized to avoid potential bias toward large pods in the middle of the search area. The selection process involved a random number (≤ 5) and a count of available pods; when the preselected number matched the number of pods viewed in the zone, a focal pod was identified. This regime of searching for a focal pod had a time limit: up to 8 min in 1997 and 10 min in 1998. If there were not enough pods to meet the selection criteria, the effort was stopped and then started again after a short break. When a focal pod was selected, a primary observer with binoculars tracked it constantly, while the other observer recorded information and watched opportunistically. The identity of the focal pod was not shared with the standard-watch observers.

Data collected in the tracking test—Sightings were recorded according to time (to the second), reticle (to the nearest 0.1), horizontal angle (to the nearest degree), pod size, and direction headed (assumed to be southbound, unless noted otherwise). Time, reticle, and angle were recorded precisely to maximize comparisons to standard-watch data. When there was confusion about a time entry, it was considered tentative (T) if the error was within 10–60 s and unknown (U) if the error was >60 s. Whale pods were tracked until they were well south of the viewing window used by observers on the standard watch. Tracking teams recorded time and location data for every focal pod surfacing in the primary viewing area, especially near 241° (where the standard-watch search effort was concentrated). "Cue counts" (the number of times there was evidence of a whale's presence, *i.e.*, a "blow," a part of the whale's dorsal surface, or ripples) were maintained so the record showed how many times each pod was seen.

A track quality code (TQ), established to record the relative degree of confusion a tracker may have had, was a combination of subjective evaluations, including visibility of the whale pod, density of whale pods in the sighting area, behavior of the pod, and distractions that occurred during the tracking event. TQ reflected how confident the tracker was that the focal pod was consistently followed: TQ1 = the focal pod was clearly distinct; TQ2 = all but a few surfacings were distinct; TQ3 = there may have been some surfacings that were confused between whale pods; TQ4 = it is uncertain whether the track record was of the focal pod only or if it included one other pod; TQ5 = the focal pod could have been confused among several other pods; or TQ6 = the tracking effort could not follow the focal pod through the primary viewing area. Records with a TQ6 were treated as "failed" tracking efforts. Tracking teams reviewed their data immediately after each tracking event, or as soon as possible, to create the best possible written record.

Paired tracking teams—In 1997, when two teams of trackers were available, they conducted concurrent tracks of the same focal pod. Operating from separate sheds, teams identified a focal pod by communicating with wireless headsets. Communication stopped when both teams were confident they were watching the same pod, after which each tracking team followed the pod independently. No information about pod size was exchanged. The paired tracking effort provided a test of the veracity of results from any one team, indicating applicability as a correction factor for pod-size estimates relative to the standard watch.

Filtering the track records—During the analysis, when establishing which of the data were "good tracks," records were eliminated if TQ was >3 (41 tracks) or if visibility was >4 (8 tracks). Records were filtered even further for the category of "best tracks," that is, track records were deleted if the focal pod was seen <8 times during the tracking session (7 tracks), if a pod was tracked for <16 min (6 tracks), or if concurrent tracking teams had different pod-size estimates (6 tracks). Further, matches between the "best" tracking records and the standard-watch records were not used as "best matches" if the difference in reticles was >0.2 (27 tracks), if the difference in bearings was >3° (25 tracks), or if the difference in sighting time was >45 s (13 tracks). These limits were established by using the 95% bound in comparisons of the concurrent tracking records (described above), where fairly high compatibility was expected. Records were eliminated in the order demonstrated here, that is, first the tracks were deleted as a function of track quality, then visibility, *etc.*; therefore, there may have been overlapping reasons for eliminating a record.

Establishing matches between standard-watch and tracking records—A critical element to the test of the standard-watch counts of gray whales was to establish which of the whale sightings recorded by standard-watch observers were of the focal pods followed by the tracking teams. Field personnel made the initial matching effort between the standard-watch and tracking records within a day or two of the sightings. Months later, when all data were computerized, accuracy was examined by comparing discrepancies in time and location. Boundaries to accepting or rejecting matches were determined through a test of the precision in data recordings when two tracking teams simultaneously followed the same whale pods (described above).

Confusion in multiple surfacings of a whale pod—Between deep dives, traveling gray whales have surfacing intervals (shallow dive series) that average 1.34 min, with 27.00 s (range = 0.27–59.97 s; 95% CI = 26.42–27.58 s) between surfacings and 2.75 surfacings per surfacing interval (Swartz *et al.* 1987). Two observers may have recorded any one of a series of surfacings by the same whale pod, but not necessarily the same surfacings. Especially deep dives (mean = 3.10 min; range = 1.00–13.08 min; 95% CI = 2.99–3.22 min; Swartz *et al.* 1987) made it difficult to recognize if two observers' sightings were of the same whale. The tests of data precision and a protocol for comparing independent sighting records helped establish logical matches of sightings made by both teams.

Observer experience—In this study of counting protocol, most observers had already participated in several seasons of shore-based counts of gray whales; in fact, two observers were involved in most of these projects since 1975. Seven of the eight trackers participated in both the 1997 and 1998 tracking projects; individual participation ranged from 13 to 33 tracks, for an average of 25 tracks each. Most observers were rotated between the standard watch and tracking efforts daily.

Results

Sample Size

Tracking efforts were in effect during 14 d between 8 and 23 January 1997 and 16 d between 7 and 23 January 1998. While trackers recorded whale groups, the standard watches were also underway: there were 63 watches (generally 3 h each) with one observer in each of the two counting sheds ("North" and "South") doing concurrent, independent standard watches; in addition, there were 11 watches by single observers when only one of the counting sheds was in use. Of 242 attempts to track individual groups of whales, 170 (70%) groups were followed long enough to have a record sufficient for analysis. However, only 120 of these groups were considered "good tracks" recorded in excellent-to-fair conditions (TQ = 1–3; VIZ = 1–4), with a match to a pod in the standard-watch records. There were 39 concurrent tracks while two tracking teams were operational. Of these, only 21 concurrent tracks had summary evaluations with good records (TQ = 1–3 and VIZ = 1–4).

Among records considered "good tracks," the average pod had 15 recorded locations (SE = 0.7; range 3–44) and was seen 39 times (SE = 2.8; range 4–200). Dividing the number of sightings per pod by the number of whales within a pod shows that, on average, a whale was seen 15 times (SE = 0.6; range 4–41 times each). Good tracks were followed for 0.5 h each (SE = 0.02; range 0.2–1.2 h). When two concurrent tracking teams followed the same pod, they generally kept track of the pod for a similar amount of time because they started the track together.

For some of the pod-size analyses, we used only "best matches" between trackers and standard-watch observers. A total of 83 pods matched between the standard watch and the tracking records met these specifications; and, because two standard watches were usually in session while the tracking effort was underway, there were 133 comparisons of pod-size estimates between the tracking and standard-watch records.

Data Precision

Time—Of 268 comparisons between standard-watch observers and trackers, discrepancies in time entries were mostly (77%) within 30 s, but credible matches could sometimes be found even 4 min apart when data were recorded on different surfacings. Discrepancies greater than 10 s probably occurred when observers recorded different surfacings of the same whale pod.

Bearing—During the sample period for evaluating data precision from binoculars with reticles and compasses (1997–2002), there were 904 recordings of the static calibration target (mean = 315.0°; SE = 0.03), with a range of 15° (308°–323°), but 95% of the values were within 1.5° of the mean (313°–317°; bearings in Table 1 are absolute values expressed relative to the overall mean of 315.0°). Therefore, we can consider magnetic bearings to generally be accurate within 2° when aimed at a static target.

The measured bearings on a calibration target were also checked as a function of observation shed (Table 1). No significant differences were found between bearings recorded at the two standard-watch sheds (North and South sheds; Wilcoxon Rank Sum Test $W^* = -1.36$; $P = 0.087$). However, differences were found between bearings recorded at the two tracking sheds ($W^* = 6.56$; $P < 0.001$), such that a correction of 0.73° was applied to bearings from one shed to make the data more comparable. Also, differences were found between bearings recorded at the tracking sheds ($n = 55$) and the standard-watch sheds ($n = 131$; $W^* = 2{,}899$; $P < 0.001$), so a correction factor of 0.78° was added to the tracking bearings to improve the comparison between sheds.

There were 314 paired records in which two observers could be compared during 40 sessions of watching a whale pod in common (Table 2). Of these, no discrepancies in bearing occurred in 85 (27%) of the records and 56 (89%) of the discrepancies were within 2°, considered here to be the limit of expected precision even when aiming at a static target (see above). The maximum discrepancy (60°) appears to have been a recording error (288° was written in a series from 239° to 227°). As a

Table 1. Calibration of magnetic bearings on a fixed point, and a comparison of bearings taken on whales, showing absolute values. The four observation sheds are listed from south to north. Distances from the southernmost shed are shown in meters.

	Distances from south shed (m)	*n*	Mean differences	SE	Variance	Maximum range
South shed – standard watch	0	309	0.81°	0.06	1.27	15
North shed – standard watch	5.0	475	0.72°	0.03	0.41	4
South shed – trackers	10.4	43	1.00°	0.07	0.24	2
North shed – trackers	15.2	77	0.22°	0.05	0.23	3
Paired observers		287	1.46°	0.21	13.49	60[a]
Paired tracking teams		391	0.98°	0.11	4.91	36[a]

[a] Including errant recordings.

Table 2. Discrepancies in magnetic compass bearings: (1) on a static target (from calibrations prior to each watch); (2) on whale pods recorded by paired observers (who agreed on which sighting to record); (3) on whale pods followed by two tracking teams simultaneously; and (4) between the standard-watch data and teams tracking individual whale pods.

Discrepancy in bearings	Static target (shore feature)		Moving target (whales followed during practice exercises)		Moving target (whales followed by paired tracking teams)		Standard watch *vs.* trackers	
(absolute values)	Frequency	Cumulative	Frequency	Cumulative	Frequency	Cumulative	Frequency	Cumulative
0	284	31%	85	27%	164	42%	43	16%
1	533	90%	137	71%	173	86%	85	48%
2	81	99%	56	89%	31	94%	57	69%
3	3	100%	19	95%	9	96%	42	85%
4	1	100%	10	98%	2	97%	17	91%
5	0	100%	1	98%	4	98%	7	94%
>5	2	100%	5	100%	8	100%	17	100%
n	904		314		391		268	
Mean (SE)	0.04	(0.03)	1.46	(0.21)	0.98	(0.11)	2.12	(0.13)
Maximum	8°		60°[a]		36°[a]		18°	
95%	1.5°		3°		3°		3°	

[a] Including errant recordings.

generalized estimate of precision, we used 95% of the values from this test, which indicates that bearings taken on a whale pod are reliable only within 3°.

In a similar test comparing bearings taken by concurrent tracking efforts when following the same whale pod ($n = 391$), many bearings (42%) were the same, and 95% were within 3°. Therefore, 3° was treated as the outer limit when establishing "best matches."

Reticle—Of the 50 records of reticles made when aimed at a static calibration target, 34 (68%) were the same (discrepancy = 0; Table 3), and the rest were within 0.1 reticles. Therefore, when aiming at a static target, we can expect a precision of 0.1 reticles.

Distances calculated by targeting a ship at known locations ($n = 29$) showed no error in seven readings (24%), seven discrepancies were within 0.1 reticles, and the remainder (15) were off by 0.2 reticles (Table 3). The average of these comparisons (mean = −0.13; SE = 0.02) suggests that observers using reticled binoculars underestimated distances; however, there was no correlation between distances and the size of discrepancies in measurements (Kendall Distribution-free Test for Independence; $K = -38$; $P = 0.246$). Therefore, the reticle readings are considered unbiased.

Reticle values were also recorded while pairs of observers conducted training exercises in logging whale sightings ($n = 40$ exercises). In nearly a third (32%) of the 314 sightings, the two observers were in perfect agreement, and nearly two-thirds (62%) of the sightings were within 0.1 reticles of each other. The remaining values (38%) had higher discrepancies. Most discrepancies (96%) were within 0.4 reticles; but one discrepancy was 1.6, perhaps the result of a recording error (Table 3). In a generalized estimate of precision (using 95% of the values), observers' reticle values for sightings of moving whales had discrepancies within 0.4 reticles.

Concurrent tracking records (using two pairs of observers) showed perfect agreement in nearly half (45%) of the matched records, and most (83%) were within 0.1 reticle precision. Almost all (95%) values were within 0.2 reticles (= 0.0057°), so that was treated as the cutoff for establishing "best matches" when comparing records between the tracking teams and the standard watch.

TQ agreement—When two teams of trackers followed the same pod ($n = 39$), most (49%) of the judgments on TQs were the same or differed by only one increment (34%), and a few (17%) discrepancies were greater than 1. This shows that observers were fairly similar in their assessments of how well they followed a whale group.

Viability of tracking—Data from the tracking efforts were examined to quantify how difficult it was to track a pod of whales passing through the viewing area. Of the 242 tracking efforts examined (excluding practice and training sessions and treating pairs of concurrent tracks as a single effort), 44 failed because there were too few whales to satisfy the selection protocol (a randomly selected pod had to be found within 8–10 min), 3 failed because the visibility was poor during pod selection, 21 failed because the observers lost track of the pod (TQ = 6), and 4 failed for other reasons (*e.g.*, killer whales present in the search area, observers lost track of the pod during pod selection, or no reason was given). If we disregard tracking efforts that failed during the pod selection process (44 + 3 failures), then 170 of 195 (87%) attempts to track whale pods were successful enough to be considered completed tracks.

Pod-size agreement between tracking teams—In 15 of the 21 instances (71%) when two tracking teams followed the same pod in good conditions (TQ < 4), there was complete concurrence in pod size; in five cases there was a discrepancy of only

Table 3. Discrepancies in reticles: (1) on a static, onshore target used to calibrate the binoculars; (2) on a U.S. Coast Guard ship positioned at known distances; (3) on whale pods recorded by paired observers (who agreed on which sighting to record); and (4) between the standard-watch data and teams tracking individual whale pods.

Discrepancy in reticles	Static target (shore feature)		Static target (ship positioned at calibration points)		Moving target (whales followed during practice exercises)		Moving target (whales followed by paired tracking teams)		Standard watch *vs.* trackers	
(absolute values)	Frequency	Cumulative	Frequency	Cumulative	Frequency	Cumulative	Frequency	Cumulative	Frequency	Cumulative
0	34	68%	7	24%	99	32%	175	45%	79	29%
0.1	16	100%	7	48%	95	62%	148	83%	102	68%
0.2	0	100%	15	100%	62	82%	50	95%	48	85%
0.3	0	100%	0	100%	30	91%	13	99%	19	93%
0.4	0	100%	0	100%	14	96%	2	99%	7	95%
0.5	0	100%	0	100%	6	97%	2	100%	2	96%
>0.5	0	100%	0	100%	8	100%	1	100%	11	100%
n	50		29		314		391		268	
Mean (SE)	0.03	(0.01)	0.13	(0.02)	0.15	(0.01)	0.08	(0.01)	0.18	(0.02)
Maximum	0.1		0.2		1.6[a]		1.5[a]		4[a]	
95%	0.1		0.2		0.4		0.2		0.4	

[a] Apparently an errant recording.

Table 4. Pod-size estimates compared between trackers (considered to have accurate pod sizes) and the standard watch (with pod sizes to be corrected for abundance estimates). There was no difference whether "good matches" or only "best matches" were used.

Discrepancy	Good matches		Best matches	
0	90	43.9%	52	44.4%
1	75	36.6%	43	36.8%
2	26	12.7%	15	12.8%
3	9	4.4%	5	4.3%
4	5	2.4%	2	1.7%
Number of tracks	120		83	
Number of matches	205		117	

1; and in one instance, there was a discrepancy of 2 (pod size 5 *vs.* 7). Observers on each team recorded from 11 to 152 sightings, *i.e.*, "cues" (mean = 60.7; SE = 6.4), before making their final determination of pod size. Pod-size discrepancies were correlated to the estimated size of the pod (Kendall Test for Independence; $P = 0.049$). However, there was no correlation between size discrepancies and the number of sightings per pod (Kendall Test; $P = 0.500$) nor was there a correlation between size discrepancies and track quality (Kendall Test; $P = 0.246$). Although the sample size does not allow for a rigorous comparison of observers (there were only 0–4 pairwise comparisons per observer, and 6–17 concurrent tracks were collected by each of the seven observers), no one observer performed very differently from the others: pod-size discrepancies occurred only 2–4 times per observer.

Pod-size comparison to the standard watch—Using "good matches," a comparison was made between pod-size estimates made by trackers ($n = 120$) relative to estimates made by observers on the standard watch ($n = 205$). Results from this test were then compared to "best matches," with 83 tracks relative to 117 pods on the standard watch (Table 4, Fig. 1). There were no differences between good and best matches whether there was perfect agreement between trackers and the standard watch or

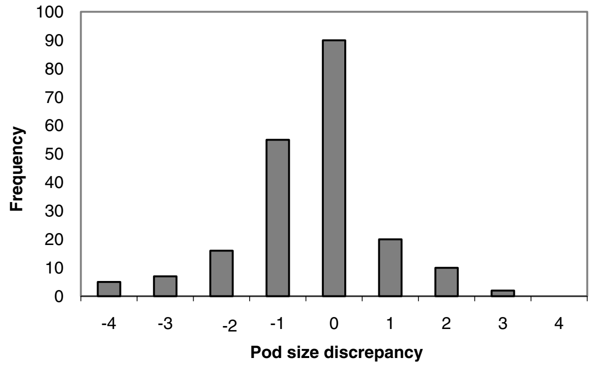

Figure 1. Discrepancies in estimates of pod sizes of gray whales migrating south past Granite Canyon, California. The discrepancies are the differences between standard-watch observers doing the census and trackers who concentrated on one pod at a time. Zero values indicate perfect agreement; negative values show underestimates made by the observers on the standard watch; positive values show overestimates.

Table 5. Comparisons of gray whale pod-size estimates by tracking teams *vs.* standard-watch observers at Granite Canyon, California. Cells indicate the number of estimates corresponding to the respective pairing (*e.g.*, in 64 instances, both methods agreed that there was only one whale in a pod). Numbers in bold are the samples in which both methods agreed on the pod size.

Trackers'	Pod sizes recorded on the standard watch									
pod sizes	1	2	3	4	5	6	7	8	9	10
1	**64**	8	1	1	0	0	0	0	0	0
2	37	**24**	1	1	0	0	0	0	0	0
3	3	10	**4**	2	1	0	0	0	0	0
4	5	11	5	**7**	2	1	0	0	0	0
5	3	7	4	2	**1**	1	0	0	0	0
6	2	3	3	1	1	**0**	4	0	0	0
7	1	1	0	0	0	0	**0**	0	0	0
8	0	0	0	0	0	0	1	**0**	1	0
9	0	0	0	0	0	0	0	0	**0**	0
10	0	0	0	0	0	0	1	1	0	**0**

there were discrepancies of 1–4 in pod-size estimates. Therefore, filtering the quality of matches did not affect the results.

In "good matches," standard-watch observers overestimated pod size in 32 cases (16%) and underestimated in 83 cases (41%) (Table 5). Accordingly, when standard-watch observers recorded a pod size of 1, trackers established that the average pod was actually 1.56; a recorded pod size of 2 should be 2.54; a recorded pod size of 3 should be 3.79; and pods recorded as being greater than 3 averaged 4.45 in size; however, because 5.03 was the average recorded pod size for pods with >3 whales, the standard-watch records indicated that large pods should be reduced by 0.58 (Table 6).

Discussion

Our first objective, to develop and test a reliable, efficient method for tracking whales, was achieved with equivocal success. Although it proved difficult to

Table 6. Pod-size estimates of gray whales migrating past a shore-based counting station, near Granite Canyon, California, compared to pod-size estimates by teams tracking the whales through the viewing area (considered here to be the "true size" of the whale pods). The rightmost column indicates the percentage of the sightings that were recorded as being in the respective pod size (from Rugh *et al.* 2005).

Pod-size estimates	Means of "true size"	SE	*n*	*t*	*P* (two-tail)	Bias	%
1	1.56	0.09	96	−6.53	<0.001	−0.56	62.0
2	2.54	0.16	52	−3.34	0.002	−0.54	24.2
3	3.79	0.34	24	−2.33	0.029	−0.79	8.3
>3[a]	4.45	0.38	33	+2.27	0.030	+0.58	5.5

[a] Mean pod size for all pods larger than 3 was 5.03.

consistently follow a pod of whales through the viewing area, this test did provide an empirical record of just how difficult it is to document whale sightings. Only half of the attempts to track a whale group were considered good enough to be used in comparisons with the standard watch, and it is the standard-watch data that are applied to abundance estimates used in management decisions. Much of the initial effort in our analysis was to filter sighting records to maintain only good-quality matches. Because the tracking sessions and standard watches were conducted from the same site by the same observers (in rotation) using the same tools (reticled binoculars), both the trackers and the standard-watch observers had the same perspective of the whale pods.

Trackers provided relatively more accurate sighting records than can be expected from observers on the standard watch because they had several advantages: (1) open communication between members of the tracking team allowed them to share sightings or opinions of links between sightings; (2) a dedicated observer was able to search for whales without the distraction of looking down to record data; (3) the dedicated recorder could increase data precision by recording accurate times and immediately reviewing the data; and (4) the trackers could focus on one pod at a time, watching each for approximately half an hour, instead of maintaining a vigilance across most of the field of view. Paired tracking teams independently conducted concurrent tracks to test the repeatability of this effort. If each whale surfacing within the viewing area had a 100% probability of being detected by a tracking team, then two tracking teams following the same whale pod should have identical records. However, there were evident differences between paired records, which indicated how difficult it was for even trained observers to track whales. Only 54% of the tracks were considered good, and, of these, only 71% of the pod-size estimates were the same between the paired teams.

The second objective—measuring precision in time and location data—was met through a variety of tests: location data taken daily on a static target and a calibration test on a vessel in the viewing area; comparisons of sighting data between paired observers who focused on one whale pod at a time; and comparisons between observers on the standard watch relative to the trackers. Precision in recording time, bearing (horizontal location), and reticles (vertical location) have provided boundaries for comparing sighting records. It is fair to assume that the tracking team collected relatively more precise data than the standard-watch observers because trackers had a dedicated recorder who made a real-time plot of each whale's track as a check of the location data, while standard-watch observers had to minimize time spent looking down at their recording sheets. Matches between the standard-watch and tracking sighting records were examined manually during the field season and were later checked and then compared to a computerized matching algorithm. The combination of these analyses made it highly probable that all appropriate matches were found. Furthermore, setting bounds on acceptable matches (eliminating tracks if TQ was >3 or VIZ was >4, and eliminating the 5% outer bound of comparisons in time, bearing, or reticle) maximized the probability that appropriate matches were made.

The third objective, a test of accuracy in pod-size estimates, was attempted by comparisons between the standard-watch and tracking records. These comparisons showed that all but large pods (>3 whales) were underestimated by standard-watch observers (Table 6). This may be a function of the demands placed on the standard-watch observers, who must search for whales, make judgments on resightings, collect sighting data, keep track of multiple pods simultaneously, and then

Table 7. Comparison of corrections to pod-size estimates of gray whales migrating past a shore-based counting station near Granite Canyon, California.

Pod-size estimates	Aerial survey[a]	*n*	Aerial survey[b]	*n*	Thermal sensor[c]	*n*	Trackers[d]	*n*
1	+0.350	225	+0.941	102	+0.36	106	+0.56	96
2	0.178[e]	101	+0.646	82	0[e]	61	+0.54	52
3	0.350[e]	28	+0.607	28	0[e]	45	+0.79	24
>3	+0.333	27	+0.250	28	+0.35	30	−0.58	33
Total matches		381		240		242		205
Total pods		62		66		242		120

[a] Reilly (1981), including results applied in Buckland *et al.* (1993).
[b] Laake *et al.* (1994).
[c] DeAngelis *et al.* (1997).
[d] This study.
[e] No significant differences.

record the data on sighting forms. During particularly busy times, standard-watch observers based their pod-size estimates on very few surfacings.

Other studies have attempted to provide more accurate estimates of gray whale pod sizes relative to estimates made by observers on the standard watch (Table 7). Reilly (1981), conducting aerial observations of whale pods, established that pod-size estimates of two or three whales as seen from shore were accurate enough on average that corrections were not necessary, while single whales or pods recorded as four or more should be corrected by +0.35 and +0.33, respectively. Laake *et al.* (1994), conducting aerial observations similar to those done by Reilly (1981), found that each pod-size estimate needed corrections, and the size of the corrections diminished as the size of the pods increased. Pod-size estimates from thermal-sensor data (DeAngelis *et al.* 1997) matched standard-watch estimates 70% of the time, considerably more than the 43% agreement between standard-watch observers and trackers in our study. These data from thermal sensors resulted in no significant differences between methods when standard-watch observers recorded pods of two or three whales, but the thermal sensors found more whales in pod sizes recorded by standard-watch observers as one whale (+0.36) or four or more whales (+0.35). Their results are nearly identical to those of Reilly (1981) but different from those collected by Laake *et al.* (1994) or this study. Each of these calibration methods (aerial, thermal sensor, and tracking teams) has advantages: aerial views give a very accurate count of whales in a pod; thermal-sensor video tapes allow for multiple reviews of each sighting and provide efficient sighting data in the same field of view that was searched by observers on the standard watch; and tracking teams can follow a single pod for over half an hour in the same viewing range as the standard watch, helping maximize the probability of matching the two records. However, each method has problems. During aerial surveys, it is hard to establish which whale pod is being circled by an aircraft relative to what is being seen from shore, and distances between whales are harder to discriminate from shore, making pod sizes more inclusive of nearby whales than would be apparent from an aerial view. Furthermore, aerial calibrations tend to draw attention to the circling aircraft and may bias upward the amount of time shore-based observers watch pods in that area. Aerial operations are increasingly more expensive, and a few days of aircraft time could cost almost half of the budget for an entire season of whale counts by shore-based observers. The accuracy of pod-size estimates made from thermal sensors was limited to a 6.8° field of view, effectively watching the average migrating whale for only 240 m (Perryman *et al.* 1999). And the tracking effort was beset with the vagaries of sighting conditions also experienced by observers on the standard watch.

Among all the studies of pod-size estimates, there has been a fairly consistent agreement that observers tend to underestimate the size of whale pods, but results from the different studies indicate a wide range of corrections. The aerial study conducted by Laake *et al.* (1994) was the source of bias corrections that have been applied in the most recent abundance calculations (Rugh *et al.* 2005). Calculations of abundance of gray whales include a sizable correction for bias in recorded pod sizes (Rugh *et al.* 2005), emphasizing the importance of finding appropriate correction factors.

Our analysis of gray whale counts made at a shore-based station should be applicable to other shore counts of migrating gray whales, such as those at Point Vicente, California (conducted by the American Cetacean Society/Los Angeles Chapter [http://www.acs-la.org/GWCensus.htm]), and at Piedras Blancas, California (conducted by NOAA's National Marine Fisheries Service (NMFS) during the northbound gray whale migration; Perryman *et al.* 2002). Our methods to check observers' whale counts could also be applied in shore-based counts of other migrating whale species, such as bowhead whales (*Balaena mysticetus*) migrating past ice-based counting sites near Point Barrow, Alaska (*e.g.*, George *et al.* 2004), or humpback whales (*Megaptera novaeangliae*) migrating past sites in the Fiji Islands (http://www.whaleresearch.org/update_006.htm) or South Africa (*e.g.*, Findlay and Best 1996). Furthermore, these methods of checking observers' sighting data may also be applicable to ship-based surveys where the ship passes by the whales instead of the whales migrating past a shore station. In conclusion, we hope that this study can be used in many other surveys to improve the accuracy of whale counts.

Acknowledgments

The National Marine Mammal Laboratory (NMML), NMFS, NOAA, provided support for this study with funds directed through the NMFS Office of Protected Resources' Marine Mammal Research Plan. Observers on the standard watch in 1997–1998 were Cynthia D'Vincent and Melissa Scillia; members of the tracking teams (who also conducted standard watches) were Lisa Baraff, Scott Hill, Rod Hobbs, Jim Lerczak, Marcia Muto, David Rugh, Kim Shelden, and Janice Waite. Doug DeMaster provided administrative overview. Max Puckett and Jim Lytle provided hospitable support for our use of the Granite Canyon research station, run by the State of California Department of Fish and Game as their Marine Pollution Laboratory. This document was improved by reviews by Kim Shelden (NMML) and Janice Waite (NMML) and input from Steve Reilly and Wayne Perryman (both from Southwest Fisheries Science Center) and an anonymous reviewer.

Literature Cited

Buckland, S. T., J. M. Breiwick, K. L. Cattanach and J. L. Laake. 1993. Estimated population size of the California gray whale. Marine Mammal Science 9:235–249.

DeAngelis, M. L., T. Martin and W. L. Perryman. 1997. Pod size estimates studied through thermal sensors. Contract report to National Marine Mammal Laboratory, National Marine Fisheries Service, NOAA, 7600 Sand Point Way NE, Seattle, WA. 7 pp.

Findlay, K. P., and P. B. Best. 1996. Estimates of the numbers of humpback whales observed migrating past Cape Vidal, South Africa, 1988–1991. Marine Mammal Science 12:354–370.

George, J. C., J. Zeh, R. Suydam and C. Clark. 2004. Abundance and population trend (1978–2001) of western Arctic bowhead whales surveyed near Barrow, Alaska. Marine Mammal Science 20:755–773.

Hobbs, R. C., and D. J. Rugh. 1999. The abundance of gray whales in the 1997/98 southbound migration in the eastern North Pacific. Document SC/51/AS10, International Whaling Commission, Impington, Cambridge, UK.

Hobbs, R., D. Rugh, J. Waite, J. Breiwick and D. DeMaster. 2004. Abundance of eastern North Pacific gray whales on the 1995/96 southbound migration. Journal of Cetacean Research and Management 6:115–120.

Kinzey, D., and T. Gerrodette. 2001. Conversion factors for binocular reticles. Marine Mammal Science 17:353–361.

Laake, J. L., D. J. Rugh, J. A. Lerczak and S. T. Buckland. 1994. Preliminary estimates of population size of gray whales from the 1992/93 and 1993/94 shore-based surveys. Document SC/46/AS7, International Whaling Commission, Impington, Cambridge, UK.

Perryman, W. L., M. A. Donahue, J. L. Laake and T. E. Martin. 1999. Diel variation in migration rates of eastern Pacific gray whales measured with thermal imaging sensors. Marine Mammal Science 15:426–445.

Perryman, W. L., M. A. Donahue, P. C. Perkins and S. B. Reilly. 2002. Gray whale calf production 1994–2000: Are observed fluctuations related to changes in seasonal ice cover? Marine Mammal Science 18:121–144.

Reilly, S. B. 1981. Population assessment and population dynamics of the California gray whale (*Eschrichtius robustus*). Doctoral thesis, University of Washington, Seattle, WA. 265 pp.

Rugh, D. J., R. C. Ferrero and M. E. Dahlheim. 1990. Inter-observer count discrepancies in a shore-based census of gray whales (*Eschrichtius robustus*). Marine Mammal Science 6:109–120.

Rugh, D. J., J. M. Breiwick, M. E. Dahlheim and G. C. Boucher. 1993. A comparison of independent, concurrent sighting records from a shore-based count of gray whales. Wildlife Society Bulletin 21:427–437.

Rugh, D. J., R. C. Hobbs, J. A. Lerczak and J. M. Breiwick. 2005. Estimates of abundance of the eastern North Pacific stock of gray whales (*Eschrichtius robustus*) 1997–2002. Journal of Cetacean Research and Management 7:1–12.

Swartz, S., M. Jones, J. Goodyear, D. Withrow and R. V. Miller. 1987. Radio-telemetric studies of gray whale migration along the California coast: A preliminary comparison of day and night migration rates. Report of the International Whaling Commission 37:295–299.

Received: 16 August 2007
Accepted: 1 May 2008