

**APPENDIX B
DATA ANOMALIES AND BASELINE ACCIDENT DATA**

To gain a perspective on the NTSB's historical workload levels and output, RAND first needed a detailed baseline of accident rates. Although the NTSB keeps detailed records, the methods the agency uses to maintain and query information have changed over the years. Currently, several data systems, maintained by various offices within the NTSB, contain information about accidents.

Accident records are maintained primarily for archival purposes at the NTSB. Because the data records are rarely used for research, there was no contiguous historical data system that RAND could query to establish trends about the character of accidents over time. Therefore, to build the required database, RAND integrated the many NTSB accident data sets listed earlier in Table 6.3. Cross-checking was also performed against accident data records maintained by the FAA.

The task of integrating these records was a difficult one. Over time, the definition and quantity of information that is gathered following an accident has changed, as have the definitions of key parameters. Because the older records are kept primarily as archived data sets, they have been subject to little quality checking and validation. In addition, the older data sets did not share a common format with the newer data sets.

Some individuals at NTSB who provided assistance with RAND's analysis reported that when a large data entry backlog develops, temporary workers are hired to enter the data into databases quickly with neither oversight nor any assurance of accuracy. Single-point errors and systematic problems are not uncovered until a later analysis is performed. Some analysis has been performed on a limited basis by NTSB staff, and when errors were found the database was corrected. However, quality checks were not extensive and usually involved a small set of records for which paper reports were available for validation.

The AADB is a collection of disparate data sets that have been built at various stages during the evolution of the NTSB's computer and

database management systems. But discontinuity is only one of the problems affecting the quality and effectiveness of the NTSB's accident records. The various problems can be grouped into three distinct areas:

- **Inconsistent data entry practices.** The data record begins when an investigator completes an accident data entry form.¹ These forms often are incomplete when received by data entry personnel, and attempts to follow up for more information seldom yield complete and consistent records. The investigation staff is inconsistent in how they fill out accident data entry forms. Some investigators are meticulous in completing an accurate record, other less so. Interpretation of parameters, such as "severity of damage," is also inconsistent. These parameters are somewhat subjective. Incomplete records pose serious problems when NTSB data resources are later used to perform safety studies. A recent FAA study that attempted to correlate crash statistics with aircraft age and operational factors noted that incomplete data records, and a lack of data fields to enable users to acquire needed information, precluded the development of the required causal relationships (Federal Aviation Administration, July 1998).
- **Missing data.** Several gaps exist in the historical data record. For example, many final report dates are missing during the years 1983 through 1987, and approximately half of the records do not report aircraft airframe hours.
- **Data anomalies.** The NTSB's data records contain some obvious errors and inconsistencies. For example, a computer problem reported during the 1991 to 1992 timeframe was resolved with estimates that leave parameters such as the Final Report issuance date suspect. In general, little correlation exists among the various NTSB databases. When the AADB was compared with other databases, such as the publications record and the

¹NTSB Form 6120.1/2 dated November 1987 is used by pilot/operators to report accidents to the Safety Board. Field investigators dispatched to an accident investigation use an expanded form to acquire data for later entry into the NTSB's electronic database.

Safety Recommendations Information System (SRIS), many discrepancies were discovered.

Some of the analysis RAND had planned to conduct could not be performed because of the limitations of the integrated data record. RAND found numerous anomalies in the integrated record. Spot checks revealed that many errors were simple clerical mistakes. In other cases, the various ways in which information was input by investigators led to inconsistencies in the database.

Another subtle inconsistency in the data relates to the use of the term "report date," which has no precise meaning in the context of the database. "Report date" sometimes means the date when the Board approved the Final Accident Report and, at other times, the date when the document was printed. Sometimes, the only date recorded is the date when the report's findings were entered into the record. These dates are all roughly equivalent, provided each one of these processes occurs within a few days of each other, but as was seen in the 1990 to 1992 period, this may not be the case.

In general, RAND's analysis of the NTSB's record-keeping indicates a need for greater attention to quality control. Individual investigators need to be more systematic in how they input information so that the data are more consistent. If properly maintained, the NTSB's accident data record can be an important information resource for analytical monitoring of trends.