RELIABILITY TESTING IN THE DERIVATION OF PHYSICAL ENVIRONMENTAL INFORMATION FROM HISTORICAL SOURCES

by

Marcia-Anne Faurer

A Thesis
presented to the University of Manitoba
in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy
in
The Department of Geography
University of Manitoba

Winnipeg, Manitoba

ISBN 0-315-63237-2

Canada

RELIABILITY TESTING IN THE DERIVATION OF PHYSICAL ENVIRONMENTAL INFORMATION FROM HISTORICAL SOURCES

BY

MARCIA-ANNE FAURER

A thesis submitted to the Faculty of Graduate Studies of the University of Manitoba in partial fulfillment of the requirements of the degree of

DOCTOR OF PHILOSOPHY

ABSTRACT

This thesis identifies and addresses a deficiency in the existing approach to the derivation of environmental information from historical documentary sources. The deficiency in question is a failure to adequately test the reliability of the methodology. Although the principles of content analysis have been applied in some studies, they are not universally applied, and impressionistic methods of interpretation are still encountered in the literature. This research applies the methodology of content analysis in the derivation of scientific information from descriptive historical sources. It incorporates recent techniques in content analysis as developed in the social sciences, and specifically reliability tests which statistically account for chance occurrences of agreements and provide a means of assessing the quality of the method of the interpretations. The tests are conducted in the context of a case study that employs the 18th and 19th century log books of the Hudson's Bay Company's sailing ships to derive sea ice information. The case study commences with a form of reliability testing that had been traditionally applied by climatologists at the time that the research began. After finding major deficiencies with this approach, the thesis then draws upon principles and techniques of reliability testing developed in the social sciences, and thus applies an approach that had not been employed in historical climatic reconstructions. Krippendorff's agreement coefficient is calculated to assess how well five independent researchers could repeatedly produce the same interpretations of the log books' sea ice descriptions. The coefficient determines the degree to which their agreements exceed what would be expected by chance, and also facilitates the modification of the methods of interpretation. This research illustrates clearly the importance of reliability testing and the close relationship that exists between the reliability of the methodology and the nature and resolution of the derived data. By adopting the concepts and procedures demonstrated in this thesis, the quality and validity of historical climatic reconstructions will be substantially improved.

ACKNOWLEDGEMENTS

My first acknowledgement must be to Alan Catchpole, my advisor, teacher, and friend. He has guided me through two graduate degrees with patience, and has never failed in his support and encouragement. It is difficult to adequately thank him for his time, advice, and wisdom, I can only say that it has been a privilege to have worked with him.

I would also like to thank the members of my committee, Wayne Moodie, Geoffrey Smith, Ross Hartsough, and Geoffrey McBoyle. They have offered freely of their time, and their advice has been greatly appreciated. It has been very encouraging to experience the interdisciplinary cooperation of this committee in achieving a common goal. Their varied areas of expertise enabled me to obtain a broad education.

Appreciation must also be extended to Mrs. Shirlee Smith, Keeper of the Hudson's Bay Company Archives for her cooperation and assistance, and to the Hudson's Bay Company for its permission to consult and quote from its archives.

Reliability tests can only be successfully accomplished by employing qualified coders. I would like to express my gratitude to Danny Blair, Alan Catchpole, Irene Hanuta, and John Teillet who formed a conscientious and hard-working team of coders who made the tests possible. I thank them for their work and valuable comments.

Throughout my studies in the Geography Department at the University of Manitoba, I have received much valuable assistance from Ed Pachanuk and Marjorie Halmarson. I am grateful for their expert advice in cartography and graphics. Acknowledgement must also go to Chris Kirby of the Machine Geography Lab at the University of Manitoba for his computer assistance in the preparation of this thesis. I would also like to thank the departmental faculty, office staff, and graduate students for their assistance and friendship.

I would like to extend my appreciation to Faye Grant for her time and hard work in typing this dissertation, and also for her encouragement, support, and many years of friendship.

Words cannot express my gratitude to my parents, Lillian and Charles Faurer, their guidance and support have provided me with the confidence necessary to undertake and complete an endeavor such as this. I am, and will always be grateful for their immeasurable contribution to this work, and it is to them that this dissertation is dedicated.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Currently one of the most pressing environmental concerns is the prediction of the climatic changes that will occur in the immediate future. These predictions are primarily derived by projecting past climatic trends into the future. In this context, the reconstruction of climates of the recent past is a significant objective in paleoclimatology because it is the recent past that contributes most to predictions of the immediate future. The information used in these reconstructions is obtained from both the direct and the proxy sources of evidence. The latter comprise the traces of past climates found in the organic and inorganic components of the environment as well as in the human record. In general, the sources of evidence used to reconstruct climatic changes in the Quaternary period, can be grouped into four broad categories, glaciological, geological, biological and historical. The strengths and limitations of the specific sources contained in each of these categories are summarized in the following table by providing the temporal, spatial, and climate-related characteristics of each. Some of the sources are described as being able to provide an unbroken, continuous record while others can only offer information on an episodic basis. The resolution of the data derived from each source can also be determined from the column that provides the minimum sampling interval, and the accuracy of the dates of the samples is given as a percent of the sample's true age. The research presented in this dissertation relates to the methods by which climatic information is retrieved from the historical sources of evidence.

SOURCES OF QUATERNARY PALEOCLIMATIC EVIDENCE

Data sources	Variable measured	Continuity of evidence	Potential geographical coverage	Period open to study (years)	Minimum sampling interval (years)	Usual dating accuracy (years)	Climate-related inferences
Ocean sediments	Isotopic composition of planktonic and benthic fossils Floral and faunal assemblages Morphological characteristics of fossils Mineralogical composition and abundance	Continuous	Global ocean, except (for carbonate fossils) deepest zones (below $CaCO_3$ compensation depths)				Global ice volume; surface temperature and salinity; bottom temperature and bottom water flux; aridity of adjacent land areas; prevailing wind direction and strength
		Sedimentation rates (cm per 1000 years)					
		<2	Favored areas	1 000 000+	1000+	±5%	
		2–5	along continental	200 000+	500+	±5%	
		>10	margins	10 000+	50+	±5%	
Ice cores	Oxygen isotope composition Trace chemistry and electrolytic conductivity Fabric	Continuous	Glaciated regions in polar and alpine areas (optimally in dry snow zones)	100 000+	Variable, but optimally 1–10 years for last 10^4 years	Variable, but optimally 0.05% for last 10^3 years	Temperature, accumulation rates, atmospheric composition and turbidity, ice thickness (height), solar output variations
Mountain glaciers	Terminal positions Glaciation levels and equilibrium line altitudes	Episodic	45°S to 70°N	50 000	—	±5–10%	Temperature, precipitation (net accumulation)
Closed basin lakes	Lake level	Episodic	Low to mid latitudes (arid and semi-arid environments)	50 000		±5%	Moisture availability ("effective precipitation")
Bog or lake sediments	Insect assemblage composition Pollen type concentration, geochemical and	Continuous	All continents	10 000+ (common) 150 000 (rare)	~50	±5%	Temperature, precipitation, soil moisture, air mass frequencies
(varved sediments)	sedimentological composition		Mid to high latitudes	10 000+	1–10	+1–10	
Tree rings	Ring width anomaly, density, isotopic composition	Continuous	Mid- and high-latitude continents	1000 (common) 8000 (rare)	1	1	Temperature, runoff precipitation, soil moisture, pressure (circulation modes)
Written records	Phenology, weather logs, sailing logs, etc.	Episodic or continuous	Global	1000+	1	1	Varied

(Bradley, 1985:6-7)

The historical evidence of climatic change derives from the strong impact of these changes on individual people and on human society. This evidence is contained in the written record of human history which includes direct weather observations and indirect sources containing references to weather-related conditions. The direct sources are those that focus specifically on the weather as descriptive accounts of the weather and include sporadic instrumental observations. The indirect sources contain accounts of a variety of

phenological phenomena such as animal migrations, harvests, and dates of blossoming that indicate the passage of the seasons. Records such as account books of crop prices can also provide indirect information about the climate by indicating the success of various crops. Different types of historical sources can also be distinguished by the regularity with which they were kept. Some sources are sporadic through time such as personal letters, occasional reports, and travel accounts. In these sources, references are made to the weather and to indicators of weather at irregular intervals. Other sources were regularly kept and may or may not be specifically devoted to the weather. These chronicles include diaries, journals, and letters written at regular intervals. There are two types of chronicle, they are non-instrumental weather chronicles for example weather diaries and the records in ships' log books, and general chronicles such as private diaries, ships' log books, and commercial and governmental journals. Some chronicles, such as weather diaries, were devoted to the recording of weather conditions, but most were of a general nature and made only passing references to the weather.

The historical sources have many advantages. Unlike the glaciological, geological, and biological sources, they are all forms of communication from one person to another so that while they do require some interpretation, they were created specifically to convey information. A second very important strength that is unique to these sources is that they can provide direct and specific dates. Finally, historical sources focus on a very crucial period of time. It is from the reconstructions of the most recent past that predictions of climatic change for the coming decades will primarily be made. Forecasts of short term climatic fluctuations have a direct impact upon decisions concerning food production, energy needs, and water availability for the immediate future. This is particularly crucial to marginal agricultural regions.

Historical sources of evidence have been used to reconstruct past climates since the early 20th century (Brooks, 1922). The traditional and most common approach to the

extraction of climatic information from these sources has involved intuitive decision-making. Typically, a body of records is read and interpreted impressionistically by individual scholars. Concern about the subjectivity of this approach has periodically been expressed, but this is usually secondary to the reconstruction. Research directed specifically to the development of an objective and standardized method for extracting environmental information from historical sources is seriously lacking. Previous discussions about methodology have related to serving the needs of the particular reconstruction rather than addressing this problem in general terms. This is unfortunate since a general approach would enable the development of basic principles and procedures that can be applied to all historical reconstructions and would probably improve the quality of the derived data and the climatic reconstruction. The research presented here is directed specifically to this methodological question. By presenting a case study which involves a developmental process designed to resolve this problem, the value of a general methodology to historical climatology will be demonstrated. This is the focus of the research presented here in which a case study is employed to demonstrate the development and application of an objective and reliable methodology.

The general procedure applied here was developed in the social sciences and is called content analysis. This research will show how content analysis can be adapted to suit the needs of historical climatology, and specifically, the case study will show how content analysis can be used to objectively derive sea ice data from descriptions contained in 18th and 19th century ships' log books. Of particular concern will be the development and application of reliability tests. These tests assess the degree to which a methodology can repeatedly yield the same results. A high degree of reliability indicates that the method facilitates the objective and systematic extraction of information from the sources. Historical sources of climatic evidence are potentially of great value for the reasons discussed above. However, the achievement of this potential depends on the viability of

the method by which the information is derived from these sources. This dissertation will therefore contribute to the field of paleoclimatology by investigating the techniques used to interpret historical sources and, in particular, to demonstrate how objective methods of interpretation can be developed.

This research is founded upon three fields of study. The social sciences, in particular psychology, have contributed to the theoretical and practical aspects by providing the general methodology, namely content analysis. History has provided both the sources and background knowledge required to understand the historical context. The physical sciences contributed the knowledge of sea ice behaviour required for an understanding of the information derived from the sources. This interdisciplinary approach is reflected in the three preliminary chapters of this dissertation. In Chapter I the principles of historical climatic reconstructions are examined using examples drawn from the substantial body of literature on this topic. These studies are discussed in order to illustrate the valuable contribution that historical sources have made to paleoclimatology, to show the diversity of sources that have been explored, and to demonstrate the need for an established systematic and objective methodology. Chapter 1 will conclude with descriptions of the sources used in this research. The second chapter presents an introductory survey of content analysis written in language that can be understood by the climatologist unfamiliar with the terminology of the social sciences. Chapter 2 presents content analysis as an application of the scientific method to the interpretation of human communications. This is followed, in Chapter 3, with a survey of the principles of reliability testing within content analysis. This is the portion of the content analysis procedure that forms the central theme of this dissertation. It is, therefore, examined separately and in greater detail. The thesis research was developed in three phases, and Chapters 4, 5, and 6 present these sequentially. The final chapter examines the reliability test results from the three phases and assesses the contributions that these tests make to the field of historical climatology.

CHAPTER 1

HISTORICAL CLIMATIC RECONSTRUCTIONS

The field of historical climatology is ever searching for new sources of evidence and a great degree of ingenuity has been applied in the exploitation of a wide variety of sources. These have ranged from sources that bear directly on climatic conditions, such as weather journals, to quite nebulous sources such as the appearance of sky conditions in landscape paintings and allusions to weather in fictional literature, both prose and poetry. These latter two examples do not comprise the mainstream of historical climatic research, but they are worth noting because they illustrate so well the pervasive influence of climate on humans and their propensity to describe the weather in all forms of communications.

Landscape paintings frequently convey clues about the weather in their portrayal of sky conditions, and the state of the ground conditions and vegetation. In one study by Hans Neuberger (1970), for instance, a sample of 12,000 paintings from the United States and Europe from the period 1400 to 1967 were analysed. The purpose was to test whether artists working in different climatic regions reflected those differences in their art. The features included in this study were the sky conditions such as blueness and cloudiness, and the presence of habitable structures for shelter. Although this is an interesting approach, there are certain obvious problems that make this source too subjective for rigourous climatic research. The blueness of the sky and cloud cover may be influenced by the practices of a particular school of art, or may simply be an aesthetic decision. Neuberger noted such problems, particularly in respect to the depiction of habitable

structures. He pointed out that 18th century European artists often depicted Roman ruins merely because it was fashionable to do so. Nevertheless, this study concluded by accepting the hypothesis that since the artist is a "conscious or subconscious chronicler of the environment" who is subject to the effects of climate, there is a correlation between the meteorological elements portrayed in paintings and prevailing weather conditions. Written sources of climatic evidence have included similarly subjective examples from fiction. Chu K'o-chen interpreted Chinese poetry for the period 1100 B.C. to 1400 A.D. in which the allusions to specific plants such as plum trees, bamboo, and orange trees provided information about past climatic conditions in the places where these poems were written (Chiao-min Hsieh, 1976).

These fictional and artistic sources are secondary, however, to the vast body of factual records that exist and which have served as the predominant source of written historical evidence of climatic change. Some of these factual sources are thousands of years old and are quite remarkable in their form as well as their contents. For example, in China during the period from 1400 to 1100 B.C. records were inscribed on tortoise shells and the bones of oxen. These are called oracle bones and they describe genealogies and also contain accounts of agricultural practises, harvests, and rain and snow predictions. The most noted researcher using the vast body of Chinese sources was Chu K'o-chen who is known as the father of meteorology in China (Hsieh, 1967). Chu K'o-chen classified the Chinese sources into four periods three of which include written historical sources, namely the phenological, gazetteer, and instrumental periods. The phenological period covers the years 1100 B.C.to 1400 A.D. This was the period before the development of meteorological instruments and includes written accounts of seasonal phenomena for instance, the dates of first snowfalls, blossoming of plants, and bird migrations. It is interesting to note here that Chu K'o-chen made climatic inferences not only on the basis of the contents of the written records, but also from the materials from which they were made.

For example, some of the earliest of these records were in books made from bamboo. Based on this information and the fact that the Chinese civilization developed along the Yellow River, he concluded that these areas must have been warmer than at present in order to have supported an abundance of bamboo. From the 15th century onwards, the inhabitants of cities, towns, and administrative counties began to keep records which contained climatic information, among other things. This was the gazetteer period from which over 5000 gazettes are available. The beginning of the 20th century marked the start of the instrumented period which is very brief when compared to the earlier three periods. The best source from this period is the Shanghai record of only 80 years, and so the non-instrumental written record has a greater potential to be of greater value in reconstructing past climatic conditions for China. In general, Chu's conclusions highlight certain key concepts concerning the longevity of types of evidence and the scales of climatic change that they detect. The first is that the oldest records provide the most general information about long term changes. In this way the source dictates the time scale and resolution of the climatic information that can be detected. He also found that there were 400 to 800 year periods of temperature fluctuations of 1° C and 2° C; within those periods were smaller fluctuations of 0.5° C to 1.0° C for periods of 50 to 100 years. This general observation can also be applied to other geographic regions and any environmental phenomenon.

Japan also has a rich record of environmental observations because many Japanese customs and traditions were closely linked to certain seasonal events. These included the blossoming of the cherry trees in Kyoto, the freezing of Lake Sewa, and the first snowfall in Tokyo. These sources can provide as much as 1000 years of proxy climatic information (Arakawa, 1957).

British climatologists have also been prolific in the field of historical climatic reconstruction throughout the 20th century. C.E.P.Brooks was one of the first climatologists to work in this field. When he first began, most paleoclimatic information

had been obtained by geologists, and the major climatic anomalies revealed by the geological record were attributed solely to astronomical phenomena such as orbital variations. As a result, his first reconstruction began before the Quaternary period (Brooks, 1922). Brooks' subsequent research, published in *Climate Through the Ages* (1926, revised 1949), includes descriptions of the sources of evidence for the historical period. In this book, he described the following five sources drawn from human history, (1) instrumental records and old weather journals, (2) literary accounts of catastrophic events, (3) religious or folkloric traditions such as the Great Flood, (4) river and lake fluctuations that could have had an historic impact, and (5) records of human migrations and the rise and fall of civilisations. Brooks examined each of these sources critically and thoroughly, and by doing so, he made historical sources more readily available and acceptable.

Brooks' categories of sources were similar to those of Chu K'o-chen. In each, the instrumental period of record was the shortest, yet it appears to be the most desirable. On the surface, thermometric measurements seem to have a much greater potential for detailed reconstructions than diary entries that describe relative temperatures, yet this is not necessarily the case. Another British climatologist, Gordon Manley (1946), published a reconstruction of temperatures for Lancashire, England for the period 1753 to 1945 using instrumental records. The majority of his report on this research was directed to descriptions about the instruments and the different observing practices of the amateurs who maintained those records. Once he had isolated the records, their locations, and their individual problems of measurement, Manley was faced with the task of having to compensate for spatial and temporal gaps and then to find a way to correct the measurements. This latter job required that the types and locations of the instruments be established for each source. Consequently, much of Manley's report contains daunting observations like the following:

> ... the 'Oldham Road' station records the temperature of the air in an enclosed yard surrounded by buildings...
> (Manley, 1946:10)

For each of his sources, the circumstances of the observations was thoroughly researched and the enormity of this process was clearly described.

> The slow process of estimating probabilities, reducing the fixed-hour means, and smoothing all these values into a consistent series took a long time it being necessary to examine each record with care. Some of them were only found after much searching. In all these early printed records there are also misprints; these again may be detected by bringing together the results from several stations.
> (Manley, 1946:13)

Considering that there were 30 stations in northwest England that had records, and that many of the larger locations had more than one recording station, this was a painstaking process. At the end of this work, Manley produced a table of monthly decadal mean temperatures in Lancashire from 1754 to 1940 (Manley, 1946:29). The problems that he confronted in this research clearly demonstrate that seemingly objective and direct sources such as instrumental records cannot always be used directly. A certain degree of interpretation and background information is necessary.

Another British climatologist, H.H.Lamb, refered to several different sources of evidence to reconstruct four Holocene climatic epochs namely, the post-glacial climatic optimum (5000 to 3000 B.C.), the post-glacial climatic revertence (900 to 450 B.C.), the secondary climatic optimum (1000 to 1200 A.D.), and the Little Ice Age (1430 to 1850 A.D.) (Lamb, 1966:59). The sources used by Lamb were primarily biological and geological, but he also included historical sources for the later periods. From these combined sources, information was obtained about temperature, precipitation, and atmospheric circulation. For the earlier periods, the accuracy of the dates and resolution of the derived data were relatively low. When the paleoclimatic evidence was examined in

conjunction with current knowledge of atmospheric processes, however, it was possible to make inferences which led to detailed reconstructions. For example, it was possible to describe the locations of high pressure belts north of the Mediterranean in 5000 B.C., and from this it was then possible to reconstruct the trade winds and monsoon rainfall for the same period (Lamb, 1966:61). As Lamb's reconstructions progressed to the more recent periods, the resolution and certainty of the derived data improved due to the contribution of the historical sources which contained firsthand descriptions and also compilations of instrumental observations. From these sources it was possible to reconstruct the mean sea level pressure for most of the world as far back in time as 1750. Lamb also turned his attention to *An Experiment in the Systematic Treatment of Documentary Weather Records Since A.D.800* (Lamb, 1966:94) in which he expressed concern about the need for techniques required to detect and correct for changes in the quality of historical sources. Another concern was for the quality of the compilations of meteorological observations. This cautious approach to the use of historical records lead to the conclusion that early manuscripts were best able to reveal the thermal conditions of winter weather and the moisture conditions of summer weather. It was presumed that the most extreme conditions in these two seasons would have had a strong enough impact on people's lives so as to prompt them to record those conditions which was done either directly by describing extreme weather, or indirectly by describing for example, crop failures. Lamb also realized that it is critical to the quality of the reconstruction to assess the type and resolution of the data that could be derived from the sources before information is extracted. He restricted his research to coldness and wetness by considering only simple indications of these relative elements "...so as not to make too great demands..." (Lamb, 1966:96). Based on the historical accounts, Lamb devised two indices, a winter severity index and a summer wetness index. The values of each were applied only to records containing evidence which Lamb considered to be clear and unambiguous. Since Lamb does not include an

explanation of the precise method by which he derived the climatic data from the historical sources, it can only be assumed that he took the traditional intuitive approach. Despite this simplicity, his results have been validated by comparison with other evidence (Hammer, Clausen, and Dansgaard, 1980). Although specific temperature and precipitation measurements were not derived, inferences about the prevailing climate could be derived by considering just the extreme conditions. The product of this research was a reconstruction of winter severity and summer wetness for the 1160 years covering the period 800 to 1960 A.D. The difficulties associated with the interpretation of different types of historical sources were discussed by Lamb (1982). Problems which he considered included the provenance of the documents, the "trustworthiness of the reporter", and the establishment of dates.

An important objective of research in historical climatology is the identification of new sources of evidence. In 1970, J.A. Kington and J. Oliver provided information about a very valuable source, namely the log books of sailing ships (Oliver and Kington, 1970) that are housed in the National Maritime Museum in London, England. This collection includes over 5000 log books for the period 1678 to 1809 covering many areas of the world. A problem that is peculiar to this source is that while at sea, the observations continually refer to different locations. This makes it difficult to produce a long term reconstruction for individual locations. Oliver and Kington dealt with this problem by using the log book information to reconstruct daily synoptic patterns rather than as measures of temporal variations at specific locations. Although the log books often contained remarks about the weather in general, wind strength and direction were chosen as the main focus of their research because of its critical importance to sailing. This required the interpretation of the subjective descriptions of wind speed in terms of the range of forces on the Beaufort wind scale. The wind directions were recorded using the 32-point compass and these could be transcribed without subjective interpretation. Beside the wind

information, comments on the state of the sky, precipitation, and visibility were also used. From all of this derived data, daily synoptic maps were produced. Even though these charts contain generalized synoptic conditions, the authors contend that these maps have useful applications including the statistical study of weather types, the reconstruction of the paths of weather systems, and the enumeration of frequencies of recurring synoptic conditions (Oliver and Kington, 1970:526). In addition to the reconstruction of these synoptic conditions, the authors also offered the following general observation of the use of historical descriptive sources.

> There is a frequent assumption that descriptive weather comments are not only imprecise, but also inaccurate. Not always justifiably it is also believed that the degree of unreliability is likely to vary directly with the time one goes back before the adoption of standard recording procedures. So far as descriptive entries are concerned such assumptions are not necessarily valid. The results of the work undertaken so far have certainly indicated a much greater uniformity and accuracy of recording than was initially anticipated.
> (Oliver and Kington, 1970:526)

As a result of work such as that described above, Kington was able to produce a remarkable series of synoptic pressure maps for each day of the period 1781 to 1785 (Kington, 1988).

Much of the work in historical climatology in Britain that followed these earlier landmark studies was done by the members of the Climatic Research Unit at the University of East Anglia. In 1977 they completed a project in which spatial temperature patterns, temperature change, and precipitation patterns were mapped for the period 1000 B.C. to 1700 A.D. in the northern hemisphere (Wigley, 1977). This reconstruction used a wide variety of sources that were grouped into two categories, proxy and historical. The proxy sources were the physical and biological sources of evidence such as glaciers, tree lines, tree rings, pollen, bog layers, river sediments, and ice cores. The historical sources

were comprised of written documents that were either existing compilations of weather descriptions or sources that had not been previously used in climatic reconstructions. The latter included Greek and Roman records (0 to 300 A.D.), Westminster Abbey manorial accounts (1250 to 1400 A.D.), and various other English sources (1450 to 1600 A.D.). As with Lamb's indices, this research also produced generalized reconstructions of warm and cold, and wet and dry periods for the northern hemisphere. This report also provided a detailed account of the physical and biological proxy and historical sources in which the disadvantages of each are discussed. The proxy sources were found to have four disadvantages. The first, and most often cited weakness, is the difficulty involved in establishing their dates. Although this can be resolved, the process of fixing a date can be complex and often establishes only an approximate year. Secondly, even though the climate affects various physical and biological components in the environment, there is usually a delay period between the climatic event and the physical or biological response. When the response time is known, then it is possible to make adjustments. Unfortunately, this is not always the case. The third weakness of the physical and biological proxy sources is that the connection between the climatic condition and the nature of the response involves a number of variables and involves complex interpretations of cause and effect. The last disadvantage cited is the possibility of human influence on the sources. Six disadvantages of the historical sources are given. The first and most common problem is that the non-instrumental weather and environmental descriptions are qualitative and subjective. Secondly, when the historical information is in the narrative form, an element of bias is introduced into the record by the individual observers as was clearly demonstrated by Manley (1946). The third disadvantage attributed to the historical sources is that they often only include relative descriptions such as 'the coldest season in memory' and this involves the individual's perception of the normal conditions. The limited geographic range is another disadvantage since historical sources are limited to areas in

which people have lived and travelled. Dating is also given as a disadvantage for historical sources although this problem is not common. The last disadvantage is that the veracity of these sources can be questionable. Historical sources of information for climatic reconstructions are often in the form of compilations from various other sources rather than firsthand accounts. Where they do consist of firsthand observations, they may be biased to serve ulterior purposes. It is therefore important to examine the origins of the sources thoroughly before extracting climatic information to ensure the veracity of their contents. In Wigley's study (Wigley, 1977), all of the proxy and historical sources were examined and used with consideration given to these questions. As a result, the maps that were produced were very general, but like the earlier studies, more complex inferences could be derived from them.

In 1984, Ogilvie published a reconstruction of the Icelandic climate and sea ice from the Medieval period to 1780 A.D. The data for this reconstruction were derived from a variety of sources including annals, traveller's accounts, official crop and weather reports, weather diaries, personal descriptions, and the Icelandic sagas. Ogilvie stressed the importance of establishing the "historical veracity" of each source. In doing this, she found occasions in which there were "errors, misconceptions, and even forgeries"(Ogilvie, 1984:134). Two noteworthy advantages of historical sources were also identified. The first is that contrary to the physical and biological sources, environmental information is comparatively easily extracted from the historical sources. Secondly, the use of tree rings is often difficult or impossible in Iceland and therefore, the historical documents were the best available sources. The large and varied volume of information that was available to Ogilvie made the problem of reconstruction very complex. This involved the translation of Medieval Icelandic texts and a careful assessment of the information contained in secondary compilations. The careful evaluation of the provenance and veracity of the sources that

Ogilvie applied in this research resulted in sea ice and thermal indices that were very general.

Historical evidence of climatic change involves not only the descriptions of past climates communicated in the historical record, but also the affect of climatic changes on human activities and in turn, human artifacts and documents contain an indirect record of the climate. It is also possible, therefore, to examine the impact of climate on human activities and thus provide a basis for making inferences about climatic changes from changes in these activities. Conversely, knowledge of historical climatic conditions can help to explain certain historical events. Of the numerous and varied studies of this nature, the following two may be used to illustrate each of these approaches. Emmanuel Le Roy Ladurie (1972) reconstructed the European climate since the year 1000 A.D. by considering the relationship between grape harvests and certain climatic variables. He maintained however that the use of historical weather conditions to explain anything other than short term agricultural events is questionable. John D. Post (1977) used existing climatic reconstructions of the 19th century to explain the extreme economic crisis of the years 1816 to 1819. The generally held explanation for this was that the postwar disruption of trade, manufacturing, and agriculture was responsible. Post argued, however, that this explanation was inadequate for such a widespread crisis and that it was crop failures due to the inclement weather of 1816 that was responsible. This anomolous year, aptly named the "year without a summer" (Stommel and Stommel, 1979), was caused by the eruption of the Indonesian volcano Mt. Tambora in the previous year.

The period of time covered by historical sources of evidence is very brief in North America when compared with the records for Egypt, Asia, and Europe. However, the temporal deficiency of the North American record is largely compensated for by the exceptional quality of the resources of the Hudson's Bay Company. In 1670, Charles II awarded a charter that established a trading company whose domain included Hudson Bay

and all of the land comprising its drainage basin. This was a vast area that included most of Canada. The Company required that meticulous records be kept of all its posts' activities and of all its ships' voyages. The majority of these documents have been preserved and are available at the Company's archives in Winnipeg, Manitoba. Although these records have been the source of numerous climatic reconstructions throughout the past two decades, they have the potential provide additional information about Canada's climatic history. Table 1.1 is a chronological summary of this research spanning a total of 200 years of climatic history for a large proportion of northern Canada. The locations of the places named in the table are shown in Figure 1.1. Because of the northern locations of most of these studies, temperature and variables related to temperature have been the major focus of the research. Most of these studies followed the same procedure as those discussed above and were faced with the same problem of subjectivity. A review of many of these studies is available in an Environment Canada report by A.E. Hoeller (1982).

The research described in this dissertation employs the log book collection of the Company's ships which is one source from the Hudson's Bay Company's records. Although the case study contained in this dissertation is directed to one specific source, the log books, and to one environmental variable, sea ice, the methodology that it develops is applicable to all historical reconstructions using documentary sources of evidence. While the published research literature generally exhibits a close concern for the quality of the sources, it also demonstrates a much weaker concern for the methods applied in the extraction of climatic information from these sources. This research therefore, is a response to the need to address the methods of historical climatology, and its findings are directly relevant to all of the studies that have preceded it and to those that will follow. To serve as a setting for the case study, a description of the Hudson's Bay Company log book collection will be given here.

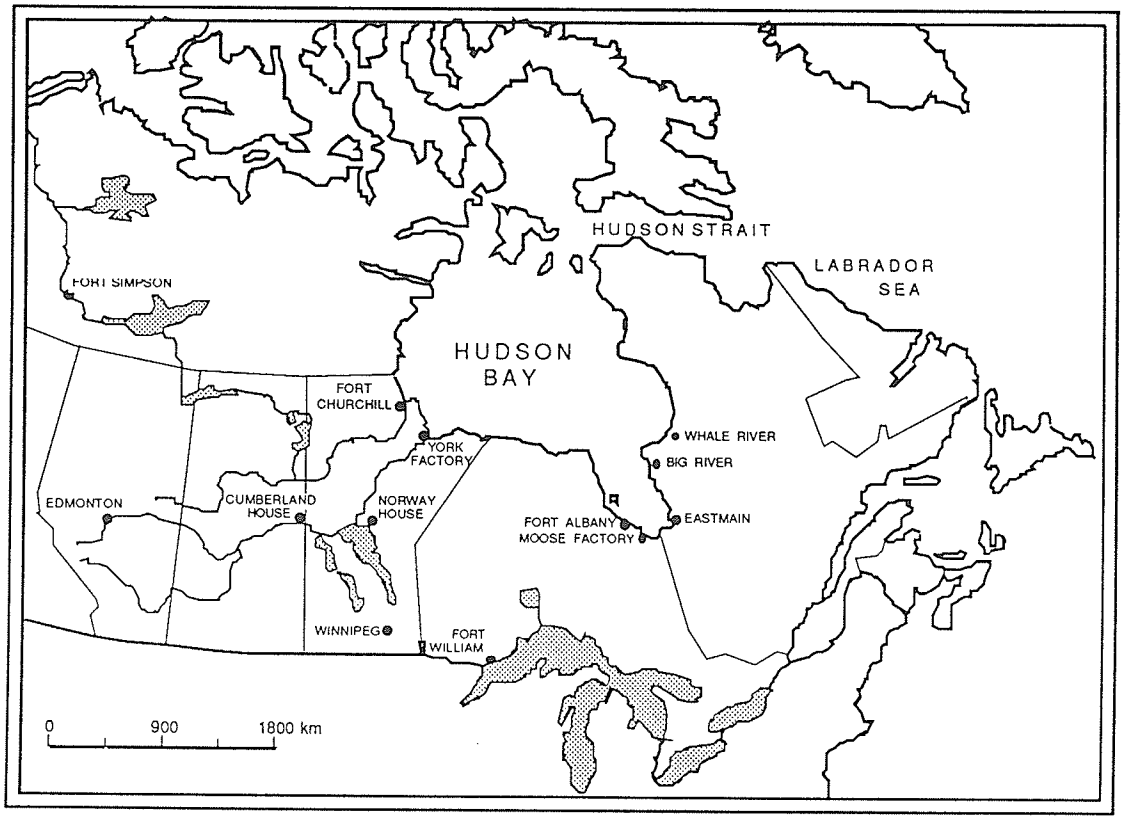
TABLE 1.1 CLIMATIC RECONSTRUCTIONS USING HUDSON'S BAY COMPANY RECORDS

Year	Author(s)	Source(s)	Period Reconstructed	Climatic Variables	Location of Sources
1965	MacKay & Mackay	Post journals	1714-1939	River freeze-up & break-up	Churchill & York
1970	Catchpole, Moodie & Kaye	Post journals	1775-1870	First freeze-up & break-up of rivers	Norway House Edmonton House & Cumberland House
1970	Minns	Post journals	1824-1851	Air mass frequency	Ft. Simpson, Edmonton House, Winnipeg, & Ft. William
1975	Moodie & Catchpole	Post journals	1714-1870	Dates of river freeze & break-up	Churchill, York Factory, Moose Factory, Ft.Albany
1981	Faurer	Log books	1751-1870	Sea ice	Hudson Strait
1981	Madison	Post journals	1705-1870	Dates of first snow & first frost	Ft.Albany & Moose Factory
1981	Magne	Post journals	1743-1940	Dates of freeze-up & break-up	Ft. Severn & Eastmain
1982	Wilson	Temperature records	1814-1821	Temperature	Great Whale R., Big R., & Eastmain
1983	Catchpole & Faurer	Log books	1751-1870	Sea ice & atmospheric circulation	Hudson Strait
1983	Rannie	Post journals	1815-1908	Dates of freeze-up & break-up	Red R. at Winnipeg
1983	Wilson	Temperature records & post journals	1814-1821	Summer temp-erature, wind, precipitation'	Great Whale, Eastmain, Big R.
1984	Ball & Kingsley	Temperature records	1768-1910	Temperature	York & Churchill
1985	Ball	Weather journals & post journals	1715-1805	# days with rain, snow, thunder & lightning, wind, cloud, & frost	York & Churchill
1986	Ball	Samuel Hearne map	1772	Boreal forest / tundra transition	Canadian treeline

TABLE 1.1 continued

Year	Author(s)	Source(s)	Period Reconstructed	Climatic Variables	Location of Sources
1987	Catchpole & Halpin	Log books	1751-1870	Summer sea ice severity	Eastern Hudson Bay
1988	Teillet	Log books	1751-1870	Summer sea ice & Icebergs	Labrador Sea
1988	Wilson	Post journals, correspondence & annual reports	1800-1900	Summer thermal & wetness indices	Great & Little Whale Rivers, Big River, Eastmain
1989	Catchpole & Hanuta	Log books	1751-1870	Summer sea ice after volcanic eruptions	Hudson Strait & Bay

FIGURE 1.1 MAP SHOWING LOCATIONS OF SOURCES IN TABLE 1.1

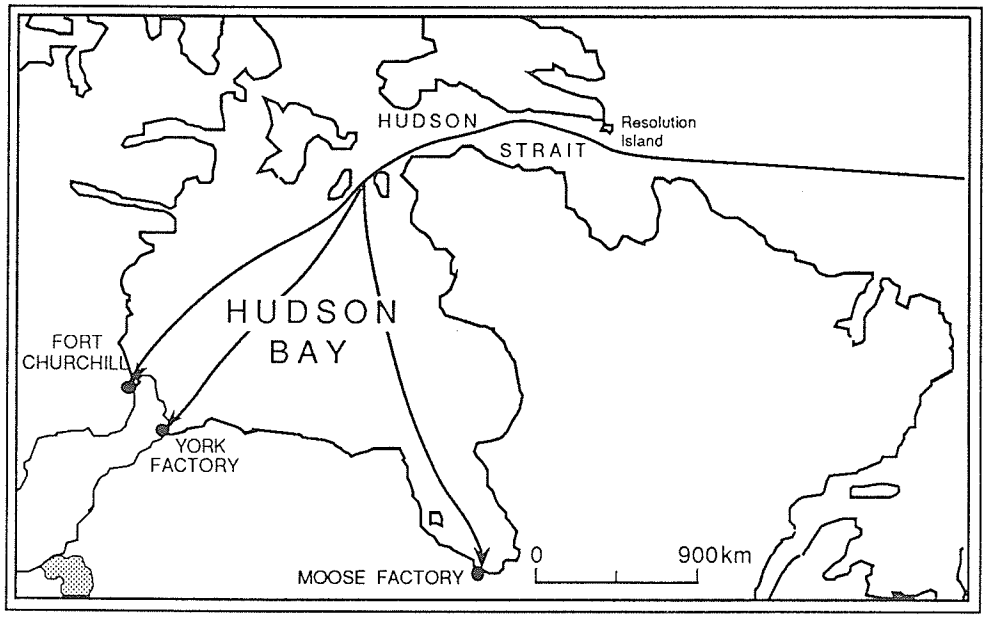


The documents used for this case study are a part of an exceptional collection of historical environmental information. The ships' log books of the Hudson's Bay Company form a complete and meticulously kept record for the period 1751-1870. The scope of this collection has been described in a number of sources (Catchpole and Moodie, 1978; Catchpole, 1980; and Faurer, 1981) but the attributes of the collection will be addressed here to provide background information about the data source for the case study which follows. and to stress the high quality and potential of this source as has been the accepted practice in previous reconstructions.

The log book observations were entered by the crew or captain on board the ship. They are therefore eyewitness accounts and are neither compilations nor second-hand descriptions. Another important characteristic is the continuity of the record for the time period and geographical area of the reconstruction. The HBC's ships' log books fulfill all of these requirements. The collection covers the period 1751 to 1870 and due to the Company's meticulous care in preserving its records, there is only one gap of three years, 1839, '40, and '41 otherwise there is at least one log book. The Company's merchant ships were dispatched each year from England to supply its trading posts on the shores of Hudson Bay, and they returned to England in the same year. In most years, more than one ship sailed at a time in a small convoy so that several posts could be served, and a log book was kept on each ship. In fact, in the period 1751 to 1870, 313 ships yielded log books (Catchpole and Moodie, 1978). Another 169 log books can be added to this number to account for those cases when more than one crew member on the ship kept a log book. In total, then, this collection contains 482 log books (Faurer, 1981). Although it is not necessary to use all of these logs for a reconstruction, the overlap is useful as a means of checking the accuracy of questionable observations and it also allows the selection of the most detailed and legible log book for each year. The ships approached Hudson

Strait by setting their sights for Resolution Island at the eastern end of the Strait. They approached the Island from about 58° N in order to avoid Cape Farewell, Greenland since its position had not yet been firmly fixed. Hudson Strait was entered just south of Resolution Island. From there, they hugged the south shore of Baffin Island as closely as possible to avoid the majority of sea ice that tends to drift to the south with the current as it heads out to sea. The western half of the Strait widens considerably, and once the ships passed Big Island they sailed toward the north shore of the Labrador Peninsula. When they arrived at Mansel Island, the convoy parted company and each ship sailed into the Bay bound for their own destinations. A generalized depiction of this route is given in Figure 1.2.

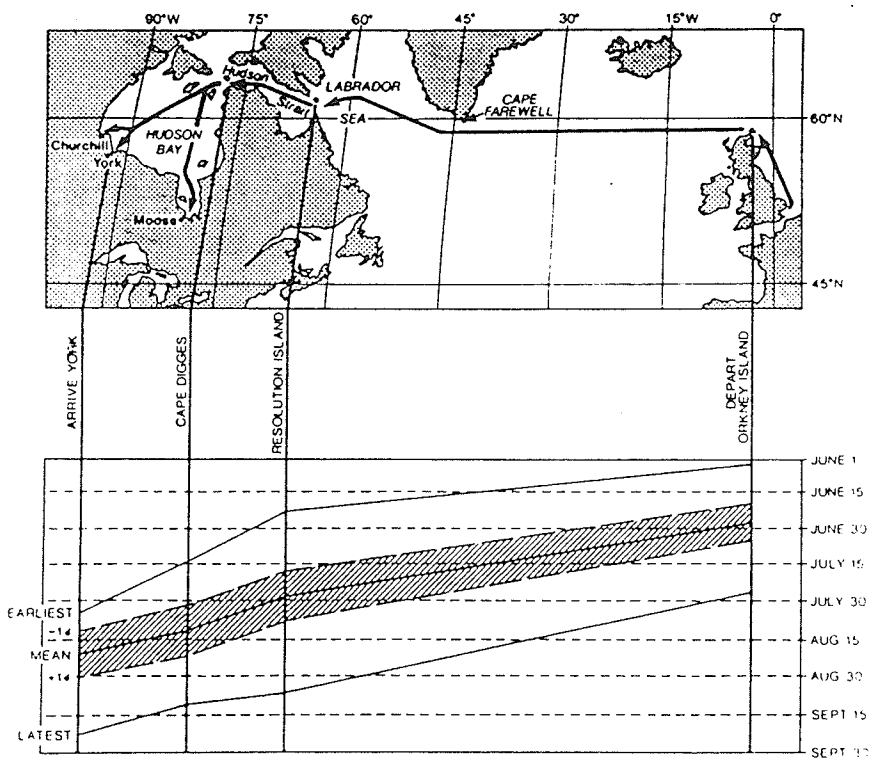
FIGURE 1.2 HUDSON'S BAY COMPANY SAILING ROUTE



(After: Catchpole and Faurer, 1985)

The time of year in which the ships sailed was also a fairly constant factor throughout the period. Their departure date from England was selected to ensure that the ships could arrive at the posts, conduct their business, and leave the area before the ice prevented a return voyage. To allow this to occur, the ships were in the Strait and Bay during a crucial period of the sea ice season on their westward voyages. The average date on which Resolution Island was sighted was July 27 (Faurer, 1981) which is during the ice clearing season. The consistency of the dates of sailing can be seen in Figure 1.3 which shows the mean dates on which the ships were at certain points along the westward portion of the voyage, the earliest and latest dates on which those dates were reached, and ± one standard deviation from the mean.

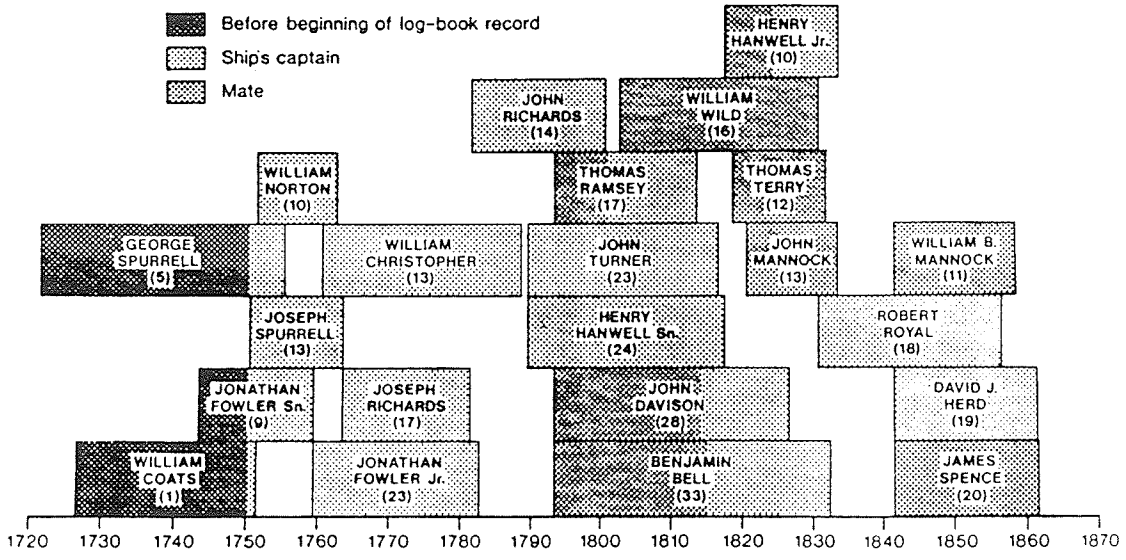
FIGURE 1.3 DATES OF ARRIVAL AT SELECTED POINTS ALONG THE SAILING ROUTE OF THE HUDSON'S BAY COMPANY SHIPS



(Catchpole, in press)

All of these factors contribute to the spatial and temporal homogeneity of this collection, but its greatest attribute for the purpose of climatic reconstructions the continuity of their form and contents. The high degree of similarity among all of the log books for the 119-year span of the collection is due to the fact that the Company guarded its records very closely due to the hostile atmosphere created by the commercial competition with the French for the fur trade in Canada. As a result of the Company's policy of secrecy, the number of ships' captains throughout the period was small, in fact 31% of the log books were kept by only six officers. In most cases, the records overlapped as shown in Figure 1.4.

FIGURE 1.4 DURATION OF RECORD-KEEPING FOR HUDSON'S BAY COMPANY OFFICERS



* Numbers in parentheses indicate the number of log books per officer.

(Catchpole, in press)

Each officer had a long record of service with the Company and, in some cases, the special navigational skills required to sail through the ice-congested waters of Hudson Strait and Hudson Bay were passed on from father to son. The log book collection, therefore, represents a closed system in which the captains worked together, and this naturally resulted in the high degree of homogeneity of the format and contents of the log books.

In summary then, the HBC log book collection is virtually a continuous record for the period 1751 to 1870, kept by captains and crewmen who recorded firsthand observations of environmental conditions. There is also a certain degree of yearly overlap, and the record has both seasonal and geographical continuity.

CHAPTER 2

CONTENT ANALYSIS

This chapter will address the definition, development, theoretical framework, and procedures of content analysis. Since the focus of this research was methodologically-oriented, it is important to discuss thoroughly the framework which formed the basis of that methodology. The Introduction alluded to the conceptual basis of content analysis and this chapter will provide a more detailed discussion of content analysis itself.

2.1 Definition of Content Analysis

The definition of content analysis (CA) seems to be one of its most perplexing aspects. Part of this difficulty may be attributed to the fact that those involved in defining CA are also involved in the analysis of communications. As a result, they have also attempted to analyze the content of the definitions.

One of the earliest and most prominent researchers using CA, Bernard Berelson, defined it as follows:

> A research technique for the objective, systematic and quantitative description of the manifest content of communication.
>
> (Berelson, 1952:18)

The first, and most significant problem with this definition is the use of the word *manifest*. By specifying that the content must be manifest, or evident, the possibility of applying CA to latent content is omitted. One of the greatest advantages and most common applications of this technique, however, is to make latent content evident. Berelson placed another unnecessary restriction on CA by requiring a quantitative description. Although

CA makes it possible to quantify verbal descriptions, this is not possible in all cases, nor is it a prerequisite for a systematic approach or for objectivity.

Common to many definitions of CA is that the technique is *systematic* and *objective* . This was included in Berelson's definition and also in the following, less restrictive definition.

> Content analysis is a research technique for making inferences by systematically and objectively identifying characteristics within a text.
>
> (Stone, et. al., 1966:5)

These two requirements are crucial elements of scientific research. Stone however, introduces the idea of *inference* which provides a reason for describing the contents of the communication. Although this definition is less restrictive and more comprehensive than Berelson's, it still lacks one important element. It does not relate this procedure to other phenomena by providing a statement referring to the purpose of the inferences. Without this aspect, the analysis would seem to exist for its own sake with no concern for the context in which the inferences exist. A later definition by Krippendorff addressed this component.

> Content analysis is a research technique for making replicable and valid inferences from data to their context.
>
> (Krippendorff, 1980:26)

Further amendments to the definition of CA paralleled the progress which took place in the technique itself. Early work in CA amounted to simple frequency counts of specific words or symbols. The premise of this approach was that a word, or words which were used frequently could be inferred as being important to the message. In this light, Berelson's definition was appropriate. As the procedure developed and the inferences which were possible became more sophisticated, the definition needed to be more general and yet more detailed. The increased generality was necessary to encompass

the widening scope of research which employed this technique, and greater detail was required to better articulate the expanding range of goals made possible by CA.

One definition which is less concise but perhaps more meaningful than those discussed above was given by T. Carney eight years prior to Krippendorff's definition.

> Content analysis, then, is a general-purpose analytical infrastructure, elaborated for a wide range of uses. It is intended for anyone who wishes to put questions to communications (pictorial and musical, as well as oral and written) to get data that will enable him to reach certain conclusions...All [content analyses] are more objective than impressionistic assessment of the same question and materials.
>
> (Carney, 1972:26)

This thesis provides another definition which is a product of those which came before, and which will serve as the definition for this research. It is as follows: *Content analysis is a procedure used to derive particular meanings from the content of various forms of communication by the application of the scientific method.* The scientific method is a process of inquiry that involves the identification of a problem, the collection of data by observation and experimentation, the testing of hypotheses, demonstrating that the method is repeatable and validating of the results. One aspect of the scientific method which has contributed to its success is that it is a general procedure and as such, it can embrace many different types of investigation.

2.2 Events in the History of Content Analysis

Although the origin of CA is often attributed to the social science analysis of World War II propaganda , Karin Dovring (1954-1955) discovered a case study using CA from the 18th century in Sweden. This study provides an interesting account of the application of a simple form of CA which was highly relevant at the time, and which was at the core of a religious controversy. This controversy was centered on the *Songs of Zion* which was a

hymnal published in Sweden in 1743. Its first publication came at a time when the established Lutheran Church felt threatened by dissenters who were turning people away from the State Church. The concern of the Swedish orthodoxy was somewhat relieved by the appearance of the Moravian Brethren whom they thought might return the dissenters to the State Church. As a result, they allowed the printing of the *Songs of Zion* even though it differed from the established hymnal. Despite the fact that the Moravians were not fulfilling their anticipated goal, the church granted permission for a second printing which appeared in 1745. Although there was some concern about this, no organized reaction occurred until unauthorized reprints of both editions appeared in 1747 and 1748 in which the wordings had been changed. These publications sparked a debate which resulted in the forced denial by many Moravians of their faith and the exile of those who refused to recant. Dovring isolated the key question of this debate:

> What did these songs say which influenced people to break the law and threaten the power of the State Church?
> (Dovring, 1954-1955:390)

The supporters of the Songs claimed that there was no real difference between them and the official hymns, but the clergy argued that there must be a difference because they elicited a different response. A secular investigation stated that although the Songs included the basic tenets of Lutheranism, the words and ideas that were stressed were contrary to the doctrine of the State Church. This examination of the words and ideas of the *Songs of Zion* was a form of CA. This was followed by a series of similar investigations of varying complexity. One of these used a frequency count of certain words in the Songs. The criticisms of this 18th century CA were also similar to the criticisms of many 20th century studies based on word frequency studies. One primary criticism was that the same message can be conveyed by the use of different words so a frequency count of specific words could not be conclusive. Criticism was also directed to the fact that the frequency count took the words out of context. The final study of this

issue cited by Dovring involved listening to sermons given in churches for revolutionary "ways of expression" and then comparing them with words in the Songs and Moravian writings. It is not surprising that a connection was found. Of this study, Dovring observed that:

> The authorities investigated everyone suspected of Moravian propaganda, using tests well-known to every expert of modern propaganda analysis.
>
> (Dovring, 1954-1955:393)

Another 18th century study, cited by Berelson (1952), was published in the *New Hampshire Spy* (November 30, 1787). It discussed the controversy over the ratification of the United States Constitution. It was noted that the opposition to the Constitution was driven by a class-bias reflected in the frequency of certain words contained in an 'Anti-Federalist' essay:

> ...Wellborn, nine times - Aristocracy eighteen times - Liberty of the press, thirteen times.
>
> (Berelson, 1952:21)

The first modern practitioners of CA were American journalism students. The earliest of these has been attributed by Krippendorff (1980) to G.J. Speed who, in 1893 published an article titled *Do Newspapers Now Give the News?*. This reflected the growing interest in public opinion which was inspired by the increase in the mass production of newsprint and newspapers. The major factor which caused these inquiries to give rise to an early form of CA was the demand for the application of ethics in the development of an empirical procedure. The result was a technique called Quantitative Newspaper Analysis. In this early form of CA, the number of column inches devoted to certain topics were counted. In most of these studies, the goal was to demonstrate quantitatively how the focus of the printed news medium had shifted from 'the news' to 'cheap yellow journalism'.

With changes in the dominant form of communication from print to radio, the form of CA became more complex. Descriptive categories were devised to assess political values, public opinion, and propaganda. World War II and the Cold War therefore created a new need for the systematic analysis of the content of communications. Further advancements were fueled by the emergence in psychology of attempts to measure and analyze attitudes and symbolism in political writing. As statistical tests and behavioral studies progressed, so did CA.

Even though CA continued to progress and to gain broader acceptance, it was not without its critics. The two most common complaints were that it was too simplistic in merely counting certain components of qualitative information and that this quantification was being confused with objectivity. Siegfried Kracauer (1952-1953) put forth the idea that there were real dangers to quantifying descriptive sources. He stated that by rejecting a qualitative approach, there would be a reduction in accuracy. Furthermore, he called for a reorientation of the methodology away from quantification because:

> The potentialities of communications research can be developed only if, as a result of such a reorientation, the emphasis is shifted from quantitative to qualitative procedures.
>
> (Kracauer, 1952-1953:631)

These criticisms persist today and have been responsible for improvements in the technique.

Throughout the 1940's and 1950's, the use of CA in the social sciences and humanities grew in frequency and scope, and in 1955 an international conference on CA drew participants from various disciplines including psychology, political science, literature, history, anthropology, and linguistics (Pool, 1959). The most recent advances in CA have come from the assistance offered by the computer. Although the analytical aspects still lie with those who design the research projects and categories, and those who

interpret the results, the computer has relieved the researcher of the clerical components of CA such as categorizing and counting. CA programs consist mainly of dictionaries, category lists, and definitions.

Continuing from this recent contribution, CA may see future advancement in two directions. First is the continued refinement of two attributes on which CA relies heavily, namely, reliability and validity testing. It is significant to note that researchers have constantly focused on theoretical and practical ways of improving CA through these two components. A second avenue for future advancement of CA comes from the continued expansion of the range of research which employs this technique. The six disciplines represented at the 1955 conference on CA can be expanded to include, for example, environmental reconstructions. With the addition of each new discipline, the technique is enhanced by new perspectives, requirements, and problems.

2.3 Content Analysis Procedure

CA is selected as a research method when the objective is to systematically and objectively obtain information from a form of communication. The procedure commences with the classification of selected segments of the communication into categories that are developed specifically to suit the purposes of each study. Ironically, the CA literature is clothed in jargon making it quite difficult for the novice to extract the basic concepts on which the technique is based, and the procedures it employs. However, when this terminology is interpreted, CA is revealed as a specific application of the scientific method. Accordingly, the procedures of CA are derived from those of the scientific method and involve a systematic method of inquiry that involves these steps, the formulation of hypotheses, the collection of data , the testing of repeatability, and the testing of the validity of the results. This section will discuss the fundamental stages of CA in light of its relationship to the scientific method.

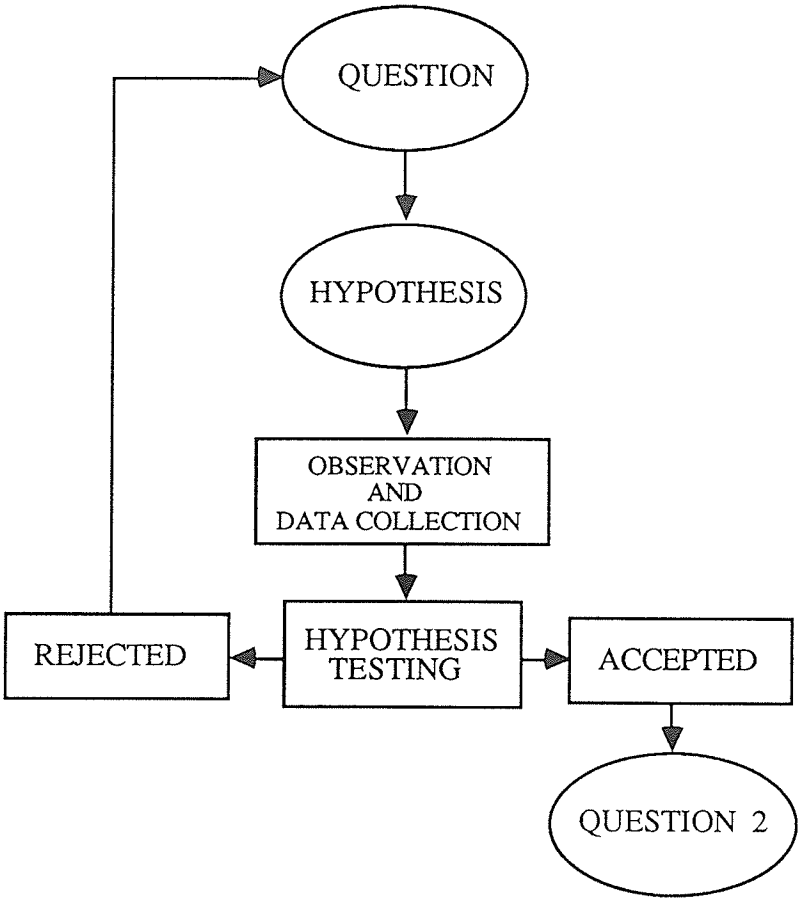
Figure 2.1 is a general depiction of the scientific method and Figure 2.2 is a depiction of the scientific method as it was applied in this research. Regardless of the terminology or the precise details of these two plans, they are the same with respect to certain basic characteristics . They are both multistage processes in which each step leads logically to the next, and they are both self-corrective processes which involve a means of detecting and correcting errors. More specifically, they are also similar because they involve the same five steps even though they may be named differently and are expressed in varying degrees of detail:

- identification of a research question
- hypothesis formulation and testing
- data collection
- new questions arising from acceptance of original hypothesis
- return to beginning to correct errors leading to rejection of hypothesis.

The procedure in Figure 2.2 can be divided into three stages, the first is the preliminary stage which is highlighted in Figure 2.3. In this stage the research question is formulated in the context of the communications to be analyzed and a body of knowledge regarding the physical phenomena under investigation. In this research, for instance, the communications are the log books of the Hudson's Bay Company, and the sea ice descriptions contained within them. The physical phenomena are the sea ice conditions during the summer period of ice dispersal in Hudson Bay and Hudson Strait. These two components contribute to the formulation of the research question. It is important to note that the scientific method requires that the research question arises from pre-existing information, and that the results must in turn contribute to that body of knowledge. In this case, there are two questions:

- Can the sea ice descriptions in the log books be reliably interpreted?
- What sea ice information can these sources provide?

FIGURE 2.1 GENERAL METHOD



(After: Leedy, 1974:8)

FIGURE 2.2 SCIENTIFIC METHOD AS APPLIED IN THIS RESEARCH

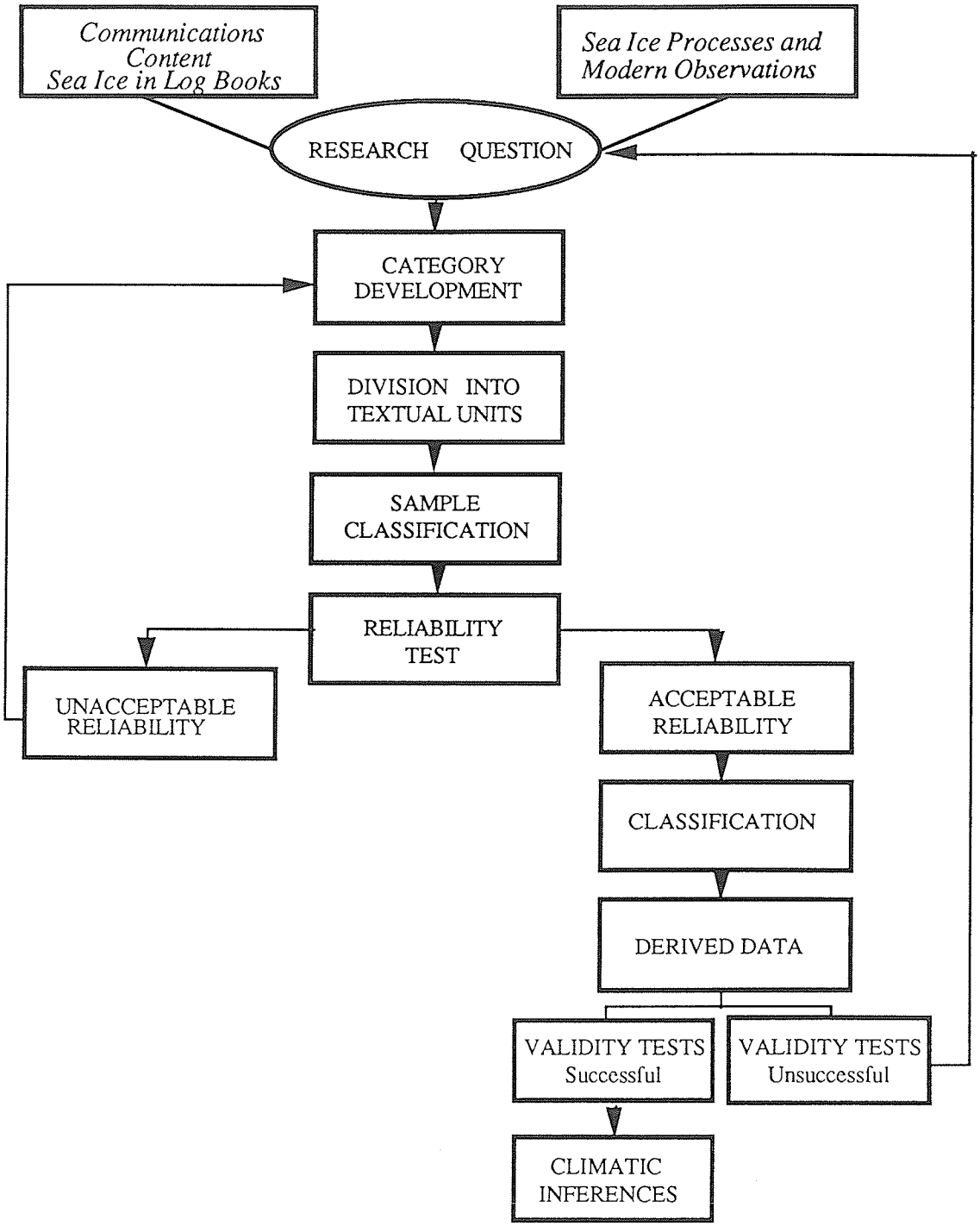
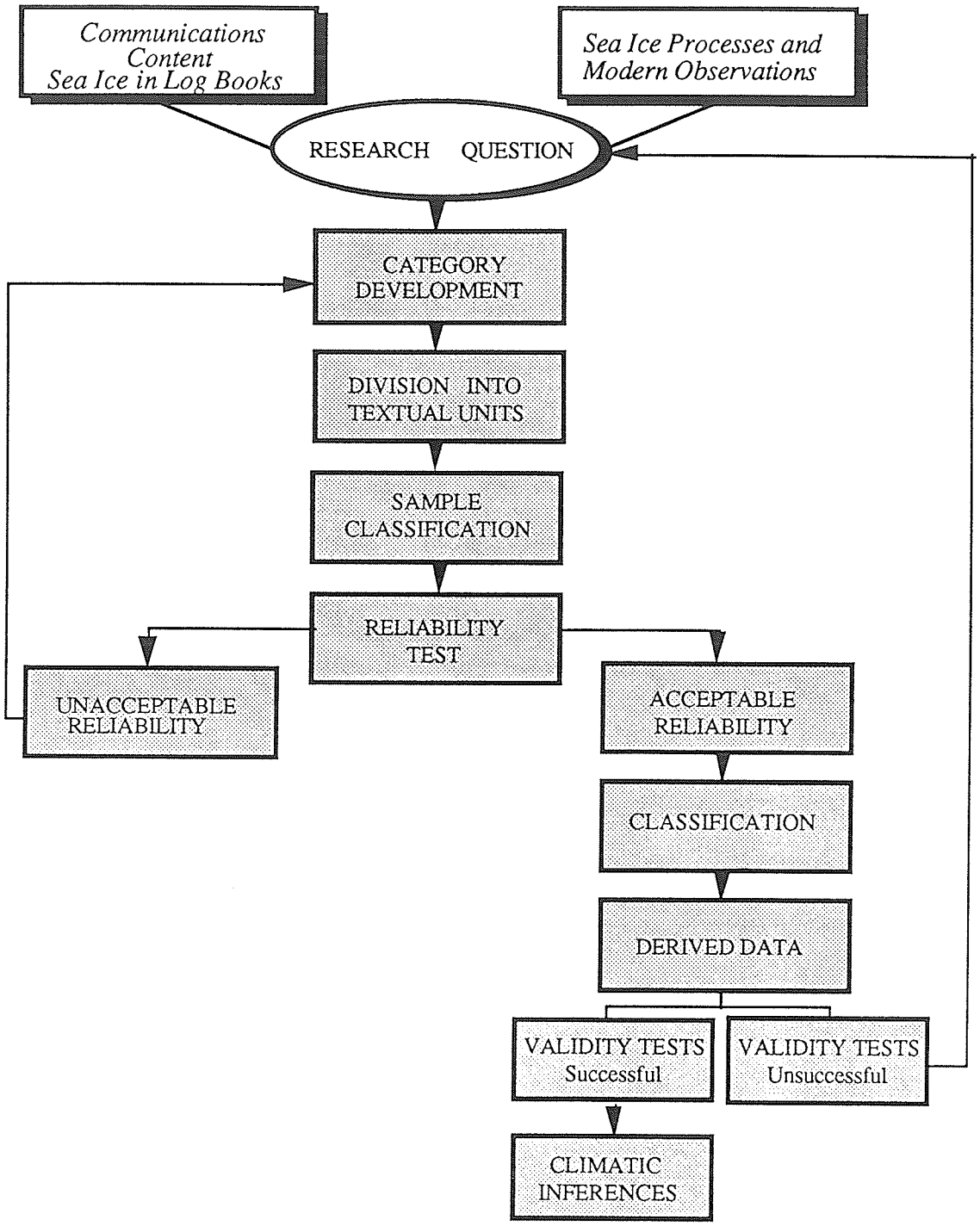


FIGURE 2.3 CONTENT ANALYSIS - PRELIMINARY STAGE



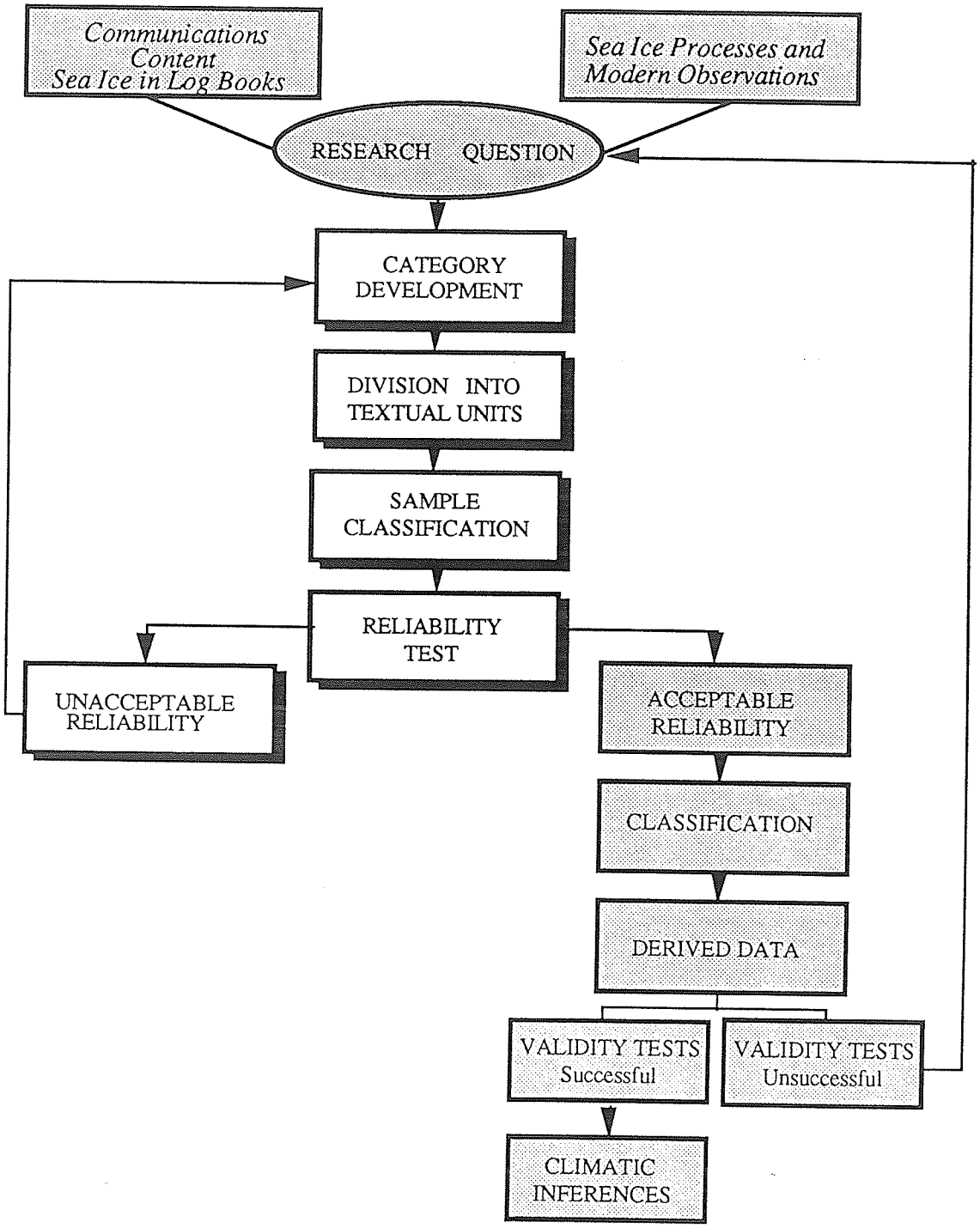
The information from the original communication and the sea ice processes are combined in the formulation of the research question, yet they are depicted separately in figure 2.3. The reason for this can be seen when the precise nature of the information that each contributes is examined separately. The log books are composed of impressionistic descriptions of sea ice as seen from the masthead or deck of sailing ships in the 18th and 19th centuries. They do not, of course, contain standardized ice observations such as those that comprise the modern record. Therefore, the log book contents to be analyzed impose limitations on the research because they do not contain the same type of information as that found in the modern record. Likewise, the standards of modern observation practices place demands on the information to be derived from the log books. Consequently, a major problem in the formulation of the research question, and one that arises throughout this research, is the reconciliation of the conflict between demands and limitations. By a series of experiments, a balance is reached so that all the information that the log books can reliably yield will be obtained.

The second stage (Figure 2.4) is the method used to resolve this conflict. The first step is to develop a set of categories into which the contents of the communication will be classified. This set of categories is formed directly from the research question and therefore contributes directly to the outcome of the research. This important role was clearly stated by Berelson:

> Content analysis stands or falls by its categories...Since the categories contain the substance of the investigation, a content analysis can be no better than its system of categories.
> (Berelson, 1952:147)

Because of its pivotal role in the CA procedure, it is worthwhile to examine the process of category development at this point.

FIGURE 2.4 CONTENT ANALYSIS - DEVELOPMENTAL STAGE

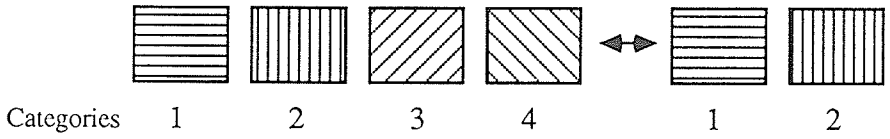


The specific categories that are developed vary with each study because they are derived from the research question. There are, however, three requirements that the categories should fulfill. First, the categories should be derived from the research question which, in turn, was based on the information contained in the log books and the modern sea ice data. Secondly, the categories must be mutually exclusive or unambiguous. That is, the contents being classified must fit into only one category. Finally, the system of categories must be exhaustive so that all relevant elements of the communications can be classified. If this is not fulfilled, then information will be lost. These last two characteristics are common to all forms of classification systems.

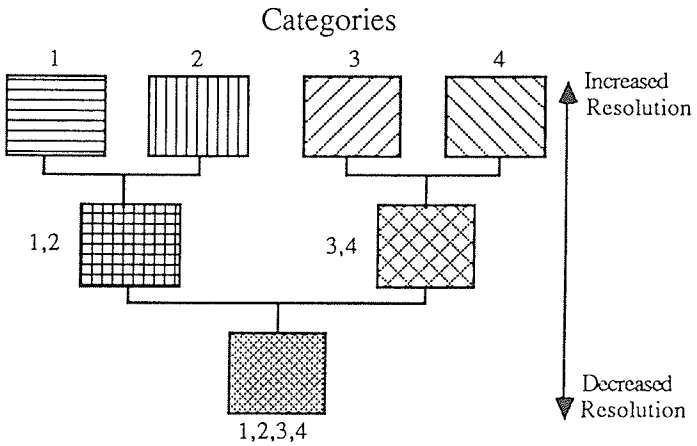
Another factor which must be considered during the process of category development is the number of categories that will be used. This has an effect on the type and resolution of the information that will be obtained. As illustrated in Figures 2.5a and b, there are two ways of altering the number of categories in the classification system.

FIGURE 2.5 MODIFICATION OF CATEGORY SETS

a) Increase or decrease number of categories by adding or deleting.



b) Increase or decrease number of categories by changing resolution.



In Figure 2.5a, new categories are added when the set of categories is not exhaustive; or they are deleted if they are found to be inappropriate. In Figure 2.5b, those categories that are found to be either ambiguous or too specific are combined to produce a smaller, more general, category set. Conversely, if the communications can provide more detail than the classification system allows, then the categories can be subdivided. In both cases (Figures 2.5a and b), increasing the number of categories increases the resolution of the derived data. In all cases, the decision to increase or decrease the number of categories is based on trial classifications using a sample of the communications. Chapters 4 to 6 provide detailed examples of the processes involved in changing the number of categories.

The next step in the second stage of this CA plan is to divide the communication into smaller units that contain the information to be categorized. In the jargon of CA, these segments are called coding units because the process of categorisation is called coding. This more generalized terminology is appropriate when CA is discussed in the broadest sense because the communication may not involve textual material but could include pictures, for example. In the context of this research, however, the communication is textual and therefore the segments are more aptly called textual units. Textual units range in size from words, phrases, lines, sentences, paragraphs, and pages, to chapters and the smaller the textual unit, the finer the resolution of the derived data. The decision regarding which textual unit is most useful is made following a process of experimentation because there are no established rules on which to base this decision. In this research, four textual units were tested at various stages. The purpose of experimenting with these four units was to objectively determine an appropriate unit of text that would contain sufficient information about sea ice to enable the reliable retrieval of useful ice data. This process illustrates the close relationship that exists between the development of the categories and the determination of the textual units. The textual units tested in this research are given in Table 2.1 in the order in which they were tested.

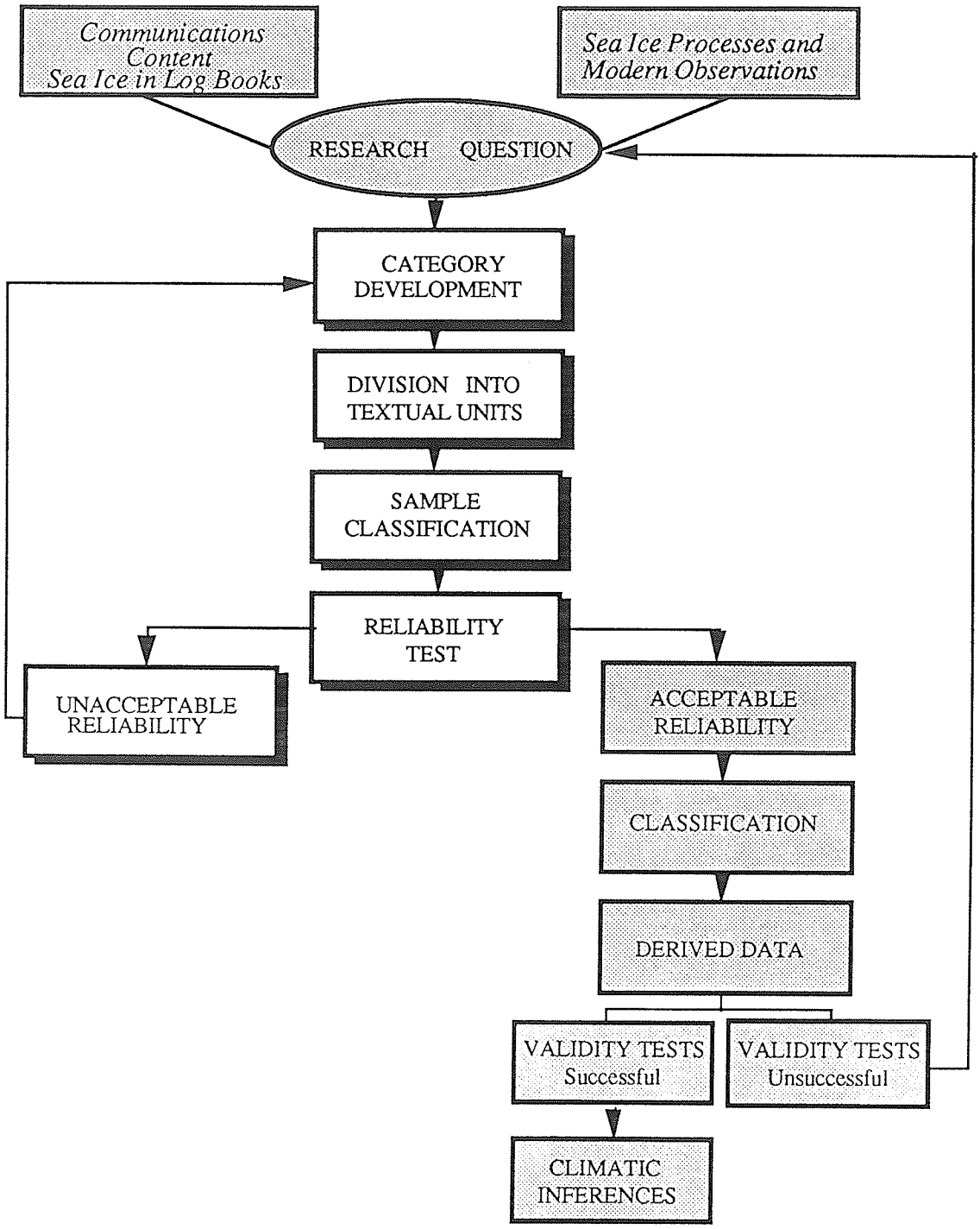
TABLE 2.1 TEXTUAL UNITS TESTED

Textual Unit	Information Contained in Unit
Each log book page	ice descriptions for a 24-hour period.
Individual entries	ice descriptions for every two hours and, more recently, every hour.
Individual words	words used to describe sea ice or the navigational activities employed to deal with the ice.
Seaman's watch	ice descriptions made during 4-hour periods, six per day.

After each of these four units was tested, it was found that the seaman's watch was the most appropriate. The factors that entered into this decision will be addressed in Chapters 4 and 5, but it is important to stress at this point that this was the outcome of experimentation and was not known at the outset.

The next two steps in Figure 2.4, following the division into textual units, are the sample classification and evaluation of reliability. These steps together form an evolutionary trial-and-error analysis which ultimately produces a set of categories that best accommodates the objectives of this research. This is done by a team of independent testers who each apply the same set of categories to the same sample of the communications in order to facilitate the evaluation of the reliability with which the categories are applied by different people. If the evaluation does not reveal an acceptable level of agreement, the categories are modified. The same sample is then reclassified by the testers using the modified categories, and the reliability is evaluated again. This process is repeated until the results of the evaluation are acceptable, and the final stage can begin. Chapter 3 - Reliability, will address in detail, the methods of evaluation and the question of acceptable levels of reliability.

FIGURE 2.6 CONTENT ANALYSIS - IMPLEMENTATION STAGE



The implementation stage illustrated in Figure 2.6 is composed of three steps. The first involves the classification of the entire body of communications using the final set of categories that was developed and tested in the previous stage. This follows the same procedure by which the sample was classified but in this case only one coder is involved. In this research, this step would involve the reconstruction of sea ice conditions using the classified descriptions contained within each seaman's watch. As is the case with any investigation that employs the scientific method, the research does not end at this point. The derived data that comprise the reconstruction must be subjected to validity tests. This is accomplished by comparing these findings with established facts and theory, and/or comparing them with similar data derived independently from other sources. If the results do not pass the validity tests, this may indicate a need to refine the research question or it may indicate a substantive weakness in the original communications. If the validity tests are passed, then this research is added to the body of background information which will fuel future research. Once the derived data, have been validated, they may then be used in paleoclimatic reconstructions.

This section has outlined the procedure of CA, and has examined some of the more important methodological concerns. More detailed descriptions of the CA procedure applied in this research are given in the following chapters.

CHAPTER 3

RELIABILITY

This chapter focuses on the developmental stage in Figure 2.4 namely reliability testing which is the central theme of this research.

3.1 Definition and Functions of Reliability

Reliability is a characteristic of a methodology that describes the extent to which a procedure can yield the same results when it is repeatedly applied. The importance of reliability testing in scientific research was clearly stated by Feibleman.

> The repeatability of scientific experiment is responsible for what has been called the self-corrective nature of the scientific method. What is proposed by an hypothesis cannot be established by a single experiment; the scientific method makes this impossible because of its demand that all experiments shall hold themselves repeatable and that crucial experiments must be repeated.
>
> (Feibleman, 1972:129)

As pointed out in Chapters 1 and 2, environmental reconstructions based on historical sources have often been described as reliable because the authors of the original communications were contemporaries of the event which they described (Oliver and Kington, 1970; Ingram, Underhill, and Farmer, 1981; and Ogilvie,1981). This is an entirely different meaning of the term than that applied in this research. Reliability has also been confused with validity which is a characteristic of the data that are derived using CA. Validity is the degree to which the procedure measured what it was intended to measure and is tested by comparing the research findings with similar data derived from independent

sources. Consequently, a study may have been conducted reliably but yet produce results that are not valid. It is unlikely, however, that valid data would be derived using an unreliable method.

Besides measuring the repeatability of the method, reliability has an important role to play in determining the amount and validity of the information that is obtained from the communications. This can be explained in the following way. In this research, data are derived following the categorisation of the log book descriptions. Therefore, the degree to which the data measure sea ice conditions in the 18th and 19th centuries is a reflection of the degree to which the categories describe sea ice. The most compelling reason for the reduction of the category resolution would be to improve the reliability with which the log book descriptions can be classified into categories. Consequently, the attainment of a high degree of reliability can limit the detail of the information obtained from the communication. Thus, the goal of maximizing both validity and reliability results in a conflict of the demands imposed by validity tests and the restrictions imposed by reliability tests. The deciding factor in resolving this conflict is the reliability of the method. This is because it is important to know that the data were objectively derived before comparing them to the external standards used in validity tests. The search for optimum levels of reliability, then, is a major factor in determining the amount of information that will be obtained in research. This will be demonstrated in the chapters which follow by showing how the application of sequential reliability tests led to the reduction of several fairly detailed category sets. Originally these included many measures of ice concentration, thickness, fragmentation, movement, and openings, but following the reliability testing, they were reduced to four general classes of ice concentration.

There are three types of reliability tests that evaluate different components of the methodology. These include intracoder, intercoder, and standardized tests. A coder is a person who applies the set of categories to the textual units from the communication.

Intracoder reliability is the degree to which one person repeatedly categorizes the textual units of the communications in the same way. The amount of time between the repetitions can be a significant factor in the test results. The optimum time between repetitions depends on the coder's level of experience, the size of the sample being classified, and the complexity of the categories. Thus the interval between tests is increased with increased coder experience, decreased sample size, and decreased complexity of the classification system. This prevents the coder from remembering the decisions of the previous session. This type of evaluation is completely internal because the reliability is determined solely by one coder repeating the process without the use of outside criteria against which the coder's performance is compared.

The objectivity with which a set of categories can be applied is also determined by testing the intercoder reliability. In this case, two or more people independently apply the same categories to the same sample from the communication. Their results are then compared to determine levels of consensus. It is important that all coders receive the same instructions on how to categorize the sample and that the instructions are clearly understood and unambiguous. If this is not the case, then confusion over the instructions may introduce disagreements that are not related directly to the categories or the sources. Intercoder reliability tests are also internal because they are based on the same descriptive sources.

Standardized reliability, unlike the previous two, involves an external measure against which the intra- and intercoder agreements are compared. This external measure is an accepted level of reliability that has been established by previous research. Standardized reliability tests, therefore, evaluate the observed agreements by comparing them with predetermined levels of agreement. Although this test is more objective than the internal tests, external standards are often not available. As a result, the best that can be expected is

that independent coders can repeatedly produce the same results using the same categories and communications.

3.2 Methods of Evaluating Reliability

There are many ways of evaluating reliability which all involve the analysis of inter- and intracoder agreements. It is important to point out here that reliability is a measure of agreement and not association. The difference between these is illustrated by the example in Table 3.1a and Table 3.1b. Table 3.1a contains a matrix of co-decisions between two coders using the same categories and textual units. There is a high degree of association between categories 1and 2, 2 and 4, 3 and 1, and 4 and 3 by the two coders but no agreement between them. Table 3.1b however, shows a situation of complete agreement.

TABLE 3.1a and b
FIGURE 3.1a CO-DECISIONS WITH HIGH ASSOCIATION

		Coder A			
	Categories	1	2	3	4
Coder B	1		8		
	2				8
	3	8			
	4			8	

FIGURE 3.1a CO-DECISIONS WITH HIGH AGREEMENT

		Coder A			
	Categories	1	2	3	4
Coder B	1	8			
	2		8		
	3			8	
	4				8

Levels of agreement are affected by three factors, namely, the coders' experience and training, the categories, and the communications being classified. Therefore, the best method of evaluating reliability would be one which not only assessed the level of agreement but which also provided information about those factors which reduced the number of agreements. This would make it possible to improve the reliability by correcting the problems detected by the tests. Of the four methods examined in this chapter, percentage agreements, Scott's pi test, Cohen's kappa, and Krippendorff's agreement coefficient, only the last has this capability.

Percentage Agreements This is a simple ratio of the number of agreements between two coders to the total number of textual units classified by each coder, as follows:

$$CR = \frac{2M}{N_1+N_2} x100 \quad (1)$$

Where: CR = coefficient of reliability
M = the number of agreements between two coders
N_1 & N_2 = number of textual units classified by each coder

(After: Holsti, 1969:140)

This calculation can only provide a general impression of the agreements between two coders, and while this can be of some value, its limitations must be noted. A major problem is that chance or random agreements are not considered. The element of chance can alter the level of reliability to a great degree. This can be illustrated using a simple example involving two coders each classifying 10 textual units into one of two categories, first year ice (I) and no ice (NI) as in the following example:

Textual Units	1	2	3	4	5	6	7	8	9	10
Coder A	NI	I	NI	NI	NI	NI	NI	NI	I	NI
Coder B	NI	I	I	NI	NI	I	NI	I	NI	NI

In the example, the number of agreements between the coders (2M) is 12, and the total number of textual units classified is 10 for each coder. According to equation (1) this results in a coefficient of reliability (CR) of 60%. This value does not, however, provide any assessment of whether 60% is a high or low level of agreement, it is simply a means of describing the number of agreements. To assess the level of these agreements, it is necessary to compare them with a situation in which each classification was made by chance. This can be done by relating the observed disagreements to the expected disagreements as follows:

$$\alpha = 1 - \frac{\text{Observed disagreements}}{\text{Expected disagreements}} \quad (2)$$

Where α = agreement coefficient

(After: Krippendorff, 1980:134)

The observed disagreements in the example above are four, the expected disagreements are determined by calculating the number that would be expected to occur by chance . With the distribution in this example of 14 'no ice' decisions and 6 'ice' decisions, the probability that these two coders will agree in determining the presence of ice is the product of their individual probabilities, 6 out of 20 for the first coder and 5 out of 19 for the second which is 0.08. Therefore the expected frequency of agreement for this category is 0.08 . Similarly, the probability that they will agree on the no ice category is 14 out of 20 for the first coder multiplied by 13 out of 19 for the second. This results in a probability of 0.48. Therefore the probability of a disagreement is 1-(0.08+0.48) = 0.44 and the expected frequency of disagreements is determined by the product of the probability of disagreement and the number of textual units. Therefore, in this example, the expected disagreements would be 4.4. When this value is used in equation (2), the reliability is described quite differently.

$$R = 1 - \frac{4}{4.4} = 0.091$$

This means that although the coders agreed on 60% of their decisions, this level of agreement is only 10% better than what would be expected to occur by chance.

Another problem with percentage agreements is that they are biased in favour of systems with fewer categories. For example, by chance alone, a system with five categories would be expected to have fewer agreements than a classification system with two categories. In an attempt to remedy the latter problem of bias towards fewer categories, an index of consistency was developed.

$$S = \frac{k}{k-1} (P_o - \frac{1}{k}) \qquad (3)$$

Where: S = index of consistency
k = number of categories
P_o = observed proportion of agreements
(Bennett, et. al., 1954:307)

While this equation does include the number of categories, it has been criticized because it assumes that each category has the same probability of being used. This problem can be illustrated by using the example given in the case of percentage agreement. By using these figures in equation (3), the index is calculated as follows:

$$S = \frac{2}{2-1} (0.6 - 0.5) = 0.2$$

If the proportion of agreements remains the same (0.6) and two more categories are added, but never used, to accommodate second year and third year ice as well as first year ice, then the index of consistency would be increased.

$$S = \frac{4}{4-1} (0.6 - 0.25) = 0.47$$

Scott's P_i Test Scott's pi test (4) accounts for the number of categories as well as for their frequency of use.

$$\pi = \frac{P_o - P_e}{1 - P_e} \qquad (4)$$

Where: π = index of intercoder agreement
P_o = observed agreement (%)
P_e =agreement expected by chance (%)
(Scott, 1955:323)

The percentage agreement expected by chance is based on the proportional frequency with which each category was used.

$$P_e = \sum_{i=1}^{k} P_i^2 \qquad (5)$$

Where: k = total number of categories
P_i = proportion of the textual units in the i^{th} category.
(Scott, 1955:324)

Therefore, in the following example, the second and third year ice categories will never be used because all of the ice identified in these examples is first year ice and π accounts for this.

CATEGORY	P_i
no ice	60%
ice	40
second year ice	0
third year ice	0

$$P_e = (0.6)^2 + (0.4)^2 + (0)^2 + (0)^2 = 0.52$$

$$\pi = \frac{0.6 - 0.52}{1 - 0.52} = 0.167$$

This provides a much lower index of agreement than Bennett's (.47) because it takes into consideration the fact that two of the categories were not used.

Cohen's Kappa This method of calculating the degree of agreement between two coders is similar to Scott's solution with the exception that the expected agreements are calculated differently. In determining the value of Pe for π, Scott assumes an equal distribution of responses in each category for the two coders. In Cohen's formula however, Pe is calculated by using the proportions in each category for each coder. The equation for Cohen's κ however, takes the same form as Scott's π.

$$\kappa = \frac{P_o - P_e}{1 - P_e} \qquad (6)$$

(Cohen, 1960:40)

Table 3.2 illustrates the differences between these two reliability coefficients by considering a hypothetical situation.

Both Scott's π test and Cohen's κ determine the percentage agreement that should be expected after chance agreement has been removed yet they produce different results. In the example given in Table 3.2, Scott's π test reveals that 37.5% of the agreements were better than chance given the particular set of categories and textual units, and two coders. Using the same information, Cohen's κ indicates that although the two coders agreed on 60% of the units classified, 41.1% exceeded chance agreement. This raises the question, what are the chances that a third or fourth coder would agree? Scott's and Cohen's tests provide measures of agreement for only two coders, or two repetitions by the same coder. It would be of greater value to the assessment of the reliability, however, if more than two coders and repetitions could be tested.

TABLE 3..2
FIGURE 3.2 COMPARISON OF SCOTT'S pi TEST AND COHEN'S k

Matrix of agreements and disagreements for 2 coders:

		CODER 1					
	Categories	No Ice	1st yr Ice	2nd yr Ice	3rd yr Ice	Totals	Category total for Coder 2
CODER 2	No Ice	8	4	0	0	12	6
	1st yr Ice	0	4	0	0	4	2
	2nd yr Ice	0	0	0	0	0	0
	3rd yr Ice	0	0	4	0	4	2
	Totals	8	8	4	0	20	
	Category total for Coder 1	4	4	2	0		

a.) Scott: $Po = \frac{8+4}{20} = 0.6$ (60% observed agreement)

$$Pe = (\frac{4+6}{20})^2 + (\frac{4+2}{20})^2 + (\frac{2+0}{20})^2 + (\frac{0+2}{20})^2$$

$$= .25+.09+.01+.01=.36$$

$$\pi = \frac{.6-.36}{1-.36} = .375$$

b.) Cohen:

		Coder 1				
	Categories	A	B	C	D	
Coder 2	A	8/20=.4	.2	0	0	.6
	B	0	.2	0	0	.2
	C	0	0	0	0	0
	D	0	0	.2	0	.2
		.4	.4	.2	0	

Po = .4+.2+0+0 = .6
Pe = (.4x.6)+(.4x.2)+(.2x0)+(0x.2) = .32

$$\kappa = \frac{.6-.32}{1-.32} = 0.411$$

Krippendorff's Agreement Coefficient This measure of reliability was derived by Krippendorff (1971). The calculations shown in this section serve to demonstrate, by example, the application of the equations used to calculate the agreement coefficient. The main purpose of the following discussion is to explain how this coefficient describes reliability but first the procedure will be demonstrated.

This method provides a means of calculating the degree to which observed agreements exceed chance occurrences for any number of categories, coders, or repetitions. This coefficient can also be applied to category sets based on nominal, ordinal, interval, and ratio scales. These scales describe the magnitude of difference between adjacent categories. In the nominal scale, there is no quantifiable difference. Each category simply describes a different variable and there is no logical order in which the categories can be arranged. A set of ordinal scale categories is one in which the categories are ranked but there is no specific quantifiable difference between them. They are, therefore, relative measures. Interval scale categories are those in which there is a specified and equal value from one category to the next. In the ratio scale, the difference between the categories is specified as a proportional value. The calculations for Krippendorff's Agreement Coefficient are given in equations 7, 8, and 9 below. The adjustments for the four scales will be addressed following the presentation of the basic calculations.

$$\alpha = 1 - \frac{D_o}{D_e} \qquad (7)$$

Where: α = agreement coefficient
D_O = observed disagreement
D_e = expected disagreement

(Krippendorff, 1980:142)

The agreement coefficient will have a value between 0 and 1.0 in which α=0 means that all agreements were made by chance and when α=1.0, none of the agreements were made by chance. Since this coefficient was used in most of the tests in this research, its calculation will be discussed in detail here.

The first step is to total all of the agreements and disagreements for each category and to summarize these totals in a symmetrical form called a coincidence matrix as in Table 3.3. This matrix and the calculations which follow are based on tests conducted in this research.

TABLE 3.3
FIGURE 3.3 INTERCODER COINCIDENCE MATRIX

Categories

Categories	1 Ice Free	2 Open Water	3 Open Ice	4 Very Close Ice	
1 Ice Free	**72**	34	8	0	114
2 Open Water	34	**658**	80	0	772
3 Open Ice	8	80	**618**	67	773
4 Very Close Ice	0	0	67	**48**	115
					1774

In this table, the diagonal entries are the number of times that there was an agreement for each category. There were 72 agreements among five coders on the classification of textual units in category 1; 658 agreements in category 2; 618 in category 3; and 48 in category 4. All of the off-diagonal entries are disagreements. The observed disagreement (D_O) is then calculated using the information in this matrix by equation (8).

$$D_o = \sum_b \sum_c \frac{x_{bc}}{x..} d_{bc} \qquad (8)$$

Where: x_{bc} = number of disagreements in a matrix of categories
$x..$ = total of all agreements and disagreements
d_{bc} = differences between all pairs of categories b and c
(After: Krippendorff, 1980:142)

The calculation of d_{bc} varies with the scale of the categories. In the following example, it will be assumed that the value of d_{bc} is 1 which is the case when a nominal scale is used. Many of the sets of categories used in this research were of the ordinal scale

and this will be discussed after all of the basic calculations for D_o and D_e have been presented. Using the frequencies in Figure 3.3, D_o is calculated as follows:

$$D_o = \frac{34+8+0+34+80+0+8+80+67+0+0+67}{1774} = .213$$

The expected disagreement (D_e) is calculated as follows:

$$D_e = \sum_b \sum_c \frac{x_{b.} x_{.c}}{x..(x..-m+1)} d_{bc} \qquad (9)$$

Where: $x_{b.}\, x_{.c}$ = sum of all marginal products for category totals
$x..$ = total of the marginal entries
m = number of coders (or repetitions for intracoder)

If the values in Figure 3.3 are used, the expected disagreement is calculated in this manner:

$x_{b.}x_{.c}$ =
114x772 = 88008
+114x773 = 88122
+114x115 = 13110
+772x773 = 596756
+772x115 = 88780
+773x115 = 88895
+115x773 = 88895
+115x772 = 88780
+115x114 = 13110
+773x772 = 596756
+773x114 = 88122
+772x114 = 88008
Total **1927342**

and $x..(x..-m+1) = 1774\ (1774-5+1) =$ **3139980**

therefore, $\mathbf{D_e} = \frac{1927342}{3139980} = .614$

These values for D_o and D_e are then used in equation (7) to calculate the agreement coefficient.

$$\alpha = 1 - \frac{.213}{.614} \times 1 = .653$$

This value (.653) means that when this set of categories is used to classify this sample of textual units, 65% of the resulting agreements will be repeatable. Krippendorff's agreement coefficient is, therefore, a more meaningful evaluation of reliability since it predicts the degree to which categories yield repeatable agreements rather than simply

providing another means of describing the agreements. It is also of value because all of the categories and coders' decisions can be evaluated together, regardless of their number, rather than in pairs. This property is unlike Scott's π and Cohen's κ which can only include two coders. Furthermore, it provides a standardized measure so that test results can be compared even though they may involve different sample sizes and numbers of coders. This coefficient can also be used to isolate the cause of disagreements by using it to compare one category with all of the others, by grouping ambiguous categories together and recalculating the coefficient, and by evaluating one coder against the others. Finally, it is possible to weight the disagreements according to the type of category set which is used, that is, if the differences between the categories are nominal, ordinal, interval or ratio. Table 3.4 shows the calculations for the differences between two categories (d_{bc}) for these four scales.

TABLE 3.4 CALCULATION OF d_{bc}

1. Nominal Scale

$$d_{bc} = \begin{cases} 0 \text{ iff } b = c \\ 1 \text{ iff } b =/= c \end{cases}$$

2. Ordinal Scale

$$d_{bc} = \left(\sum_{k>b} \frac{n_k}{rm} - \sum_{k<b} \frac{n_k}{rm} + \sum_{k<c} \frac{n_k}{rm} - \sum_{k>c} \frac{n_k}{rm} \right)^2$$

where: n_k = the frequency with which category k is used
r_m = total for matrix

3. Interval Scale

$$d_{bc} = (b-c)^2$$

4. Ratio Scale

$$d_{bc} = \left(\frac{b-c}{b+c} \right)^2$$

In many of the sets of categories tested in this research, each individual category indicated an amount of ice which was greater than the previous category by an unspecified amount,

therefore the d_{bc} calculation for the ordinal scales was applied. Although the nominal scale could have been used, it would have provided an evaluation of reliability that was less accurate because the disagreements are not of equal value for categories on the ordinal scale. For example, a disagreement involving the lowest (no ice) and highest (very close ice) categories should receive a greater weighting than a disagreement between two adjacent categories. Table 3.5 shows two hypothetical matrices which depict opposite patterns of disagreement with the same number of agreements.

TABLE 3.5

COMPARISON BETWEEN TWO DISAGREEMENT PATTERNS

A

Low Ordinal Disagreement

Categories	1	2	3	4	
1	**6**	4	3	2	15
2	4	**6**	4	3	17
3	3	4	**6**	4	17
4	2	3	4	**6**	15
					64

$\alpha = .165$

B

High Ordinal Disagreement

Categories	1	2	3	4	
1	**6**	2	3	4	15
2	2	**6**	2	3	13
3	3	2	**6**	2	13
4	4	3	2	**6**	15
					56

$\alpha = .292$

In this example, matrix A shows a pattern in which most disagreements are in adjacent categories while the opposite situation is given in B. The agreement coefficients (α) are for the nominal scale in which all disagreements are of an equal value and therefore $d_{bc} = 1$. They indicate that A ($\alpha = .165$) is less reliable than B ($\alpha = .292$). However, when the ordinal scale calculation is used with these values, the resulting agreement coefficients are quite different. To calculate the agreement coefficient for ordinal scale categories, another matrix which contains the d_{bc} values is used to weight the disagreements. This is demonstrated in Table 3.6 for matrix A.

TABLE 3.6 CALCULATION OF d_{bc} FOR ORDINAL SCALES

$$d_{bc} = \left(\sum_{k>b} \frac{n_k}{rm} - \sum_{k<b} \frac{n_k}{rm} + \sum_{k<c} \frac{n_k}{rm} - \sum_{k>c} \frac{n_k}{rm}\right)^2$$

where: n_k = the frequency with which category k is used
r_m = total for matrix

	Categories 1	2	3	4	
n_k	15	17	17	15	rm = 64

b , c	k>b	-	k<b	+	k<c	-	k>c	nk/rm	(nk/rm)^2	dbc
1 , 2	49*	-	0	+	15	-	32	32/64	$.5^2$	= .25
1 , 3	49	-	0	+	32	-	15	66/64	1.03^2	= 1.06
1 , 4	49	-	0	+	49	-	0	98/64	1.53^2	= 2.34
2 , 3	32	-	15	+	32	-	15	34/64	$.53^2$	= .28
2 , 4	32	-	15	+	49	-	0	66/64	1.03^2	= 1.06
1 , 2	15	-	32	+	49	-	0	32/64	$.5^2$	= .25

* k > b: when b = category 1, n_k for category 2 = 17
+ n_k for category 3 = 17
+ n_k for category 4 = 15
Total for k > b = 49

Table 3.6 continued

MATRIX A

Categories	1	2	3	4
1	6	4	3	2
2	4	6	4	3
3	3	4	6	4
4	2	3	4	6

X

MATRIX OF dbc VALUES

Categories	1	2	3	4
1		.25	1.06	2.34
2	.25		.28	1.06
3	1.06	.28		.25
4	2.34	1.06	.25	

$\alpha = .165$

MATRIX A - ORDINAL SCALE

Categories	1	2	3	4	
1	6	1	3.18	4.68	14.86
2	1	6	1.12	3.18	11.30
3	3.18	1.12	6	1	11.30
4	4.68	3.18	1	6	14.86
					52.32

$\alpha = .330$

* In the case of the d_{bc} matrix, only disagreements are involved, agreements remain unchanged.

When the same procedure is applied to matrix B, the ordinal agreement coefficient is .222. In adjusting for the type of disagreement therefore, the coefficient for A is increased from .165 to .330, whereas the coefficient for B is reduced from .292 to .222 even though the agreements remained the same. These adjustments reflect the degree to which the category requirements discussed in Chapter 2 have been met. Disagreements between categories of the lowest and highest values should not be expected to occur when the set of categories is unambiguous and exhaustive. Frequent occurrences of this type of disagreement must therefore reduce the reliability accordingly.

Intracoder Reliability The discussion of reliability and its evaluation to this point has been directed primarily to the degree to which a group of coders agree with each other on the categorisation of textual units. It is also important to determine the consistency of each coder by calculating the reliability with which she or he can repeat the categorisation. This information can be used to improve the intercoder reliability in two ways, by isolating one coder who has a lower level of consistency than the others, and by identifying a

category or categories which caused inconsistencies for each of the coders. In order to screen-out the ineffective coders, it is important that the coders are carefully selected so that their backgrounds in relation to the communication and the methodology are similar. It is also of great importance to ensure that all of the coders receive the same training and instructions because if this is not standardized, then the reliability tests cannot be objective.

To evaluate the intracoder reliability, each coder must repeat the classification at least once. Then, the agreement coefficient is calculated in the same way as the intercoder coefficient with two exceptions. The first difference is that 'm' is the number of repetitions rather than the number of coders, and secondly, the number of agreements and disagreements for each category are totalled for all of the repetitions rather than for the coders. This produces an agreement coefficient for each coder. An example of this process is given in Table 3.7; the values used in this example are from Phase II of this research.

TABLE 3.7 CALCULATION OF ORDINAL AGREEMENT COEFFICIENT
Phase II - Ice Concentration - Coder A - 3 Repetitions

	1	2	3	4	5	6	7		
1	60	48 (.07)*	2 (.46)	2 (1.36)		3 (2.39)	0 (3.2)	0 (3.37)	115
2	48 (.07)	22	3 (.72)	4 (.8)		9 (1.64)	1 (2.31)	0 (2.46)	87
3	2 (.46)	3 (.72)	198	16 (.23)		7 (.75)	0 (1.23)	0 (1.34)	226
4	2 (1.36)	4 (.8)	16 (.23)	96		20 (.15)	0 (.39)	0 (.46)	138
5	3 (2.39)	9 (1.64)	7 (.75)	20 (.15)		94	18 (.06)	5 (.08)	156
6	0 (3.2)	1 (2.31)	0 (1.23)	0 (.39)		18 (.06)	2	5 (.002)	26
7	0 (3.37)	0 (2.46)	0 (1.34)	0 (.46)		5 (.08)	5 (.002)	0	10
									758

* d_{bc}

Continued...

Table 3.7 Continued

Ordinal Scale Values (Nominal x (d_{bc}))

	1	2	3	4	5	6	7	
1	60.00	3.36	.92	2.72	7.17	0	0	74.17
2	3.36	22.00	2.16	3.20	14.76	2.31	0	47.79
3	.92	2.16	198.00	3.68	5.25	0	0	210.01
4	2.72	3.20	3.68	96.00	.30	0	0	105.90
5	7.17	14.76	5.25	.30	94.00	108.00	.40	122.96
6	0	2.31	0	0	1.08	2.00	.01	5.40
7	0	0	0	0	.40	.01	0	.41
								566.64

$$D_o = \frac{94.64}{566.64} = .167 \qquad D_e = \frac{242828}{319947.61} = .759$$

$$\alpha = 1 - \frac{.167}{.759} = .778$$

This is a fairly high level of reliability which means that only 22% of the consistent categorisations made by this coder could be attributed to chance. Therefore, 78% of coder A's agreements can be relied on to be repeatable.

3.3 Acceptable Levels of Reliability

Once the reliability has been tested and evaluated, the decision to alter the categories depends on whether or not the level of reliability is acceptable, as illustrated in Figure 2.3. This is a very crucial question that places limitations on the data that will be derived. There are two ways of approaching this problem, by defining a specific level of agreement which must be attained or by continually adjusting the categories on the basis of the test results until the reliability cannot be raised any further. Two examples of the first approach will be discussed here. William Schutz (1952) devised a method of optimizing percentage agreements by using a dichotomous decision method of classification. In this approach, the coders are given two categories followed by a second pair that are conditional upon the previous decision. Percentage agreements can be calculated at each decision level. To determine whether or not these agreements were sufficiently high, Schutz developed confidence interval tables for the number of units being categorized and the percentage agreements necessary to attain the particular confidence interval. Two of these tables are given in Table 3.8. The function of these tables can be shown by applying them to a sample from this research in which 780 decisions were made by the coders. Table 3.8a shows that for 800 textual units (N = number of classifications), there must be a 93% agreement (A = percentage agreement) to ensure that 90% of those agreements were not made by chance.

TABLE 3.8 AGREEMENT LEVELS

a

AGREEMENT LEVEL = .90

N	A	N	A	N	A	N	A
2	100	35	99	100	97	260	95
	100		98		96		94
3	100	40	99	110	97	280	95
	100		98		96		94
4	100	45	99	120	97	300	95
	100		97		95		94
5	100	50	99	130	96	350	95
	100		97		95		94
6	100	55	98	140	96	400	94
	100		97		95		94
7	100	60	98	150	96	450	94
	100		97		95		93
8	100	65	98	160	96	500	94
	100		97		95		93
9	100	70	98	170	96	600	94
	100		97		95		93
10	100	75	98	180	96	700	94
	100		96		95		93
15	100	80	97	190	96	800	93
	100		96		95		93
20	100	85	97	200	95	900	93
	99		96		95		93
25	100	90	97	220	95	1000	93
	99		96		94		93
30	100	95	97	240	95		
	98		96		94		

b

AGREEMENT LEVEL = .85

N	A	N	A	N	A	N	A
2	100	35	98	100	94	260	92
	100		96		93		91
3	100	40	97	110	94	280	92
	100		96		93		91
4	100	45	97	120	94	300	92
	100		95		92		91
5	100	50	97	130	94	350	91
	100		95		92		90
6	100	55	96	140	93	400	91
	100		95		92		90
7	100	60	96	150	93	450	91
	100		94		92		90
8	100	65	96	160	93	500	91
	100		94		92		90
9	100	70	96	170	93	600	90
	100		94		92		90
10	100	75	95	180	93	700	90
	100		94		92		89
15	100	80	95	190	93	800	90
	99		93		91		89
20	99	85	95	200	93	900	90
	98		93		91		89
25	99	90	95	220	92	1000	90
	97		93		91		89
30	98	95	95	240	92		
	96		93		91		

N = number of classifications
A = percentage agreements

(Schutz, 1952:125)

Table 3.8b is the 85% agreement level and as a result of this lowered expectation, an agreement of only 89 - 90% for the 800 decisions is required. These tables are not applicable to this research, however, because the decisions made here were not

dichotomous, there being as many as seven categories. Another approach to defining acceptable levels is used by Krippendorff. He calculated correlations between sets of categories of several content analyses. He then determined the agreement coefficients for the same categories and compared the correlation values with the agreement coefficients. From this he found that when the agreement coefficients were less than 0.7, the correlations were not statistically significant (Krippendorff, 1980:147). As a result, he adopted the practice of rejecting agreement coefficients that are lower than 0.8.

The research presented here took the second approach in which levels of acceptability were not predetermined. The reliability was tested, the categories were adjusted accordingly, and this was repeated until the reliability could not be improved without the loss of a considerable amount of information. The advantage of this method is that the information contained in the communications is not lost due to *insufficient* reliability. If Schutz's tables or Krippendorff's 0.8 agreement level were applied to this research, then the sea ice descriptions contained in the log books could not be used at all. By using this more flexible method, the agreement coefficients will be the best that can be expected from this particular communication without lowering the resolution of the derived data to the point that they are meaningless.

This chapter has described those aspects of reliability that were necessary for the research that will be described in the following chapters. There are other factors to be considered when conducting reliability tests such as the selection of coders, coder training, and the design used for the classification process. These aspects will be discussed in the following chapters.

CHAPTER 4

PHASE I

4.1 Background

This research began in 1985, exactly 20 years after the publication of the first climatic reconstruction based on the Hudson's Bay Company Archives (MacKay and Mackay, 1965). Since then, several climatic reconstructions have been published and are given in Table 1.1. These were primarily based on the post journals but they also made use of the ships' log books and meteorological records. From these sources, a variety of data were derived including the dates of occurrence of phenological indicators, monthly frequencies of meteorological events, indicies of summer sea ice severity, and evaluations of mean monthly temperatures. Thus, by the time this research began, a substantial body of paleoclimatic research using the Hudson's Bay Company archives had accumulated. However, a characteristic of this body of research is that quality testing did not figure prominently within it and this generalization applies to both reliability and validity testing. Two exception to this were the studies by Cynthia Wilson (1982 ; 1983) and Moodie and Catchpole (1975). Cynthia Wilson studied early 19th century temperature observations made at the posts on the eastern coast of Hudson Bay and applied to them meticulous corrections for the eccentricities of instrument design, exposure, and observing routines of the time. Moodie and Catchpole (1975) applied rudimentary reliability tests in their reconstructions of dates of first freezing and first breaking of river estuaries using the descriptions in the Company's post journals. These tests involved a number of coders who

each derived river ice dates from the same sample of post journals so that the intercoder reliability, reliability of dating categories, and the reliability of dating places could be evaluated. The techniques used to assess these types of reliability were conducted at the most rudimentary level by expressing the numbers of agreements as a percentage of the total numbers of decisions. Following the completion of these tests in 1975, there was no further development or application of reliability tests until the present research began in 1985. Thus, this project originated in a research environment where reliability testing was conceived of in a rudimentary way. It then gradually evolved as the more sophisticated principles of reliability testing developed in the social sciences were discovered and assimilated. This origin and evolution is reflected in the three phases of reliability testing that form the core of this thesis.

Phase I reflects the state of CA and reliability testing practiced in paleoclimatology at the time when this work began. It employed impressionistically-derived categories, and tested reliability by using an approach that was similar to that applied by Moodie and Catchpole (1975). Upon completion of this phase, the more advanced techniques of the social sciences, principally in psychology, became available and were incorporated in Phases II and III. In Phase II, these techniques were employed to assess the relevance of the categories to the textual units in the log books and to evaluate and improve the reliability of the categories. Phase III assessed the corrective measures suggested by the results of the previous phase and evaluated the technique that was introduced in Phase II. In doing so, Phase III resulted in providing the most reliable set of categories for the purpose of deriving sea ice information from the log book descriptions.

4.2 Phase I Parameters

Data A sample of 56 log book pages was used in the first phase, each page contained the entries for a 24-hour period. These pages were chosed from 21 randomly selected log books, but the selection of the sample of log book pages was not completely random so as to obtain a wide range of different types of ice conditions. Table 4.1 lists the log books used in Phase I.

TABLE 4.1 LOG BOOK SAMPLE FOR PHASE I

Day #	Year	Date	Archive Call #	Ship
1	1836	Aug 25	930	Prince Rupert
2		Sep 29		
3	1843	Aug 3	670	Prince Albert
4		Aug 10		
5		Aug 11		
6		Sep 11		
7	1855	Aug 18	846	Prince of Wales
8		Sep 1		
9		Sep 2		
10	1859	Aug 28	719	Prince Arthur
11		Sep 18		
12	1848	Aug 6	683	Prince Albert
13		Aug 7		
14	1803	Aug 8	756	Prince of Wales
15		Aug 13		
16	1809	Aug 28	772	Prince of Wales
17		Aug 29		
18	1796	Aug 5	741	Prince of Wales
19		Aug 7		
20		Aug 15		
21	1810	Aug 1	774	Prince of Wales
22		Aug 4		
23		Aug 5		
24	1824	Aug 18	224	Camden
25		Aug 25		
26		Aug 27		
27	1829	Aug 5	818	Prince of Wales
28		Aug 24		
29		Aug 25		
30	1827	Aug 5	813	Prince of Wales
31		Aug 6		
32		Aug 10		
33		Aug 16		

(...Continued)

Table 4.1 Continued

Day #	Year	Date	Archive Call #	Ship
34		Aug 22		
35	1821	Aug 6	795	Prince of Wales
36		Aug 13		
37		Aug 16		
38		Aug 19		
39	1815	Aug 15	783	Prince of Wales
40		Aug 25		
41	1808	Aug 3	420	King George
41	1808	Aug 3	420	King George
42		Aug 7		
43		Aug 9		
44	1808	Aug 12	420	King George
45	1795	Aug 22	398	King George
46	1782	Aug 7	386	King George
47		Aug 12		
48	1775	Aug 4	898	Prince Rupert
49		Aug 6		
50		Aug 9		
51	1774	Aug 16	378	King George
52		Sep 15		
53	1761	Aug 13	1031	Seahorse
54		Aug 14		
55	1755	Aug 15	873	Prince Rupert
56		Aug 18		

The first column (Day #) lists the identification number given to each page so that August 25, 1836 was Day 1 and August 18, 1755 was Day 56. The reason for this type of identification was to prevent the coders from using the dates to infer ice conditions. This might have been possible because each coder had some knowledge of the seasonal variations in ice conditions in Hudson Bay. Each log book page was transcribed in such a way that the hourly layout of the page was retained but only the sea ice comments were transcribed as illustrated in Table 4.2.This layout was preserved even though some hours had no ice descriptions. These were used to test how the coders interpreted these blanks, and to give a more realistic impression of how the conditions changed throughout the day.

TABLE 4.2 EXAMPLE OF A LOG BOOK PAGE TRANSCRIPTION FOR PHASE I

Hour	
1p.m.	*Heaving ship off a piece of ice*
2	
3	*saw streams of water to the Southward*
4	*Ship beset in close ice*
5	
6	*Ice close ship beset*
7	
8	
9	
10	
11	
12	
1a.m.	
2	*Ice begins to open*
3	*Ship began to move*
4	*Ice open in lanes*
5	
6	
7	
8	*Ice closing*
9	*Forcing in heavy ice*
10	
11	*Ice more open at times*
noon	

Categories: In this phase the set of categories was intuitively derived, and were developed to facilitate the extraction of detailed ice information. These categories were objective in the sense that they were not based on *a priori* interpretations of the log book descriptions. This was particularly important because the primary researcher was also one of the five coders and any attempt to base the categories on a preliminary interpretation would have therefore resulted in a bias toward the preconceptions of one coder.

The phase I system of categories consisted of five category sets, each of which required the coders to make a dichotomous decision about the sea ice information contained within the log book description. In addition, each of these five pairs included the option of *not enough information* to be used when a coder decided that a particular entry did not

permit the dichotomous decision to be made. The phase I set of categories is given in Table 4.3. It was anticipated that this classification system would reveal those general ice properties that could be identified in the log books. The main purpose of these categories, therefore, was to establish a framework on which to base more functional category sets and was not intended to serve as a system on which to base a final reconstruction.

TABLE 4.3 PHASE I CATEGORIES AND DEFINITIONS

GENERAL CLASSIFICATION	CATEGORY NUMBER	CATEGORIES
A. Presence	0	Ice not present in vicinity of ship
	1	Ice present in vicinity of ship
	*	not enough information
B. Concentration	2	Small area covered by ice (<50%)
	3	Large area covered by ice (>50%)
	*	not enough information
C. Fragmentation	4	Ice cover highly fragmented
	5	Ice cover NOT highly fragmented
	*	not enough information
D. Thickness	6	Thin layer of ice
	7	Thick layer of ice
	*	not enough information
E. Motion	8	Ice in motion
	9	Ice NOT in motion
	*	not enough information

Textual Units At this early stage in the research, it was anticipated that three particular textual units would be appropriate for testing, namely, the day, the hour, and the individual word. These units will be discussed here in the order in which they were tested.

The day unit required that the entire body of information in each of the 56 sample pages would be classified by the system given in Table 4.3. To accomplish this, the coders read all of the ice descriptions on the page and then assigned the category numbers which they decided best described the ice conditions for the day as a whole. A sample of the form used by the coders to facilitate this is given in Table 4.4.

TABLE 4.4 CLASSIFICATION FORM FOR THE 'DAY' TEXTUAL UNIT

DAY 1 LEVEL_____

Coder_____ Date_____ Time (hrs)_____ (min)_____

1	
2	Forcing this close heavy Ice under all sail
3	
4	
5	Ice opening
6	Leaving (?) thro' open Ice
7	
8	
9	
10	
11	
12	
1	
2	
3	
4	
5	
6	Sailing thro' open Ice
7	
8	
9	Ice very close
10	
11	
noon	No open water in sight

Day Code_____

Code based on (give basis for your decision):

At the top of this form, each coder was to provide her or his name, the date on which the classification was done, and the length of time taken to complete the classification. These tests were not done anonymously so that the intracolder reliability could be evaluated. Each coder repeated each test at least twice and a period of 24 hours between tests was prescribed to prevent the coders from remembering their previous decisions and so invalidating the intracoder tests. The space left for the 'Level' will be discussed in section 4.3 - Procedure. The body of the form was comprised of the hourly transcriptions of the sea ice comments and observations. The column of numbers represents the hour of the day on which the observation was made so that the first number represents 1:00 p.m. since the seaman's day began at noon. Those hours which did not have sea ice observation were left blank and the coders were to decide whether or not this indicated that the ice conditions from the previous entry prevailed to the next entry. At the bottom of each form, a space was given for the category code number assigned by the coder to that particular day. A space was also given for the coder to explain the basis for the classification. It was anticipated that these comments might help in the interpretation of the results of the reliability tests. It should also be noted here that although days were given consecutive identification numbers, the coders were informed that they were not actually consecutive days. The coders were not, therefore, permitted to use the previous or following days to help make their decisions, each day was to be classified in isolation. The coders were also not allowed to refer back to previous decisions as this, of course, would have produced spuriously high intracoder agreements.

The hour textual unit required that the coders provide a classification, from the same set of categories, for each hourly entry of the 56-day sample. The form used for this was essentially the same as for the day unit, and is given in Table 4.5. The coders entered the category numbers beside each hourly entry and a space was provided at the bottom for comments. This procedure was also repeated with at least 24 hours between each test.

TABLE 4.5 CLASSIFICATION FORM FOR THE HOUR TEXTUAL UNIT

DAY 1 LEVEL_____

Coder_____ Date_____ Time (hrs)_____ (min)_____

1	
2	Forcing this close heavy Ice under all sail
3	
4	
5	Ice opening
6	Leaving (?) thro' open Ice
7	
8	
9	
10	
11	
12	
1	
2	
3	
4	
5	
6	Sailing thro' open Ice
7	
8	
9	Ice very close
10	
11	
noon	No open water in sight

COMMENTS:

To test the 'word' textual unit, the coders were given a list of 81 individual words of which 24 were descriptions of the ice itself, and 57 related to the navigational activities employed to deal with the ice. These words were presented out of the context of the log book page and were classified individually. The forms used for this are given in Tables 4.6a and b (a complete list of the navigational words is given in section 4.4).

Each of these three textual units was tested sequentially, and so all of the repetitions for each unit were completed before the set of tests for the next unit were begun.

TABLE 4.6 CLASSIFICATION FORMS FOR THE 'WORD' TEXTUAL UNIT

a) Ice

Level(s)_____

Coder_____Date_____Time (hrs)_____(min)_____

TERM	CODE	TERM	CODE
1 Body			
2 Broken			
3 Close			
4 Fast			
5 Field			
6 Floe			
7 Heavy			
8 Ledge			
9 Loose			
10 More Clear			
11 No Water			
12 Open			
13 Packed			
14 Patch			
15 Piece			
16 Ridge			
17 Shattered			
18 Skim			
19 Slack			
20 Small			
21 Straggling			
22 Thick			
23 Thin			
24 Unbroken			

b) Navigation Level (s)_____

Coder_____Date_____Time (hrs)_____(min)____

TERM	CODE	TERM	CODE
1 Along		17 Ease From	
2 Alter Course		18 Enter	
3 Among		19 Fall In With	
4 Anchored		20 Fall Thro'	
5 Appear		21 Fast	
6 Bear Away		22 Force	
7 Beset (Set)		23 Forge Thro'	
8 Bore Thro'		24 Get Into	
9 Break Thro'			
10 Bring To			
11 Came Up To			
12 Cannot Move			
13 Close To			
14 Cast Off			
15 Detained			
16 Drive Thro'			

4.3 Test Procedures

All of the coders met before the first unit was tested. At this meeting they were given a set of 56 classification forms for the 'day' textual unit, a card with the categories, code numbers and definitions, and the written list of instructions given below.

TABLE 4.7 CODER INSTRUCTIONS - DAY UNIT

1. Fill in your name as CODER '_____'

2. Fill in date (month and day) only on the first day which you code. If you code days 1 to 30 on one day, and days 31 to 56 on another day, then only days 1 and 31 will have a date entered.

3. When you have finished coding for one sitting, note the hour(s) and minutes taken. Therefore in the example above, only days 1 and 31 will have the time recorded.

4. Read the entries for each entire day (noting the spacing between entries) and assign a code for the entire day from the code card. Therefore:

   -for level I there will be 1 to 2 code numbers (enter ** if necessary)
   -for levels I & II there will be 3 code numbers (enter *** if necessary)
   -for levels I II & III there will be 4 code numbers (enter **** if necessary)
   -for levels I II III & IV there will be 5 code numbers (enter ***** if necessary)

TABLE 4.7 Continued

5. Indicate the primary factor which influenced your decision.

6. You may provide extra comments on the back of the form.

7. When you have coded all 56 days, please return the code card and forms one level at a time.

8. You may take as long as you need to code, but you will not be given the next level until 24 hours after the last date of coding.

9. Please do not discuss coding with anyone else.

10. It is advisable to code in pencil because you will probably change your mind.

11. Once you have coded a day, do not go back to it either to change a code or to compare it with other days.

As noted in these instructions, the coders were not given the next set of forms until 24 hours after the previous session was completed. Therefore, each coder had only one set of 56 classification forms at a time. The coders were allowed to work at their own speed and could complete each test in more than one sitting. This rarely occurred however, as the coders preferred to complete the test at one time which ranged from 40 minutes to one-and-one-half hours. The coders worked individually in their own time rather than at a group session and were instructed to refrain from discussing any aspect of the tests with anyone else. This was to ensure that the tests reflected each coder's interpretation without any outside influences. There was very little discussion at the first meeting other than to make sure that the instructions were completely understood. There was no discussion about the meaning of the categories so that the tests would truly reflect how the coders interpreted the log book entries and the categories. The main purpose of the meeting was simply to ensure that all of the coders received the same instructions, both written and oral, and not to arrive at a concensus about the meanings of the categories and log book contents. This would have yielded spuriously high agreements which would only reflect the coders' clerical abilities. The group did not meet again until all of the phase I tests were complete.

As each coder completed the day unit tests, they were given the instructions for the 'entry' unit as well as the first set of 56 classification forms for this unit. This list of instructions is given in Table 4.8.

TABLE 4.8 CODER INSTRUCTIONS - HOUR UNIT

1. Fill in your name as CODER '_____'

2. Fill in date (month and day) only on the first day on which you code. If you code days 1 to 30 on one day, and days 31 to 56 on another day, then only days 1 and 31 will have a date entered.

3. When you have finished coding for one sitting, note the hour(s) and minutes taken. Therefore, in the example above, only days 1 and 31 will have the time recorded.

4. Divide each entry into phrases and separate them with a slash (/). If the entire entry is a phrase, place two slashes at the beginning of the entry (//).

5. At the end of each entry, place a code number from the code card such that:

   -for level I there will be 1 to 2 code numbers (enter ** if necessary)
   -for levels I & II there will be 3 code numbers (enter *** if necessary)
   -for levels I II & III there will be 4 code numbers (enter **** if necessary)
   -for levels I II III & IV there will be 5 code numbers (enter ***** if necessary)

   eg. Every / good boy / deserves / fun. (code)

6. A space is provided for your comments. You may use the back if necessary.

7. When you have coded all 56 days, please return the code card and forms one level at a time.

8. You may take as long as you need to code, but you will not be given the next level until 24 hours after the last date of coding.

9. Please do not discuss coding with anyone else.

10. It is advisable to code in pencil because you will probably change your mind.

11. Once you have coded a day, do not go back to it either to change a code or to compare it with other days.

These instructions were essentially the same as for the day unit with the exception that the coders were required to divide each hourly entry into phrases. This was intended to help assess those types of information which had low levels of agreement. This was not used however, because it was ultimately determined to be beyond the scope of these tests. The classification number was assigned to each entire entry regardless of the phrase subdivisions. The range of time taken to complete a test for this unit was from 30 minutes to one hour and 50 minutes.

The instructions and classification forms for the word unit were given to each coder on the completion of the last 'entry' test. Table 4.9 provides the instructions for this unit. The average time taken to complete a test for this unit was 15 minutes.

TABLE 4.9 CODER INSTRUCTIONS - WORD UNIT

1. Fill in your name as CODER '_____'

2. Fill in the date (month and day) and note the hour(s) and/or minutes taken.

3. Write a code number from the code card next to the word number so that:

   -for levels I II III & IV there will be 4 code numbers (enter **** if necessary)

   NOTE: 0 & 1 are not used for word coding.

4. Do not write under the heading 'Interpretation/Comments'.

5. A space is provided for your comments, you may use the back if necessary.

6. When you have coded all the words, please return the code card and forms one level at a time.

7. You may take as long as you need to code, but you will not be given the next level until 24 hours after the last date of coding.

8. Please do not discuss coding with anyone else.

9. It is advisable to code in pencil because you will probably change your mind.

On each of the instruction and classification forms, a space was provided for the coder to enter the level of the classification. This was because of the hierarchical system on which the repititions were based. At each level, one category set was added so that the resolution was increased with each test. Therefore, all 56 days were classified at Level I, then all were classified at Level II, and so on for each test so that at the end of Phase I each of the five coders had tested the 56 days five times, the 261 hour entries five times, and the 81 words twice. A summary of this system is given in Tables 4.10a and b.

TABLE 4.10 PHASE I CLASSIFICATION SYSTEM

a.) Categories for Each Level

Level	Category Set	Code Number	Definition
I	Presence and Concentration	0	Ice not present in vicinity of ship
		1	Ice present in vicinity of ship
		*	Not enough information
		2	Small area covered by ice
		3	Large area covered by ice
		*	Not enough information
II	Fragmentation	4	Ice cover highly fragmented
		5	Ice cover not highly fragmented
		*	Not enough information
III	Thickness	6	Thin layer of ice
		7	Thick layer of ice
		*	Not enough information
IV	Motion	8	Ice in motion
		9	Ice not in motion
		*	Not enough information

b.) Test Levels

Test Number	Textual Unit	Category Level(s)
1	Day	I
2	Day	I II
3	Day	I II III
4	Day	I II III IV
5	Day	I II III IV
6	Entry	I
7	Entry	I II
8	Entry	I II III
9	Entry	I II III IV
10	Entry	I II III IV
11	Word	I II III IV
12	Word	I II III IV

According to this scheme, each level was tested at least twice so that the intracoder reliability could be assessed. As the levels increased, one more number was added to the classification code given by the coder to each textual unit of the sample. Level I, which included ice presence and concentration, was repeated most often because it included the most basic information required for a reconstruction and as a result, the contents of the log books required the most rigorous testing at this level. Table 4.11 provides a few examples of the interpretation of this system.

TABLE 4.11 SAMPLE CLASSIFICATIONS FOR PHASE I

Level	Code Numbers	Interpretation
I	1, 2	Small area of ice in vicinity of ship
II	1, 2, 4	Small area of fragmented ice in vicinity of ship
III	1, 2, 4, 6	Small area of thin, fragmented ice in vicinity of ship
IV	1, 2, 4, 6, 8	Small area of thin, fragmented ice in motion in vicinity of ship
II	1, 3, *	Large area of ice in vicinity of ship - not enough information for fragmentation
III	1, 3, *, 6	Large area of thin ice in vicinity of ship - not enough information for fragmentation
IV	1, *, *, *, *	Ice present in vicinity of ship - not enough information for remaining categories

As shown in Table 4.11 the word textual unit was treated differently. By the time the coders had finished the day and entry units they had applied the classification system 10 times and this was considered sufficient to establish their consistencies for this set of categories. Therefore, the word units were classified only twice so that a check of their intracoder reliabilities could be tested for comparison with the other two units at a later

point. This unit was also different in that categories 0 and 1 were not used because the presence of ice was implied by the words themselves.

When the final tests were returned, each coder was given a questionnaire. It was anticipated that the answers to these questions would help in the preparations for phase II. The questionnaire is given below.

TABLE 4.12 PHASE I QUESTIONS

GENERAL

1. Did you visualize the descriptions in your mind before assigning a code?
   Always_____ Often _____ Rarely_____ Never_____
2a. When you had information related to both navigational activities and to sea ice (day and hour only), were you more likely to base your decision on one than the other? Y / N
   b. Why?

3. Did you draw on experience from earlier sessions to help in later decisions? Y / N
   If yes, please explain:

4. Were you working on any other related project during this time? Y / N
   If so, please explain:

DAY

1a. You were given individual days to classify without knowing if they were consecutive or their dates. Do you think that it would have made a difference if you were given:
   i. Dates? Y / N
   ii. Consecutive days? Y / N
   b. If you answered yes to ii, how many consecutive days would you have required?_____

2a. Did you find a change in the degree of difficulty as more levels were added? Y / N
   b. If yes, please explain:

ENTRY

1a. After having completed all 4 levels for the DAY unit, did you anticipate the higher level codes as you classified the entries? Y / N
   b. If yes, please explain:

2. Did you use previous entries during each day to classify the entries?
   Always_____ Often _____ Rarely_____ Never_____

(...Continued)

Table 4.12 Continued

3. Did you find that you recognized certain repeated entries, and that you classified them automatically?
   Always_____ Often _____ Rarely_____ Never_____

4. Did you divide the entries into grammatical subdivisions?
   Always_____ Often _____ Rarely_____ Never_____

WORD

1. Which did you find easier to classify?
   Ice words_____ Navigation words_____ No real difference_____

2. Which gave the greater amount of information?
   Ice words_____ Navigation words_____ No real difference_____

Since the primary researcher was also a coder, many of these questions were derived from personal experience, and the first of the general questions arose in this way. All of the descriptions being classified were of an element of the physical environment which is visible. It was, therefore, anticipated that the coders would visualize the ice descriptions while they classified them and, in fact, all of the coders indicated that they always visualized the descriptions. This proved to be valuable information for the next phase by supporting the decision to use diagrams to accompany the category definitions. The balance of the general questions revealed that, for the most part, the coders used the ice and navigational terms roughly equally in making their decisions, that experience with the earlier sessions was used in subsequent tests, and only one coder was working on a similar project at the same time that these tests were being conducted. These responses helped to provide a general idea of how the coders approached this system of testing. The second set of questions related to the day coding unit specifically. The first question was the most revealing in this section. First, it was surprising to find that four of the five coders did not feel that it would have been helpful to know the dates of the days that were classified. The second part of this question produced the unanimous response that consecutive days, particularly the previous day, would be useful because often there was no entry for the first part of the day. Conversations with the coders revealed that they assumed the blanks to

mean that the conditions described by the previous entry prevailed until the next entry. As a result of this interpretation, it would be essential to have access to the previous log book page when there was no entry for the first part of the day being classified. Most coders did not find that the addition of levels altered the difficulty of classifying this unit. The hour unit was classified under different circumstances than the day unit. In the first place, with each advancement in the classification level, the coders were faced with a new and unanticipated category for the day unit, whereas the coders were already familiar with the entire classification system for the hour unit. Secondly, on many occasions, a particular entry was repeated on one page with the result that a certain amount of consistency was guaranteed for this unit. In response to the first question for the hour unit only two of the coders responded that they anticipated how they would classify the entries for the levels that were to follow. The second question revealed that three of the coders attempted to view the entries on each page as if they were isolated descriptions even though they were not specifically instructed to do so. The third question elicited the unanimous response that the coders often recognized certain repeated entries and classified them consistently. On the basis of these answers and individual conversations with the coders it is interesting to note that they were consciously attempting to maintain a level of consistency within the confines of the instructions. The first question of the word unit provided curious results when viewed together with the answers to the second of the general questions. When asked whether they based their classifications mainly on navigational or ice words, three coders replied that they were not likely to base their decisions on one more than the other, one preferred navigational terms, and one used ice words. When they were asked which they found easier to classify in the word unit, however, two found there was no difference, two chose the ice words, and one the navigational words. The last question, regarding the relative potential for the two types of words to yield detailed information, served to confirm the answers to the previous question by yielding the same responses.

At the end of this questionnaire, the coders were invited to give their comments concerning phase I, and three responded as follows:

> *- Have we paid too little attention to the role of the watches?*
> *- Perhaps day started with 12 noon - 4 p.m. watch. That watch had to open log page with an ice description.*
> *- Other watches kept the page going but perhaps they*
> *(a) only added new information if this really changed*
> *(b) [were] greatly influenced by night time and so on.*
> *- If my hunch is vaguely right, then the basic ice description was made at 12 - 4 in the afternoon.*
>
> (Coder A)

> *- I thought that if distance sailed on a particular day was available, then more information for a more confident decision of a code would be present.*
> *- The wording of the code levels might have been a little more detailed. I was particularly uncomfortable with the ice cover information (<50% and 50%+). Not knowing any distances made this decision difficult.*
>
> (Coder I)

> *-Navigation words gave more detailed information on ice conditions but were useless on their own; i.e. all information was valuable in the context of the log entries. As "words" - very little information.*
>
> (Coder J)

All of these comments were very valuable in preparing for phase II as will be shown in Chapter 5. The observations of coders I and J are reflected in the reliability tests and the frequency with which the *not enough information* category was used.

4.4. Evaluation Of Agreements

At this early stage of the research, the agreements were assessed by the use of percentage agreements. At the end of Phase III, however, Krippendorff's agreement coefficient was determined for selected components of Phase I for comparative purposes. These results are discussed in Chapter 7. The evaluation of the agreements for each textual unit will be given separately in this section and then a comparison of the three units will follow.

4.4a Day

The agreements for this unit were examined from four perspectives, namely, intercoder agreements, intracoder agreements, analysis of the categories, and analysis of particular days.

Intercoder Agreements The first step was to tabulate all of the category numbers assigned to each day by each coder for every level. An example of such a summary is given in Table 4.13.

TABLE 4.13 EXAMPLE OF AN INTERCODER SUMMARY TABLE
Phase I, Day, Level IV

	Coder					Level			
Day #	A	D	I	J	M	I	II	III	IV
1	1247*	124*9	134**	135**	1*4**	2/2/1	4	4	4
2	*****	*****	*****	*****	*****	4	5	5	5
3	12478	12479	1*478	13478	134**	2/2/1	5	4	3/1/1
4	13578	13579	1257*	13578	13578	4	5	5	3/1/1
5	13578	13579	1*578	13578	13578	4	5	5	4
51	0	124*9	124**	12468	124**	4	4	3/1/1	2/1/1/1
52	1****	1*579	1*4**	12578	1247*	3/2	2/2/1	3/2	3/1/1
53	1****	0	1****	1****	0	3/2	3/2	3/2	3/2
54	1****	12479	124**	12***	12***	4	3/2	3/2	4
55	135**	135*9	1*4**	13468	1*57*	3/2	3/1/1	3/1/1	3/1/1
56	0	12479	124**	124*8	1247*	4	4	2/2/1	2/1/1/1

The letters A, D, I, J, and M are the first initials of the coders and the five-digit numbers below the initials are the categories that the coders assigned to each day of the sample. For

example, coder J assigned day 52 the code numbers 12578. This means that day 52 was interpreted by coder J as indicating that a small area of thick consolidated ice was present in the vicinity of the ship and this ice was not moving. Coder A, on the other hand, assigned 1**** to the same day which means that this coder agreed that ice was present in the vicinity of the ship, but the entries for day 52 did not provide enough information for the determination of the other categories. The right hand side of Table 4.14 shows the agreement pattern amongst all five coders for each day and level. The patterns of agreement are different combinations ranging from complete agreement (5) to complete disagreement (1,1,1,1,1). The various patterns that are possible and the corresponding numbers of agreement are given in Table 4.14a, and Table 4.14b shows the intercoder agreements for each level based on these patterns.

TABLE 4.14a PATTERNS AND NUMBERS OF AGREEMENT

Pattern of Agreement	Number of Agreements*	Percentage Agreement out of Total Possible
5	10	100
4, 1	6	60
3, 2	4	40
3, 1, 1	3	30
2, 2, 1	2	20
2, 1, 1, 1	1	10
1, 1, 1, 1, 1	0	0

* $\frac{N(N-1)}{2}$ Where N = number of coders in agreement.

(After: Holsti, 1969:141)

TABLE 4.14b INTERCODER AGREEMENT PATTERNS PER LEVEL

Pattern of Agreements	Number of Agreements	Levels I	II	III	IVa	IVb
5	10	8	15	24	1	1
4, 1	6	13	21	16	26	26
						50%
3, 2	4	22	10	13	9	6
3, 1, 1	3	11	3	3	10	12
2, 2, 1	2	2	7	0	9	9
2, 1, 1, 1	1	0	0	0	1	2
1, 1, 1, 1, 1	0	0	0	0	0	0
	>50% Agreement	21	36	40	27	27
	%	37.5	64.3	71.4	48.2	49.2

It is interesting to note in Table 4.14b that the fewest agreements were for Level I in which the most basic decisions were made regarding ice presence and concentration. The fact that there were relatively few agreements for this level demonstrates the necessity for reliability testing. When a reconstruction is based on historical documents that are interpreted by one researcher without testing the reliability of that interpretation, even the most basic decisions may not be repeatable. The major cause of disagreements for Level I was the decision regarding the spatial extent of the ice coverage (Categories 2 and 3). The average percentage of agreements for each of the five repetitions of these categories were 51% for the first test, 47% for the second and 44%, 49% and 46% for the remaining three. Therefore, the coders agreed on less than 50% of the decisions for this classification. The number of agreements for Levels IVa and IVb (motion) were also below 50%. This was due to a systematic disagreement on the part of coder D who assumed that if the movement of the ice was not specifically mentioned in the transcription, then the ice was not in motion (Category 9). The other coders however, independently decided that this meant that there was not enough information (*). Table 4.15 is a summary of the intercoder agreements and the average frequencies with which each coder was in agreement with each of the others are given in Tables 4.16a-e.

TABLE 4.15 AVERAGE INTERCODER AGREEMENTS (ALL LEVELS)

Coder	Average Agreement	Percent
A	134.6	60.0*
D	92.6	41.3
I	122.0	54.5
J	124.4	55.5
M	128.8	57.5

*$\frac{134.6}{224}$ = 60% (224 = 56 days x 4 coders)

TABLE 4.16 INTERCODER AGREEMENTS PER CODER

a. Coder A

	D	I	J	M
I	44	16	46	31
II	36	33	29	38
III	40	45	38	41
IVa	4	44	30	37
IVb	3	46	35	37

b. Coder D

	A	I	J	M
I	44	17	39	23
II	36	33	33	35
III	40	41	39	40
IVa	4	6	11	7
IVb	3	2	3	7

c. Coder I

	A	D	J	M
I	16	17	15	24
II	33	33	34	38
III	45	41	32	39
IVa	44	6	33	39
IVb	46	2	35	38

d. Coder J

	A	D	I	M
I	46	39	15	28
II	29	33	34	30
III	38	39	32	38
IVa	30	11	33	36
IVb	35	3	35	38

e. Coder M

	A	D	I	J
I	31	23	24	28
II	38	35	38	30
III	41	40	39	38
IVa	37	7	39	36
IVb	37	7	38	38

These averages include all levels of the category system and as a result, coder D has a low average agreement due to the discrepancy involving Category 9 as mentioned above. To test this, the average agreements were recalculated for all of the coders without Categories 8 and 9. The results are given in Table 4.17.

TABLE 4.17 AVERAGE INTERCODER AGREEMENTS (LEVELS I - III)

Coder	Average Agreement	Percent
A	145.7	65.0
D	140.0	62.5
I	122.3	54.6
J	133.7	59.5
M	135.0	60.3

As a result, all of the agreements were increased with the exception of coder I which showed a difference of only 0.1percent. The agreements for coder D however were increased by 21.2%. This type of problem will be addressed in more detail when the categories are examined.

Intracoder Agreements To evaluate the intracoder agreements or the degree of consistency for each coder, the repeated classifications were first summarized in tabular form in much the same way as they were for the intercoder agreements. In this case however, the tables summarized the classifications for each of the 56 days and four levels for a particular coder as shown in Table 4.18.

TABLE 4.18 EXAMPLE OF AN INTRACODER SUMMARY TABLE
Phase I Day Coder D

	Sessions					Agreements Per Level			
	1	2	3	4	5				
Day #	I	I-II	I-III	I-IVa	I-IVb	I	II	III	IV
1	13	124	1247	124*9	124*9	4	4	2	2
2	**	***	****	*****	*****	5	4	3	2*
3	13	134	1247	12479	12479	3\2	4	3	2
4	13	135	1357	13579	13579	5	4	3	2*
5	13	135	1357	1357*	13579	5	4	3	0
51	12	124	124*	124*9	124*9	5	4	3	2*
52	1*	1*5	1*57	1*579	1*579	5	4	3	2*
53	0	1**	124*	1****	0	2\2\1	2\1\1	2	0
54	12	124	1247	12479	12479	5	4	3	2*
55	13	135	135*	135*9	135*9	5	4	3	2*
56	12	124	124*	124*9	12479	5	4	2	2

In Table 4.18, the first column lists the day numbers, and the next five columns give the categories assigned by Coder D for those days at each level. The four columns that follow are the agreement patterns for each day. Because of the scheme that was used, Level I was repeated five times, and therefore a complete consistency for this level is 5, for Level II it is 4, for Level III it is 3, and for Level IV it is 2. In the sample given in Table 4.18, Coder D was completely consistent for six days as indicated by an asterisk. In fact, Coder D displayed the highest overall intracoder reliability determined by counting the number of days for which the pattern of agreement was 5, 4, 3, 2. In this case, this was achieved on 28 of the 56 days (50%). The results for the other coders were as follows, A-13 (23%), I-17 (30%), J-12 (21%), and M-5 (9%). A less demanding assessment of the coders' consistencies is to examine the frequency with which each coder attained a complete consistency for each level rather than for all levels together as above. These results are given in Table 4.19, and are considerably higher than those given above.

TABLE 4.19 FREQUENCY OF COMPLETE CONSISTENCY PER LEVEL (%)

	Coder					
Level	A	D	I	J	M	Average
I	30	42	40	39	18	33.8
	(54)*	(75)	(71)	(70)	(32)	(60.4)
II	36	42	25	37	29	33.9
	(64)	(75)	(45)	(66)	(52)	(60.4)
III	50	45	30	36	34	39.0
	(89)	(80)	(54)	(64)	(61)	(69.6)
IV	52	54	34	35	51	45.2
	(93)	(96)	(61)	(63)	(91)	(80.7)
Average	42.0	45.8	32.3	36.8	33.0	
	(75.0)	(81.8)	(57.6)	(65.6)	(58.9)	

* $\frac{\text{frequency of consistency}}{\text{total possible}} \times 100$

This table also shows that certain categories presented a greater challenge to the coders than others as will be examined in the following section.

Categories The categories were assessed by examining the frequencies with which they were used by each coder. Table 4.20 shows the average number of times that each coder used each category.

TABLE 4.20 AVERAGE CATEGORY FREQUENCY PER CODER

	Categories														
	Level I						Level II			Level III			Level IV		
	0	1	*	2	3	*	4	5	*	6	7	*	8	9	*
A	2	53	2	19	**23**	2	20	20	12	**2**	23	29	10	**3**	42
D	2	53	1	15	**21**	9	21	23	9	**1**	22	32	1	**52**	2
I	0	55	1	16	**5**	**34**	31	14	11	**1**	17	38	12	**3**	42
J	0	56	0	27	**25**	4	22	24	10	**11**	20	25	17	**5**	34
M	1	54	1	19	**15**	21	24	18	13	**2**	21	32	7	**5**	43
Total	5	271	5	96	89	70	118	99	55	17	103	156	47	68	163
%	2	97	8	34	32	25	42	35	20	61	37	56	17	24	58

Three anomalous situations are evident in this table shown in bold type. The first, and most obvious, concerns Coder D and Category 9 - *ice not in motion.* This has also been noted in the examinations of inter- and intracoder agreements. In this case, Coder D made an interpretive decision while the others classified the entries more literally. The second observation involved Coder I and Category 3 - *large area covered by ice > 50%.* This is the opposite situation in which this coder most often decided that there was not enough information, this was also reflected in this coder's comments on the questionnaire:

> I was particularly uncomfortable with the ice cover information (<50% and 50%+). Not knowing any distances made this decision difficult.
>
> (Coder I)

The other coders however used Categories 2 and 3 almost equally. The third discrepancy was with the use of Category 6 - *thin ice* by coder J who used this category considerably more often than did the others. This may not however, be reflective of the same systematic disagreement as that which occured with Coder D because the discrepancy in this case is not as clearly illustrated in either of the other categories in this level (7 and *). Therefore the explanation for this is not readily available.

The ability of the categories to translate the meanings of the ice descriptions was, to a large extent, reflected in the frequency with which the coders decided that there was not enough information at each level. These frequencies are given in Table 4.21. The percentage frequencies are also shown for comparative purposes and were calculted on the basis of the maximum frequency possible for each level such that for Level I, the maximum was 56 days x 5 repetitions = 280; for Level II - 56 days x 4 repetitions = 224; for Level III - 56 days x 3 = 168; and for Level IV - 56 days x 2 = 112.

TABLE 4.21 FREQUENCIES OF 'NOT ENOUGH INFORMATION' CATEGORY

	Level						
	Ia	Ib	II	III	IV		
Coder	*(0,1)*	*(2,3)*	*(4,5)*	*(6,7)*	*(8,9)*	Total	Average
A	9	59	46	86	83	283	56.6
	3	21	21	51	74		
D	5	43	34	95	4	181	36.2
	2	15	15	57	4		
I	5	172	43	114	83	417	83.4
	2	61	19	68	74		
J	2	21	41	76	68	208	41.6
	1	8	18	45	61		
M	5	106	53	97	86	347	69.4
	2	38	24	58	77		
Total	26	401	217	468	324		
Average	5.2	80.2	43.4	93.6	64.8		

As anticipated, this category was used least for Level Ia (ice NOT present in vicinity of ship) since it was apparent that ice was present on most days. It was used for some units, however when it was not clear whether or not the ice was present *in the vicinity of the ship* as stated by the definition of this category. Level Ib (ice concentration) and Level III (ice thickness) showed the most frequent use of this category but apparently for different reasons. It was often difficult to determine the amount of ice present in the vicinity of the ship because the quantities were specified (less than, or greater than 50%). This illustrates the degree of difficulty that can arise when attempts are made to translate these types of subjective descriptions into numerical values. Another reason for difficulties with Level Ib was that the coders did not have information pertaining to the distances that the ships

traveled in one day, as was noted by Coder I. The asterisk was used most often for level III simply because ice thickness was rarely mentioned in the log books. Table 4.21 also shows that each coder required different amounts of information to make her or his decisions. Coder D for instance, was able to make a decision on an average of 77% of the units classified, while coder I decided that more information was required for 53% of the decisions.

Days Table 4.22 summarizes the inter- and intracoder agreements for each day. An 'X' indicates complete agreement for the level and/or coder for that day. In general, three observations can be made from Table 4.22:

1. There were more intracoder agreements than intercoder agreements.
2. There were no days with complete inter- and intracoder agreements.
3. There were no agreements at all for days 21 to 25.

The first observation was expected since the coders were free to devise their own schemes by which they could classify certain recurring descriptions but they were not allowed to collaborate and thereby increase the number of intercoder agreements.

TABLE 4.22 INTER - AND INTRACODER AGREEMENTS PER DAY

	INTERCODER					INTRACODER				
Day	LI	LII	LIII	LIVa	LIVb	A	D	I	J	M
1										
2			x			x	x	x	x	
3								x		
4							x		x	
5									x	
6										
7							x			
8										
9										
10										
11							x	x		
12							x	x		
13						x				
14	x									
15						x				
16						x	x	x		
17						x	x			
18								x		
19								x		
20						x	x			
21										
22										
23										
24										
25										
26	x									
27										
28										
29								x	x	
31										
32	x	x	x			x	x	x	x	x
33							x			
34							x			
35							x			
36										
37						x	x	x		
38							x			x
39										
40							x		x	
41						x	x			
42						x	x			x
43						x	x			
44							x		x	x
45	x								x	
46	x								x	
47										
48										
49							x	x		
50	x	x				x	x		x	x
51	x						x	x		
52							x			
53		x						x	x	
54							x			
55						x	x	x		
56								x	x	
Total	7	3	2	0	0	13	28	17	12	5

4.4b Hour

This textual unit gave rise to the unanticipated problem of yielding an exceptionally large amount of derived data. As a result, it was beyond the scope of this preliminary phase to examine the 261 entries with the same degree of detail as were the 56 days. A sample of 30 representative descriptive phrases is given below to show their diversity. Furthermore, these phrases did not exist in isolation but were often used in combination with each other. These combinations often created a new meaning.

REPRESENTATIVE DESCRIPTIVE PHRASES

Forcing close heavy ice	Ice close but much smaller	Striking heavy
Ice opening	Sailing among heavy straggling ice	Heaving off
Leaving thro' open ice	At grapple in heavy ice	Piece of ice
Sailing thro' open ice	Ice inclined to open	Ice close
Ice very close	Ship fast ice close	Ice begins to open
No open water in sight	Clear water	Boring thro' close ice
Ship is a complete iceberg	Thick and heavy	Passing thro' open sailing ice
Ice open and heavy	Ice in motion	Passing thro' slack ice
Beset in heavy ice	Ship rolling and striking	Passing thro' a deal of sailing ice
Ice more open but heavy	Streams of water	Between close ice and land

The hour unit was assessed by examining the intracoder agreements and the frequencies with which the categories were used after tabulating the classifications in the same way as the day unit . The total agreements in Table 4.23 show that Coder I had the highest overall consistency (89.4%). However, when each level is examined, it is evident that the reason for this is the high frequency with which Coder I used the *not enough information* category. This is particularly apparent in Levels Ib and II where the use of this category by the other coders was comparatively infrequent. In order to further assess the role played by this category in the intracoder agreements, the agreements for the *not enough information* category were subtracted from the total agreements and this produced the values given in Table 4.24.

TABLE 4.23 AVERAGE INTRACODER AGREEMENTS FOR THE 'HOUR' TEXTUAL UNIT

		A	D	I	J	M	Average percentage for level
Level Ia	0	5	8	5	7	5	
	1	220	234	228	243	232	
	*	13	1	7	0	1	
	%	91	93	92	96	91	*Ice Presence = 93*
Level Ib	2	49	58	3	84	5	
	3	88	73	5	95	70	
	*	42	59	223	1	23	
	%	69	73	89	69	38	*Ice Coverage = 68*
Level II	4	78	70	69	35	86	
	5	14	86	6	67	58	
	*	97	41	121	38	27	
	%	72	75	75	54	66	*Ice Fragmentation= 68*
Level III	6	2	5	1	36	2	
	7	6	44	41	62	43	
	*	220	182	204	86	136	
	%	87	89	94	70	69	*Ice Thickness = 82*
Level IV	8	19	4	27	47	15	
	9	0	179	0	8	0	
	*	220	49	227	156	210	
	%	92	89	97	81	86	*Ice Motion = 89*
Total		1073	1093	1167	965	913	
Percentag		82	84	89	74	70	

Percent agreements for each level = $\frac{\text{Total agreements for level}}{261} \times 100$

TABLE 4.24 INTRACODER AGREEMENTS WITH AND WITHOUT 'NOT ENOUGH INFORMATION' (*) CATEGORY

Coder	With * Agreements	%	Without * Agreements	%
A	1073	82.2	481	36.9
D	1093	83.8	761	58.3
I	1167	89.4	385	29.5
J	965	73.9	684	52.4
M	913	70.0	516	39.5

As Table 4.24 shows, 60% of the agreements for Coder I were in the category *not enough information,* whereas this category accounted for less than 50% of the agreements for the other four coders. Also evident in Table 4.23 is the same conflict involving Coder D and Level IV as was seen in the day unit. This coder continued to use Category 9 (ice NOT in motion) while the other coders decided that in most entries there was not enough information to determine whether the ice was in motion. Another repeated observation in Table 4.23 is the high number of agreements for Coder J in category 6 (thin layer of ice). The average for this category among the other coders was 2.5 while Coder J exceeded this by four times. In general, the highest consistencies were in Category Ia (presence or absence of ice) and the lowest was in Category Ib (areal extent of ice coverage). Coder I displayed the highest consistency in this category, however this was because 97% of these agreements were in the *not enough information* category.

4.4 c Word

The following is the list of words that were given to the coders to test the applicability of this textual unit. As indicated, most of these words were derived from a study by Catchpole and Halpin (1987).

ICE WORDS	**Body*	**Loose*	**Shattered*		
	Broken	*More Clear*	*Skim*		
	**Close*	**No Water*	**Slack*		
	**Fast*	**Open*	**Small*		
	**Field*	**Packed*	**Straggling*		
	**Floe*	*Patch*	**Thick*		
	**Heavy*	**Piece*	*Thin*		
	**Ledge*	*Ridge*	*Unbroken*		
NAVIGATION	*Along*	**Cannot Move*	**Forge Thro'*	**Meet*	**Saw*
WORDS	**Alter Course*	**Close To*	*Get Into*	**Moored*	*Stand Thro'*
	Among	**Cast Off*	*Get Thro'*	*No Opening*	*Steer Clear*
	**Anchored*	*Detain*	*Got Head From*	**Observe*	**Steer Thro'*
	**Appear*	**Drive Thro'*	*Got Past*	**Passed*	**Stopped*
	**Bear Away*	*Ease From*	**Grapple*	**Pass Thro'*	*Surrounded By*
	**Beset (set)*	*Enter*	**Haul Away (up)*	**Ply Between*	**Tacking*
	**Bore Thro'*	**Fall In With*	*Haul Thro'*	*Progress Thro'*	**Traversing*
	**Break Thro'*	**Fall Thro'*	*Heaving In*	**Rounding*	**Turning*
	**Bring To*	**Fast (in)*	*Hove To*	**Row And Tow*	**Warp*
	**Came up To*	**Force*	*Make Sail*	**Run In*	**Wore*
				**Sail Thro'*	**Work Among*

(* From Catchpole and Halpin, 1987:240)

Those words not marked with an asterisk appeared in the transcriptions used in this research but not in the Catchpole and Halpin study, those which are marked were used in both. The words were divided into two groups, those which described the navigational activities employed to deal with the sea ice and those that described the ice directly. Within each group the words were listed alphabetically to avoid biasing the coders' decisions. The coders were to provide a category number from each level (2 to 9*) excluding Categories 0 and 1 since ice presence was implied. This was repeated twice and the responses were then tabulated for inter- and intracoder agreements, and for the assessment of the categories.

TABLE 4.25 EXAMPLE OF A SUMMARY TABLE FOR THE 'WORD' TEXTUAL UNIT

a. Intercoder

	Coder					Agreements										
Word #	A	D	I	J	M	2	3	*	4	5	*	6	7	*	8	9	*
1	*5**	35**	*	357*	357*		3	2		4	1		2	3			5
2	24**	*4**	*4**	*468	*4**	1		4		5		1		4	1		4
3	3***	35**	*	3***	35**		4	1		2	3			5			5
4	35*9	35**	*	35*9	*5**		3	2		4	1			5		2	3
5	*9**	35**	*	3***	35**		3	2		3	2			5			5
53	24**	*	*4**	346*	*4**	1	1		4		1	1		4			5
54	24**	*	*	346*	24**	2	1	2	3		2	1		4			5
55	3*7*	*	*4**	357*	3*7*		3	2	1	1	3		3	2			5
56	24**	*	*	*	3*7*	1	1	3	1		4		1	4			5
57	34**	35**	*4**	346*	347*		4	1	4	1		1	1	3			5

b) Intracoder

	Session		Agreements											
Word	A	B	2	3	*	4	5	*	6	7	*	8	9	*
1. Body *5**	*5**			2		2				2			2	
2. Broken	*4**	24**	1		1	2					2			2
3. Close	3***	3***		2				2			2			2
4. Fast 35*9	35*9		2			2				2		2		
5. Field *5**	*5**			2		2				2			2	
53.Traversing	24**	24**	2			2					2			2
54.Turning	2***	24**	2			1		1			2			2
55.Warp	3*7*	3*7*		2				2		2				2
56.Wore	24**	24**	2			2					2			2
57.Work Among	34**	34**		2		2					2			2

Intercoder Agreements The numbers of complete four-digit intercoder agreements for each session are given below in Tables 4.26a and b.

TABLE 4.26 COMPLETE INTERCODER AGREEMENTS - WORD

a) Session A

		Coders					
		A	D	I	J	M	Average
	A		26	14	9	24	18.3
	D	26		26	17	34	25.8
Coders	I	14	26		6	11	14.3
	J	9	17	6		15	11.8
	M	24	34	11	15		21.0

b) Session B

		Coders					
		A	D	I	J	M	Average
	A		18	17	14	29	19.5
	D	18		22	12	25	19.3
Coders	I	17	22		10	15	16.0
	J	14	12	10		10	11.5
	M	29	25	15	10		19.8

The overall intercoder agreements were quite low, the averages were 22.5% for Session A and 21.3% for Session B. This was in part due to the fact that these figures are based on complete four-digit agreements. Tables 4.27a and b provide the intercoder agreements for each category by giving the number of coders in agreement. For example in Session A,

Category 2:
5 coders agreed on this category 2 times
4 coders agreed on this category 7 times
3 coders agreed on this category 3 times
2 coders agreed on this category 11 times
None of the coders agreed on this category 12 times.

TABLE 4.27 NUMBER OF INTERCODER AGREEMENTS PER CATEGORY

a.) Session A

Coders in Agreement	Categories 2	3	*	4	5	*	6	7	*	8	9	*
5	2	2	7	9	2	9	2	1	31	0	0	62
% (/81)	2	2	9	11	2	11	2	1	38	0	0	77
4	7	13	9	5	8	15	0	2	28	0	0	15
%	9	16	11	6	10	19	0	2	35	0	0	19
3	3	9	21	9	12	8	0	3	13	1	0	3
%	4	11	26	11	15	10	0	4	16	1	0	4
2	11	15	17	7	7	18	3	7	3	1	1	1
%	14	19	21	9	9	22	4	9	4	1	1	1
0	12	13	20	15	13	15	17	17	2	8	7	0
%	15	16	25	19	16	19	21	21	2	10	9	0

a.) Session B

Coders in Agreement	Categories 2	3	*	4	5	*	6	7	*	8	9	*
5	1	2	5	5	2	13	2	0	32	0	0	65
% (/81)	1	2	6	6	2	16	2	0	40	0	0	80
4	8	12	10	14	8	15	0	3	27	0	0	12
%	10	15	12	17	10	19	0	4	33	0	0	15
3	2	11	22	9	5	7	0	4	13	1	0	3
%	2	14	27	11	6	9	0	5	16	1	0	4
2	12	13	20	6	5	10	5	5	4	1	2	1
%	15	16	25	7	6	12	6	6	5	1	2	1
0	15	16	19	10	18	21	21	14	3	6	7	0
%	19	20	23	12	22	26	26	17	4	7	9	0

The highest number of agreements was in the *not enough information* category for Level IV which is not surprising when the list of words is re-examined. The overall low number of intercoder agreements for this textual unit illustrates the importance of context in CA. When these words were isolated from the entry in the log books, the coders no longer had enough information to make a decision about their meanings.

Intracoder Agreements Table 4.28 shows the intracoder agreements for the word textual unit.

TABLE 4.28 INTRACODER AGREEMENTS

Coder	Ice	Navigation	Total
A	23(96*)	49(86)	72(89)
D	23(96)	49(86)	72(89)
I	22(92)	47(82)	69(85)
J	8(33)	19(33)	27(33)
M	17(71)	27(47)	44(54)

* Percentages

The most notable observation here is that the general level of consistency is considerably higher than that of the intercoder agreements. This is due to the fact that there were only two repetitions and also most of the decisions were in the *not enough information* category. This is also evident in Table 4.29 which shows the frequencies and numbers of agreements for each category.

TABLE 4.29 INTRACODER AGREEMENTS PER CATEGORY

	Categories											
	2	3	*	4	5	*	6	7	*	8	9	*
A	40	70	52	59	16	87	4	14	144	2	2	158 # of decisions
	36	68	48	54	16	82	4	14	144	2	2	158 # agreements
	90	97	92	92	100	94	100	100	100	100	100	100 % agreements
D	23	49	100	23	47	92	6	10	146	0	0	162
	20	48	96	18	46	86	6	10	146	0	0	162
	87	98	96	78	98	93	100	100	100	0	0	100
I	13	5	144	65	15	82	4	4	154	2	0	160
	6	4	138	60	12	76	4	4	154	2	0	160
	46	80	96	92	80	93	100	100	100	100	0	100
J	59	65	38	57	60	45	48	39	78	17	15	130
	48	52	26	46	40	24	34	27	58	14	9	120
	81	80	68	81	67	53	71	69	74	82	60	92
M	29	87	46	51	50	61	11	37	114	1	1	160
	24	80	38	38	36	44	8	28	102	1	1	158
	83	92	83	75	72	72	73	76	89	100	100	99

4.5 Comparison Of Agreements

Intercoder/Intracoder Agreements Tables 4.30a and b give the percentage inter- and intracoder agreements. It should be noted that the percentage intercoder agreements given in Table 4.30a are the percentage number of times that each coder agreed with all of the other four coders. For comparative purposes, intercoder agreements were calculated for the last repetition of the hour textual unit.

TABLE 4.30a PERCENTAGE INTERCODER AGREEMENTS

Coder	Day	Hour	Word
A	60	18	24
D	41	10	24
I	54	13	20
J	56	10	14
M	58	15	24
Average	54	13	21

TABLE 4.30b PERCENTAGE INTRACODER AGREEMENTS

Coder	Day	Hour	Word
A	75	82	89
D	82	84	89
I	58	89	85
J	66	74	33
M	59	70	54
Average	68	80	70

It was surprising to find that there were proportionally more intercoder agreements for the day unit than for the others since there was a greater degree of difficulty involved in synthesizing all the information from one day into one set of categories. The intracoder agreements however, were the highest for the hour unit. To a large degree, this must be attributed to the fact that one entry was often repeated on a given day. It is important to note that the intracoder agreements were high for all three textual units. This indicates that the coders were well trained and that they participated conscientiously in these tests. Their

consistencies also demonstrated that they each had their own clear interpretations of the communications and categories.

Not Enough Information Instead of providing a category-by-category comparison for each textual unit, this section will focus on the *not enough information* category since it was the only one which was common to all four levels. When untested categories are used, as in this case, it cannot be assumed that all of the units can be classified into one, and only one category. It is, therefore, important to give the coders the opportunity to indicate those textual units that cannot be assigned to any of the categories at this early stage of development. Tables 4.31 a-c summarize the frequency with which this category was used.

TABLE 4.31 FREQUENCY OF 'NOT ENOUGH INFORMATION' CATEGORY

a.) Day

	Level					
	Ia	Ib	II	III	IV	Total
(Total possible)	280 (56x5)	280 (56x5)	224 (56x4)	168 (56x3)	112 (56x2)	1064
Coders						
A	9	59	46	86	83	283
%	3	21	21	51	74	27
D	5	43	34	95	4	181
%	2	15	15	57	4	17
I	5	172	43	114	83	417
%	2	61	19	68	74	39
J	2	21	41	76	68	208
%	1	8	18	45	61	20
M	5	106	53	97	86	347
%	2	38	24	58	77	33
Total Classified	26	401	217	468	324	436

b.) Hour

	Level					
	Ia	Ib	II	III	IV	Total
(Total possible)	1305 (261x5)	1305 (261x5)	1044 (261x4)	783 (261x3)	522 (261x2)	4959
Coders						
A	103	343	501	707	450	2104
%	8	26	48	90	86	42
D	17	449	218	573	122	1379
%	1	34	21	73	23	28
I	80	147	548	619	457	1851
%	6	11	53	79	88	37
J	1	60	475	349	349	1234
%	0	5	46	45	67	25
M	27	444	200	510	441	1622
%	2	34	19	65	85	33
Total Classified	228	1443	1942	2758	1819	8190

c.) Word

	Level					
	Ia	Ib	II	III	IV	Total
(Total possible)	N/A	162	162	162	162	648
Coders						
A		52	87	144	158	441
%		32	54	89	98	68
D		100	92	146	162	500
%		62	57	90	100	77
I		144	82	154	160	540
%		89	51	95	99	83
J		38	45	78	130	291
%		23	28	48	80	45
M		46	61	114	160	381
%		28	38	70	99	59
Total Classified		380	367	636	770	2153

The frequency with which this category was used was greater than 50% in Levels III and IV for all three textual units. It is possible, therefore, to conclude that there was not enough information in the log books to make a decision about ice thickness or motion. It is also possible to conclude that the word unit was the least useful since a single word, taken out of context, provides the least amount of information. Conversely, this category was used the least for the day unit because it provides the greatest amount of information in a useful context.

4.6 General Conclusions

This phase proved to be of critical importance in determining the direction that the remainder of the research would take. As a result of the tests conducted in this phase, it was clear that the textual units, the categories, and the method of determining reliability were inappropriate. Experimentation with the three textual units revealed that the resolution of the information contained within the Hour and Word units was too fine for a yearly climatic reconstruction, and that the resolution of the Day unit was too coarse for the coders to reliably interpret. Therefore, a new textual unit was required. The categories tested in this phase also posed a problem because they were based on a set of simple, preconceived geometrical and mechanical ice properties, and were derived neither from the log book contents nor from modern categories of ice observations. As a result, they could not be clearly defined and the coders frequently had difficulties in making classification decisions. The use of percentage agreements was found to be an inadequate method of measuring reliability. This approach simply gives a generalized description of the numbers of agreements and, therefore, does not provide a measure of the degree to which a procedure can be expected to repeatedly yield the same results using the same data.

The method employed in this phase derived from the rudimentary approach to CA and testing procedures that had been applied at that time in the field of paleoclimatology. Because the focus of this research was turned from the reconstruction to the methodology from which the data are derived, the problems inherent in this particular approach were revealed and the research was redirected towards their solution. This, then, became the objective of Phase II.

CHAPTER 5

PHASE II

5.1 Introduction

The experience gained in Phase I led to major refinements in the procedure that were implemented and tested in Phase II. The three elements of the methodology that needed to be changed as revealed in Phase I were the textual units, the categories, and the method of measuring reliability. Thus, Phase II began with a critical assessment of how these factors could be improved, and with an examination of the state of reliability testing in the social sciences.

The amount of information contained in the Hour and Word textual units had a resolution that was too fine for a yearly reconstruction of sea ice. The Day textual unit often contained so much information that it could not readily be digested into one classification representative of the whole day. A major weakness of the categories in Phase I was that they were impressionistically derived and, therefore, were not clearly defined. In Phase II the research adopted the principle that the categories should be directly comparable with the modern categories of ice observations and this decision was based on three considerations. The categories employed for the modern observations have been developed from experience and represent the ice information that can be detected by a person in the field. Conceivably, therefore, this body of ice information may be obtained from the log books with the greatest degree of reliability. Secondly, if these categories can be found to be applicable to the log book descriptions, then the derived historical

information will be directly comparable with the modern records of observations. Finally, diagrams and detailed definitions are available for the modern categories, and these can be used by the coders as aids to the standardization of their interpretations.

Throughout Chapter 4, the word *agreement* was deliberately used in place of the word *reliability* because percentage agreements are not true measures of reliability.As discussed in Chapter 3, this approach is biased and merely describes the number of agreements that resulted from a specific test. What is needed, is a means of comparing the observed number of agreements with the expected agreements to account for the element of chance. This gives the level of agreement that can be expected if the test is repeated, thereby providing a true measure of reliability.This problem was resolved by examining the current research on reliability testing being conducted in the social sciences. As a result, Krippendorff's agreement coefficient (α) was adopted as the means of assessing reliability in Phase II.

5.2 Phase II Parameters

Log Book Sample In this phase, the sample of transcriptions was reduced from 56 days to 26 by using every second page beginning with Day 3 (August 3, 1843). The sample of log book days used in Phase II is given in Table 5.1.

TABLE 5.1 PHASE II SAMPLE OF LOG BOOK PAGES

DAY # PHASE I	DAY # PHASE II	YEAR	DATE	ARCHIVE CALL NUMBER	SHIP
3	1	1843	Aug 3	670	Prince Albert
5	2	1843	Aug 5	670	Prince Albert
7	3	1855	Aug 18	846	Prince of Wales
9	4	1855	Sep 2	846	Prince of Wales
11	5	1859	Sep 18	719	Prince Arthur
13	6	1848	Aug 7	683	Prince Albert
15	7	1803	Aug 13	756	Prince of Wales
17	8	1809	Aug 29	772	Prince of Wales
19	9	1796	Aug 7	741	Prince of Wales
21	10	1810	Aug 1	774	Prince of Wales

(...Continued)

TABLE 5.1 Continued

DAY# PHASE I	DAY # PHASE II	YEAR	DATE	ARCHIVE CALL NUMBER	SHIP
23	11	1810	Aug 5	774	Prince of Wales
25	12	1829	Aug 25	224	Camden
27	13	1829	Aug 5	818	Prince of Wales
29	14	1829	Aug 25	818	Prince of Wales
31	15	1827	Aug 6	813	Prince of Wales
33	16	1827	Aug 16	813	Prince of Wales
35	17	1821	Aug 6	795	Prince of Wales
37	18	1821	Aug 16	795	Prince of Wales
39	19	1815	Aug 15	783	Prince of Wales
41	20	1808	Aug 3	420	King George
43	21	1808	Aug 9	420	King George
45	22	1795	Aug 22	398	King George
47	23	1782	Aug 12	386	King George
49	24	1775	Aug 6	898	Prince Rupert
51	25	1774	Aug 16	378	King George
53	26	1761	Aug 13	1031	Seahorse

Having tested the 56 days and 261 entries five times each in Phase I, it was concluded that the sample could be reduced without losing a significant amount of information. Furthermore, upon completion of the previous phase, all of the coders were in agreement that the 56-day sample was excessive and this could potentially affect the coders' ability to maintain their level of concentration throughout a test.

Textual Units. As revealed by the tests in Phase I, it was necessary to adopt a new textual unit. The primary reason for this was the resolution of the Phase I units. This has an impact on the reliability with which the sea ice information can be obtained from the log books by the coders, and on the reconstruction that will ultimately result from the derived data. In this regard, the day unit was an appropriate unit of time over which to calculate an ice index. It was found, however, that the ice descriptions written in one day could not be reliably interpreted when the ice conditions changed throughout the day. Therefore, the textual unit has to be less than one day for categorisation purposes. The hour unit, however, was too fine in resolution for reconstruction purposes. Thus, a major innovation in Phase II was the adoption of a new textual unit, namely the seaman's watch. Every four

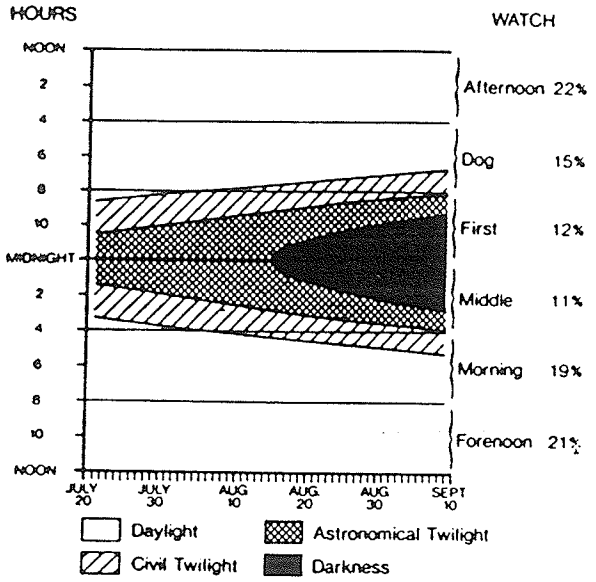
hours a new set of crew members assumed the ship's watch, and the first task of the officer of the watch was to assess the sailing conditions. Table 5.2 shows the times of the six daily watches.

TABLE 5.2 SHIP'S WATCHES

Watch	Time
Afternoon	noon - 4:00 p.m.
Dog	4:00 - 8:00 p.m.
First	8:00 - midnight
Middle	midnight - 4:00 a.m.
Morning	4:00 - 8:00 a.m.
Forenoon	8:00 - noon

Of primary importance was the safety of the ship and her crew, and the greatest threat to safety was from sea ice. Therefore, the watch would commence with an assessment of the sea ice conditions and a review of the entries of the previous watch to ensure that the conditions were described faithfully. The system of watches encompassed the entire 24-hour day, and so some of the watches fell during the night. While this was responsible for reducing the frequency of ice comments, they were not omitted. Furthermore, at this latitude during the summer months, the duration of daylight hours is extended. Figure 5.1 illustrates the times of the ships' watches; their corresponding hours of daylight, civil twilight, and astronomical twilight; and the proportions of ice descriptions for each watch.

FIGURE 5.1 HOURS OF DAYLIGHT AND PROPORTIONS OF ICE DESCRIPTIONS PER WATCH



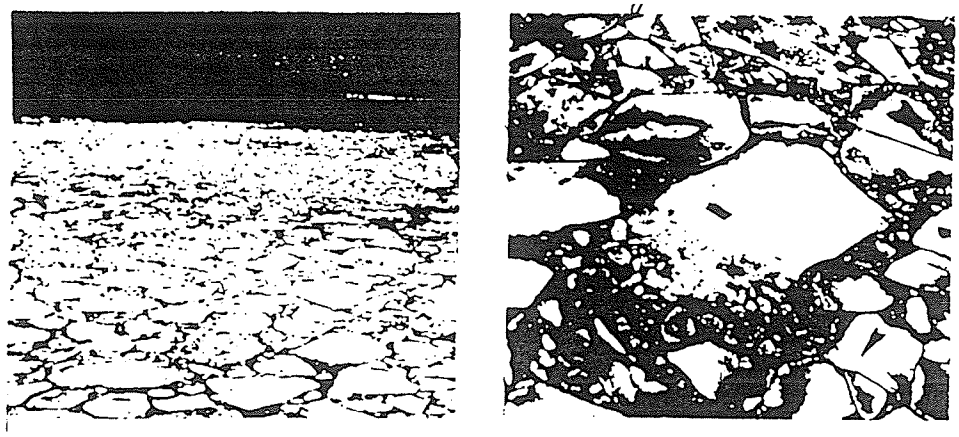
(Catchpole and Halpin, 1987:235)

As Figure 5.1 shows, although there were fewer descriptions during the First and Middle watches, they did contain sea ice information. In Phase II then, the watch was tested to determine whether it would be the most reliable unit even though the resolution of this unit was also too fine. Should the tests conducted in this phase reveal the watch unit to be the most reliable, then it would ultimately be necessary to combine the classifications for the six daily watches to provide an ice index for each day.

Categories The simple set of preconceived categories used in Phase I proved to be inappropriate as indicated by the frequency with which the *not enough information* category was used. The decision was made at the beginning of this Phase to develop new categories related to the ice observation categories used by modern observers. The Phase II categories were thus obtained from the *Ice Observer Extension Course - Training Manual - Ice Terminology and Symbology* (Environment Canada, 1984). This book is used to train people in observing and recording sea ice conditions. It provides ice observers with a

standardized method of identifying various ice parameters such as ice concentration, age, forms of floating ice, arrangement of sea ice, openings in the ice, and pack ice motion. The terminology, definitions, and observing practices described in this manual were derived from experience that has shown that these represent the ice information that can be detected by a person in the field. Therefore, these categories might also be the information that can be most reliably extracted from the log books. Another advantage of this approach to category construction is that the historical data will be directly comparable with the modern observations. This manual also provided detailed definitions for each category that could be given to the coders to help them in making their classification decisions which was not possible in Phase I. One adjustment that would have to be made, however, because the vantage point from the masthead of a sailing ship offered an oblique view while the modern observations are made vertically from an aircraft. This factor would primarily affect the observations of ice concentration as illustrated in Figure 5.2.

FIGURE 5.2 APPEARANCE OF SEA ICE FROM 75m (a) AND FROM AERIAL VIEW (b)



a b

(Bauer and Martin, 1980:2052-2053)

As these illustrations show, the ice appears to be more consolidated towards the horizon when viewed obliquely. From the masthead of a ship which would be 35-40m lower than the vantage point in Figure 5.2a, this affect would be more exaggerated. Therefore, although the range of vision from the top of the masthead has been reported to be from 12.5 to 21 nautical miles (Catchpole, 1989), the distortions due to the perspective would reduce the range of accurate observations. The coders were alerted to this problem and instructed to adjust their interpretations of the log book descriptions to take this distortion into consideration.

The categories, as derived from the training manual, are given in Table 5.3. The definitions and diagrams which accompanied the categories will be discussed in Section 5.3 - Phase II Procedure.

TABLE 5.3 PHASE II CLASSIFICATION SYSTEM

GENERAL CLASSIFICATION	CATEGORY NUMBER	CATEGORY
Concentration	A. 1	Ice free
	2	Open water
	3	Very open ice
	4	Open ice
	5	Close ice
	6	Very close ice
	7	Consolidated/compact ice
Floe Size	B. 1	Giant floe
	2	Vast floe
	3	Big floe
	4	Medium floe
	5	Small floe
	6	Ice cake
	7	Small ice cake
Openings	C. 1	Crack
	2	Open lead
	3	Blind lead
	4	Shore lead
	5	Flaw lead

(Continued...)

TABLE 5.3 Continued

Arrangement	D. 1	Ice field
	2	Belt
	3	Tongue
	4	Strip
	5	Ice edge (compacted)
	6	Ice edge (diffuse)
	7	Concentration boundary
Motion	E. 1	Diverging
	2	Converging
	3	Shearing

As in Phase I, it was possible for the coders to indicate those textual units which could not be categorized. It was anticipated that many of these categories would be too precise to be of use with the log book descriptions, but a major reason for conducting the reliability tests was to isolate those categories that could be used.

5.3 Phase II Procedure

The Phase II tests were also preceded by a meeting of the five coders, however, at this meeting a certain amount of discussion was allowed regarding the categories, definitions, and diagrams. At this meeting, the coders were given the information contained in Table 5.4 .

TABLE 5.4 PHASE II NOTES AND INSTRUCTIONS TO CODERS

This phase will include more detail, information, and 'possibly' less work. You have been provided with the following:

1. 26 days to be classified (divided into watches)
2. 1 day before and after the code day
3. code forms
4. definitions
5. diagrams

There are five general classifications (A-E) each with a number of categories (#):

A. Concentration (7)
B. Floe size (7)
C. Openings (5)
D. Arrangement of sea ice (7)
E. Motion (3)

(Continued...)

TABLE 5.4 Continued

NOTES:
- Each watch will have a code, if there is a significant change during a watch, apply the code to the most predominant (the longest in duration).
- The blue pages are the days before (#a) & after (#b) the day to be classified. DO NOT PROVIDE CODES FOR THESE DAYS.
-Where there is no blue page (a,b, or neither) there was no ice mentioned so these days were not transcribed.
-You will be asked to repeat this process 3 times with one full day between each.
-Please do not discuss this with each other.
-Please keep a log of your difficulties, thoughts, and suggestions as you code.
-Assume grappling is to ICE
-Assume an optimum field of vision to be 13 nautical miles on all sides from the top of a 120-foot masthead. Remember, some of the watches were at night which of course would diminish the visibility.

The most noteable difference in this phase is that the coders were given considerably more information on which to base their decisions. One such addition was the incorporation of the adjacent days which came before and after the day being classified. This came about as a direct result of the responses to the Phase I questionnaire. The purpose of this was to aid in the classification of those days which did not have an entry for the first or last watch of the day. As is noted in the instructions, some days were lacking one or both of the adjacent days because there were no ice comments given on those days. The format of the transcriptions of days to be classified and of the adjacent days was the same as those for Phase I except that the hourly entries were separated into the six four-hour watches.

Another difference between this phase and Phase I was that the coders were given a separate set of forms in which to enter their classifications for each watch, a sample of this form is given in Table 5.5.

TABLE 5.5 SAMPLE CLASSIFICATION FORM

Category

Day 1	X	A	B	C	D	E	Y
Watch 1							
2							
3							
4							
5							
6							
Day 2							
Watch 1							
2							
3							
4							
5							
6							
Day 3							
Watch 1							
2							
3							
4							
5							
6							
Day 4							
Watch 1							
2							
3							
4							
5							
6							

Category

Day 5	X	A	B	C	D	E	Y
Watch 1							
2							
3							
4							
5							
6							
Day 6							
Watch 1							
2							
3							
4							
5							
6							
Day 7							
Watch 1							
2							
3							
4							
5							
6							
Day 8							
Watch 1							
2							
3							
4							
5							
6							

Using the classification system given in Table 5.3, the coders were to enter the category numbers that best described the ice observations for each watch. In this phase, the coders were given the complete set of categories at the beginning rather than following the hierarchical system used in Phase I. In those cases where there was not enough information for a particular category, the coders were to enter a dash (-). When it was not possible to provide a category number for any of the five general classifications, the coder entered an 'O' in the column labeled 'Y'; and when there was no entry for the watch but a classification could be inferred, the coder placed an 'O' in the 'X' column followed by the

classification numbers. For those watches in which there was no entry and a classification was not possible, a line was drawn from column 'X' to 'Y' to indicate that the watch was not accidentally overlooked. An example of this process is given in Table 5.6.

TABLE 5.6 EXAMPLE OF PHASE II CLASSIFICATION PROCEDURE

Day *eg*	X	A	B	C	D	E	Y
Watch 1		3	6	–	4	–	
2	0	4	6	2	–	1	
3							0
4							
5	0	1	–	–	–	–	
6		5	5	4	–	–	

INTERPRETATION

Watch

1	Very open ice, Ice cake, Strip
2	No entry, Open ice, Ice Cake, Open lead, Diverging
3	No Classification possible but there was an entry
4	No entry, No Classification
5	No entry, Ice free
6	Close ice, Small floe, Shore lead

To help the coders in making their decisions in a more standardized manner than in Phase I, they were issued the set of definitions for each category given in Table 5.3 .

TABLE 5.7 CATEGORY DEFINITIONS

A. CONCENTRATION

1. *Ice Free:* No ice present. If ice of any kind is present this term should not be used.
2. *Open Water:* A large area of freely navigable water in which sea ice is present in concentrations less than 1/10
3. *Very Open Ice:* Ice in which the concentration is 1/10 to 3/10 and water preponderates over ice.
4. *Open Ice:* Ice in which the concentration is 4/10 to 6/10 with many leads and the floes are generally not in contact with one another.
5. *Close Ice*: Ice in which the concentration is 7/10 to 8/10 composed of floes mostly in contact.
6. *Very Close Ice:* Ice in which the concentration is 9/10 to less than 10/10.
7. *Consolidated/Compact Ice:* Ice in which the concentration is 10/10, no water is visible and the floes are frozen together.

B. FLOE SIZE

1. *Giant Floe:* Over 10km across
2. *Vast Floe:* 2-10 km across
3. *Big Floe:* 500 - 2,000 m across
4. *Medium Floe:* 100 - 500 m across
5. *Small Floe:* 20 - 100 m across
6. *Ice Cake:* Any relatively flat piece of sea ice less than 20 m across.
7. *Small Ice Cake:* An ice cake less than 2 m across.

C. OPENINGS

1. *Crack:* A small, unnavigable, narrow break in sea ice that may reveal the sea water surface. Cracks are usually caused by tides, temperature change, current and/or wind.
2. *Open Lead:* A long navigable break in pack ice which may vary in width from approximately 50 feet to a few miles and which provides passage to vessels for an indeterminate distance. When describing this type of lead, the lengthwise dimension should be indicated. An open lead can abruptly terminate inside the pack ice.
3. *Blind Lead:* The same width dimension as an open lead except that one end of the lead narrows and ends inside the pack ice.
4. *Shore Lead:* A lead between pack ice and the shore or between pack ice and an ice front.
5. *Flaw lead:* A passage-way between pack ice and fast ice which is navigable by surface vessels.

D. ARRANGEMENT

1. *Ice Field:* Area of pack ice consisting of any size of floes, which is greater than 10 km across. An ice field is so called because of its size only. The effects of pressure, erosion or age have no part in the definition.
2. *Belt:* A large feature of pack ice arrangement, longer than it is wide, from 1 km to more thatn 100 km in width.
3. *Tongue:* A projection of the ice edge up to several km in length, caused by wind or current.
4. *Strip:* A long, narrow area of pack ice, about 1 km or less in width, usually composed of small fragments detached from the main mass of ice, and run together under the influence of wind, swell or current.

5 & 6. *Ice Edge:* The demarcation at any time between the open sea and sea ice of any kind, whether fast or drifting. It may be termed compacted or diffuse.

7. *Concentration Boundary:* A line approximating the transition between two areas of pack ice with distinctly different concentrations.

E. MOTION

1. *Diverging:* Ice fields or floes in an area are subjected to diverging or dispersive motion, thus reducing ice concentration and/or relieving stress in the ice.
2. *Compacting:* Pieces of floating ice are said to be compacting when they are subjected to a converging motion, which increases ice concentration and/or produces stresses which may result in ice deformation.
3. *Shearing:* An area of pack ice is subject to shear when the ice motion varies significantly in the direction normal to the motion, subjecting the ice to rotational forces. These forces may result in phenomena similar to flaws.

(Environment Canada, Atmospheric Environment Service, *Ice Observer Extension Course - Training Manual - Ice Terminology and Symbology.* 1984)

It should be noted that although the definitions included measurements of ice concentrations and floe sizes, the categories were treated as ordinal rather than interval scales for the purpose of identifying the various ice conditions.

As indicated by the responses given in the questionnaire at the end of Phase I, all of the coders visualized the ice descriptions while they classified them. In doing this, each coder had in mind two sets of visual interpretations, one for the log book descriptions and another for their interpretations of the categories. While it is not possible to standardize the ways that the log book descriptions were pictured, it was possible to provide the coders with the same visual image of the definitions. This was facilitated by the set of diagrams depicting certain categories for four of the five general classifications, concentration, floe size, openings, and arrangements given in Figures 5.3 a-d. These diagrams were obtained from the *Ice Observer Training Manual.*

The instructions for this phase also included guide-lines for instances in which there was a significant change in the ice conditions during one watch. In such a case, the coders were to base their decision on the condition which lasted the longest in the four hour period. In previous studies, the decisions were based on the most severe conditions for the day (Faurer, 1981, Catchpole and Halpin, 1987), even though that situation may have only lasted for one or two hours. This was an appropriate procedure for those studies which incorporated the duration of the voyage as a variable in calculating an ice index because severe sea ice conditions were the primary factors responsible for protracted voyages (Faurer, 1981). In this research, the goal is to test the reliability with which the descriptions could be interpreted and to determine the textual unit which would best facilitate that process. Since duration was not a factor, it was decided that the greatest amount of information could be obtained by using the most prevalent conditions for each watch on those occasions when there was a significant change in the conditions within a watch.

FIGURE 5.3a PHASE II DIAGRAMS - ICE CONCENTRATION

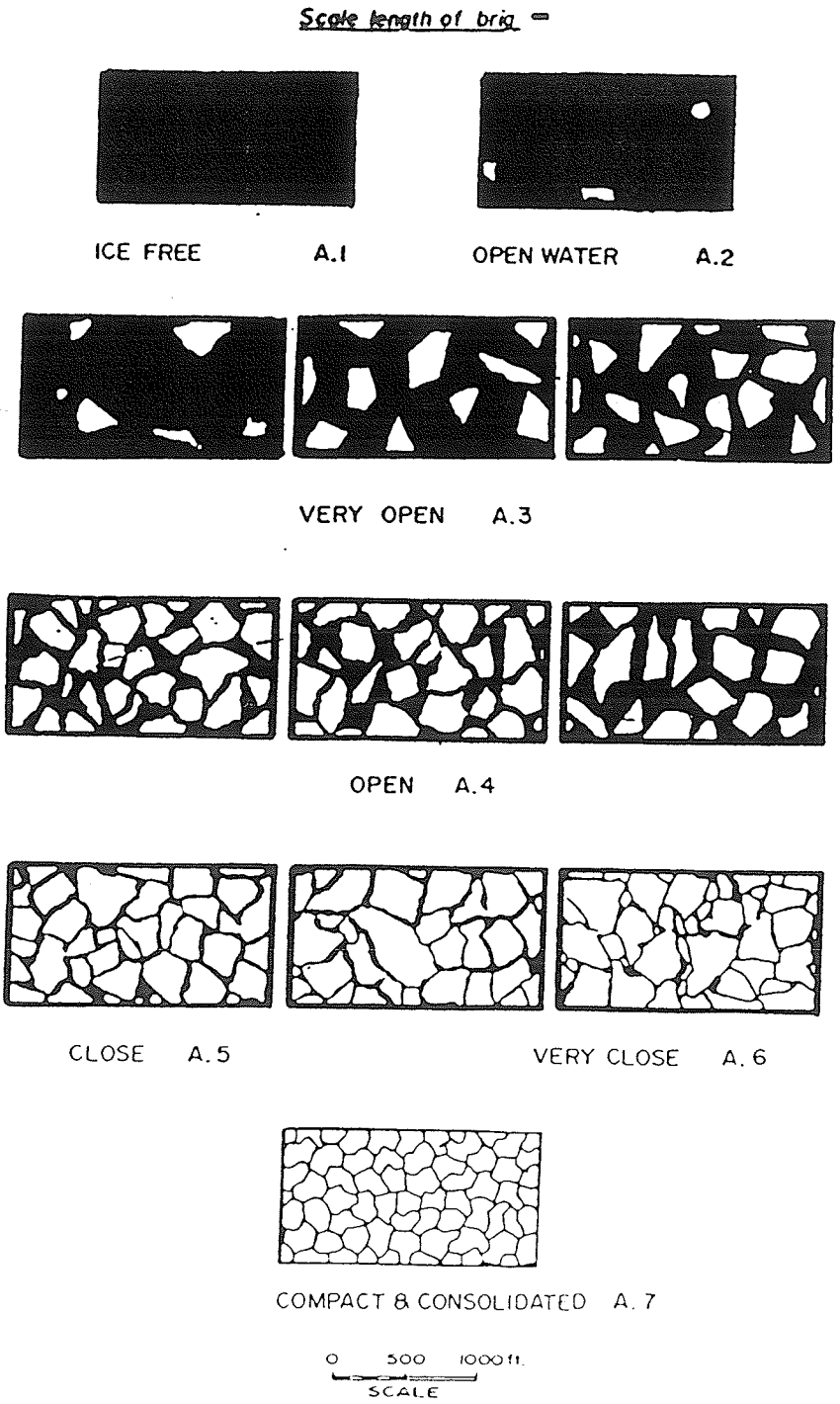


FIGURE 5.3b PHASE II DIAGRAMS - FLOE SIZE

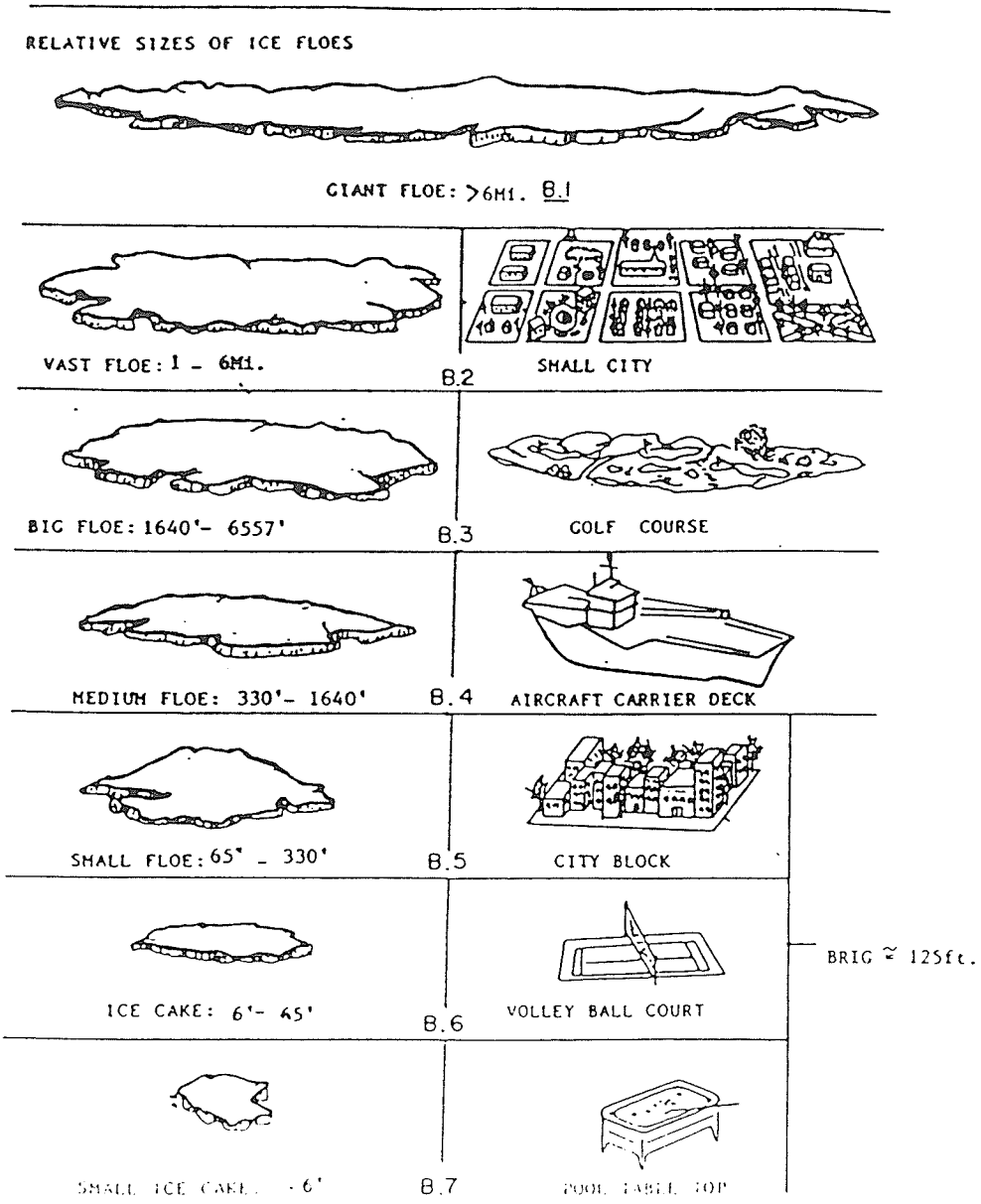


FIGURE 5.3c PHASE II DIAGRAMS - ICE OPENINGS

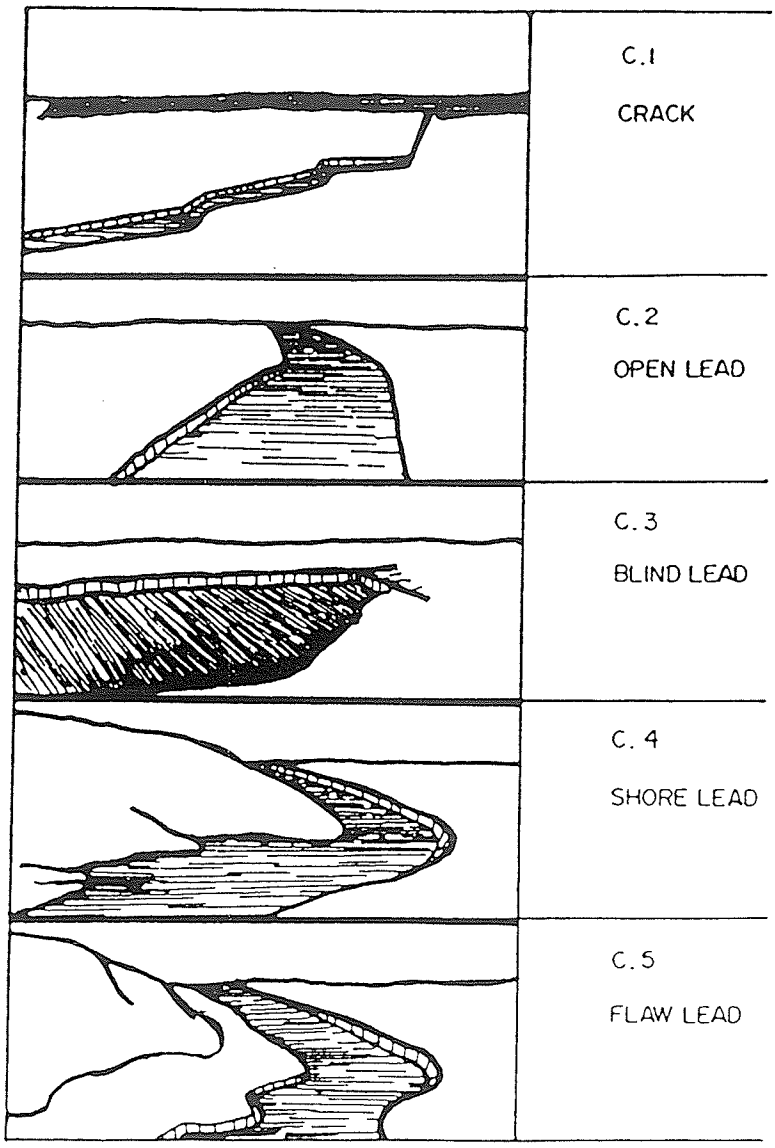
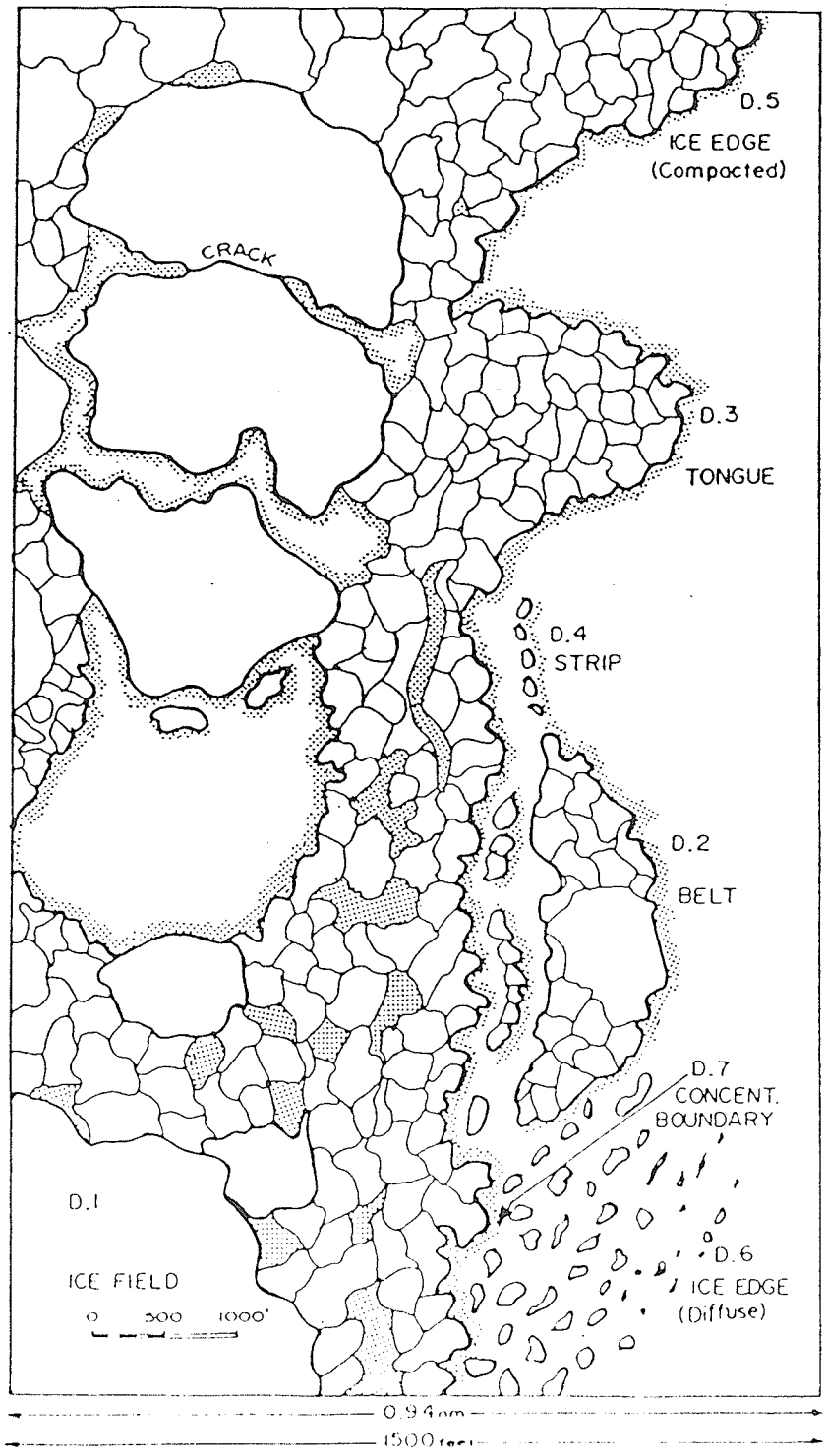


FIGURE 5.3d PHASE II DIAGRAMS - ICE ARRANGEMENT



5.4 Evaluation of Reliability

The test results were evaluated by using Krippendorff's agreement coefficient (α) as described in Chapter 3. This coefficient was used in the assessment of both intra- and intercoder reliabilities as well as in the modification of the category sets.

Intracoder Reliability The classification procedure described in the preceding section was repeated three times by each coder so that the intracoder reliabilities could be assessed. Once the coders had completed three sessions of tests, their decisions were transfered to a table to facilitate the calculation of α. An example of this is given in Table 5.8. Even before calculating the coefficients, it was apparent from these tables that the categories for the 'Ice Openings' classification were used so infrequently (Table 5.9) that they should be omitted from the evaluation process.

TABLE 5.8 EXAMPLE OF INTRACODER SUMMARY TABLE PHASE II - CODER A

General Classification: A- Concentration Coder: A

		Session			Category							
Day	Watch	1	2	3	1	2	3	4	5	6	7	-
1	1	3	3	3			3					
	2	3	3	3			3					
	3	3	3	3			3					
	4	3	3	3			3					
	5	5	5	5					3			
	6	4	3	4			1	2				
2	1	5	6	5					2	1		
	2	5	5	4				1	2			
	3	4	4	4				3				
	4	4	4	4				3				
	5	4	4	4				3				
	6	4	4	5				2	1			
26	1	3	1	-	1		1					1
	2	-	1	-	1							2
	3	-	1	-	1							2
	4	-	1	-	1							2
	5	-	1	-	1							2
	6	-	1	-	1							2

This table shows the category assigned to each watch by Coder A by the three columns under the Session heading. The next eight columns contain the frequency with which the categories were used for each watch, for example, for Watch 1 of Day 2, Category 5 was assigned twice in Sessions 1 and 3 and Category 6 once in Session 2. From this table, the numbers of agreements and disagreements were counted.

TABLE 5.9 FREQUENCY OF ICE OPENINGS CLASSIFICATION

	Coders				
Session	A	D	I	J	M
1	1	3	3	1	1
2	1	1	1	1	1
3	4	1	1	2	2
Total	6	5	5	4	4
%*	1	1	1	0.9	0.9

* % = frequency / total possible x 100. Total possible = 26 days x 6 watches x 3 sessions = 468.

The Ice Concentration classification was used most frequently as shown in Table 5.10.

TABLE 5.10 FREQUENCY OF ICE CONCENTRATION CLASSIFICATION

	Coders				
Session	A	D	I	J	M
1	144	148	69	107	99
2	130	148	86	106	135
3	142	148	99	106	132
Total	416	444	254	319	366
%	89	95	54	68	78

Once all of the agreements and disagreements were counted from the summary tables, they were entered into coincidence matrices from which the agreement coefficients for each coder and the four general classifications could be calculated. The matrix for Coder A, classification A is given in Table 5.11 as an example.

TABLE 5.11 INTRACODER COINCIDENCE MATRIX FOR ICE CONCENTRATION CLASSIFICATIONS MADE BY CODER A

Category	1	2	3	4	5	6	7	-	
1	60	48	2	2	3	0	0	25	140
2	48	22	3	4	9	1	0	19	106
3	2	3	198	16	7	0	0	18	244
4	2	4	16	96	20	0	0	6	144
5	3	9	7	20	94	18	5	3	158
6	0	1	0	0	18	2	5	0	26
7	0	0	0	0	5	5	0	0	10
-	25	19	18	6	3	0	0	35	108
									936

In this matrix, the diagonal entries are the agreements for each category and the off-diagonal figures are the disagreements. Using Equations 7, 8, and 9 from Chapter 3, the agreement coefficient for this example was calculated to be 0.451 which means that only 45% of the agreements are repeatable. It should be pointed out here that this type of table is also of value in illustrating the patterns of disagreement. For example, in Table 5.11 there is some confusion between Categories 1 and 2, and Categories 4 and 5. Furthermore, it shows that the highest number of agreements are found in Category 3. The intracoder reliabilities for this phase were actually calculated twice, once using the blank entries and once without. However, when the blanks are included, the set of categories is no longer of an ordinal scale and, therefore, in calculating α, d_{bc} was equal to one. For comparative purposes, the nominal scale was also used when the intracoder reliabilities were calculated without those watches for which the coder could not supply a classification. Table 5.12

provides the results of these calculations. It should be noted that this table includes only those general classifications for which the coder classified more than 5% of the watches.

TABLE 5.12 SUMMARY OF INTRACODER AGREEMENT COEFFICIENTS α

CODER	GENERAL CLASSIFICATION	A INCLUDING BLANKS	B EXCLUDING BLANKS	DIFFERENCE (A - B)
A	Concentration	.451	.529	-.078
	Floe Size	.511	.361	.150
	Arrangement	.667	.506	.161
	Motion	.836	.669	.167
D	Concentration	.657	.650	.007
	Arrangement	.479	.592	-.113
I	Concentration	.502	.714	-.212
	Arrangement	.446	.500	-.054
	Motion	.690	1.000	-.310
J	Concentration	.908	.828	.080
	Floe Size	.960	.488	.472
	Arrangement	.733	1.000	-.267
M	Concentration	.435	.601	-.166
	Floe Size	.540	.569	-.029
	Arrangement	.492	.073	.419
	Motion	.617	.951	-.334
Average		.620	.627	-.007

These coefficients indicate the degree to which the agreements were better than chance. The average coefficient which included the blanks was marginally lower than that which did not include the blanks. In general, the intracoder reliabilities were quite high, only seven were below 0.5 which means that in 25 cases the coders' consistencies were at least 50% better than would be expected by chance. In the case of the average value (0.627) for example, only 37% of the agreements would be expected by chance while 63% of the consistencies could be repeated. On two occasions, there were coefficients of 1.0.

However, this only occurred when the blanks were not included. This indicates that the only inconsistencies in these decisions were those that involved blanks. Since the inclusion of the blanks did not greatly change the overall intracoder reliability (average difference = -0.007), the coefficients could be recalculated omitting the blanks and using the ordinal scale. This was not done, however, for floe size (B) because it was not used in more than 5% of the watches by all of the coders. The arrangement (D) and motion (E) classifications were not included because their categories were not ranked. The nominal and ordinal intracoder coefficients for the ice concentration classification for each coder are given in Table 5.13.

TABLE 5.13 NOMINAL AND ORDINAL INTRACODER AGREEMENT COEFFICIENTS - ICE CONCENTRATION

Coder	A Ordinal	B Nominal	Difference (A-B)
A	.787	.529	.258
D	.920	.650	.270
I	.897	.714	.183
J	.959	.828	.131
M	.906	.601	.305
Average	.894	.664	.229

In every case, the coefficients were increased by using the ordinal scale calculations which means that most of the disagreements involved adjacent categories. This increase was anticipated when the coincidence matrices were examined and revealed that most of the disagreements were between adjacent categories. This was a very important observation for those categories that are ranked because a disagreement between two adjacent categories should not be given the same weighting as a disagreement between two polar categories. The calculation of the d_{bc} values for the ordinal scale assigns higher weightings as the distance between the categories increases. Therefore, when the nominal scale calculations

were applied, all disagreements were counted as one ($d_{bc} = 1$) and the observed disagreements were higher than when the ordinal scale was applied.

In general, all of the coders were able to repeat the tests with a high degree of reliability. The average coefficient for the ordinal scale (0.894) means that on average, 89% of the classifications were repeatable. The fact that all of the coefficients were improved by the use of the ordinal values of d_{bc} is very significant in terms of modifying the category structure. It suggests that the distinctions between the categories might be too fine for the information contained in the log book descriptions. Therefore, the coders were consistent in determining whether the ice concentration was increasing or decreasing but they did not agree on the amount of ice coverage as defined by the seven categories.

Intercoder Reliability The intercoder reliability was evaluated in basically the same way as the intracoder reliability. The agreements among the coders were counted and tabulated as shown in Table 5.14.

TABLE 5.14 SAMPLE INTERCODER SUMMARY TABLE PHASE II
ICE CONCENTRATION - SESSION 1

		Coders					Categories							
Day	Watch	A	D	I	J	M	1	2	3	4	5	6	7	-
1	1	3	3	3	4	3			4	1				
	2	3	3						2					3
	3	3	2	2	3	2		3	2					
	4	3	3	3	4	3			4	1				
	5	5	5	5	5	4				1	4			
	6	4	4	4	4	4				5				
						.								
						.								
						.								
26	1	3							1					4
	2													5
	3													5
	4													5
	5													5
	6		2		3			1	1					3

A table such as this was completed for the three sessions of each general classification. In calculating the coefficient for the intercoder reliability, the value for m was the number of coders as defined in Equation 9 (Chapter 3) rather than the number of repetitions which was used to calculate the intracoder agreement coefficients. The 'ice openings' classification was also omitted from the intercoder calculation due to infrequent usage. Table 5.15 gives the agreement coefficients for the remaining four classifications including and excluding the blank watches for each.

TABLE 5.15 SUMMARY OF INTERCODER AGREEMENT COEFFICIENTS (NOMINAL)

		α		
SESSION	GENERAL CLASSIFICATION	A INCLUDING BLANKS	B EXCLUDING BLANKS	DIFFERENCE (B - A)
1	Concentration	.302	.423	.121
	Floe Size	.110	.510	.400
	Arrangement	.086	.127	.041
	Motion	.348	.709	.361
2	Concentration	.317	.451	.134
	Floe Size	.011	.317	.306
	Arrangement	.104	.146	.042
	Motion	.288	.796	.508
3	Concentration	.326	.468	.142
	Floe Size	.014	.201	.187
	Arrangement	.111	.203	.092
	Motion	.247	.550	.303
Average		.175	.408	.220

For the same reasons as given for the intracoder reliabilities, the ordinal scale coefficients were again only calculated for the ice concentration classification and are given in Table 5.16.

TABLE 5.16 INTERCODER NOMINAL AND ORDINAL COEFFICIENTS FOR ICE CONCENTRATION

Session	A Nominal	B Ordinal	Difference (B-A)
1	.423	.733	.311
2	.451	.786	.325
3	.468	.860	.392
Average	.450	.793	.343

Again, the ordinal calculations produced higher coefficient values which indicates a confusion between adjacent categories.

As discussed in Chapter 2, the number of categories can be determined by either increasing a small number of general categories or by decreasing a large number of specific categories. This phase followed the second approach. In this way, Phase II began with 29 categories in five general classifications. As will be discussed in the next section, this was reduced to 14 categories and four general classifications by examining the frequency and reliability with which each of the original categories was used.

5.5 Modification of the Category Set

In the context of this research, the purpose of the reliability tests is to determine those categories which can be applied most reliably to the log book descriptions. As illustrated in Figure 2.3, the classification system is repeatedly tested and modified until the reliability cannot be improved further or until further simplification causes the loss of too much information and very vague uninformative categories are produced. In this stage of Phase II, tests were conducted to determine whether the agreement coefficients given in Tables

5.15 and 5.16 could be improved by grouping the categories. Although the categories for all of the general classifications were grouped and tested, the first part of this discussion will focus on the ice concentration categories.

Since the agreement coefficients for the three repetitions were increased by an average of 34% (0.343) when the ordinal scale was used, it was anticipated that a further improvement could be obtained by amalgamating certain adjacent categories. These tests were based on the original coincidence matrices for Phase II since it is not necessary for the coders to repeat the classification process for each new combination of categories. To facilitate these evaluations, the values in the cells of the original matrix are simply added together as specified by each grouping. This is illustrated in the hypothetical example given in Figure 5.17.

FIGURE 5.17 EXAMPLE OF GROUPING CATEGORIES IN A MATRIX

Original Matrix

Categories	A	B	C	D	E	F	
A	8	3	1	0	2	0	14
B	3	5	2	0	0	0	10
C	1	2	6	4	1	1	15
D	0	0	4	2	2	0	8
E	2	0	1	2	7	1	13
F	0	0	1	0	1	4	6
							66

Grouped Matrix

Categories

	AB	CD	EF	
AB	19	3	2	24
CD	3	16	4	23
EF	2	4	13	19
				66

SUMMARY OF AGREEMENT COEFFICIENTS

	Original Matrix (O)	Grouped Matrix (G)	Difference (G-O)
Nominal (N)	.409	.807	.398
Ordinal (Or)	.559	.689	.13
Difference (Or-N)	.150	- .118	

In the original matrix above, the disagreements involved adjacent categories which implies that these categories are ambiguous. By combining all of the adjacent categories in which the majority of disagreements are found, the agreement coefficient is raised for both the nominal and ordinal scales. The increase is higher for the nominal scale, however, because the calculation of d_{bc} for the ordinal scale compensated for some of the disagreement. In this research, five different groupings of the ice concentration categories all produced higher agreement coefficients than the ordinal scale value for the original matrix. Tables 5.18a-g show all of the matrices and agreement coefficients used to modify the ice concentration category set and Table 5.19 which follows is a summary of the category groupings and coefficients.

FIGURE 5.18 GROUPING OF PHASE II ICE CONCENTRATION CATEGORIES

a. Original Matrix - Nominal

Categories *	1	2	3	4	5	6	7	
1	168	105	2	0	0	0	0	275
2	105	138	86	10	0	0	0	339
3	2	86	434	47	1	0	0	570
4	0	10	47	218	107	16	6	404
5	0	0	1	107	208	43	11	370
6	0	0	0	16	43	12	11	82
7	0	0	0	6	11	11	0	28
				$\alpha = .468$				2070

b. Original Matrix - Ordinal

Categories	1	2	3	4	5	6	7	
1	168	9.45	1.08	0	0	0	0	178.53
2	9.45	138	16.34	8.3	0	0	0	172.09
3	1.08	16.34	434	10.34	.71	0	0	462.47
4	0	8.3	10.34	218	14.98	5.6	2.52	259.74
5	0	0	.71	14.98	208	2.15	.77	226.61
6	0	0	0	5.6	2.15	12	.033	19.78
7	0	0	0	2.52	.77	.033	0	3.32
				$\alpha = .860$				1322.54

c. Grouping A - Ordinal

Categories	1	2/3	4/5	6/7	
1	168	10.53	0	0	178.53
2/3	10.53	604.68	19.35	0	634.56
4/5	0	19.35	455.96	11.04	486.35
6/7	0	0	11.04	12.07	23.11
			$\alpha = .900$		1322.54

(Continued...)

(Figure 5.18 Continued)

d. Grouping B - Ordinal

Categories	1	2-4	5/6	7	
1	168	10.53	0	0	178.53
2-4	10.53	859.96	21.29	2.52	894.3
5/6	0	21.29	224.3	.8	246.39
7	0	2.52	.8	0	3.32
					1322.54

$\alpha = .892$

e. Grouping C - Ordinal

Categories	1	2-4	5-7	
1	168	10.53	0	178.53
2-4	10.53	859.96	23.81	894.3
5-7	0	23.81	225.9	249.71
				1322.54

$\alpha = .894$

f. Grouping D - Ordinal

	1	2-6	7	
1	168	10.53	0	178.53
2-6	10.53	1126.84	3.32	1140.69
7	0	3.32	0	3.32
				1322.54

$\alpha = .910$

g. Grouping E - Ordinal

Categories	1	2-7	
1	168	10.53	178.53
2-7	10.53	1133.48	1144.01
			1322.54

$\alpha = .932$

* The categories for this classification are:
1 = Ice free, 2 = Open water, 3 = Very open ice, 4 = Open ice, 5 = Close ice, 6 = Very close ice, 7 = Consolidated ice.

TABLE 5.19 SUMMARY OF CATEGORY GROUPINGS AND INTERCODER AGREEMENT COEFFICIENTS FOR ICE CONCENTRATION

Matrix	Category Grouping	Number of Categories	α (Ordinal)
Original (nominal)	1,2,3,4,5,6,7	7	.468
Original (ordinal)	1,2,3,4,5,6,7	7	.860
A	1,2/3,4/5,6/7	4	.900
B	1,2-4,5/6,7	4	.892
C	1,2-4,5-7	3	.894
D	1,2-6,7	3	.910
E	1,2-7	2	.932

Before discussing these results, it should be noted that all of the above figures were based on the results of the third test session. By that time, the coders had the most experience with the classification system and, therefore, conducted the tests with the greatest degree of confidence. One of the most useful aspects of this type of reliability evaluation is that experimental category groupings can be evaluated without having the coders apply them in a series of tests. Once the best grouping has been determined, then that one set can be tested by the coders. The assessment of the trial category sets in Figure 5.18 was the last stage of Phase II, and the testing of the best grouping by the coders comprised Phase III which will be discussed in Chapter 6.

The agreement coefficient for the original matrix was increased by 0.392 to 0.860, which means that only 14% of the agreements among the coders were made by chance. Although 0.860 represents a high degree of reliability, the category groupings were assessed to determine whether this value could be improved, and as Table 5.19 shows, the 0.860 coefficient was increased in every case. It is interesting to observe that reducing the number of categories does not necessarily increase the coefficient even though grouping the categories decreases the number of disagreements. For example, Grouping A has a value of 0.900 with four categories while C has a coefficient of 0.894 with three categories. The decision to omit or to include a category is not based solely on the

agreement coefficient, it also depends on the amount of information that can be obtained from the category set. The two highest α values were for Groupings D (0.910) and E (0.932) yet they were both omitted because they reduced the resolution of the potential sea ice information to a very low level. This was not acceptable because the coefficients only exceeded that of Grouping A by 0.01 and 0.32 respectively. In this case, the substantial loss of information by the vague categories in D and E was not compensated for by a comparable increase in reliability. Grouping C was rejected for both reasons. Its coefficient was the second lowest of the five groupings and it was comprised of only three very general categories. The final decision, between A (0.900) and B (0.892) was the most difficult to make. Both groupings yielded high coefficients which differed by only 0.02, and both provided a fairly high degree of information with four categories each. It was decided, however, that Grouping A would be retained because it more accurately represented the coders' responses. Since Category 7 (consolidated ice) was rarely used by the coders it was reasonable to group it with category 6 (very close ice). In all of the groupings, Category 1 - ice free was kept as a separate category because it is distinct from the others which describe the ice. As illustrated in Tables 5.18a and b, most of the disagreements were in adjacent categories and therefore it was logical that they should be grouped together. In Grouping A, all of the six remaining categories were grouped in pairs, in B Category 7 was left separate and the three 'open' categories were grouped together as were the two 'closed' categories. The major difference between A and B was in the basis for the grouping. In A, the categories were grouped according to the coders' agreements and disagreements while B was based on the wording of the categories such that the 'open ice' categories were grouped together and the 'close ice' categories were grouped. In summary, there were three major reasons for the decision to keep A. The first is that Category 7 was rarely used, and when it was used it was a part of a disagreement. Therefore, there was no reason to keep it as a separate category. Secondly, there were

twice as many disagreements between Categories 2 and 3 as there was between 3 and 4 which means that the coders could distinguish between the definitions for 'very open ice' and 'open ice' more clearly than between the definitions for 'open water' and 'very open ice' so there was no reason to sacrifice resolution by grouping these three 'open' categories together. Finally, even though Category 4 was termed 'open ice' and Category 5 was 'close ice' there was considerable confusion between these categories among the coders. This is shown in Tables 5.18 a and b by the high number of disagreements among the coders. It was logical therefore, to group them together.

Although the remaining three general classifications (B-floe size, D-arrangement, E-motion) were used less frequently than ice concentration, they were also grouped. The results of these groupings are given in Table 5.20.

TABLE 5.20 AGREEMENT COEFFICIENTS OF CATEGORY GROUPS FOR FLOE SIZE, ARRANGEMENT, AND MOTION

General Classification	α (Original)	Category Grouping	α (Grouped)	Difference (Grouped-Original)
B. Floe Size*	.310	1-3,4/5,6,7	.592	.282
D. Arrangement	.203	1&5, 2,3,4/6/7	.736	.533
	.203	1&5,2/3,4/6/7	.734	.531
E. Motion	.550	1&3, 2	.667	.117

*Agreement coefficients for floe size were calculated for the ordinal scale; arrangement and motion used the nominal scale because their categories were not ordered.

5.6 General Conclusions

The innovations introduced in Phase II were the direct result of the tests conducted in Phase I. They represented a departure and an advancement in the methodology from the impressionistic approach that was employed in Phase I and which was typical of paleoclimatic reconstructions based on the CA of written historical sources.

In arriving at a new textual unit, the record-keeping procedure of the log books was considered rather than their format. In doing so, the Watch textual unit was found to be the most useful for the coders. This was to be expected because the ice descriptions contained within each watch form a cohesive unit of information that was reported by a single observer. This unit offered enough information for a classification decision to be made without providing a wide range of ice conditions. When there was a variation within the watch, it was possible to determine the most prevalent condition by refering to the previous and/or following watch.

In Chapter 2, the requirements for the development of categories were discussed. The three requirements are that the categories should be unambiguous, exhaustive, and directly related to the log book descriptions and the modern sea ice data. The degree to which these criteria are met by the categories is largely a reflection of the reliability with which they can be used. In Phase I, the categories were impressionistically derived and, therefore, they were not directly related to the log book contents or to the modern data, and because they were not clearly defined, they were often ambiguous. These problems were resolved in Phase II by adopting the categories that are used for modern observations. This source also made it possible to provide the coders with detailed descriptions and diagrams in order to reduce the ambiguity of the categories. The results of the tests of this set of categories, however, indicated that there was still a degree of uncertainty regarding the distinction between some of the categories.

The technique adopted in Phase II to assess reliability made it possible to resolve the problems of ambiguity. Krippendorff's agreement coefficient was the most recent contribution to reliability testing in CA at the time that Phase II commenced. It made it possible to determine the proportion of agreements that could be expected when tests of a given set of categories were conducted. In the process of calculating this value, the factors that lowered the agreement coefficient were revealed making it possible for them to be

resolved. One factor that was responsible for lowered coefficients was due to the fact that the Concentration and Floe Size categories were ranked but the agreement coefficient was calculated for the nominal scale. Once the coefficient was calculated for the ordinal scale using the same data, the coefficients were improved. It was also clear from the coincidence matrices, that the resolution of the categories was often too high for the information contained in the textual units. This created confusion for the coders between certain adjacent categories. Using this technique, the resolution of the categories can be reduced by grouping similar categories together . This process relates to the question of the number of categories that are required. This was addressed in Chapter 2 in which two approaches to changing the number of categories were given. In one, categories are simply added or deleted, and in the other, they are either subdivided to give a higher resolution or grouped together to create fewer categories and lower resolution. Once these categories were grouped together, the coefficients were recalculated, producing higher levels of reliability. One of the advantages of Krippendorff's agreement coefficient is that it facilitates the evaluation of the reliability of the grouped categories without requiring the coders to recategorize the textual units. This makes it less difficult and time consuming to experiment with the category groupings. In this research, several different category groupings were tested and the observed and expected disagreements (D_o and D_e) and the agreement coefficients (α) are given in Appendix II. At the conclusion of Phase II, a new, condensed set of categories had been created. However, the ability of the agreement coefficient to accurately predict reliability had not yet been formally tested in this research. Although the average value for the four sets of categories that include Concentration, Floe Size, Arrangement, and Motion, had increased from 0.468 for the original categories at the beginning of Phase II to 0.724 for the final set of grouped categories, these categories could not be accepted until the coders had tested them by using them to categorize the textual units. This, then, became the function of Phase III, to apply the new sets of

categories to test two factors. The first was the reliability of the categories and the second factor was the ability of the agreement coefficient to predict reliability based on recalculation without the reapplication of the categories by the coders , so that in the future, this step may not be necessary.

CHAPTER 6

PHASE III

6.1 Introduction

By the time that Phase III commenced, the approach to the research and all of its central components had been modified considerably. The procedure had advanced from being impressionistic and subjective to one that followed the general principles of the scientific method, and which was focused on objectivity and reliability. This was the result of an evolutionary process of experimentation and modification that was largely conducted in Phase II. This second phase left two elements to be tested. The first was the new set of categories for the ice concentration, floe size,arrangement, and motion classifications that were developed by grouping the categories in PhaseII. While agreement coefficients had already been calculated for the grouped categories in Phase II, they had not been applied to the textual units by the coders. Secondly, the success with which agreement coefficients could predict the reliability of a particular set of categories had not yet been tested in this research. Therefore, in Phase III, the coders applied the Phase II grouped categories to the textual units to test both of these elements. The agreement coefficients for the new categories would be determined and if they corresponded closely with those calculated in Phase II, then Krippendorff's agreement coefficient could be accepted as a prognostic device for category reliability making this final test unnecessary in future research.

In Phases I and II, during the processes of tabulating, examining, and evaluating the coders' decisions, the question of the reliability of relative ice concentrations arose.

That is, perhaps the coders the coders were in agreement on whether the ice concentration was increasing or decreasing even though they may not have agreed on the precise concentration categories. This question did not stem from a systematic or deliberate investigation, but simply from cursory observations of the results of the repeated classifications, and yet the answer to this question could lead to important conclusions regarding the resolution of the derived data that can be reliably obtained from the log books. The notion to investigate the reliability of relative ice concentration categories originated in Phase I and was incorporated into the research at the end of Phase II by which point there was some suggestion that this level of resolution might be the most reliable because the calculation of the ordinal agreement coefficients produced increased values. One final test was therefore added to assess the reliability of relative ice concentration categories.

6.2 Phase IIIa Parameters and Procedure

The tests conducted in Phase III were based on the same 26-day sample of log book pages and textual units as those used in Phase II. The major difference between the two phases was the category set. Phase III employed the categories that resulted from the grouping procedures of Phase II, and therefore, they had a lower resolution. In the process of grouping the original Phase II categories, their definitions were also grouped in the same way so that they were not altered, but simply broadened. The categories and definitions used in this phase are given in Table 6.1.

TABLE 6.1 PHASE III CATEGORIES AND DEFINITIONS

General Classification	Category	Definition
A.Concentration	1. Ice free	- No ice of any kind present.
	2. Open water/ Very open ice	- Concentration ≤30%. More water than ice.
	3. Open ice/ Close ice	- Concentration 40% to 80%. Floes may be in contact with each other.
	4. Very close/ Consolidated/ Compact ice	- Concentration 90% to 100%. 100% - no water is visible and floes are frozen together.
B.Floe Size*	1. Small ice cake	- An ice cake <2m across.
	2. Ice cake	- Any relatively flat piece of sea ice <20m across.
	3. Small/Medium floe	- 20-500m across.
	4. Big floe	- 500-2000m across.
C. Arrangement	1. Strip/Diffuse ice edge/Concentration boundary	- A long narrow area of ice (1km or less in width) mostly of small fragments run together by wind or currents / an irregular line limiting an area of dispersed ice, usually on the leeward side of an area of pack ice/a line approximating the transition between two areas of pack ice with distinctly different concentrations.
	2. Belt	- A large feature of pack ice arrangement, longer than it is wide,from 1km to more than 100 km in width.
	3. Tongue	- A projection of the ice edge up toseveral km in length, caused by wind or current.
	4. Ice field/Compacted ice edge	- An area of pack ice consisting of any size of floes which is >10km across/a clear-cut line compacted by wind or current usually on the windward side of an area of pack ice.
D. Motion	1. Diverging	- Ice floes subjected to diverging or dispersive motion thus reducing ice concentration and/or relieving stress in the ice.
	2. Compacting	- Pieces of floating ice are compacting when they are subject to a converging motion, which increases the concentration and/or produces stresses which may result in ice deformation.

* The order of the categories for Floe Size was reversed from that used in Phase II, due to the observation by Coder J that all other classifications increased from category 1 while in the Floe Size classification, the largest floe size was category 1.

(After: Environment Canada, 1984)

The procedure for classifying the 156 watches was also the same as that of Phase II except that there were fewer general classifications and categories, and only two sessions. The most notable difference was that for Phase III the coders were required to provide an ice

concentration category number for every watch whether or not there was an entry. The coders were again provided with a set of diagrams which were modified to correspond to the new category definitions. These diagrams are given in Figures 6.1 a to c. The forms used by the coders for the Phase IIIa tests were basically the same as those for Phase II as shown in Table 6.2. The column labeled 'X' in Table 6.2 was checked by the coder when there was no entry given for the particular watch.

TABLE 6.2 PORTION OF PHASE IIIa CLASSIFICATION FORM

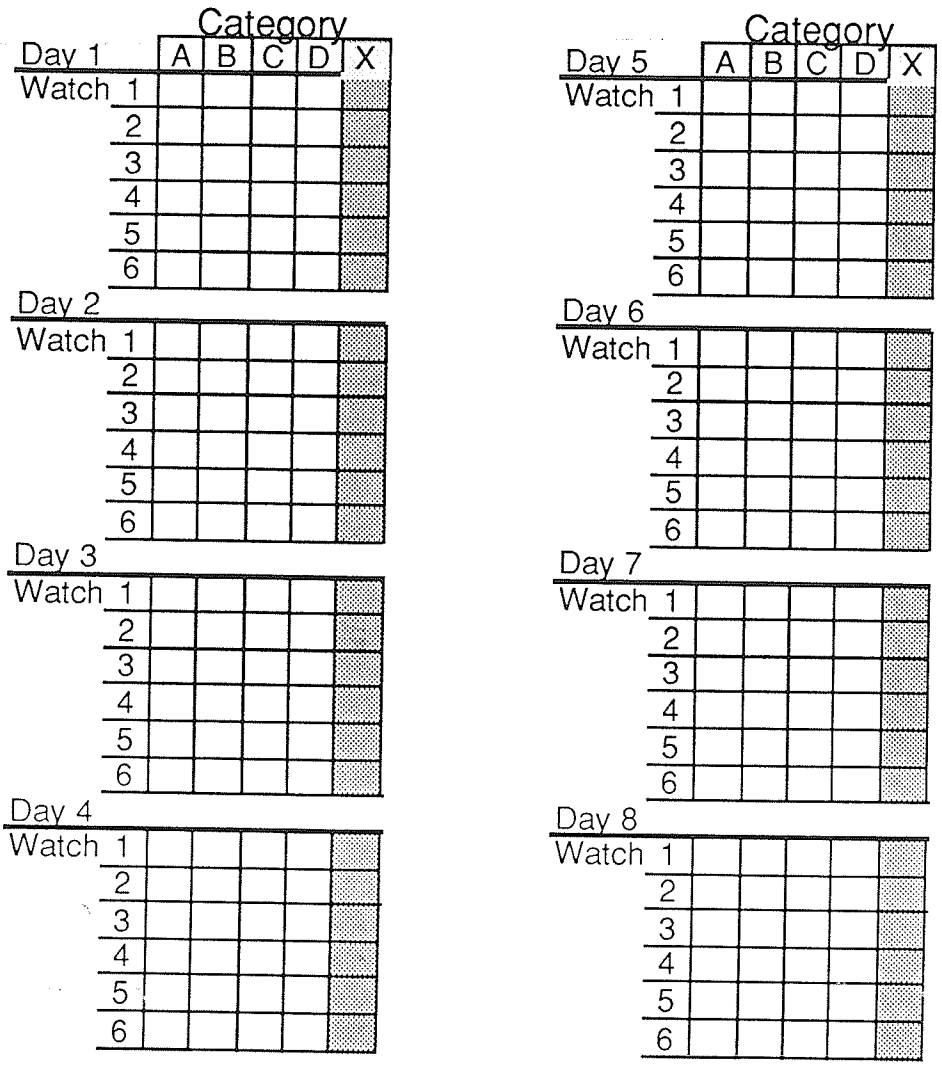
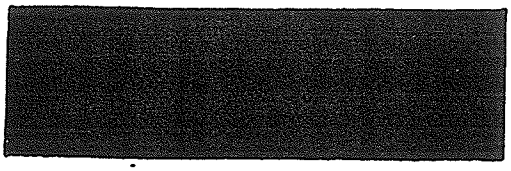
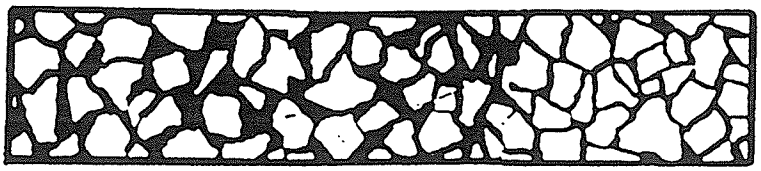


FIGURE 6.1a DIAGRAMS FOR ICE CONCENTRATION CATEGORIES

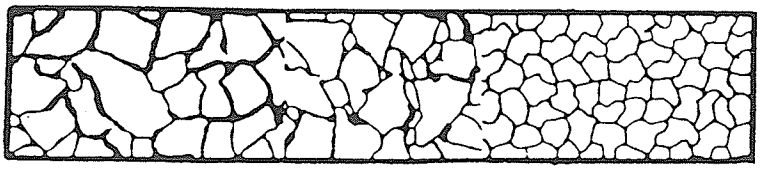


1 ICE FREE (*no ice*)

2 OPEN WATER / VERY OPEN ICE (*<10%, to 30%*)

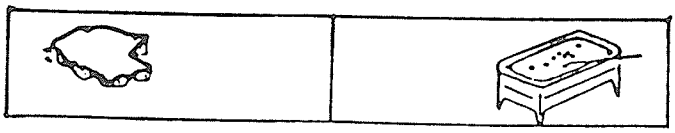


3 OPEN ICE / CLOSE ICE (*40% to 80%*)

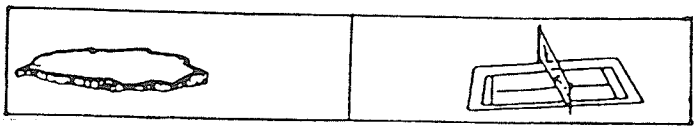


4 VERY CLOSE / CONSOLIDATED / COMPACT ICE (*90% to 100%*)

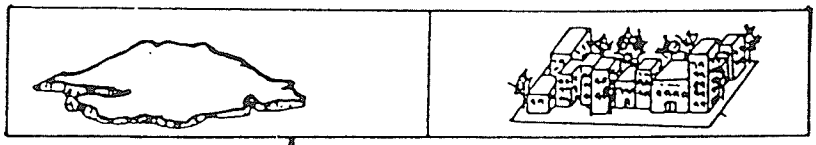
FIGURE 6.1b DIAGRAMS FOR FLOE SIZE CATEGORIES



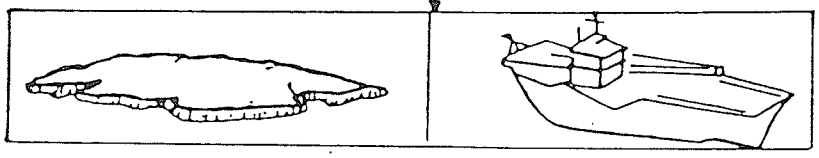
1 SMALL ICE CAKE (< 2m across)

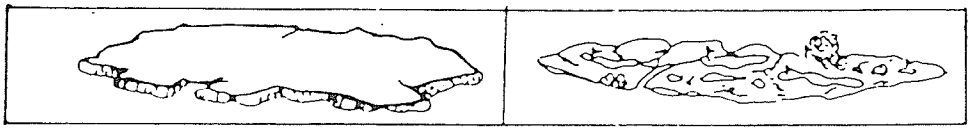


2 ICE CAKE (2m to 20m)



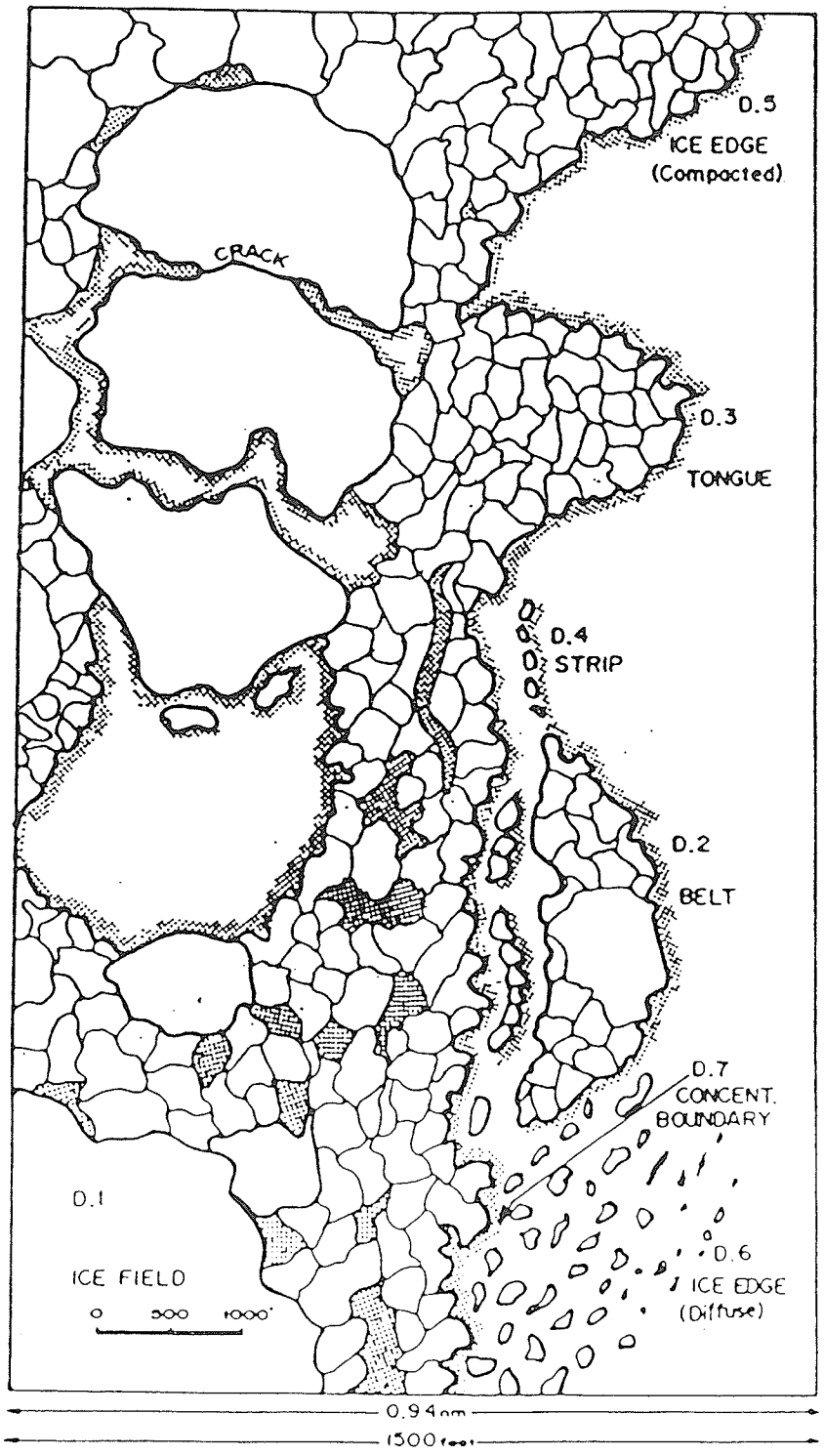
3 SMALL / MEDIUM FLOE (20m to 500m)





4 BIG FLOE (500m to 2,000m)

FIGURE 6.1c DIAGRAMS FOR ICE ARRANGEMENT CATEGORIES



Once the coders had completed the two sessions, their decisions were tabulated as they were for Phase II so that the intra- and intercoder reliabilities could be evaluated. Examples of these tables are given in Tables 6.3a and b.

TABLE 6.3 EXMPLE OF PHASE IIIa SUMMARY TABLES

a) Intracoder

Coder A

Session

Day	Watch	1	2
1	1	2	2
	2	2	2
	3	2	2
	4	2	2
	5	3	3
	6	2	2
	.		
	.		
	.		
26	1	1	1
	2	1	1
	3	1	1
	4	1	1
	5	1	1
	6	2	2

Classification A - Concentration

Category

1	2	3	4
	2		
	2		
	2		
	2		
		2	
	2		
2			
2			
2			
2			
2			
	2		

b) Intercoder

Session 1

Coder

Day	Watch	A	D	I	J	M
1	1	2	2	2	3	2
	2	2	2	2	2	2
	3	2	2	2	2	2
	4	2	2	2	3	2
	5	3	4	3	4	4
	6	2	3	2	3	3
	.					
	.					
	.					
26	1	1	1	1	1	1
	2	1	1	1	1	1
	3	1	1	1	1	1
	4	1	1	1	1	1
	5	1	1	1	1	1
	6	2	2	2	2	1

Classification A-Concentration

Category

1	2	3	4
	4	1	
	5		
	5		
	4	1	
		2	3
	2	3	
5			
5			
5			
5			
5			
1	4		

Before calculating the agreement coefficients, the frequency with which each classification was used was determined with the exception of Classification A - concentration which was compulsory for each watch according to the instructions. These frequencies are given in Table 6.4.

TABLE 6.4 FREQUENCY OF CLASSIFICATION USE BY EACH CODER

	B. Floe Size Session				C. Arrangement Session				D.Motion Session			
Coder	1		2		1		2		1		2	
A	103	*66.0**	106	*68.0*	15	*9.6*	18	*12.0*	18	*12.0*	20	*13.0*
D	5	*3.0*	5	*3.0*	27	*17.0*	21	*13.0*	11	*7.0*	6	*4.0*
I	4	*3.0*	3	*2.0*	7	*5.0*	4	*3.0*	22	*14.0*	13	*8.0*
J	25	*16.0*	25	*16.0*	1	*0.6*	0	*0.0*	2	*1.0*	1	*0.6*
M	46	*29.0*	42	*27.0*	24	*15.0*	27	*17.0*	8	*5.0*	8	*5.0*
Average	37	*23.5*	36	*23.0*	15	*9.6*	14	*9.0*	12	*7.8*	10	*6.2*

* Numbers in italics are percentages.

At this point, it was decided that the above three classifications would not be included in the reliability evaluation as a result of their infrequent use with the exception of the high frequency with which Coder A applied the floe size classification. This decision was of significance because it provided a partial answer to the question of the resolution of information that the log books could reliably yield. Although the reliability for these classifications was not actually assessed, their infrequent use by the coders indicated that they felt there was not enough information in the log book descriptions to determine the size of ice floes, the arrangement of the ice, or the type of motion. The concentration classification had a 100% response in Phase III because it was compulsory, however, the decision to make this mandatory for every watch was based on the high frequency with which it was used in Phase II (average for all five coders = 75.8% - see Table 5.10).

6.3 Phase IIIa Evaluation of Reliability

Intracoder Reliability The high degree of consistency with which the coders classified the log book descriptions was also demonstrated in Phase III even though they were not given the option of leaving blank those watches that they felt did not provide enough information on ice concentration. Table 6.5 gives the intracoder agreement coefficients for the ice concentration classification.

TABLE 6.5 PHASE IIIa INTRACODER AGREEMENT COEFFICIENTS FOR ICE CONCENTRATION

Coder	Nominal	Ordinal	Difference (O-N)
A	.869	.940	.071
D	.700	.917	.217
I	.889	.960	.071
J	.922	.980	.058
M	.782	.899	.117
Average	.832	.939	.107

An important feature of these results emerges when they are compared with the intracoder reliabilities of the grouped categories in Phase II. This is shown below in Table 6.6.

TABLE 6.6 COMPARISON OF PHASE II AND IIIa ORDINAL COEFFICIENTS FOR ICE CONCENTRATION

Coder	A Phase II (Grouped)	B Phase IIIa	Difference (B-A)
A	.743	.940	.197
D	.925	.917	-.008
I	.919	.960	.041
J	.988	.980	-.008
M	.941	.899	-.042
		Average	.036

This table shows that in two cases, the values increased and in three, they decreased, but of greater importance is the fact that the changes were very minor. The average difference was 0.036 with a standard deviation of 0.085, and if the anomalous value for Coder A is

omitted, the average difference is reduced to 0.004 (standard deviation = 0.03). The reason for the abnormally high increase for Coder A was that this coder had a higher number of inconsistencies between categories that were not adjacent such as between Categories 1 and 3, 1 and 4, 1 and 5, whereas the inconsistencies of the other four coders involved adjacent categories. This has an affect on the coefficient that is calculated for the grouped categories particularly when the ordinal scale is used. The first step in this process is to determine the d_{bc} values for the disagreements of the original seven-category matrix thereby weighting each disagreement proportionally to the relative distance between the categories. The categories are then grouped using the weighted inconsistencies. When there are many disagreements between non-adjacent categories, the value for the observed disagreements (D_o) is increased and this results in a lower agreement coefficient. This would account for the low intracoder coefficient obtained by Coder A for Phase II (grouped). The high coefficient for this coder in Phase IIIa was due to the fact that the classification decisions were based on a smaller, more generalized set of categories which removed the source of inconsistency for this coder.

In general, the intracoder results between the two phases were very similar. This implies that the coefficients that were calculated in Phase II were predictive of the consistencies which were obtained when the grouped categories were tested in Phase IIIa.

Intercoder Reliability The intercoder results from the Phase IIIa tests were also tabulated in the same way as they were in Phase II and an example of this is shown in Table 6.3b. From these tables, the ordinal intercoder agreement coefficients for the ice concentration classification were calculated for each of the two sessions. These results are shown in Table 6.7 together with the corresponding data from Phase II.

TABLE 6.7 ORDINAL INTERCODER AGREEMENT COEFFICIENT

Session	A Phase II	B Phase II (grouped)	C Phase IIIa	Difference (C-B)
1	.733	.773	.854	.081
2	.786	.813	.831	.018
3	.860	.900	*	
			Average	.05

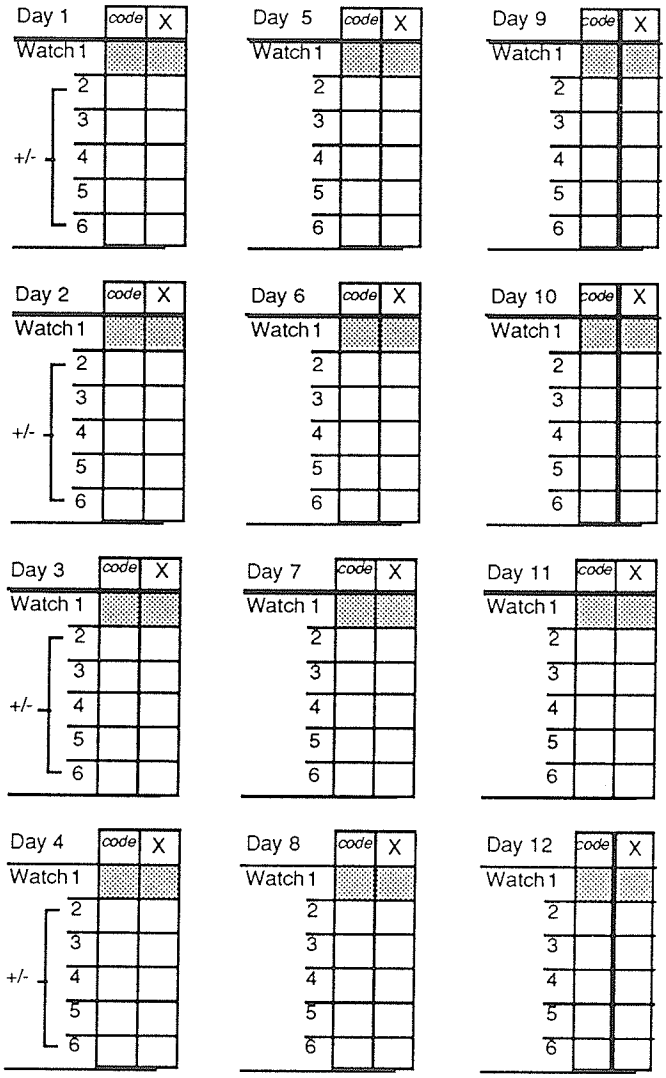
* There were only two repetitions in phase IIIa.

As with the intracoder tests, the differences between the two phases as shown in Table 6.7 were less than five per cent. Therefore, the intercoder coefficients calculated for the grouped categories in Phase II were able to predict the test results of Phase IIIa. This one test alone does not necessarily prove that these results can always be expected to occur. The similarity of the results between the two phases, however, does support the decision made in Phase II to group the categories in this way since the Phase II coefficients were calculated to be much higher than the ungrouped categories, and this was confirmed when the modified categories were tested by the coders in Phase IIIa.

6.4 Phase IIIb - Relative Sea Ice Concentration

This portion of Phase III was undertaken to determine the degree to which the coders could agree on whether the ice concentration increased, decreased, or remained the same from one watch to the next. Casual observations of the summary tables appeared to indicate that while the coders might not always have agreed on the absolute concentration given by a specific category, they might, nevertheless, agree on the relative changes in concentration between the watches. Phase IIIb, therefore, was intended to explore this possibility. To facilitate this test, the coders used the same 26-day sample as was used in Phases II and IIIa, and the form illustrated in Table 6.8.

TABLE 6.8 PORTION OF PHASE IIIb CLASSIFICATION FORM



This form is slightly different from the form used in Phase IIIa. There are no general classification headings (A, B, C, D) because this test applied only to Classification A - ice concentration. Using the same definitions as in phase IIIa, the coders entered a category number for the first watch in the shaded box labeled *code* , and as in the first part of Phase III, they placed a check mark under the heading 'X' when there was no entry for the watch. The instructions to the coders for this test are given in Table 6.9.

TABLE 6.9 PHASE IIIb NOTES AND INSTRUCTIONS TO CODERS

In this phase, you will be using Classification A (Concentration) only. The instructions for phase IIIb are:

1. Provide a category number (1-4) from Classification A for the first watch only (in the shaded box).
2. If Watch 2 shows an increase in concentration from Watch 1 enter +.
3. If Watch 2 shows a decrease in concentration from Watch 1 enter -.
4. If Watch 2 shows no change in concentration from Watch 1 enter **0.**
5. Compare Watch 3 to Watch 2 in the same way.

Example:

Day	Watch	Classification	X	Interpretation
1	1	**3**		*Close ice*
	2	**+**		*more than Watch 1*
	3	**+**		*more than Watch 2*
	4	**0**		*same as Watch 3*
	5	**0**	*x*	*same as Watch 3*
				no entry
	6	**-**		*less than Watch 3*

Note:

- There must be a classification for every watch.
- Even though there are only 4 ice concentration categories, you can still enter + for more than one successive watch even if watch 1 was assigned category 3.
- The number assigned to Watch 1 is to provide a standardized starting point only.
- **Do NOT refer to the Classification A categories for anything but the first watch.**

This test was not repeated because the intracoder reliabilities had been tested for 17 sessions by this point and had consistently produced acceptable results. Once the coders had completed this test, their decisions were synthesized in a table a sample of which is shown in Table 6.10.

TABLE 6.10 SAMPLE OF PHASE IIIb INTERCODER SUMMARY TABLE

Day	Watch	A	D	I	J	M	1 -	2 0	3 +	4
13	1	1	1	1	1	1	5			
	2	0	0	0	0	0		5		
	3	0	0	0	0	0		5		
	4	+	+	+	+	+			5	
	5	+	+	+	+	+			5	
	6	+	+	+	+	+			5	
14	1	3	4	3	3	3			4	1
	2	+	+	+	+	+			5	
	3	0	0	0	0	0		5		
	4	0	0	0	0	0		5		
	5	-	-	-	-	-	5			
	6	0	0	0	0	0		5		
15	1	2	2	2	2	2		5		
	2	+	+	+	+	+			5	
	3	+	+	+	0	+		1	4	
	4	0	-	0	0	-	2	3		
	5	0	+	0	+	+		2	3	
	6	0	+	0	0	0		4	1	

In this table, the category number (1 to 4) assigned by each coder was entered for the first watch of each day. The remaining five watches were given either a '+', '-', or '0' as assigned by the coders. The next four columns show the agreement patterns for each category in such a way that the top row of numbers was applied to Watch 1 and the bottom row of three symbols was applied to Watches 2 to 6. Therefore, the fourth column was used only for the first watch of each day.

Days 13, 14, and 15 were selected for the sample in Table 6.10, rather than Days 1 and 26 as in the previous samples, in order to demonstrate three types of agreement patterns. Day 13 shows a situation in which there was complete agreement for all six watches. This means that the five coders agreed on the category number (1) as well as the

relative changes in concentration throughout the day. By comparing these decisions for Day 13 with the transcription for that day, it is clear why this was possible.

TABLE 6.11 DAY 13 TRANSCRIPTION AND PHASE IIIb CLASSIFICATION

Watch	Category # (# of coders)	Hour	Log Book Entry
1	1 (5)	1 p.m.	
		2	
		3	
		4	
2	0 (5)	5	
		6	
		7	
		8	
3	0 (5)	9	
		10	
		11	
		12	
4	+ (5)	1 a.m.	
		2	
		3	
		4	*Passed some Ice*
5	+ (5)	5	
		6	*Sailing amongst small Ice*
		7	
		8	*Ice heavy and close*
6	+ (5)	9	
		10	
		11	
		12 noon	*The Ice very close*

The second day of the sample (Day 14) illustrates the type of situation on which Phase IIIb was based. In this case, there was a disagreement about the concentration category number. There was, however, a complete agreement among the coders on the relative changes in the ice concentration as described in the log book.

TABLE 6.12 DAY 14 TRANSCRIPTION AND PHASE IIIb CLASSIFICATION

Watch	Category # (# of coders)	Hour	Log Book Entry
1	3 (4) 4 (1)	1 p.m.	*Forcing through thick Ice*
		2	
		3	
		4	
2	+ (5)	5	
		6	*Ice close*
		7	*Fast in a close body of Ice*
		8	
		9	
3	0 (5)	9	
		10	
		11	
		12	
4	0 (5)	1 a.m.	
		2	
		3	
		4	
5	- (5)	5	
		6	*Ice opening*
		7	*Ungrappled*
		8	*Ice heavy*
6	0 (5)	9	
		10	
		11	
		12 noon	

As can be seen in these two samples, all of the coders all interpreted blank watches as meaning that the conditions of the previous watch prevailed. There were, however, only 39 cases for which a classification was given when there was no entry. Day 15 was included in Table 6.10 to illustrate a third possibility in which the coders were in agreement on the classification for Watch 1 but not on the relative changes after that point. The

reasons for the agreements or disagreements are not always as apparent as is shown when these classifications are compared with the corresponding transcription.

TABLE 6.13 DAY 15 TRANSCRIPTION AND PHASE IIIb CLASSIFICATION

Watch	Category # (# of coders)	Hour	Log Book Entry
1	2 (5)	1 p.m.	
		2	
		3	
		4	*Traversing open Ice*
2	+ (5)	5	
		6	
		7	
		8	*Traversing and forcing among close small ice*
3	0 (1) + (4)	9	
		10	*The Ice very close*
		11	
		12	*The Ice more open*
4	0 (3) - (2)	1 a.m.	
		2	
		3	
		4	
5	0 (2) + (3)	5	
		6	*Forcing through a ledge of Ice*
		7	
		8	
6	0 (4) + (1)	9	
		10	*Traversing and forcing the Ice to the W'ward*
		11	
		12	

It was clear to the coders that in Watch 1 there was a small amount of ice and that there was more ice by Watch 2 as implied by the words *forcing* and *close*. Watch 3 however, presented a situation in which there was a change from *very close* at 10:00pm to *more open* at midnight. This was interpreted by four coders as an increase in ice concentration over

the conditions in Watch 2, but one coder interpreted this same information as meaning that there was no change.

Two agreement coefficients were calculated for Phase IIIb. The first was for the intercoder reliability for Watch 1 only, and the second was the intercoder coefficient for Watches 2 to 6. It was necessary to calculate two coefficients because they were two separate sets of categories. The objective here was not to examine the changes from watch to watch but to evaluate the reliability among the coders as was done in Phases II and IIIa. Another objective was to compare the reliability of the category decisions for Watch 1 with the reliability of the relative decisions for Watches 2 to 6 . In both cases, the ordinal scale values for d_{bc} were used. The reason for this is probably more obvious for Watch 1 than for Watches 2 to 6. Although there were no defined categories for the second coefficient, a disagreement involving an increase and decrease in ice concentration required a higher weighting than one involving an increase or decrease and no change. Tables 6.14 a to d are the nominal and ordinal coincidence matrices and agreement coefficients for Watch 1 (a and b) and Watches 2 to 6 (c and d).

TABLE 6.14 PHASE IIIb COINCIDENCE MATRICES

a. Watch 1 - Nominal Categories

Categories	1	2	3	4	
1	104	8	4	4	120
2	8	180	32	0	220
3	4	32	50	26	112
4	4	0	26	38	68
	$\alpha = .599$				520

b. Watch 1 - Ordinal Categories

	1	2	3	4	
1	104	3.42	6.64	10.76	124.82
2	3.42	180	13.12	0	196.54
3	6.64	13.12	50	3.12	72.88
4	10.76	0	3.12	38	51.88
	$\alpha = .761$				446.12

c. Watches 2-6 - Nominal Categories

Categories	-	0	+	
-	316	202	34	552
0	202	1220	82	1504
+	34	82	428	428
	$\alpha = .575$			2600

d. Watches 2-6 - Ordinal Categories

	-	0	+	
-	316	126.45	85	527.45
0	126.45	1220	50.92	1397.37
+	85	50.92	544	563.92
	$\alpha = .642$			2488.74

In both cases, the ordinal coefficients were higher than the nominal coefficients indicating that most of the disagreements involved adjacent categories. The ordinal coefficient for Watch 1 was slightly lower than the coefficients for Phase IIIa with a difference of 0.093 for Session 1 and 0.070 for Session 2. It was surprising however, to find that the intercoder reliability for the relative concentration (Watches 2 to 6) was the lowest ordinal coefficient for Phases II and III as shown in Table 6.15.

TABLE 6.15 ORDINAL SCALE INTERCODER AGREEMENT COEFFICIENTS FOR PHASES II, II GROUPED, IIIa, AND IIIb

	Phases				
Session	II	II Grouped	IIIa	IIIb (1)	IIIb (2-6)
1	.733	.773	.854	.761	.642
2	.786	.813	.831	N/A	N/A

In general, the results of this test of relative ice concentrations showed that to reduce the resolution of the categories to the lowest possible level, does not necessarily improve the reliability. In fact, the results of tests IIIb (Watches 2 to 6) were lower than those calculated for the ungrouped categories in Phase II.

6.5 General Conclusions

Two separate sets of tests were conducted in this phase. In the first (IIIa), the grouped categories from Phase II were applied to the textual units by the coders to evaluate the reliability of those categories. Secondly, these values were then compared to the coefficients calculated directly from the coincidence matrices in Phase II to see how well the agreement coefficient could predict the agreements when the categories were applied by the coders. In both sessions of Phase IIIa, the coefficients were higher than the values calculated in Phase II, however, they differed by only 0.081 for the first session, and by 0.018 for the second. It was decided, therefore, that this set of categories for ice

concentration could be accepted as yielding the highest level of reliability without substantially reducing the resolution of the derived data. Secondly, because the results of the grouped calculations in Phase IIIa differed so little from the test results in Phase II, it was concluded that this last test may not be necessary in the future. The similarity of these results indicates that coefficients which have been calculated directly from the grouped coincidence matrix are sufficient indicators of reliability to preclude repeated categorizations by the coders. This test was, in fact, conducted a third time in Phase IIIb. Here, the first watch of each day of the sample was classified according to the same ice concentration categories as in Phases II. Although this was a smaller sample, the coefficient for Watch 1 in Phase IIIb (0.761) differed from those calculated in Phase II by only 0.012 for the first session and by 0.052 for the second.

The first set of categories tested in Phase II represented the highest resolution of all the category sets tested in this research, and Phases IIIb tested the lowest. In this test, the coders classified textual units as describing increasing, decreasing, or constant ice concentrations from one watch to the next. Intuitively, this might be expected to produce the highest levels of reliability, however, this was not shown to be the case. In fact, this test yielded the lowest coefficient (0.642) with the exception of Phase I (see Appendix II). Table 6.16 is a summary of the key coefficients from Phases II, IIIa, and IIIb.

TABLE 6.16 SUMMARY OF ORDINAL SCALE INTERCODER ICE CONCENTRATION AGREEMENT COEFFICIENTS FOR PHASES II, IIIa, AND IIIb

Phase	Session 1	Session 2	Session 3	Average
II	0.733	0.786	0.860	0.793
II Grouped	0.773	0.816	0.900	0.829
IIIa	0.854	0.831	*	0.843
IIIb Watch 1	0.761	*	*	0.761
IIIb Watches 2-6	0.642	*	*	0.642

* Less than 3 sessions.

The results of this final test revealed a very important point regarding the interrelationship among three central components of CA, the number of categories, category resolution, and reliability. It shows that reducing the number of categories and their resolution to the lowest level does not necessarily guarantee a higher level of reliability. Therefore, resolution should not be sacrificed for the sake of improving reliability. In light of this discovery, it is to the best advantage of the research to begin with the highest category resolution and to reduce it as revealed by the coincidence matrices and agreement coefficients. This, then, marked the final stage of the research in which the testing procedure itself was assessed and both this technique and the categories were accepted.

CHAPTER 7

SUMMARY AND CONCLUSIONS

7.1 Context and Relevance

Written historical sources of climatic change have made possible substantial contributions to the field of paleoclimatic reconstructions by providing environmental information on the most recent portion of the earth's history. In those cases where the historical sources are specifically devoted to instrumental meteorological observations, attention has been paid to standardizing and correcting the measurements to ensure that the resulting reconstructions are valid (Manley, 1946; Wilson, 1982 and 1983). When the sources contain subjective, descriptive accounts, the quality of the sources is often examined in detail, but the reliability of the methods employed to derive the data from those descriptions has rarely been addressed. This thesis has investigated the consequences of this failure by examining ways of testing the reliability with which climatic data could be derived from an historical source. The results of these tests were used to improve the content analysis methodology in the context of historical climatic reconstructions.

The first phase of the case study in this dissertation was based on the rudimentary, impressionistic approach that prevailed among historical climatologists at the time that the study began. Although this phase employed more extensive reliability tests than those applied in previous reconstructions, a critical evaluation of the methodology revealed that major adjustments were required. The lack of research in historical climatology related to methodology required that the adjustments be based on techniques developed in other disciplines, in this case, the social sciences and humanities. This interdisciplinary approach led to innovations in the selection of the most suitable textual units, the development of

categories, and the implementation of the most recently developed technique for assessing and measuring reliability. This last innovation was then validated in the third, and final, phase of the research.

This research contributes to the field of paleoclimatology by providing a research strategy that is objective and is based on the principles of the scientific method. Furthermore, it demonstrates the relevance and importance of reliability testing. While historical sources have often been noted for the valuable information that they provide for a critical time period, the recent past, very little attention has been paid to the methods by which the descriptions in the sources have been interpreted. By employing the scientific method, in the form of CA, in this context, this research has demonstrated how the quality of reconstructions can be greatly enhanced by providing additional information about the interpretive methods and reliability of the derived data. The methods employed to interpret the geological, glaciological, and biological sources have a long history of research and development, and their reliabilities have been exhaustively tested. Consequently, reconstructions based on these sources are generally accepted as methodologically valid. This practice of applying a scientific approach to the development and testing of a methodology has not, to this point, been a routine part of historical climatic reconstructions. The methodology that emerged from the three-phase evolutionary procedure of this research, and the general reasoning on which it was based, are directly relevant to any descriptive source and environmental variable. This is so because this research represents more than the practical development of a methodology. It supports the adoption, in historical climatology, of a new conceptual approach such that the method of each reconstruction be critically examined.

One of the most important components in scientific research is the reliability of the procedure. The assessment and improvement of reliability was the major theme of this research. This provides the only means of systematically and objectively evaluating the

procedure before the reconstruction is completed. These tests also determine the resolution of the derived data by evaluating how well the CA categories can be applied to the textual units. Through the technique adopted in Phase II, it was possible to determine those categories that were ambiguous or that had a resolution that was too fine for the log book contents, and to make the changes indicated by these findings.

7.2 Summary

This case study began by employing the impressionistically-based approaches to CA and rudimentary reliability tests that were customarily applied in the research that preceded it. The difference, however, was that in this research, Phase I represented the beginning of what was to become the evolutionary development of an objective technique for deriving climatic data from historical documents. In this way, Phase I served two purposes. The first was to place this research into the context of the state of CA and reliability testing in paleoclimatology at the time that the research began. Secondly, the results of Phase I revealed those elements of the research plan that needed modification. It was in this way that this phase was of great importance, and established the subsequent direction that the research would take.

Although the method of describing the agreements resulting from the tests of Phase I was very rudimentary, it was still possible to determine a basic assessment of reliability. In Phase I, three textual units were tested, days, hourly entries, and individual words, all of which were found to be inappropriate. This was an important observation because the textual unit is a crucial contributing factor to the reliability. It also represents the amount of information that will be extracted from the source, and, therefore, it ultimately determines the resolution of the categories. The more information that is contained in the textual unit, the more general the categories must be to encompass the wide range of information that the textual unit will contain. It was also decided, at the end of Phase I, that the categories were

too simplistic. The role of the categories is of great importance because the categories define the derived data. In Phase I, the categories were developed intuitively and, therefore, they were not clearly defined. Without specific definitions, the coders' decisions did not have a common basis and this lowered the reliability of the CA. The third conclusion drawn from Phase I was that a new technique for evaluating reliability was needed. In the past, the few studies that tested reliability employed either percentage agreements (Moodie and Catchpole, 1975) or Scott's π test (Baron, 1980), however, both these approaches are inadequate. Scott's π test is biased in terms of the number of categories, and percentage agreements simply summarize the number of agreements without distinguishing those that occurred by chance. Furthermore, in both cases, it is only possible to conduct the calculations for two coders at a time. The search for a more appropriate measure of reliability lead to a reexamination of the CA literature.

In general, then, Phase I was based on the traditional approach, but contributed to the research by prompting a critical appraisal of the methodology. To this point, this type of examination of the procedure itself had not been done in other historical climatic research. Basically, this appraisal raised questions central to the methodology and created a new context for the next phase.

While Phase I served an evaluative function to determine the specific areas that needed to be changed, Phase II was comprised of the innovations developed to facilitate those required modifications. To devise a new textual unit, the log books were refered to, and this lead to the realization that the most logical unit would be the seaman's watch which was essentially the time unit in which the ice observations were originally recorded. Phase II also involved the development of a new set of categories based on the definitions contained in the contemporary manual for observing and describing sea ice. From this source, a new set of categories was devised that was accompanied by diagrams and detailed definitions. The classification headings were similar to those applied in Phase I. They

included ice concentration, floe size, and motion, but added categories that described arrangement of the ice floes and ice openings. In addition to the improvement that the new categories made to the study, was a more fundamental contribution made by the reasoning behind this modification.This was the decision that linked the historical descriptions to the modern sea ice records. This is of particular importance to historical reconstructions in which the period of the historical record does not overlap the modern period of record. In this situation, the derived data cannot be calibrated against the modern data, and can only be linked in a general way to the modern observations by employing the contemporary definitions. To resolve the third problem, a review of the most recent CA literature was necessary to arrive at a more appropriate measurement of reliability. This was found in Krippendorff's agreement coefficient (α) which provides a measure of the proportion of agreement that can be expected from repeated tests, excluding chance agreements. An important contribution of this technique is that it can illustrate those categories that are inappropriate and, in the process of calculating this coefficient, direction is given for the modification of the categories. This can be done by grouping the categories and recalculating α without reclassification by the coders using the grouped categories. This procedure led however, to new questions about the ability of α to predict future agreements by simply recalculating the coefficient. Therefore, one more test remained after the completion of Phase II.

By the completion of Phase II, a new, more generalized set of categories had been formed as a result of experiments with several different groupings. In Phase III, the coders applied this new set of categories to the textual units and this yielded agreement coefficient values that were very close to those calculated in Phase III. The success of these tests demonstrated that this step would not be necessary in future reconstructions. Another strength of Krippendorff's coefficient is that it reveals those categories that create problems for the coders, illustrates possible solutions, and describes the reliability of the modified

categories. Therefore, for the purposes of reliability testing in CA, Krippendorff's coefficient is an important contribution.

7.3 General Conclusions

This research has made several conceptual and practical contributions to the field of historical climatology by providing a thorough examination of methodology in this area. In order to resolve the problems identified in the initial phase of the research, it was shown that an interdisciplinary approach must be adopted. In this research, contributions from the social sciences, humanities, and physical sciences were integrated to more adequately interpret historical descriptions of sea ice. This dissertation also shows that the reliability of the methodology is as important as the quality of the sources. Regardless of how trustworthy the sources are, the quality of the resulting reconstruction is dependent upon the method employed to interpret the sources, and so greater attention should be paid to the reliability of the method. Most importantly, therefore,this research demonstrates that reliability tests of the methodology play a critical role in improving the resolution, quality, and credibility of historical climatic reconstructions using subjective documentary sources.

REFERENCES CITED

Arakawa, H. 1957. "Climatic Change as Revealed by the Data from the Far East", WEATHER, Vol.12:46-47.

Ball, T. 1985. "A Dramatic Change in the General Circulation on the West Coast of Hudson Bay in 1760 A.D.: Synoptic Evidence Based on Historic Records", In: C.R. Harington, Climatic Change in Canada. SYLLOGEOUS, Vol. 55: 219 - 245.

_____ 1986. "Historical Evidence and Climatic Implications of a Shift in the Boreal Forest Tundra Transition in Central Canada", CLIMATIC CHANGE, Vol. 8: 121 - 134.

_____ and Kingsley, R.H. 1984. "Instrumental Temperature Records at Two Cities in Central Canada: 1768 to 1910", CLIMATIC CHANGE, Vol.6: 39 - 56.

Baron, Wm. 1980. "Tempests, Freshets and Mackeral Skies: Climatological Data from Diaries Using Content Analysis", Ph.D. dissertation, University of Maine, Orono, Maine.

Bauer, Jane and Martin, Seelye. 1980. "Field Observations of the Bering Sea Ice Edge Properties During March 1979", MONTHLY WEATHER REVIEW, Vol.108 No.1:2045-2056.

Bennett, E.M., Alpert,R., and Goldstein,A.C. 1954. "Communications Through Limited-Response Questioning", PUBLIC OPINION QUARTERLY, Vol.18 No.3 (Fall): 303-308.

Berelson, B. 1952. Content Analysis in Communications Research. New York:Free Press.

Bradley, R.S. 1985. Quaternary Paleoclimatology: Methods of Paleoclimatic Reconstruction. Boston: Allen and Unwin, Inc.

Brooks, C.E.P. 1922. The Evolution of Climate. London: Benn Brothers Ltd.; reprint edition, New York: AMS Press, 1978.

_____ 1926. Climate Through the Ages. London: Benn Brothers, Ltd.; republication of 1949 revised edition, New York: Dover Publications, Inc., 1970.

Carney, T.F. 1972. Content Analysis: A Technique for Systematic Inference from Communications. Winnipeg: University of Manitoba Press.

Catchpole, A.J.W. 1980. "Historical Evidence of Climatic Change in Western and Northern Canada", In: C.R. Harington (ed.) Climatic Change in Canada. SYLLOGEUS, Vol. 26:17-60.

REFERENCES CITED

_____ (in press) "Hudson's Bay Company Ships' Log-Books as Sources of Sea Ice Data, 1751-1870". In, R.S.Bradley ans P.D.Jones (eds.) Climate Since A.D.1500. London and Boston: Unwin Hyman.

_____ Moodie, D.W. and Kaye, B. 1970. "Content Analysis: A Method for the Identification of Dates of First Freezing and First Breaking from Descriptive Accounts", THE PROFESSIONAL GEOGRAPHER, Vol. XXII No. 5:252 - 257.

_____ and Moodie, D.W. 1978. "Archives and the Environmental Scientist", ARCHIVARIA, Vol.6:113-136.

_____ and Faurer, M.A. 1983. "Summer Sea Ice Severity in Hudson Strait, 1751-1870", CLIMATIC CHANGE, Vol. 5:115-139.

_____ and Faurer, M.A. 1985. "Ships' Log-Books, Sea Ice and the Cold Summer of 1816 in Hudson Bay and Its Approaches", ARCTIC, Vol. 38 No.2:121-128.

_____ and Halpin, J. 1987. "Measuring Summer Sea Ice Severity in Eastern Hudson Bay", THE CANADIAN GEOGRAPHER, Vol. 31 No.3:233-244.

_____ and Hanuta, I. 1989. "Severe Summer Ice in Hudson Strait and Hudson Bay Following Major Volcanic Eruptions, 1751 to 1889 A.D.", CLIMATIC CHANGE, Vol. 14: 61 - 79.

Cohen, J. 1960. "A Coefficient of Agreement for Nominal Scales", EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT, Vol. XX No.1:37-46.

Dovring, K. 1954-55. "Quantitative Semantics in Eighteenth Century Sweden", PUBLIC OPINION QUARTERLY, Vol. 18 No.4:389-394.

Environment Canada. 1984. "Ice Observer Extension Course, Training Manual, Ice Terminology and Symbology", Atmospheric Environment Service, Ice Analysis and Forecasting Division. Ottawa, Ontario.

Faurer, M.A. 1981. "Evidence of Sea Ice Conditions in Hudson Strait, 1751-1870, Using Ships' Logs", M.A.Thesis, University of Manitoba, Winnipeg, Manitoba.

Feibleman, J.K. 1972. Scientific Method: The Hypothetico-Experimental Laboratory Procedure of the Physical Sciences. The Hague: Martinus Nijhoff.

Hammer, C.V., Clausen, H.B., and Dansgaard, W. 1980. "Greenland Ice Sheet Evidence of Post-Glacial Volcanism and its Climatic Impact", NATURE, Vol. 288: 230 - 235.

REFERENCES CITED

Hoeller, A.E. 1982. "The Role of Environmental and Historical Evidence in Climatic Reconstruction: A Preliminary Review and Appraisal", Unpublished Manuscript. Report # 82 - 3. Canadian Climate Centre, Atmospheric Environment Service, Downsview, Ontario.

Holsti, O. 1969. Content Analysis for the Social Sciences and Humanities. Reading: Addison-Wesley.

Hsieh, Chaio-Min 1976. "Chu K'O-Chen and China's Climatic Changes", GEOGRAPHICAL JOURNAL, Vol. 142:248 - 256.

Ingram, M.J., Underhill, D.J., and Farmer,G. 1981. "The Use of Documentary Sources for the Study of Past Climates", In: Wigley, T.M., Ingram, M.J., and Farmer, G. (eds.) Climate and History: Studies in Past Climates and their Impact on Man. Cambridge: Cambridge University Press. 180-213.

Kington, J.A. 1988. The Weather of the 1780's Over Europe. Cambridge: Cambridge University Press.

Kracauer, S. 1952-53. "The Challenge of Quantitative Content Analysis", PUBLIC OPINION QUARTERLY, Vol. 16 No.4:631-642.

Krippendorff, K. 1971. "Reliability of Recording Instructions: Multivariate Agreement for Nominal Data", BEHAVIORAL SCIENCE, Vol. 16 No. 3:222 - 235.

_____ 1980. Content Analysis: An Introduction to Its Methodology. Beverley Hills: Sage Publication, Inc.

Lamb, H.H. 1966. The Changing Climate. London: Methuen and Co., Ltd..

_____ 1982. Climate, History and the Modern World. London: Methuen and Co., Ltd..

Leedy, Paul D. 1974. Practical Research: Planning and Design. New York: Macmillan Publishing Co., Inc.

Le Roy Ladurie, Emmanuel 1972. Times of Feast, Times of Famine: A History of Climate Since the Year 1000. London: George Allen and Unwin Ltd..

Madison, G.N. 1981. "Reconstruction and Testing of Historical Dates of First Frost and First Snow, James Bay, Ontario 1705 - 1870", M.A. Thesis, University of Manitoba, Winnipeg, Manitoba.

REFERENCES CITED

Magne, M.A. 1981. "Two Centuries of River Ice Dates in Hudson Bay Region from Historical Sources", M.A. Thesis, University of Manitoba, Winnipeg, Manitoba.

Manley, Gordon 1946. "Temperature Trends in Lancashire, 1753 - 1945", QUARTERLY JOURNAL OF THE ROYAL METEOROLOGICAL SOCIETY, Vol. 72 No. 311: 1 - 31.

McKay, D.K. and Mackay, J.R. 1965. 'Historical Records of Freeze-up and Break-up on the Churchill and Hayes Rivers", GEOGRAPHICAL BULLETIN, Vol.7:7 - 16.

Minns, R. 1970. "An Air Mass Climatology of Canada During the Early Nineteenth Century: An Analysis of the Weather Records of Certain Hudson's Bay Company Forts", M.A. Thesis, University of British Columbia, Vancouver, British Columbia.

Moodie, D.W. and Catchpole, A.J.W. 1975. Environmental Data from Historical Documents by Content Analysis. Manitoba Geographical Studies No.5. Winnipeg: University of Manitoba.

Neuberger, Hans 1970. "Climate in Art", WEATHER, Vol. 25: 46 - 56.

Ogilvie, A.E.J. 1981. "Climate and Society in Iceland from the Medieval Period to the Late 18th Century", Ph.D. Dissertation, University of East Anglia.

_____ 1984. "The Past Climate and Sea Ice Record from Iceland. Part 1: Data to A.D. 1780", CLIMATIC CHANGE, Vol. 6:131-152.

Oliver, J. and Kington, J.A. 1970. "The Usefulness of Ships' Log-Books in the Synoptic Analysis of Past Climates", WEATHER, Vol. 25:520-528.

Pool, I.DeS. (ed.). 1959. Trends in Content Analysis. Urbana: University of Illinois Press.

Post, John D. 1977. The Last Great Subsistence Crisis in the Western World. Baltimore: The Johns Hopkins University Press.

Rannie, W.F. 1983. "Break-up and Freeze-up of the Red River at Winnipeg, Manitoba, Canada in the 19th Century and Some Climatic Implications", CLIMATIC CHANGE, Vol.5: 283 - 296.

Schutz, Wm.C. 1952. "Reliability, Ambiguity, and Content Analysis", PSYCHOLOGICAL REVIEW, Vol. 59:119-129.

REFERENCES CITED

Scott, Wm.A. 1955. "Reliability of Content Analysis: The Case of Nominal Scale Coding", PUBLIC OPINION QUARTERLY, Vol. 19:321-325.

Stommel, H. and Stommel, E. 1979. "The Year Without a Summer", SCIENTIFIC AMERICAN, Vol. 240: 176 - 186.

Stone, P.J., Dunphy, D.C., Smith, M.S., and Ogilvie, D.M. 1966. The General Inquirer: A Computer Approach to Content Analysis. Cambridge: MIT Press.

Teillet, J.V. 1988. "A Reconstruction of Summer Sea Ice Conditions in the Labrador Sea Using Hudson's Bay Company Ships' Log-Books, 1751 - 1870", M.A. Thesis, University of Manitoba, Winnipeg, Manitoba.

Wigley, T.M.L. 1977. "Geographical Patterns of Climatic Change: 1000B.C. - 1700 A.D.", Interim Final Report to NOAA. Climatic Research Unit, University of East Anglia.

Wilson, C. 1982. "The Summer Season Along the East Coast of Hudson Bay During the Nineteenth Century", Unpublished Manuscript, Report No. 82-4. Canadian Climate Centre, Atmospheric Environment Service, Downsview, Ontario.

_____ 1983. "The Little Ice Age on Eastern Hudson Bay: Summers at Great Whale, Fort George, Eastmain,1814 - 1821", Unpublished Manuscript, Report No. 83-9. Canadian Climate Centre, Atmospheric Environment Service, Downsview, Ontario.

_____ 1988. "The Summer Season Along the East Coast of Hudson Bay During the Nineteenth Century. Part III Summer Thermal and Wetness Indicies. The Indicies, 1800 - 1900", Unpublished Manuscript, Report No. 88-3. Canadian Climate Centre, Atmospheric Environment Service, Downsview, Ontario.

APPENDIX 1

The following pages are the log book transcriptions that the coders employed for the reliability tests. The day numbers marked with an asterisk comprise the 26-day sample used in Phases II, IIIa, and IIIb.

DAY 1

LEVEL

Coder _____ Date_____ Time(hrs)____ (min)

1

2 Forcing this close heavy Ice under all sail

3

4

5 Ice opening

6 Leaving (?) thro' open Ice

7

8

9

10

11

12

1

2

3

4

5

6 Sailing thro' open Ice

7

8

9 Ice very close

10

11

noon No open water in sight

COMMENTS

DAY 2

LEVEL

Coder _____ Date_____ Time(hrs)____ (min)

1

2

3

4

5

6

7

8 Ship has been a complete Iceberg the last two days

9

10

11

12

1

2

3

4

5

6

7

8

9

10

11

noon

COMMENTS

DAY 3 * LEVEL__

Coder_____ Date_____ Time(hrs)____ (min)

1 Ice open and heavy

2

3

4

5

6

7

8

9

10 Sailing through straggling Ice

11

12

1

2

3

4 Ice open and heavy

5

6

7

8 Ship beset in heavy Ice

9 Ice more open but heavy

10

11

noon Sailing among heavy straggling Ice

COMMENTS

DAY 4 LEVEL_____

Coder_____ Date_____ Time(hrs)____ (min)_____

1

2 Ship at grapple in heavy Ice

3

4

5

6

7

8 Ice inclined to open

9

10

11

12

1

2

3

4 Ship still fast Ice close

5

6 Clear water to NNE - Ice more open

7

8

9

10 ungrappled ... Forcing to the SSE Ice thick and heavy ship beset at times

11

noon

COMMENTS

DAY 5* LEVEL_

Coder_____ Date_____ Time(hrs)____ (min)

1 Ship close beset in close and heavy Ice
2
3
4
5
6
7 The Ice in motion
8
9
10 ... wind made the ship strike the Ice heavy at times
11
12
1
2
3
4 ...Ship rolling and striking the Ice heavy at times
5
6
7
8 Less sea ship still striking heavy
9
10 Less sea Ice very close but much smaller
11
noon Ice very close from East to South and West

COMMENTS

DAY 6 LEVEL_____

Coder_____ Date_____ Time(hrs)____ (min)_____

1 Heaving ship off a piece of Ice
2
3 saw streams of water to the Southward
4 Ship beset in close Ice
5
6 Ice close ship beset
7
8
9
10
11
12
1
2 Ice begins to open
3 Ship began to move
4 Ice open in lanes
5
6
7
8 Ice closing
9 Forcing in heavy Ice
10
11 Ice more open at times
noon

COMMENTS

DAY 7*

LEVEL

Coder_____ Date_____ Time(hrs)____ (min)

1

2 Ice close no clear water in sight

3

4 Ice slackening a little

5

6 Boring thro close Ice

7

8 Passing thro open sailing Ice

9

10

11

12 passing thro slack Ice

1

2

3 Passing thro a deal of sailing Ice

4

5

6

7

8

9

10

11

noon Boring thro close Ice

COMMENTS

DAY 8

LEVEL

Coder_____ Date_____ Time(hrs)____ (min)_____

1 Sailing in for land thro slack Ice

2

3

4

5 Sailing thro open Ice between close Ice and land

6

7 Much Ice lying ashore

8

9

10

11

12

1 Passing thro close Ice

2

3

4

5

6

7

8 Ice close

9

10

11

noon Close heavy Ice all around

COMMENTS

DAY 9* LEVEL_

Coder_____ Date_____ Time(hrs)____ (min)

1 Sailing in for land thro slack Ice
2
3
4
5 Sailing thro open Ice between close Ice and land much..
6 ... Ice lying aground on shoals
7
8
9
10
11
12
1
2
3
4
5
6 Clear water in sight
7
8
9
10
11
noon Ship ... lying fast

COMMENTS

DAY 10 LEVEL_____

Coder_____ Date_____ Time(hrs)____ (min)____

1 Running along the weather edge of a heavy pack of Ice
2
3
4 Tacked from the Ice, the extreme in sight bearing NW
5
6
7 Tacked from the Ice
8
9 Tacked to Eastward from Ice
10
11 Tacked to Westwadr from Ice
12
1
2 Tacked from the Ice
3
4
5 Heavy close packed Ice to leeward
6
7
8 Running along the edge of a large pack of Ice
9
10
11
noon

COMMENTS

DAY 11*

LEVEL

Coder_____ Date_____ Time(hrs)____ (min)

1
2
3
4
5
6 Saw some Ice to Northward
7
8
9
10
11
12
1
2
3
4
5
6
7
8
9
10
11
noon

COMMENTS

DAY 12

LEVEL

Coder_____ Date_____ Time(hrs)____ (min)

1
2 Ship at grapple
3
4
5 Ungrappled and made all requisite sail
6
7
8
9
10
11
12
1
2
3
4
5
6
7
8 The Ice close but small
9
10
11
noon Ship beset

COMMENTS

DAY 13*

LEVEL

Coder_____ Date_____ Time(hrs)____ (min)

1
2
3
4
5 Grappled to a large floe of Ice
6
7
8
9
10
11
12
1
2
3
4
5
6
7
8
9
10 Ship still at grapple
11
noon

COMMENTS

DAY 14

LEVEL

Coder_____ Date_____ Time(hrs)____ (min)

1
2
3
4
5
6 Some loose pieces of Ice in sight
7
8 A ledge of Ice in sight and saw clear water over it. ...
9 ... Altered course
10
11
12
1
2
3
4
5
6 Saw Ice ahead at 1/2 past 5 enter'd it
7 Got into open water
8
9
10
11
noon

COMMENTS

DAY 15* LEVEL_

Coder_____ Date_____ Time(hrs)____ (min)_

1
2 Amongst open heavy Ice
3
4 Ice more open
5
6
7
8 Grappled for the night
9
10
11
12
1
2
3
4 Ungrappled and made sail, Ice open
5
6
7
8 Ice more open
9
10 Traversing the Ice
11
noon Ice open

COMMENTS

DAY 16 LEVEL

Coder_____ Date_____ Time(hrs)____ (min)_____

1
2
3
4
5
6 Saw Ice ahead
7 Bore away amongst loose Ice
8
9
10
11
12
1
2
3
4
5 Running amongst open Ice
6
7
8 Ice more open
9
10
11
noon Amongst open Ice

COMMENTS

DAY 17*

LEVEL

Coder_____ Date_____ Time(hrs)____ (min)

1

2 Sailing amongst open Ice

3

4 Ice more open but very heavy

5

6

7

8 Amongst open Ice

9

10

11

12 Amongst loose Ice

1

2

3

4 Clear of the Ice

5

6

7

8

9

10

11

noon

COMMENTS

DAY 18

LEVEL

Coder_____ Date_____ Time(hrs)____ (min)

1

2 Working to windward amongst open heavy Ice

3 The Ice being more close Grappled to a large piece

4

5

6

7

8

9

10

11

12

1

2

3

4

5

6

7 Cast off...and fell to leeward of several large pieces...

8 ...of Ice, then bore away to the SoWard amongst open Ice

9 Fell upon a large patch of Ice, warpt the ship ahead ...

10 ...to get her under sail on the other tack

11

noon ...made sail again amongst heavy Ice

COMMENTS

DAY 19* LEVEL

Coder _____ Date _____ Time(hrs) ____ (min)

1
2
3
4 Several pieces of Ice in sight
5
6
7
8
9
10
11
12
1
2
3
4
5
6
7
8 Past some straggling Ice
9
10
11
noon More Ice to the Westward

COMMENTS

DAY 20 LEVEL

Coder _____ Date _____ Time(hrs) ____ (min)

1
2 Grappled to a large piece of Ice finding we lost ground, ...
3 ... the Ice being so heavy and close
4
5
6
7
8
9
10
11
12
1
2
3
4 Ungrappled and made sail. The Ice a little more open
5
6 Working to Windward in a large pond of water
7
8 grappled to a piece of Ice finding we lost ground
9
10
11
noon The Ice open but very heavy

COMMENTS

DAY 21*

LEVEL

Coder _____ Date _____ Time(hrs) ____ (min)

1
2 Rounding a point of Ice
3
4
5
6
7
8
9
10
11
12
1
2
3
4
5
6
7
8
9
10
11
noon

COMMENTS

DAY 22

LEVEL

Coder _____ Date _____ Time(hrs) ____ (min)

1 Traversing the Ice
2
3
4
5
6 Amongst Ice
7
8
9
10 Got alongside a field of Ice
11
12 Got clear of it
1
2
3 Bore away among open Ice
4
5
6
7
8 The Ice more close. Bro't too
9
10 Fell alongside a field of Ice. Grappled to it
11 Cast off and made sail again
noon

COMMENTS

DAY 23*

LEVEL

Coder_____ Date_____ Time(hrs)____ (min)

1
2 Past several ledges of Ice
3
4
5
6
7
8
9
10
11
12
1
2
3
4
5
6
7
8
9
10
11
noon

COMMENTS

DAY 24

LEVEL

Coder_____ Date_____ Time(hrs)____ (min)

1
2
3
4
5
6
7
8
9
10
11
12 Pass'd a large piece of Ice
1
2 Running through heavy straggling Ice
3
4
5
6 Running among open Ice
7 Ice closed - forcing the Ice in a SSW direction
8
9
10 Ice close and heavy
11 Forcing the Ice
noon

COMMENTS

DAY 25 *

LEVEL____

Coder _____ Date_____ Time(hrs)____ (min)____

1

2

3

4 Some appearance of waterfrom SE by E to S from masthead

5

6

7

8

9

10

11

12

1

2

3

4

5

6

7

8

9 Finding the Ice open a little, ungrappled and try to force

10 ... to the SSW

11 Fast beset among close small Ice can't move

noon

COMMENTS

DAY 26

LEVEL_____

Coder _____ Date_____ Time(hrs)____ (min)____

1 Running 5 miles from the land and a close body of Ice

2

3

4 The land in sight and a body of Ice to the Eastward

5

6 Running along a body of Ice to the Et.ward of us

7 A body of Ice ahead

8 Traversing loose Ice. Shortened sail for the Ice

9

10 Running along straggling Ice

11

12

1 Got clear of the Ice - in an open sea

2

3

4

5

6

7

8

9

10

11

noon

COMMENTS

DAY 27*

LEVEL

Coder_____ Date_____ Time(hrs)____ (min)

1
2
3
4
5
6
7
8
9
10
11
12
1
2
3
4 Passed some Ice
5
6 Sailing amongst small Ice
7
8 Ice heavy and close
9
10
11
noon The Ice very close

COMMENTS

DAY 28

LEVEL_____

Coder_____ Date_____ Time(hrs)____ (min)_____

1
2
3
4
5
6
7
8
9
10
11
12
1
2
3
4 Ungrappled
5 Amongst heavy Ice
6
7
8
9
10 Passed much muddy Ice
11
noon Ice close and heavy

COMMENTS

DAY 29* LEVEL

Coder_____ Date_____ Time(hrs)____ (min)

1 Forcing through thick Ice
2
3
4
5
6 Ice close
7 Fast in a close body of Ice
8
9
10
11
12
1
2
3
4
5
6 Ice opening
7 Ungrappled
8 Ice heavy
9
10
11
noon

COMMENTS

DAY 30 LEVEL

Coder_____ Date_____ Time(hrs)____ (min)_____

1
2 Traversing loose Ice
3
4
5
6
7
8 The Ice close
9 Grappled ship
10
11
12
1
2
3
4 Ungrappled ship
5
6 Forcing through small Ice
7
8
9
10
11 Ice more open
noon Traversing the Ice to the W / N

COMMENTS

DAY 31*

LEVEL

Coder_____ Date_____ Time(hrs)____ (min)

1
2
3
4 Traversing open Ice
5
6
7
8 Traversing and forcing among close small Ice
9
10 The Ice very close
11
12 The Ice more open
1
2
3
4
5
6 Forcing through a ledge of Ice
7
8
9
10 Traversing and forcing the Ice to the Wt.ward
11
noon

COMMENTS

DAY 32

LEVEL

Coder _____ Date_____ Time(hrs)____ (min)_____

1
2 Forcing among close heavy Ice
3
4 Grappled the Ice close and heavy
5
6
7
8
9
10
11
12 The Ice very close
1
2
3
4 Ungrappled but the Ice so close & heavy the ship unmoveable
5
6
7
8 No water to be seen from the masthead in any direction
9
10
11
noon Continues beset no water to be seen

COMMENTS

DAY 33* LEVEL

Coder_____ Date_____ Time(hrs)____ (min)

1

2 Finding the Ice close to the So.ward hauled in to the .

3 ...Wt.ward to get into open water

4

5

6 In open water ... the Ice expanding towards the shore .

7 ... hauled out to the SE

8 Forcing through close Ice

9

10

11

12 Forcing through close Ice

1

2

3

4 The ship nearly fast with all sails set

5

6

7

8

9

10

11

noon Forcing but the Ice very close

COMMENTS

DAY 34 LEVEL

Coder_____ Date_____ Time(hrs)____ (min)____

1

2

3

4 Working to windward between the land and the Ice

5

6

7

8

9

10

11

12

1

2

3

4

5

6

7

8

9

10 Ice in sight to the NE

11

noon

COMMENTS

DAY 35* LEVEL

Coder_____ Date_____ Time(hrs)____ (min)

1
2
3
4
5
6
7
8
9
10
11
12
1
2
3
4 Pass'd some straggling Ice
5
6 Pass'd through a stream of Ice
7
8 A great quantity of Ice ahead and extending nearly ...
9 ... all round us
10 Traversing among close heavy Ice. Tacked occaisionally
11 ... to get into a little open water
noon

COMMENTS

DAY 36 LEVEL

Coder_____ Date_____ Time(hrs)____ (min)

1
2
3
4 Saw Ice on the Larboard bow
5
6
7
8 Saw Ice ahead
9
10
11
12 Traversing heavy Ice
1
2
3
4 The Ice more close ahead
5
6
7
8
9
10 Traversing among heavy Ice
11
noon

COMMENTS

DAY 37* LEVEL_

Coder_____ Date_____ Time(hrs)____ (min)_

1
2
3
4
5
6
7
8
9
10
11
12
1
2
3
4
5
6
7 Ungrappled ship warping and forcing amongst heavy fields
8 ... of Ice
9
10
11 Got clear of the heavy
noon Sailing among loose Ice

COMMENTS

DAY 38 LEVEL_____

Coder_____ Date_____ Time(hrs)____ (min)_____

1
2
3
4 Tack among straggling Ice
5
6
7
8
9
10 Passing straggling Ice
11
12
1
2
3
4
5
6
7
8
9
10
11
noon

COMMENTS

DAY 39*

LEVEL

Coder_____ Date_____ Time(hrs)____ (min)

1

2 Among heavy straggling Ice

3

4

5

6 Forcing through heavy Ice

7 The Ice close

8

9

10

11

12 The Ice beginning to open

1

2

3

4 Ungrappled

5

6 Among straggling Ice

7

8 In open sea

9

10 Saw Ice to the So.ward

11

noon

COMMENTS

DAY 40

LEVEL ____

Coder_____ Date_____ Time(hrs)____ (min)_____

1

2 The Ice close and heavy

3

4 Finding the Ice very close and heavy grappled

5

6

7

8

9

10

11

12 The Ice continues very close

1

2

3

4

5

6 Ungrappled. Set sail but the Ice so very close could not ...

7 ... force it

8 Ice a little open

9

10

11

noon

COMMENTS

DAY 41*

LEVEL

Coder _____ Date_____ Time(hrs)____ (min)

1

2 The Ice open

3

4

5

6

7

8

9 The Ice close

10

11 Tacked for Ice

12

1

2 Tacked for Ice

3

4 Tacked for Ice

5

6

7 Tacked. The Ice open

8

9

10 Tacked. "

11

noon

COMMENTS

DAY 42

LEVEL_____

Coder _____ Date_____ Time(hrs)____ (min)_____

1

2 The ship close beset with heavy Ice

3

4 "

5

6

7

8

9

10

11

noon The Ice close and heavy

COMMENTS

DAY 43* LEVEL

Coder _____ Date_____ Time(hrs)____ (min)

1

2 The Ice close

3

4 "

5

6 "

7

8 "

9

10 "

11

12 "

1

2

3

4

5

6

7

8 The Ice open to the SoWard

9 Ungrappled and made sail

10 The Ice open

11

noon The ship close beset

COMMENTS

DAY 44 LEVEL_____

Coder _____ Date_____ Time(hrs)____ (min)____

1

2 Falling the Ice very hard

3 Running. The Ice more open

4 Falling. The Ice close and heavy. Made fast to piece of Ice

5

6

7 The Ice close and heavy

8

9

10

11

12

1

2

3

4

5

6

7

8

9

10

11

noon Grappled to avoid hard blows. Ice close

COMMENTS

DAY 45* LEVEL

Coder_____ Date_____ Time(hrs)____ (min)

1
2
3
4
5
6
7
8
9
10
11
12
1
2 Fell in with Ice
3
4
5 Haul'd to the SoWard in order to round the Ice
6
7
8
9 Rounded the Ice and bore away
10
11
noon

COMMENTS

DAY 46 LEVEL

Coder_____ Date_____ Time(hrs)____ (min)_____

1
2
3
4
5
6
7
8
9
10
11
12
1
2
3
4 Fell in with Ice, haul'd to the NoWard 2 points
5
6
7
8
9
10 Rounded the Ice
11
noon

COMMENTS

DAY 47* LEVEL

Coder_____ Date_____ Time(hrs)____ (min)

1
2
3
4
5
6 Tacked from a body of ice
7
8
9
10
11
12
1
2
3
4
5
6
7
8
9
10
11
noon

COMMENTS

DAY 48 LEVEL_____

Coder_____ Date_____ Time(hrs)____ (min)_____

1 Rowing and towing amongst straggling Ice with the sloop in...
2 ...tow. Ice ahead extending from W to E. Tacked ship off to...
3 ...work round the EtMost part of it. Seeing no prospect of...
4 ...going any other way then getting round it between it and...
5 ...Southhampton shore
6
7
8
9
10
11
12
1
2
3
4
5
6 Tacked ship from the Ice
7
8 Working amongst Ice till 8 when we struck to the WNW Ice...
9 ...being open that way
10
11
noon Rowing and towing amongst stragling Ice

COMMENTS

DAY 49 * LEVEL___

Coder_____ Date_____ Time(hrs)____ (min)__

1
2 Rowing and towing amongst open Ice
3 Got 2 casks of water from a piece of Ice
4
5
6 Amongst stragling Ice
7
8 "
9 Altered our course - least Ice that way
10
11 Falling in with Ice haul'd the wind to the EtWard
12
1
2
3 Made sail to the NoWard. Least Ice that way
4
5
6 Running amongst stragling Ice
7 Bore away amongst "
8
9
10
11 Running as before. Stragling Ice to be seen from the ...
noon ...masthead in every direction but thickest to the SoWard

COMMENTS

DAY 50 LEVEL_____

Coder_____ Date_____ Time(hrs)____ (min)_____

1
2 Ice ahead
3
4 Run through a skim of Ice into clear water
5
6 Tacked ship from a body of Ice
7
8
9
10
11
12
1
2
3
4
5
6
7
8 Working amongst open Ice
9
10
11 Running amongst open Ice till 1/2 past 11
noon

COMMENTS

DAY 51* LEVEL

Coder_____ Date_____ Time(hrs)____ (min)

1
2 Enter'd a body of shatter'd Ice
3
4
5 Got through it into a clear sea
6
7
8
9
10
11
12
1
2
3
4
5
6
7
8
9
10
11
noon

COMMENTS

DAY 52 LEVEL_____

Coder_____ Date_____ Time(hrs)____ (min)_____

1
2
3
4
5
6
7
8
9
10 We was fortunate in not meeting one Isle of Ice all night...
11 ...but daylight had only just appeared before we bore up, ...
12 ...for a very large one; and many more were in sight. Our...
1 ...situation from 8PM to 4AM was a very desperate one...In-
2 ...debted to a Providential Guide in keeping us clear of ...
3 ...Isles of Ice last night
4
5
6
7
8
9
10
11
noon

COMMENTS

DAY 53 *

LEVEL

Coder_____ Date_____ Time(hrs)____ (min)

1
2
3
4
5
6
7
8
9
10
11
12
1
2
3
4
5
6
7
8
9
10
11

noon Much Ice from the WSW to NNE

COMMENTS

DAY 54

LEVEL_____

Coder_____ Date_____ Time(hrs)____ (min)_____

1
2 Entered the Ice
3
4 Sailing among "
5
6 Got quite clear of Ice tho' I take it to be very thick to ...
7 ...the SoWard
8
9
10
11
12
1
2
3
4
5
6
7
8
9
10
11

noon

COMMENTS

DAY 55 LEVEL

Coder_____ Date_____ Time(hrs)____ (min)

1

2 Sailing through very open Ice

3

4 Ice "

5

6 Sailing among shattered Ice

7 Ice close - Grappled

8

9

10

11

12

1

2

3

4

5

6

7

8

9 Ungrappled and kept forcing

10

11

noon

COMMENTS

DAY 56 LEVEL

Coder_____ Date_____ Time(hrs)____ (min)____

1

2 Sailing through broken ledges of Ice

3

4 "

5

6 Past several pieces of Ice

7

8 Got into a clear sea

9

10

11

12

1

2

3

4

5

6

7

8

9

10

11

noon

COMMENTS

APPENDIX II

This appendix contains all of the observed (D_o) and expected (D_e) disagreement values and agreement coefficients (α) calculated for this research. The categories and definitions for each phase precede the tables. Equations for the calculation of these variables are discussed in Chapter 3.

PHASE I

GENERAL CLASSIFICATION	CATEGORY NUMBER	CATEGORIES
A. Presence	0	Ice not present in vicinity of ship
	1	Ice present in vicinity of ship
	*	not enough information
B. Concentration	2	Small area covered by ice (<50%)
	3	Large area covered by ice (>50%)
	*	not enough information
C. Fragmentation	4	Ice cover highly fragmented
	5	Ice cover NOT highly fragmented
	*	not enough information
D. Thickness	6	Thin layer of ice
	7	Thick layer of ice
	*	not enough information
E. Motion	8	Ice in motion
	9	Ice NOT in motion
	*	not enough information

INTERCODER

NOMINAL

General Classification / Coder	Categories in Matrix	D_o	D_e	α
Day - IVb (26days) Concentration	0,2,3,*	.558	.688	.189
Day - IVb (56days) Concentration	0,2,3,*	.541	.680	.204
Day - IVb (26days) Concentration	0,2,3	.289	.533	.458
Day - IVb (56days) Concentration	0,2,3	.250	.524	.523
Hour - IVb (26days) Concentration	0,2,3,*	.520	.675	.230
Hour - IVb (56days) Concentration	0,2,3,*	.506	.673	.248
Hour - IVb (26days) Concentration	0,2,3	.212	.517	.590
Hour - IVb (56days) Concentration	0,2,3	.169	.522	.676
Word II - Concentration	2,3,*	.502	.631	.204

INTERCODER

ORDINAL

Day - IVb (26days) Concentration	0,2,3	.262	.525	.501
Day - IVb (56days) Concentration	0,2,3	.227	.515	.559
Hour - IVb (26days) Concentration	0,2,3	.163	.490	.667
Hour - IVb (56days) Concentration	0,2,3	.123	.500	.754

INTRACODER

NOMINAL

Coder A Concentration	2,3,*,0	.077	.663	.844
Coder D Concentration	2,3,*,0	.167	.616	.729
Coder I Concentration	2,3,*,0	.051	.548	.907
Coder J Concentration	2,3,*,0	.103	.574	.821
Coder M Concentration	2,3,*,0	.397	.704	.436
Coder A Fragmentation	4,5,*,0	.154	.639	.759
Coder D Fragmentation	4,5,*,0	.141	.781	.645
Coder I Fragmentation	4,5,*,0	.128	.542	.764
Coder J Fragmentation	4,5,*,0	.295	.659	.552
Coder M Fragmentation	4,5,*,0	.449	.674	.334

PHASE II

GENERAL CLASSIFICATION	CATEGORY NUMBER	CATEGORY
Concentration	A. 1	Ice free
	2	Open water
	3	Very open ice
	4	Open ice
	5	Close ice
	6	Very close ice
	7	Consolidated/compact ice
Floe Size	B. 1	Giant floe
	2	Vast floe
	3	Big floe
	4	Medium floe
	5	Small floe
	6	Ice cake
	7	Small ice cake
Openings	C. 1	Crack
	2	Open lead
	3	Blind lead
	4	Shore lead
	5	Flaw lead
Arrangement	D. 1	Ice field
	2	Belt
	3	Tongue
	4	Strip
	5	Ice edge (compacted)
	6	Ice edge (diffuse)
	7	Concentration boundary
Motion	E. 1	Diverging
	2	Converging
	3	Shearing

INTERCODER

NOMINAL

General Classification / Coder	Categories in Matrix	D_o	D_e	α
Session 1 - A. Concentration	1,2,3,4,5,6,7,--	.574	.822	.302
Session 1 - B. Floe Size	1,2,3,4,5,6,7,--	.363	.408	.110
Session 1 - D. Arrangement	1,2,3,4,5,6,7,--	.243	.266	.086
Session 1 - E. Motion	1,2,3,--	.116	.178	.348
Session 1 - A. Concentration	1,2,3,4,5,6,7	.453	.785	.423
Session 1 - B. Floe Size	1,2,3,4,5,6,7	.317	.647	.510
Session 1 - D. Arrangement	1,2,3,4,5,6,7	.733	.840	.127
Session 1 - E. Motion	1,2,3	.159	.547	.709

Session 2 - A. Concentration	1,2,3,4,5,6,7,--	.572	.837	.317
Session 2 - B. Floe Size	1,2,3,4,5,6,7,--	.344	.348	.011
Session 2 - D. Arrangement	1,2,3,4,5,6,7,--	.249	.278	.104
Session 2 - E. Motion	1,2,3,--	.104	.146	.288
Session 2 - A. Concentration	1,2,3,4,5,6,7	.445	.811	.451
Session 2 - B. Floe Size	1,2,3,4,5,6,7	.342	.501	.317
Session 2 - D. Arrangement	1,2,3,4,5,6,7	.563	.659	.146
Session 2 - E. Motion	1,2,3	.098	.481	.796
Session 3 - A. Concentration	1,2,3,4,5,6,7,--	.567	.841	.326
Session 3 - B. Floe Size	1,2,3,4,5,6,7,--	.356	.361	.014
Session 3 - D. Arrangement	1,2,3,4,5,6,7,--	.281	.316	.111
Session 3 - E. Motion	1,2,3,--	.122	.162	.247
Session 3 - A. Concentration	1,2,3,4,5,6,7	.430	.809	.468
Session 3 - B. Floe Size	1,2,3,4,5,6,7	.433	.542	.201
Session 3 - D. Arrangement	1,2,3,4,5,6,7	.506	.635	.203
Session 3 - E. Motion	1,2,3	.250	.556	.550
Session 3 - D. Arrangement	1&5,2,3,4/6/7	.089	.337	.736
Session 3 - D. Arrangement	1&5,2/3,4/6/7	.089	.334	.734
Session 3 - E. Motion	1&3,2	.023	.069	.667

INTERCODER

ORDINAL

Session 1 - A. Concentration	1,2,3,4,5,6,7	.205	.769	.733
Session 2 - A. Concentration	1,2,3,4,5,6,7	.169	.788	.786
Session 3 - A. Concentration	1,2,3,4,5,6,7	.109	.777	.860
Session 1 - A. Concentration	1,2/3,4/5,6/7	.138	.609	.773
Session 2 - A. Concentration	1,2/3,4/5,6/7	.121	.648	.813
Session 3 - A. Concentration	1,2/3,4/5,6/7	.062	.618	.900
Session 3 - A. Concentration	1,2-4,5/6,7	.053	.491	.892
Session 3 - A. Concentration	1,2-4,5-7	.052	.490	.894
Session 3 - A. Concentration	1,2-6,7	.021	.239	.910
Session 3 - A. Concentration	1,2-7	.016	.234	.932
Session 3 - B. Floe Size	1-3,4/5,6,7	.136	.333	.592

INTRACODER

NOMINAL

Coder A Concentration	1,2,3,4,5,6,7,--	.457	.832	.451
Coder A Concentration	1,2,3,4,5,6,7	.377	.800	.529
Coder A Floe Size	1,2,3,4,5,6,7,--	.325	.664	.511
Coder A Floe Size	1,2,3,4,5,6,7	.329	.515	.361
Coder A Arrangement	1,2,3,4,5,6,7,--	.045	.135	.667
Coder A Arrangement	1,2,3,4,5,6,7	.042	.085	.506
Coder A Motion	1,2,3,--	.034	.207	.836
Coder A Motion	1,2,3	.163	.493	.669
Coder D Concentration	1,2,3,4,5,6,7,--	.284	.829	.657
Coder D Concentration	1,2,3,4,5,6,7	.284	.812	.650
Coder D Arrangement	1,2,3,4,5,6,7,--	.162	.311	.479
Coder D Arrangement	1,2,3,4,5,6,7	.346	.848	.592
Coder I Concentration	1,2,3,4,5,6,7,--	.363	.729	.502
Coder I Concentration	1,2,3,4,5,6,7	.212	.741	.714
Coder I Arrangement	1,2,3,4,5,6,7,--	.303	.547	.446
Coder I Arrangement	1,2,3,4,5,6,7	.348	.690	.500
Coder I Motion	1,2,3,--	.081	.261	.690
Coder I Motion	1,2,3	0	.434	1.000
Coder J Concentration	1,2,3,4,5,6,7,--	.071	.768	.908
Coder J Concentration	1,2,3,4,5,6,7	.086	.501	.828
Coder J Floe Size	1,2,3,4,5,6,7,--	.013	.326	.960
Coder J Floe Size	1,2,3,4,5,6,7	.022	.043	.488
Coder J Arrangement	1,2,3,4,5,6,7,--	.043	.161	.733
Coder J Arrangement	1,2,3,4,5,6,7	0	.065	1.000
Coder M Concentration	1,2,3,4,5,6,7,--	.468	.829	.435
Coder M Concentration	1,2,3,4,5,6,7	.320	.803	.601
Coder M Floe Size	1,2,3,4,5,6,7,--	.184	.400	.540
Coder M Floe Size	1,2,3,4,5,6,7	.278	.645	.569
Coder M Arrangement	1,2,3,4,5,6,7,--	.090	.177	.492
Coder M Arrangement	1,2,3,4,5,6,7	.345	.372	.073
Coder M Motion	1,2,3,--	.088	.230	.617
Coder M Motion	1,2,3	.026	.535	.951

INTRACODER

ORDINAL

Coder A Concentration	1,2,3,4,5,6,7	.162	.759	.787
Coder A Concentration	1,2/3,4/5,6/7	.159	.616	.742
Coder D Concentration	1,2,3,4,5,6,7	.064	.798	.920
Coder D Concentration	1,2/3,4/5,6/7	.048	.643	.925
Coder I Concentration	1,2,3,4,5,6,7	.074	.716	.897
Coder I Concentration	1,2/3,4/5,6/7	.047	.579	.919
Coder J Concentration	1,2,3,4,5,6,7	.029	.711	.959
Coder J Concentration	1,2/3,4/5,6/7	.006	.489	.988
Coder M Concentration	1,2,3,4,5,6,7	.074	.784	.906
Coder M Concentration	1,2/3,4/5,6/7	.041	.691	.941

PHASE III

GENERAL CLASSIFICATION	CATEGORY	DEFINITION
A.Concentration	1. Ice free	- No ice of any kind present.
	2. Open water/ Very open ice	- Concentration <10% to 30%. More water than ice.
	3. Open ice/ Close ice	- Concentration 40% to 80%. Floes may be in contact with each other.
	4. Very close/ Consolidated/ Compact ice	- Concentration 90% to 100%. 100% - no water is visible and floes are frozen together.
B.Floe Size	1. Small ice cake	- An ice cake <2m across.
	2. Ice cake	- Any relatively flat piece of sea ice <2m across.
	3. small/medium floe	- 20-500m across.
	4. Big floe	- 500-2000m across.
C. Arrangement	1. Strip/Diffuse ice edge/ boundary	- A long narrow area of ice concentration (1km or less in width) mostly of small fragments run together by wind or currents/an irregular line limiting an area of dispersed ice, usually on the leeward side of an area of pack ice/a line approximating the transition between two areas of pack ice with distictly different concentrations.

	2. Belt	- A large feature of pack ice arrangement, longer than it is wide, from 1km to more than 100 km in width.
	3. Tongue	- A projection of the ice edge up to several km in length, caused by wind or current.
	4. Ice field/Compacted ice edge	- An area of pack ice consisting of any size of floes which is >10km across/a clear-cut line compacted by wind or current usually on the windward side of an area of pack ice.
D. Motion	1. Diverging	- Ice floes subjected to diverging or dispersive motion thus reducing ice concentration and/or relieving stress in the ice.
	2. Compacting	- Pieces of floating ice are compacting when they are subject to a converging motion, which increases the concentration and/or produces stresses which may result in ice deformation.

INTERCODER

NOMINAL

General Classification / Coder	Categories in Matrix	D_o	D_e	α
IIIa Session 1 - A. Concentration	1,2,3,4	.275	.824	.666
IIIa Session 2 - A. Concentration	1,2,3,4	.290	.731	.603
IIIb Watch 1 - A. Concentration	1,2,3,4	.285	.710	.599
IIIb Watches 2 - 6 Concentration	-,0,+	.245	.577	.575

INTERCODER

ORDINAL

IIIa Session 1 - A. Concentration	1,2,3,4	.104	.713	.854
IIIa Session 2 - A. Concentration	1,2,3,4	.120	.710	.831
IIIb Watch 1 - A. Concentration	1,2,3,4	.166	.694	.761
IIIb Watches 2 - 6 Concentration	-,0,+	.211	.589	.642

INTRACODER

NOMINAL

IIIa Coder A Concentration	1,2,3,4	.090	.686	.869
IIIa Coder D Concentration	1,2,3,4	.224	.747	.700
IIIa Coder I Concentration	1,2,3,4	.077	.696	.889
IIIa Coder J Concentration	1,2,3,4	.058	.743	.922
IIIa Coder M Concentration	1,2,3,4	.160	.733	.782

INTRACODER

ORDINAL

IIIa Coder A Concentration	1,2,3,4	.041	.685	.940
IIIa Coder D Concentration	1,2,3,4	.061	.732	.917
IIIa Coder I Concentration	1,2,3,4	.028	.693	.960
IIIa Coder J Concentration	1,2,3,4	.015	.742	.980
IIIa Coder M Concentration	1,2,3,4	.074	.730	.899