Information Theory Based Measures of Association Applied to Opinion Polls

Dr. Philip T. Eles
CEFCOM Operational Research Team

Dr. Etienne Vincent
CEFCOM Operational Research Team

DRDC CORA TM 2008-022
July 2008

Defence R&D Canada
Centre for Operational Research & Analysis

National Defence Défense nationale

Canada

Information Theory Based Measures of Association Applied to Opinion Polls

Dr. Philip T. Eles
CEFCOM Operational Research Team

Dr. Etienne Vincent
CEFCOM Operational Research Team

Defence R&D Canada – CORA

Technical Memorandum
DRDC CORA TM 2008-022
July 2008

Principal Author

Original signed by Dr. Philip T. Eles

Dr. Philip T. Eles

Defence Scientist

Approved by

Original signed by Dr. Dean Haslip

Dr. Dean Haslip

SH Land & Operational Commands

Approved for release by

Original signed by Jocelyn Tremblay

Jocelyn Tremblay

Chief Scientist

Defence R&D Canada – Centre for Operational Research and Analysis (CORA)

Abstract

A methodology is described for analyzing opinion poll data and, in particular, for measuring associations between poll questions. The methodology is based on information theory and relies on calculating the information content of each question and the information shared between questions. In this context, shared information is defined as the amount of information revealed about an individual's response to one poll question once his/her response to another question is known. Information diagrams are introduced to facilitate interpretation of results. The paper focuses on defining a process for identifying the most significant associations between questions from polls that have a large number of questions. The key results of this work are: the definition of information theory based measures of association, the introduction of a step-by-step process for identifying relevant associations from large datasets, insights into the interpretation of calculated measures of association, a summary of lessons learned from practical application of the technique, and a demonstration of the technique with a real-world example taken from an opinion poll of the Afghan population in Kandahar Province commissioned by the Canadian Forces in February 2008.

Résumé

Il s'agit de la description d'une méthodologie lors de l'analyse des données d'un sondage d'opinion, notamment pour mesurer les associations entre les questions du sondage. La méthodologie se base sur la théorie de l'information et repose sur le calcul de la quantité d'information de chaque question ainsi que sur l'information échangée entre les questions. Dans ce contexte-ci, l'information échangée se définit par la quantité d'information tirée de la réponse d'un individu à une question du sondage dès que sa réponse à une autre question est connue. La présentation d'information à l'aide de schémas est adoptée afin de faciliter l'interprétation des résultats. L'orientation de ce présent document porte sur l'élaboration d'un processus servant à la mesure des associations les plus pertinentes entre les questions de sondages qui contiennent un grand nombre de questions. Les résultats clés de cet ouvrage sont les suivants : la définition des mesures d'association basées sur la théorie de l'information; la mise en place d'un processus systématique pour relever des associations pertinentes provenant de grands ensembles de données; l'aperçu de l'interprétation des mesures calculées d'association; un résumé de leçons apprises à partir de l'application de la technique; la démonstration de la technique à l'aide d'un exemple concret provenant d'un sondage d'opinion effectué auprès de la population afghane de la province de Kandahar, réalisé à la demande des Forces canadiennes en février 2008.

This page intentionally left blank.

Executive summary

Information Theory Based Measures of Association Applied to Opinion Polls:

Eles PT; Vincent E; DRDC CORA TM 2008-022; Defence R&D Canada – CORA; July 2008.

Introduction:

Defence Research & Development Canada – Centre for Operational Research and Analysis (DRDC CORA) has an operational research (OR) team located at Canadian Expeditionary Forces Command (CEFCOM) Headquarters. The CEFCOM OR Team has recently been involved in the analysis of opinion polls of the Afghan population in Kandahar Province in support of effects measurement efforts by Joint Task Force-Afghanistan J5 Effects. In addition to performing statistical analysis of polling results, the CEFCOM OR Team has been exploring new polling analysis techniques. The work presented herein is a result of these efforts.

This technical memorandum introduces a novel approach to measuring associations between sets of questions using concepts from information theory. This technical memorandum is written for analysts who are interested in implementing these techniques for analyzing associations within datasets.

Methodology:

The investigation described in this paper borrows concepts from the field of information theory and develops new concepts to define a measure of association between sets of questions in an opinion poll. Based on experience from applying these concepts to real-world polling data, a process is proposed for identifying, summarizing, visualizing and interpreting sets of questions which show the highest degree of association. Software tools were developed to facilitate the analysis process, and lessons learned were identified. The concepts were applied to a set of real-world polling data for the purposes of illustrating the concepts developed.

Results:

The key result demonstrated by this work is that information theory concepts can be successfully applied to the analysis of polling data. More specifically, this report:

- demonstrates that information theory concepts such as information entropy, mutual information, and information diagrams, coupled with new concepts such as information overlap and information normalization are sufficient to identify the strength of association between sets of questions;
- proposes a step-by-step process to facilitate the analysis of large datasets and a method for visualizing the results for the purpose of identifying and extracting the strongest and most relevant associations between sets of questions;

- provides a preliminary interpretation of the information theory measures that are calculated in terms of the degree of association between questions, and suggests approximate cut-off values for what constitutes a significant association;
- provides lessons learned based on practical experience gained from applying the technique to polling data; and
- demonstrates the techniques with a real-world example taken from an opinion poll of the Afghan population in Kandahar Province commissioned by the Canadian Forces in February 2008.

Significance:

The techniques introduced in this paper augment and enrich the toolset available to analysts who are interested in finding associations, correlations or co-dependencies within large datasets such as those obtained from opinion polls. The techniques described here, and the software tools that have been developed based on this work, will allow new analysts to quickly apply the techniques to other datasets. Though the methodology is presented in the context of polling, it is generally applicable to other situations where correlations are sought in large datasets. The technique has certain advantages over other techniques, such as being able to identify associations between sets of questions rather than just pairs of questions, with no restrictions on the number of possible answers that each question may have.

Future Work:

The processes, measures, and tools presented in this work represent a starting point for the application of information theory concepts to the analysis of polling data; there is ample room to further develop this capability. Recommendations for further work include:

- further investigation of the information measures and how they correspond to degree of association;
- investigation of how response aggregation affects information measures, and how aggregation may be optimized to increase the chance of identifying sub-groups in the population;
- investigation of error calculation, for example, to determine confidence intervals for the information measures based on sampling error; and
- development of tools to automate both the analysis and interpretation of the results to further improve the capability, potentially making comparisons between large numbers of questions feasible.

Sommaire

Information Theory Based Measures of Association Applied to Opinion Polls:

Eles PT; Vincent E; DRDC CORA TM 2008-022; R & D pour la défense Canada – CORA; Juillet 2008.

Introduction:

Le Centre d'analyse et de recherche opérationnelle de Recherche et développement pour la défense Canada (CARO RDDC) possède une équipe de recherche opérationnelle (RO) située au QG du Commandement de la Force expéditionnaire du Canada (COMFEC). L'équipe de RO COMFEC a récemment participé à l'analyse de sondages d'opinion auprès de la population afghane de la province de Kandahar, à l'appui des efforts de mesure des effets du J5 Effets de la Force opérationnelle interarmées – Afghanistan. En plus d'effectuer l'analyse statistique des résultats du sondage, l'équipe de RO COMFEC a exploré de nouvelles techniques d'analyse de sondage. Le travail présenté ci-dessous est un résultat de ces efforts.

Cette synthèse technique présente une nouvelle approche pour mesurer les associations entre les ensembles de questions qui font appel aux concepts de la théorie de l'information. Ce document technique vise les analystes qui ont un intérêt pour la mise en œuvre de ces techniques dans le but d'analyser les associations entre les ensembles de données.

Méthodologie :

La recherche exposée dans ce présent document emprunte des concepts du domaine de la théorie de l'information et élabore de nouveaux concepts pour définir un degré d'association entre les ensembles de questions d'un sondage d'opinion. S'inspirant de l'expérience tirée de l'application de ces concepts aux données d'un sondage réel, un processus a été proposé pour relever, résumer, visualiser et interpréter les ensembles de questions qui exposent le degré d'association le plus élevé. Des outils logiciels ont été élaborés afin de faciliter le processus d'analyse et des leçons apprises ont été dégagées. Les concepts ont été mis en œuvre dans un ensemble de données recueilli lors d'un sondage réel afin d'illustrer les concepts élaborés.

Résultats :

Illustré dans ce présent ouvrage, le résultat clé est le suivant : les concepts de la théorie de l'information peuvent être appliqués avec succès à l'analyse des données d'un sondage. D'une manière plus précise, cet ouvrage :

- indique que les concepts de la théorie de l'information tels que l'entropie de l'information, la transinformation et l'information représentée dans des schémas, combinés à de nouveaux concepts comme le chevauchement et la normalisation de l'information, sont suffisants pour déceler le degré d'association entre les ensembles de questions;

- présente un processus systématique pour faciliter l'analyse de grands ensembles de données et une méthode pour visualiser les résultats dans le but d'établir et de relever les associations les plus importantes et les plus pertinentes entre les ensembles de questions;
- propose une interprétation préliminaire des mesures de la théorie de l'information calculées en fonction du degré d'association entre les questions et suggère des valeurs seuils approximatives qui constituent une association pertinente;
- met à profit les leçons apprises fondées sur une expérience en milieu de travail acquise lors de l'utilisation de la technique sur des données de sondage;
- expose les techniques à l'aide d'un exemple concret tiré d'un sondage d'opinion effectué auprès de la population afghane de la province de Kandahar, réalisé à la demande des Forces canadiennes en février 2008.

Signification :

Les techniques exposées dans le présent document accroissent et enrichissent l'ensemble des outils à la disposition des analystes intéressés à établir des associations, des corrélations ou une interdépendance à l'intérieur des grands ensembles de données tels que ceux tirés de sondages d'opinion. Les techniques ci-présentes et les outils de logiciel élaborés à partir de cet ouvrage permettront aux nouveaux analystes d'appliquer rapidement les techniques à d'autres ensembles de données. Bien que la méthodologie soit présentée dans le contexte des sondages, elle peut aussi être généralement appliquée à d'autres situations où les corrélations sont recherchées parmi de grands ensembles de données. La technique possède certains avantages par rapport à d'autres techniques, tels que la possibilité de relever des associations entre les ensembles de questions plutôt qu'entre les paires de questions seulement, et ce, sans limite du nombre de réponses possibles pour chaque question.

Travaux de recherche à venir :

Les processus, les mesures et les outils exposés dans ce présent ouvrage un point de départ à la mise en pratique de concepts de la théorie de l'information pour l'analyse de données d'un sondage. Il y a place à une élaboration de cette capacité. Parmi les recommandations de travaux ultérieurs, notons :

- une recherche plus poussée sur les mesures de l'information et sur leur correspondance au degré d'association;
- une recherche sur la manière dont le regroupement de réponses affecte les mesures de l'information et sur la manière dont celui-ci peut être optimisé afin d'augmenter la possibilité d'établir des sous-groupes parmi la population;
- une recherche sur le calcul de l'erreur. La tâche d'établir les intervalles de confiance en ce qui a trait aux mesures de l'information à partir d'une erreur d'échantillonnage en serait un exemple;
- l'élaboration d'outils afin d'automatiser l'analyse et l'interprétation des résultats en vue d'améliorer la capacité, ce qui pourrait permettre d'établir des comparaisons entre un grand nombre de questions.

Table of contents

Abstract …….. i
Résumé …..... i
Executive summary iii
Sommaire v
Table of contents vii
List of figures ix
List of tables x
1....Introduction..... 1
1.1 Background 1
1.2 Aim and Scope 2
2....Background on Information Theory 3
2.1 Information Entropy 3
2.2 Mutual Information 4
2.2.1 Pairs of random variables..... 4
2.2.2 More than two random variables..... 5
2.3 Information Diagrams 6
2.4 Information Theory and Data Compression 7
3....Information Theory Applied to Polling Results..... 8
3.1 Calculating Information Measures 8
3.2 A Measure of Association Between Questions 8
3.3 Normalization of Information Overlap..... 10
4....Implementation and Practical Considerations 11
4.1 A Process for Information Theory Analysis of Polls..... 11
4.2 Lessons Learned 12
5....Analysis of a Poll of Afghans Conducted in Kandahar Province in February 2008..... 16
5.1 Background 16
5.2 Two-Question Comparisons..... 17
5.3 Three-Question Comparisons..... 20
6....Conclusion 24
References 26
Annex A .. Some Instructive Examples 27
A.1 Example 1: Calculating The Mutual Information Between Two Questions 27
A.2 Example 2: Mutual Information Between Three Questions Can be Positive or Negative..... 29
Annex B .. Interpreting Information Measures..... 31

B.1 Interpreting Bits as Units of Information 31
B.2 Interpreting Mutual Information in Terms of Degree of Association 31
List of symbols/abbreviations/acronyms/initialisms 35
Distribution list 37

List of figures

Figure 1: Information entropy for a coin toss 4

Figure 2: Generic information diagram for three random variables 7

Figure 3: The amount of information about the result of question A contained in the result of question B, and in the results of question B and C 9

Figure 4: Opinion of ISAF versus a) ISAF troop professionalism and b) ISAF troop respectfulness, and c) ISAF providing reconstruction. 19

Figure 5: Opinion of ISAF versus ISAF Troops Respectful and ISAF Troops Professional. 22

Figure 6: Opinion of ISAF versus Development Projects and Geographic Location. 22

Figure 7: Opinion of Taliban versus Development Projects and Geographic Location. 23

Figure 8: Information diagram derived from hypothetical poll data in Table 1. 28

Figure 9: An information diagram for 3 questions with a positive MI. 29

Figure 10: A sample information diagram for 3 questions where the MI is negative. 30

Figure 11: MI (in bits) for two binary questions as a function of the size of the population in the associated group. 32

Figure 12: MI versus percent of the population in the associated group for questions with varying number of possible answers. 33

List of tables

Table 1: Normalized IO between pairs of questions 18
Table 2: Normalized IO with ISAF Opinion as the QoI..... 20
Table 3: Contingency table for example..... 27
Table 4: Percentage of responses to two questions. 32

1 Introduction

Defence Research & Development Canada – Centre for Operational Research and Analysis (DRDC CORA) has an operational research (OR) team located at the Canadian Expeditionary Forces Command (CEFCOM) Headquarters. The CEFCOM OR Team has recently been involved in the analysis of opinion polls of the Afghan population in Kandahar Province in support of effects measurement efforts by Joint Task Force-Afghanistan J5 Effects. In addition to performing statistical analysis of polling results, the CEFCOM OR Team has been exploring new polling analysis techniques. The work presented herein is a result of these efforts.

This work came about from a desire to understand how answers to poll questions were related. Would it be possible to predict a respondent's answer to a question based on their answers to other questions? Were there sub-populations within the surveyed population who shared a certain set of opinions? The identification of such sub-populations could allow for targeted Information Operations campaigns[1], or provide an understanding of the opinions shared by groups such as Taliban supporters. From the perspective of designing future polls, it should also be possible to determine which questions were redundant.

In general, contingency tables can be used to give a qualitative assessment of how responses to a set of poll questions are related. Contingency tables summarize the number of times each possible combination of answers was given for a set of questions. However, in polls where a large number of questions are asked, the set of possible comparisons between pairs of questions (not to mention triplets or more) becomes large and such an approach becomes infeasible. A measure of the degree of association between questions is also desirable, as is an automated approach to highlight the sets of questions that are most associated and to what degree they are associated.

1.1 Background

There is a significant amount written in the scientific literature on the subject of determining whether variables (i.e. questions in the context of polls) are independent or not [1-3]. The traditional approach involves calculating a X^2 statistic, which is normally used to determine whether independence is statistically significant, and deriving measures of association from this statistic [1,2]. There are certain problems with this approach, as well as some other common approaches, not the least of which is their inability to consider associations between more than two variables, and their inability to handle variables that can take on more than two possible values [1].

For the more generic case of associations between a set of variables, with no restrictions on the number of values that each can take on, the literature is less abundant. Some of the more theoretical work in this area includes that done by Renyi, who gives a set of postulates which a measure of association must satisfy [4], and Goodman & Kruskal who propose a number of generalized measures of association [3].

[1] Identifying and targeting sub-populations is commonplace in the world of marketing where advertisers identify and target certain groups such as "soccer moms".

The notion of using concepts from information theory, namely of using "mutual information" as a measure of association, was first proposed by McGill[2] [5] and further developed by Bell [6] among others. Not only is an information theory based approach amenable to handling variables that can take on any number of values, they provide an intuitive way of dealing with associations between three or more variables simultaneously, as will be seen later in this paper. The use of information theory based measures of association is not common in the literature, though it does appear occasionally with applications in a variety of fields, for example in meteorology [7] and cluster analysis [8]. The present paper further develops this approach by applying mutual information as a measure of associations between sets of questions in opinion polls. To our knowledge, this is the first such application of this technique to polling.

1.2 Aim and Scope

This technical memorandum is written for analysts who are interested in implementing these techniques for analyzing associations within datasets. Though the methodology is presented in the context of polling, it is generally applicable to any datasets in which associations may be of interest.

The paper begins with an introduction to information theory, followed by a description of how information theory can be applied to polling data, and concludes with a real-world example of an analysis of a subset of data from an opinion poll of the Afghan population in Kandahar Province commissioned by the Canadian Forces in February 2008. Annex A contains some instructive examples that describe details of the calculations, as well as highlighting some of the subtleties of the information theory approach. Annex B explores the interpretation of information measures.

[2] Interestingly, McGill conducted the work with support from the U.S. Air Force Human Factors Operations Research Lab.

2 Background on Information Theory

An attempt can be made to predict the outcome of a random event based on the probabilities for the various outcomes. The accuracy of the prediction depends on the values of those probabilities: if one outcome has a much higher probability than the others then an accurate prediction can be made, whereas if all probabilities are roughly equal, then the prediction will be poor. Therefore, one can say that the probability distribution contains *some* information about the outcome of a random event. The *amount* of information contained in the outcome of a random event is defined by a field of mathematics called Information Theory. In fact, the definition of information content of a random variable forms the basis of Information Theory [9].

2.1 Information Entropy

The information content of a random variable is defined mathematically by its so-called information entropy (or Shannon entropy), *H*, given by:

$$H = -\sum_i p_i \log p_i \tag{1}$$

Here p_i is the probability that outcome *i* will occur (with summation over all possible outcomes). The information entropy (IE)[3] is a measure of the average amount of information that an individual is missing when he/she does not know the outcome of the random variable. By definition, IE is in units of *bits* when the logarithms are base 2, as is the case in this paper.

As an example, consider the toss of a fair coin which can land either heads or tails with equal probability. There is no prior information regarding what the outcome will be. Therefore, revealing the result of the coin toss carries information (1 bit of information in fact since there are two possible outcomes[4]). However, consider some hypothetical unfair coin that always lands heads. Revealing the result of the toss of such a coin carries no information (0 bits[5]) because the answer is already known beforehand. Figure 1 shows the dependence of the information entropy of a coin toss on the probability of landing heads.

[3] In the text, the term information entropy is abbreviated IE, whereas *H* is the mathematical symbol used to represent information entropy in equations. Thus, IE and *H* refer to the same concept, but are used in different contexts. Similarly for MI and *I* that will be introduced shortly.

[4] $H=-[\frac{1}{2}\log_2(\frac{1}{2}) + \frac{1}{2}\log_2(\frac{1}{2})] = -[-\frac{1}{2} - \frac{1}{2}]=1$

[5] $H=-[1\log_2(1)+0\log_2(0)] = -[0 + 0] = 0$

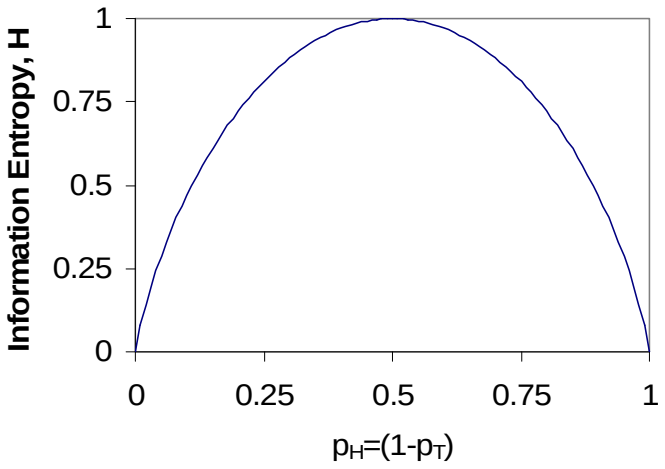


Figure 1: Information entropy for a coin toss as a function of the probability of landing heads (p_H). The information entropy is the amount of information (in bits) transmitted when the result of the toss is disclosed.

2.2 Mutual Information

2.2.1 Pairs of random variables

When considering pairs of random variables, it is often the case that their outcomes are not independent. It is possible to calculate the amount of information that is revealed about the outcome of one variable when the outcome of the other is revealed. The amount of information shared between two random variables is called the *mutual information* (*MI)*.[6] Taken together MI, IE and other values that have units of bits are typically referred to collectively as *information measures*.

For a pair of random variables *X* and *Y*, the pair-wise MI is denoted $I(X;Y)$ and is given by:

$$I(X;Y) = H(X) + H(Y) - H(X,Y), \tag{2}$$

where $H(X)$ is the entropy for the variable *X* independent of *Y* (from Equation (1)), and $H(Y)$ is the entropy for *Y* independent of *X*. The term $H(X,Y)$ is called the *joint entropy*, and is calculated from the joint probability $p(x,y)$: [7]

[6] The concept of MI forms the basis of the methodology presented in this work. The random variables are the answers to poll questions, and the probabilities are the fraction of respondents that gave each answer.

[7] The joint probability is the probability that outcomes *x* and *y* will occur together. It is typically expressed as a probability matrix with entries in each row and column representing the possible outcomes for the two variables. Generally speaking, $p(x,y) = p(x)\,p(y|x) = p(y)\,p(x|y)$ where $p(y|x)$ is the probability of obtaining outcome *y* given that *x* was obtained. For independent variables, $p(y|x) = p(y)$ and $p(x|y) = p(x)$ so that $p(x,y) = p(x)p(y)$.

$$H(X,Y) = -\sum_{x,y} p(x,y) \log p(x,y) . \tag{3}$$

Here x and y are the possible outcomes for the random variables X and Y, and summation is over all possible combinations of outcomes.

The mutual information, $I(X;Y)$, is a measure of the degree to which two variables are not independent (i.e. the degree to which $p(x,y) \neq p(x)\ p(y)$) [10]. This can be demonstrated by considering the two limiting cases: when X and Y are completely independent, and when they are fully co-dependent.

For the case of two independent variables, $p(x,y) = p(x)\ p(y)$. Then,

$$\begin{aligned} H(X,Y) &= -\sum_{x,y} p(x)\,p(y)\big(\log p(x) + \log p(y)\big) \\ &= -\sum_{x} p(x) \sum_{y} p(y) \log p(y) - \sum_{y} p(y) \sum_{x} p(x) \log p(x) . \\ &= -H(X) - H(Y) \end{aligned} \tag{4}$$

Substituting the result into Equation (2) yields $I(X;Y)=0$. Therefore, when two variables are independent, the mutual information is zero.

For the case of two co-dependent variables, $p(x,y) = p(x) = p(\mathrm{y})$ (assuming for simplicity that each has the same number of possible outcomes). Then,

$$\begin{aligned} H(X,Y) &= -\sum_{x} p(x) \log p(x) \\ &= -H(X) \end{aligned} . \tag{5}$$

Substituting into Equation (2) yields $I(X;Y)=H(X)= H(Y)$. Therefore, when two variables are co-dependent, the mutual information equal the IE. It can be shown that this represents the maximal value of IE [10]. This makes intuitive sense: the most information that two variables can share is the information contained in either one of the variables.

2.2.2 More than two random variables

The definition of MI can be extended to more than two variables and in general it describes the information shared between a set of random variables. For three random variables, *X*, *Y*, and *Z*, the MI is:

$$
\begin{aligned}
I(X;Y;Z) &= H(X,Y) + H(X,Z) + H(Y,Z) \\
&\quad - H(X) - H(Y) - H(Z) - H(X,Y,Z) \\
&= \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y)\,p(x,z)\,p(y,z)}{p(x)\,p(y)\,p(z)\,p(x,y,z)}
\end{aligned}
\tag{6}
$$

For larger numbers of variables, the equations become increasingly complicated but have the same general form.[8]

Another useful measure that will appear in this paper is the amount of information about one variable (call it *A*) that is contained in a set of other variables (call them *B, C, D,*…). This can be interpreted as the amount of information revealed about the result of *A* when the results of *B, C, D,*…are revealed. Denoted as *I*(*A;*(*B,C,D,*…)), it is given by:

$$
I(A;(B,C,D,...) = H(A) + H(B,C,D,...) - H(A,B,C,D,...)\,. \tag{7}
$$

Here, *H*(…) is the joint entropy of multiple variables calculated as in equation (6). In the text, *I*(*A;*(*B,C,D,*…)) is referred to as the *information overlap*, or *IO*.

2.3 Information Diagrams

It is often useful to represent information measures in terms of information diagrams. In these diagrams, each circle represents a random variable, the area of each circle is proportional to its IE, and the overlap area between two circles represents the MI between those variables. The combined area of two overlapping circles is the joint entropy. Figure 2 depicts a generic information diagram for three variables. From such diagrams it is easy to see the origin of Equation (2).

[8] For more than 3 random variables, the interested reader is referred more complete works on information theory such Ref. [9].

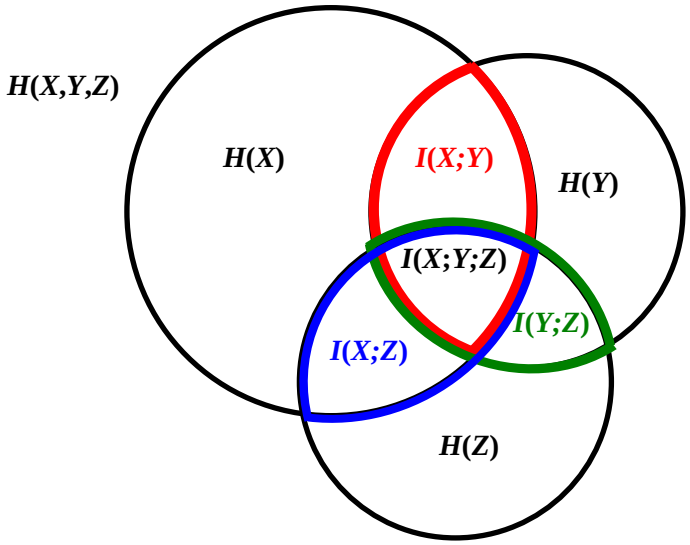


Figure 2: Generic information diagram for three random variables. The area of each circle represents the IE of the variable. The intersection of the circles represents their MI. The area of each circle that is not included in the intersection represents the amount of information that remains unknown once the other variable is revealed.

2.4 Information Theory and Data Compression

Compression of computer files (commonly referred to as "zipping") is intuitive for many people in today's computer age. Data compression, whose theoretical description has its origins in information theory, is discussed here to illustrate the concepts of information theory that were presented in the previous section.

Consider a random stream of numbers that is generated by sampling a random variable many times. The amount of information carried by the data stream is exactly the IE of the random variable. By definition, if there is any redundancy in the data, then the redundancy contains no information. Thus, the IE describes the amount of non-redundant information in the data stream. Data compression works on the principle that the redundancy in a data stream (i.e. a data file) can be eliminated through efficient coding. It can be shown that the IE of a data stream defines the theoretical limit to which the data can be compressed. Similarly, when considering two data streams together, if there is information about one data stream contained in another, then there is redundancy, and the data can be further compressed. MI is, by definition, the amount by which the data stream can be further compressed compared to compression of each variable independently.

In a similar fashion, the question being addressed in this paper is the following: "By how much can the answers to poll questions be compressed on their own, and by how much more can they be compressed when sets of questions are considered together?" Those questions that can be compressed significantly more when considered together have a high degree of redundancy, or rather a high degree of association.

3 Information Theory Applied to Polling Results

In an opinion poll, a random sample of the population is chosen with the assumption that the set of answers provided is representative of the entire population. Thus each question asked of each respondent is akin to sampling a random variable, and the answer provided by each respondent is akin to the outcome of that variable. Each question in the poll can be thought of as a separate variable. Because each respondent is asked multiple questions, there is a natural link between the outcomes of multiple variables and it is natural to look for associations between the answers.

From the perspective of analyzing polling data, the question being addressed is: "How much can be inferred regarding a respondent's answer to a question, given their answers to other questions?" This is exactly the domain of information theory, and the remainder of the paper is concerned with applying information theory concepts to polling results.

3.1 Calculating Information Measures

The results of a poll can be expressed in contingency tables that summarize the number of times each possible answer (or combination of answers when considering multiple questions simultaneously) was given. Dividing each entry in the table by the total number of respondents gives the frequency of each response, which is the most likely estimate for the probability of each response. The contingency tables can be used to calculate information measures directly from Equations (1), (2), (3) and (6).[9] A sample calculation of information measures is given in Annex A.1. The interpretation of MI values and how they relate to the degree of association between questions is discussed in Annex B.

3.2 A Measure of Association Between Questions

We propose that the measure that is most relevant in the context of polling is *the amount of information that is contained in the answer to one question about the answer to another question.* For a pair of questions, this is just the MI (see Figure 3a).

However, for more than two questions, the MI is the amount of information shared by all questions. This is not generally an adequate measure of the degree to which a set of questions is associated since it does not take into account the pair-wise MI. Thinking back to the information theory diagram in Figure 2, the MI of three variables is the area where all three circles overlap, and does not take into account the overlap between pairs of questions.[10]

Therefore, for three or more questions we propose a new information measure that we call the *information overlap* (*IO*). We define the IO as the information about one question of interest (QoI) contained in the answers to the other questions. On the information theory diagram, the IO is the area of overlap between the QoI's information circle and the remaining circles, as depicted

[9] In practice, contingency tables can be generated directly from poll results using most statistical software packages. In MS Excel, they can be generated using pivot tables.

[10] Another complicating factor is that the MI can take on negative values. This is discussed in Annex A.2.

in Figure 3b for three questions. The larger the IO, the more information about the QoI is contained in the other questions. As more and more questions are considered, the IO increases and the area of the QoI's information circle becomes more and more overlapped because the remaining questions contain more and more information about the QoI. For pairs of questions, the IO is just the MI.

By defining the IO as the relevant measure of association between sets of questions, we eliminate the need to create information diagrams for comparing questions. Though information diagrams represent the entire relationship between a set of questions, the IO is all that is required to determine whether they are associated. This is particularly important because the number of overlapping areas on an information theory diagram becomes large when looking at three or more questions, and it becomes infeasible to generate and interpret large numbers of diagrams for polls containing many questions. [11]

One added advantage of using the IO is that the calculations are simple. The IO is given by Equation (7) in Section 2.2. From a computational perspective, the calculation proceeds exactly as it would for pair-wise MI from a two-question contingency table. However, in this case, the contingency table is constructed with the possible answers to the QoI along one side, and all possible combinations of answers to the other questions along the other side.

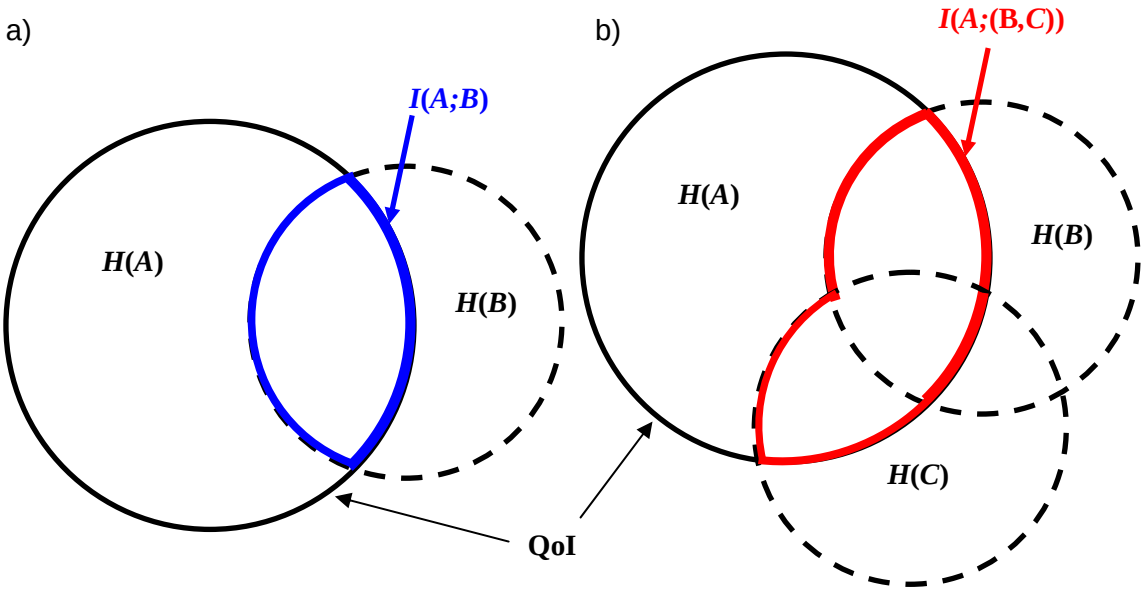


Figure 3: The amount of information about the result of question A contained in the result of question B (in blue outline on the left), and in the results of question B and C (in red outline on the right). I(A,(B,C)) is the information overlap (IO).

[11] In principle, to create a 3-question information diagram, one would need to calculate the IE of each question, the MI between all pairs of questions, and the MI between all three questions.

3.3 Normalization of Information Overlap

Section 3.2 introduced a way to measure the association between sets of questions. What remains is to describe a process for applying this to a large set of questions to extract the strongest and most relevant associations. Such a process is described in Section 4. However there remains one concept that must be introduced beforehand, namely normalization.

Normalization allows for comparisons of information measures across all questions and makes it possible to determine the relative strength of associations between disparate sets of questions. Without normalization, it would not be possible to determine which sets of questions were the most strongly associated.

The information content (i.e. the IE) of a question depends on the number of possible answers, and the probability of each answer. The most information about a QoI that any set of questions can contain is the QoI's IE (i.e. all of its information content). It is therefore both intuitive and appropriate to normalize the IO by dividing it by the IE of the QoI to obtain the fraction (or percentage) of the QoI's information that is contained in the answers to the other questions. On an information theory diagram such as that of Figure 3b, this is the fraction of the QoI's information circle that is overlapped by other questions; in other words, it represents the percentage of the information about question *A* that is contained in the answers to questions *B*, *C*, *D*, … The maximum possible normalized IO is 100%, meaning that all information about the QoI is contained in the other questions, whereas the minimum 0% means that the questions contain no information about the QoI.[12]

Calculating a normalized IO allows for comparisons across different QoI's to identify which questions share the most information with other questions. The validity of using the normalized IO to estimate the degree of association between questions is explored further in Annex B.2.

In summary, Section 3 described a theoretical approach to examining polling data to determine which sets of questions are the most strongly associated. A detailed step-wise implementation of this process that incorporates calculating the IO of a QoI for sets of questions and normalizing by the IE is outlined in Section 4.1. Section 4 also presents lessons learned from the application of this process that may be useful to an analyst interested in conducting such an analysis. Section 5 further illustrates the application of the technique by way of a real-world example. We end this section with a brief description of a practical systematic approach to finding associations within large datasets (i.e. opinion polls with many questions).

[12] The maximum possible normalized IO may not necessarily be 100%, depending on exactly how the answers are distributed. Consider a QoI with two possible answers, and a 50:50 split among respondents. A second question with a 25:75 split among respondents may contain only up to 31% of the information about the QoI. Only a question with a 50:50 split could contain 100% of the information about the QoI. This is one reason why the normalized IOs seen in practice are rarely near 100%.

4 Implementation and Practical Considerations

This section provides a detailed step-by-step guide to implementing the concepts presented in Section 3. These directions are based on the theoretical concepts, coupled with practical experience attained from applying information theory analysis of real-world data. Specifically, the process presented below was developed in the course of applying information theory analysis to data from opinion polls of the Afghan population of Kandahar Province, Afghanistan commissioned by Canadian Forces in support of Operation Athena.

4.1 A Process for Information Theory Analysis of Polls

In summary, what we propose is a process in which one question is initially selected as a QoI. The normalized pair-wise IO between the QoI and each other question is then calculated. Next, the normalized IO between the QoI and all possible pairs of remaining questions is calculated, then all possible triplets, and so on. At each step the set of questions that contains the most information about the QoI is identified. This is the set of questions that takes the "biggest bite" out of the QoI's information circle on an information diagram and represents the questions that are most associated with the QoI.

The following steps were found to be useful for implementing an information theory analysis of polls with a large number of questions.

1. Choose a question from the list of all poll questions. Call it the QoI.

2. Consider whether the QoI is a potentially interesting one. If not, choose another question as the QoI, and repeat step 2.

3. Calculate the IE of the QoI using Equation (1).

4. Choose a question from the remaining set of questions.

5. Calculate the MI between the QoI and the other question using Equations (2) and (3), and divide the result by the IE of the QoI that was calculated in step 3.

6. Repeat steps 4 and 5 for the remaining questions, filling out a table of normalized MI's for all possible questions with the QoI.

7. Next, choose a pair of questions from the set of all questions, excluding the QoI.

8. Calculate the IO between the QoI and the pair of questions using equation (7), and divide the result by the IE of the QoI that had been calculated in step 3.

9. Repeat steps 7 and 8 for all possible pairs of questions, filling out a table of normalized IO's for all possible pairs of questions with the QoI.

10. Repeat steps 2 to 9 with a new QoI until all questions have been considered as a QoI.

11. Apply highlighting to the pair-wise MI table from step 6 to identify the pairs of questions that have the strongest association (i.e. highest normalized IO).

12. Apply highlighting to the three-way IO tables from step 9 to identify the pairs of questions with the strongest association with the QoI.

13. Identify and record the largest associations that are actually interesting (since many are not interesting as will be discussed later).

14. For the interesting associations, go back to the contingency tables to investigate the nature of the relationship (i.e. which responses for one question were associated with which responses for the other question?)

4.2 Lessons Learned

Experience gained while applying the process outlined in Section 4.1 to the Kandahar opinion polling data has given insight into some practical issues that an analyst may consider when using information theory analysis to analyze polling data. Some of these "lessons learned" are discussed below.

1. Generally, Excel is well equipped to perform the analysis. Excel's pivot table tool is suited for producing contingency tables, as well as for generating charts from the tables. Scripts can be written to automate many of the calculations.[13]

2. It has been observed that not all questions make appropriate QoI's. As an example, consider that it is not interesting to know how much information about someone's gender is contained in their political preference. However, it *is* interesting to know how much information about someone's political preference is contained in their gender. Therefore, political preference would be a suitable QoI whereas Gender would not. Limiting the QoIs decreases the size of the output that has to be analyzed and increases the likelihood of finding useful/relevant associations.

3. As already discussed, it was found useful to normalize the IO by the IE to obtain the percentage of the QoI's information contained in the other questions. Normalization allows for comparison across QoI's to identify which questions shared the most information with other questions.

4. In practice, comparisons between four or more questions have not been performed. These calculations are time-consuming, due to the large number of possible combinations of four or more questions. Interpretation of the results becomes difficult and sampling error may become non-negligible (as will be discussed shortly). At some point in the future, automation of the analysis, interpretation, and error calculation will render comparisons of four or more questions feasible and efficient. However, as long as these steps are performed predominantly manually, we recommend limiting the analysis to two- and three-way comparisons.

[13] An Excel VBA script to automate the process outlined in Sections 4.1 for any generic dataset is available from the authors. Interested parties are encouraged to contact the authors to obtain copies.

5. Once the pair-wise and three-way IO table has been generated for all QoI, it is useful to highlight the largest values. In Excel this can be achieved using conditional formatting of the cell background colour. Typically it is useful to highlight the top 10-20% in each table, though this may depend on the degree of correlation between questions being considered. Excel's "percentile" function facilitates the formatting by calculating which values in the table represent a given percentile. Highlighting would not be useful without the normalization discussed in the previous point because without normalization, different QoI's would have different maximum possible values of IO and could not be compared.

6. It is also useful to consider the least associated sets of questions in each table because some useful results can be gleaned from associations that do not exist between variables where association would be expected. One may consider highlighting the entries in each table that have some of the smallest values (e.g. bottom 5-10%).

7. There is no clear answer regarding what value of IO constitutes a strong or weak association. Is 10% a strong or weak association? Annex B explores this issue further by examining the normalized IO obtained from polling of a population with two sub-populations: one in which the answers to two questions are completely correlated, and the other where the answers are uncorrelated. To summarize the results, it appears that even small values of normalized IO represent significant fractions of the population in the correlated group. A 10% normalized IO corresponds to almost 40% of the population being in the correlated group, while a 5% IO represents 25% of the population being in the correlated group. Our intuition is that 5% is roughly the cut-off for significant association, with 10% representing very significant associations. However, this is a preliminary observation based on this work alone. As it has not yet been validated, it may be specific to this dataset and may not hold true for other datasets.

8. It is often the case that in two-way comparisons, one question in particular is very strongly associated with the QoI. When that question is included with any other question as part of a three-way comparison, the result will also indicate a strong association independent of whether the third question contributed much in addition to the two-way association. As a result, the highlighting in the IO tables identifies entire columns or rows that are significantly associated. This does not usually provide any additional insight beyond what was obtained from the two-way table. Therefore, an alternate approach is to also look for pairs of questions that, taken together, reveal significantly more about the QoI than would be expected from the pair-wise comparisons. That is, the total information revealed by the two questions is more than the sum of the information revealed by each individually (i.e. $I(A;(B,C)) > I(A,B)+I(A,C)$). This occurs if, and only if, I(A;B;C) is negative [10]. Triplets of questions that have a high IO and a high negative MI would be of particular interest.

9. Related to the previous point is the fact that, in principle, it is impossible to predict which three-way comparisons will give the highest IO based on the results of all possible pair-wise comparisons. This is because, in principle, the answer to one question on its own may contain little information about the QoI, but when taken in combination with another question (which may also carry little information about QoI on its own), the answers to both questions may reveal a lot about the QoI. Annex A.2 gives an example of this. In practice, however, this seems rarely to be the case.

10. When an association between questions is identified using the information theory approach, it is important to go back to the contingency tables to identify the nature of the trend (i.e. to determine which answers are associated with each other). Information measures reveal which questions are associated, but do not describe *how* they are associated.

11. It is most useful to examine contingency tables in graphical form as "stacked column" histograms. This is easily accomplished with most statistical software packages, as well as in Excel. In a "stacked column" histogram, each column represents one possible combination of answers to all questions except the QoI. Each column is further divided into several "stacked" columns representing the responses to the QoI. A "100% stacked column" can also be generated, where each column is rescaled so that all columns are of equal height (see Figure 4 for an example).The latter type of histogram gives the clearest interpretation of the association between questions, though the former should also be viewed to verify that there are sufficient number or for each possible response to support the conclusions. Any column that has fewer than 10 responses should be viewed with caution.

12. Caution should be used when interpreting the results. An association between questions should not be confused with a causal relation. The fact that many individuals had similar responses to two or more questions does not mean that one opinion is a result of another opinion. Most times, the opinions are connected due to some other underlying factor. It is fair to state that "peoples' opinions regarding question X and question Y were linked/correlated", or "people who thought X also thought Y", but it is not correct to say "peoples' opinion regarding question X was a result of their opinion of Y" or "people's opinion regarding question X was because they shared characteristic Y".

13. For any comparison of questions, if the value of any entry in the contingency table is small (less than ten), then it may be unadvisable to apply the information theory approach.[14] This would occur if too few individuals were polled, if too many questions were compared simultaneously, if the questions being compared had too many possible answers, or if the answers to a question are heavily skewed. Small numbers in the contingency tables are subject to high sampling error that can unfairly bias information measures. For this reason, and for reasons of increased complexity of interpretation, comparisons between more than three questions are not currently performed. In principle, an approach to error analysis could be developed to give confidence intervals for the information measures.

14. When calculating information measures from contingency tables, it is advisable to remove those respondents who answered "Don't know", or those that refused to answer the question, unless there is an important reason to keep them. Oftentimes, there are only a few individuals to remove, and they are removed for the reasons stated in the previous point (i.e. low counts within contingency tables skew information measures).

15. When questions have multiple answers, it may be useful to aggregate the responses. For example, a question may ask about the respondent's view of a particular group or entity, with possible answers being: "very favourably", "somewhat favourably", "somewhat unfavourably", and "very unfavourably". If the aim is to identify sub-groups that view the entity favourably, or unfavourably, then it may prove useful to group the "very" and

[14] This is also a typical rule of thumb for the Pearson Chi-squared test.

"somewhat" categories together to create "favourable" and "unfavourable" categories. Alternately, one may be interested in identifying those sub-groups that have extreme views one way or the other. In this case, it would be useful to aggregate the "somewhat favourably" and "somewhat unfavourably" categories. In general, the aim of applying information theory is to identify sub-groups within the sample population who share similar characteristics, which are different from the rest of the sample population. Creating too few categories (or categories divided incorrectly) might result in the relevant sub-groups being lumped in with (or diluted by) other individuals who do not belong in the sub-group. Alternately, leaving too many categories may split a sub-group unnecessarily, thus reducing the chance that the sub-group will be identified.

16. Related to the previous point is the fact that often only one sub-group may be of interest (e.g. very strong supporters of some political party). If this sub-group represents a small portion of the sample population, then even if strong associations for that sub-group are present, weak associations for the rest of the population may dilute the information theory measures resulting in the association not being found. One potential solution is to repeat the information theory analysis for only the sub-group of interest. In statistics, this is referred to as identifying "partial association" [1]. Another potential solution may be to aggregate the responses so that respondents fall into the "sub-group of interest" category, or the "not the sub-group of interest" category. Further work is required to understand this fully.

5 Analysis of a Poll of Afghans Conducted in Kandahar Province in February 2008

5.1 Background

In this section, the information theory approach is applied to a real-world example. We consider the results of a poll commissioned by the Canadian Forces and conducted in Kandahar Province in February 2008. The poll asked 1226 Afghans their opinion on a large set of issues along the general lines of security, development and governance. The following example is an analysis of a very small subset of the questions to illustrate the methodology described in the previous sections. The full set of results and analysis can be found in a separate publication [11].

Ten questions were considered for the analysis presented here. Names in brackets are used in the rest of this document to refer to each question. The questions were:

1. Do you have very favorable, somewhat favorable, somewhat unfavorable or very unfavorable opinion of ISAF? (Opinion of ISAF)
2. Do you have very favorable, somewhat favorable, somewhat unfavorable or very unfavorable opinion of the Taliban? (Opinion of Taliban)
3. How would you rate the security situation in your area at present: very safe, safe, neither safe nor unsafe, unsafe, or very unsafe? (Security Situation)
4. Is your family becoming much more prosperous, more prosperous, the same, less prosperous or much less prosperous? (Prosperity)
5. Are there projects being implemented in your community to address any of the following concerns in your area or not? If so, what are they. (Development Projects)
6. Is ISAF involved in providing security/fighting the Taliban in your area? (ISAF Providing Security)
7. Is ISAF involved in reconstruction in your area? (ISAF providing reconstruction)
8. Do you strongly agree, agree somewhat, disagree somewhat or strongly disagree that ISAF soldiers are capable and professional? (ISAF Troops Professional)
9. Do you strongly agree, agree somewhat, disagree somewhat or strongly disagree that ISAF soldiers treat Afghans with respect? (ISAF Troops Respectful)
10. In the past 3 months how often have you seen the Taliban in your area? Was it almost every day, a few times a week, several times a month, less often or never? (Taliban Presence)

In addition, the respondent's gender (Gender) was recorded, as was the district within the Province of Kandahar in which the respondent resided (Geographic Location).

For the analysis presented here the responses were binned. Questions 1 and 2 were binned into *Favourable/Unfavourable* categories, question 3 into *Safe/Neutral/Unsafe*, and question 4 into *Less/Same/More*. Question 5 was binned into *None* for those who identified no projects, *Some* for those who identified up to four projects, and *Many* for those who identified five or more. Questions 6 and 7 were *Yes/No* questions, while questions 8 and 9 were binned into

Disagree/Agree, and question 10 was binned into *Regularly* if Taliban were seen at least several times a month, and *Rarely_or_Never* if they were seen less often or never. For Geographic Location, districts were grouped into operationally relevant areas of *Kandahar City* (including *Dand*), *Zhari/Panjwayi*, *Spin Boldak*, *Arghandab*, and *Other*, where the latter refers to all other districts in Kandahar Province except for the ones that were not polled due to security concerns. Not polled were Sha Wali Kot, Mianeshin, Taktha Pul, and Shorabak districts.

5.2 Two-Question Comparisons

Table 1 shows the normalized IO calculated for some pairs of questions from the Kandahar poll. In the table, values are expressed as a percentage of IE of the QoI along the top row. Therefore, the table should be read vertically by column. For example, in the first column for the question regarding Opinion of ISAF, 9.5% of that question's information is contained in the response to whether ISAF troops are professional, and 9.0% is contained in the response to whether ISAF troops are respectful. The top 5%, 10% and 20% of values are highlighted in green, orange and yellow respectively for emphasis. By examining the table, some noteworthy results can be found. These include:

1. Opinion of ISAF troops is much more linked with perception that ISAF troops are professional and respectful than with the perception that they provide security or reconstruction;[15]
2. Opinion of the Taliban is linked to Taliban presence but not to the perceived security situation or prosperity, and although the presence of Taliban is related to Geographic Location, the opinion of Taliban is not;
3. Feelings on improving/worsening prosperity are related to perception of a safe/unsafe security situation, but largely unrelated to the implementation of development projects in the respondents' communities.

Graphs of the contingency tables for Opinion of ISAF with ISAF Troops Professional, ISAF Troops Respectful and ISAF Providing Reconstruction are shown in Figure 4. From Figure 4a, it is apparent that individuals who agree that ISAF troops are professional are much more likely (61%) to have a favourable opinion of ISAF compared to those who disagree (23%). Similarly, from Figure 4b, those who agree that ISAF troops are respectful were more likely to have a favourable opinion of ISAF (65%) than those that disagree (30%). However, there is a smaller difference in opinion of ISAF between those who thought ISAF was providing reconstruction (55%) and those who did not (41%) as shown in Figure 4c.

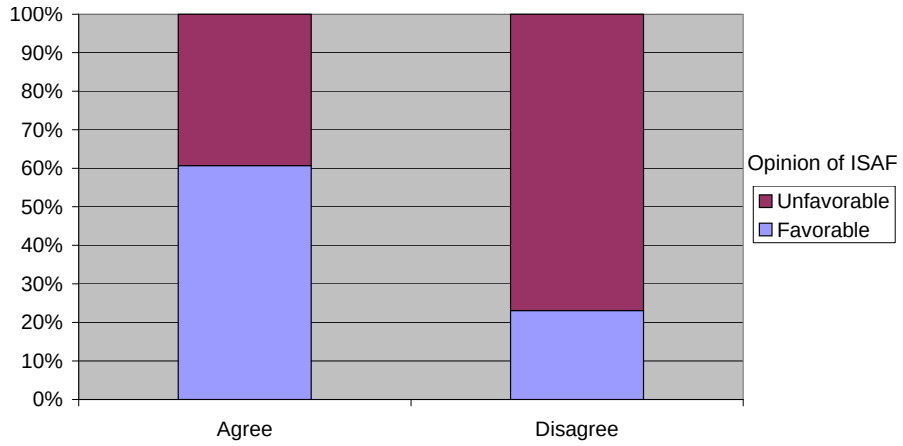
[15] Care must be taken in interpreting this result. A person's perception that ISAF troops are professional and respectful may not be related to whether they have actually observed ISAF troops being professional or respectful. Rather it could be a proxy for their approval of ISAF, hence the high correlation with ISAF Opinion.

Table 1: Normalized IO between pairs of questions

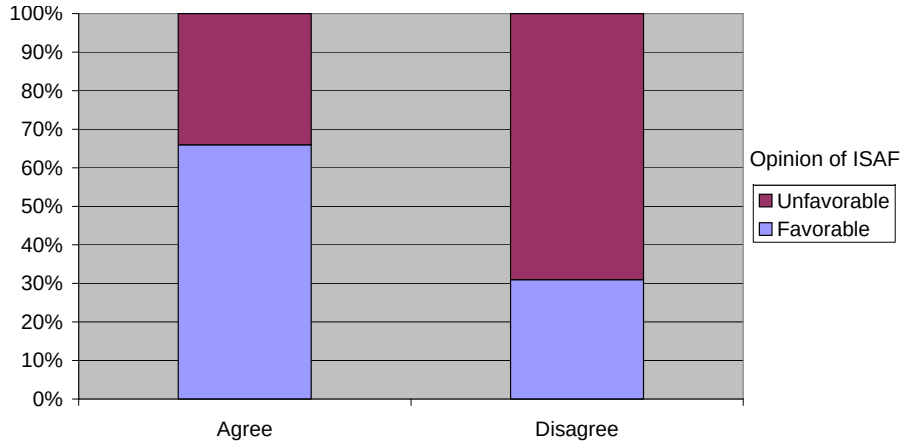
	QoI				
	Opinion of ISAF	**Opinion of Taliban**	**Security Situation**	**Prosperity**	**Taliban Present**
Opinion of ISAF		3.1	2.0	0.8	1.6
Opinion of Taliban	2.3		1.2	0.2	10.3
Security Situation	3.1	2.5		3.8	5.8
Prosperity	1.2	0.5	3.9		1.1
Development Projects	2.9	3.8	1.2	1.3	3.7
ISAF Providing Security	2.5	0.9	1.1	0.2	0.8
ISAF Providing Reconstruction	1.5	0.0	0.8	0.3	0.5
ISAF Troops Professional	9.5	4.7	0.6	0.8	4.7
ISAF Troops Respectful	9.0	0.7	2.3	1.6	2.2
Taliban Present	1.4	11.9	3.2	0.6	
Gender	0.2	0.9	2.1	0.3	0.7
Geographic Location	2.2	2.1	2.3	1.3	9.5

Note: Values are expressed as a percentage of IE of the QoI along the top row and the table should be read vertically by column. The top 5%, 10% and 20% of values are highlighted in green, orange and yellow respectively for emphasis.

a)



b)



c)

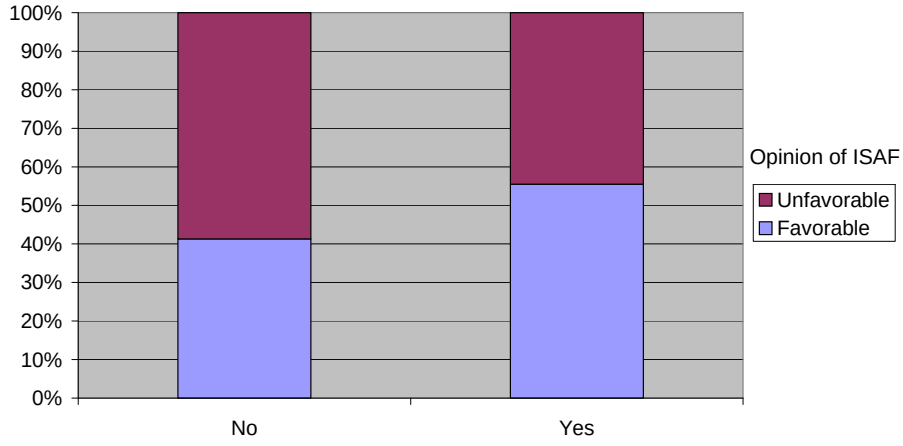


Figure 4: Opinion of ISAF versus a) ISAF troop professionalism and b) ISAF troop respectfulness, and c) ISAF providing reconstruction.

5.3 Three-Question Comparisons

Table 2: Normalized IO with ISAF Opinion as the QoI

QoI: ISAF Opinion	Security Situation	Prosperity	Development Projects	ISAF Providing Security	ISAF Providing Reconstruction	ISAF Troops Professional	ISAF Troops Respectful	Taliban Present	Gender	Geographic Location
Opinion of Taliban	4.9	3.9	4.5	4.6	3.7	10.6	11.2	2.6	2.6	3.8
Security Situation		4.4	5.8	4.8	3.8	11.9	10.6	4.0	3.2	6.2
Prosperity			5.1	4.2	2.7	10.5	9.9	2.5	1.3	5.2
Development Projects				5.1	3.9	11.6	12.8	3.6	4.0	8.0
ISAF Providing Security					3.8	11.1	11.1	4.7	2.8	4.8
ISAF Providing Reconstruction						11.7	11.4	2.6	1.6	4.6
ISAF Troops Professional							14.6	10.0	9.7	11.5
ISAF Troops Respectful								10.0	10.0	11.4
Taliban Present									1.6	3.2
Gender										6.7

Note: Values represent the percentage of ISAF Opinion question's IE contained in the answers to pairs of questions. Values above 12%, 10% and 6% are coloured green orange and yellow for emphasis.

Taking ISAF Opinion as the initial QoI, the IO was calculated between it and all pairs of questions. The results are summarized in Table 2. Note that ISAF troop professionalism and respectfulness both had high values in the two-question comparison, so it is not surprising that that pair has the highest IO with the QoI in the three-way comparison. Also, either question, when coupled with any other question has a high IO. Notably, Geographic Location with Development Projects, with Gender, and with Security Situation also have a significant IO despite the fact that their pair-wise IO were low. This is a case where knowing the answer to Geographic Location

plus Gender, Development Projects, or Security Situation gives more information about ISAF Opinion than the sum of knowing the answers individually.

Figure 5 shows a graph of the contingency table of ISAF Opinion with ISAF troop professionalism and respectfulness. The graph indicates that individuals who agree that ISAF troops are professional and respectful are much more likely (72%) to have a favourable opinion of ISAF compared to those who disagree with both (19%). Of those who agree with one statement but not the other, roughly 40% have a favourable view of ISAF. Indeed, opinions of whether ISAF troops are professional and respectful are associated with ISAF opinion. However, this does not give much more information than was derived from the two-way comparison in Table 1.

The more interesting example from Table 2 seems to be the relationship between ISAF Opinion, Development Projects and Geographic Location. Neither Development Projects nor Geographic Location was strongly associated with ISAF Opinion in the two-way comparison, but the three-way comparison reveals a significant association between the three questions. Figure 6 depicts the contingency table for these three questions. It is apparent that in “Zhari/Panjwayi” and “Other” districts of Kandahar Province, and to a lesser extend in “Spin Boldak” district, those individuals who reported many development projects in their area viewed ISAF much more favourably than those who reported few or no development projects. By contrast, in Kandahar City, there seemed to be little or no correlation between Development Projects and ISAF Opinion. In Arghandab, those who saw many Development Projects had a worse opinion of ISAF than those who saw few or none. It is tempting to conclude that outside of Kandahar City, Development Projects are responsible for a positive view of ISAF, whereas within Kandahar City, they do not affect ISAF opinion. However, there may be some other factor which explains this association.

The three-way IO table that has Taliban Opinion as the QoI (not shown) suggests an interesting association between Taliban Opinion, Geographic Location and Development Projects. A graph of the associated contingency table is presented in Figure 7. This figure indicates that, except for in Arghandab, support for Taliban decreases as the number of Development Projects in the area increases. This is especially true outside of Kandahar City. This result is similar to the one in the previous paragraph. It can therefore be stated that the data suggests that, outside of Kandahar City, ISAF is viewed more favourably and Taliban less favourably the more development projects are reported.

A similar approach can be applied to obtain further results. This example was meant to be illustrative rather than exhaustive; the reader is encouraged to contact the author for additional information if required.

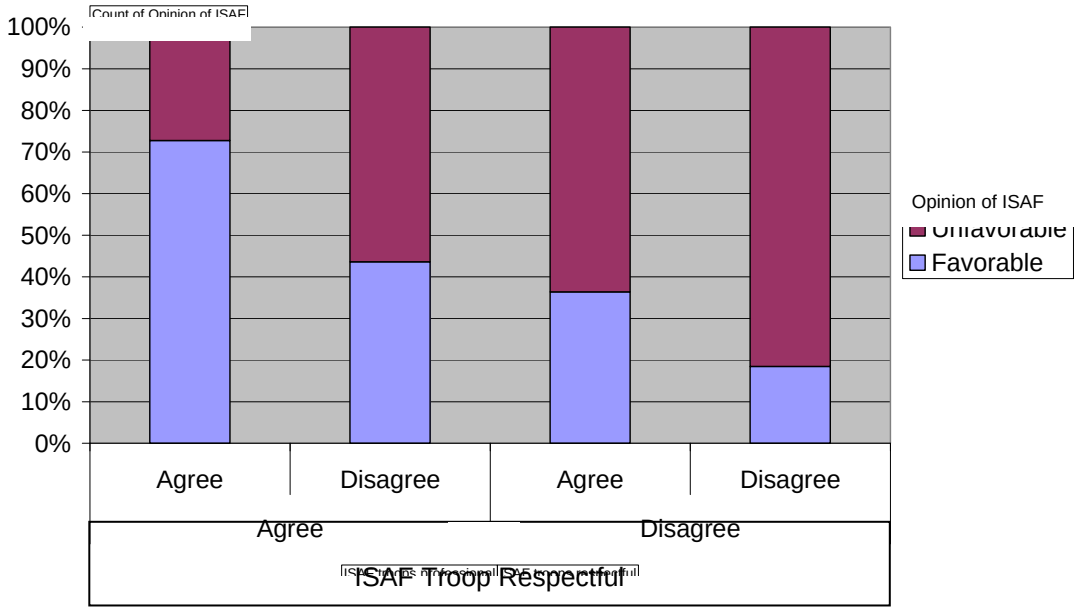


Figure 5: Opinion of ISAF versus ISAF Troops Respectful and ISAF Troops Professional. The vertical axis is a percentage of the total number of individuals giving the same answer for ISAF troop respectfulness and professionalism.

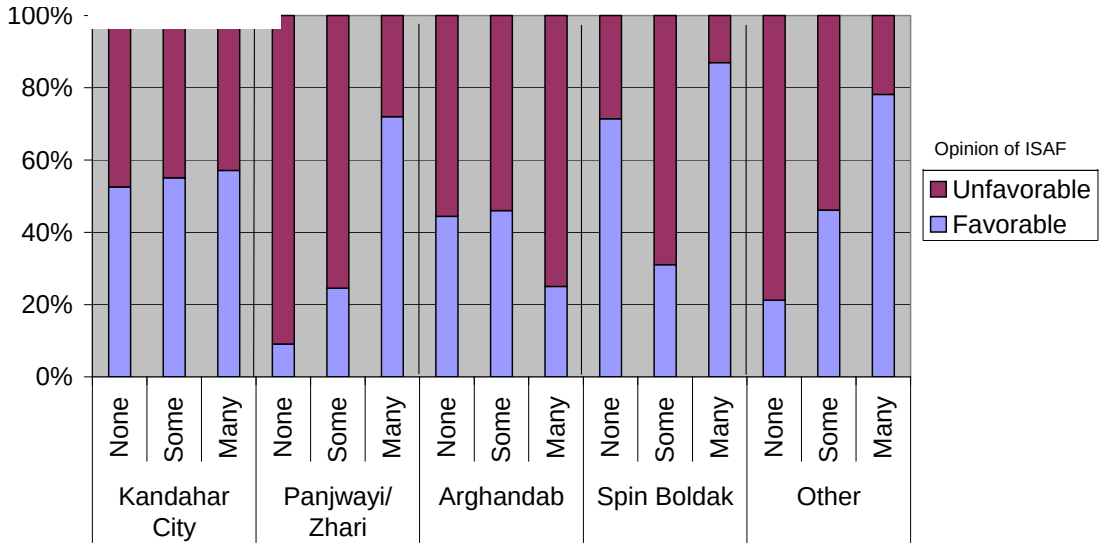


Figure 6: Opinion of ISAF versus Development Projects and Geographic Location.

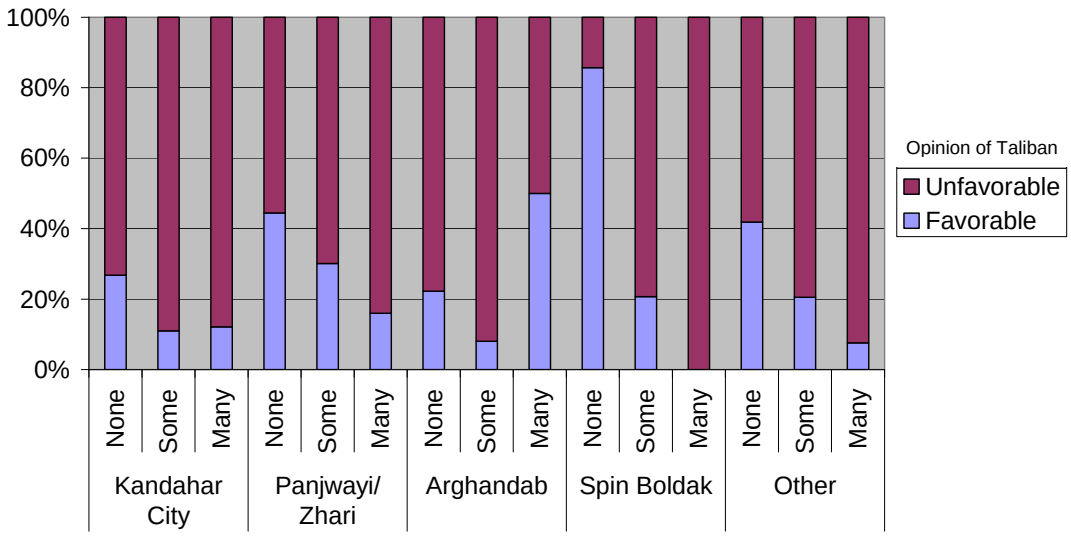


Figure 7: Opinion of Taliban versus Development Projects and Geographic Location.

6 Conclusion

This technical memorandum introduced a novel approach to measuring associations between sets of questions using concepts from information theory. It was written for analysts who are interested in implementing these techniques for analyzing associations within datasets.

The investigation described in this paper borrowed concepts from the field of information theory and developed new concepts to define a measure of association between sets of questions in an opinion poll. Based on experience from applying these concepts to real-world polling data, a process was proposed for identifying, summarizing, visualizing and interpreting sets of questions which show the highest degree of association. Software tools were developed to facilitate the analysis process, and lessons learned were identified. The concepts were applied to a set of real-world polling data for the purposes of illustrating the concepts developed.

The report:

- demonstrated that information theory concepts such as information entropy, mutual information, and information diagrams, coupled with new concepts such as information overlap and information normalization are sufficient to identify the strength of association between sets of questions;
- proposed a step-by-step process to facilitate the analysis of large datasets and a method for visualizing the results for the purpose of identifying and extracting the strongest and most relevant associations between sets of questions;
- provided a preliminary interpretation of the information theory measures that are calculated in terms of the degree of association between questions, and suggested approximate cut-off values for what constitutes a significant association;
- provided lessons learned based on practical experience gained from applying the technique to polling data; and
- demonstrated the techniques with a real-world example taken from an opinion poll of the Afghan population in Kandahar Province commissioned by the Canadian Forces in February 2008.

The techniques introduced in this paper augment and enrich the toolset available to analysts who are interested in finding associations, correlations or co-dependencies within large datasets such as those obtained from opinion polls. The techniques described here, and the software tools that have been developed based on this work, will allow new analysts to quickly apply the techniques to other datasets. Though the methodology is presented in the context of polling, it is generally applicable to other situations where correlations are sought in large datasets. The technique has certain advantages over other techniques, such as being able to identify associations between sets of questions rather than just pairs of questions, with no restrictions on the number of possible answers that each question may have.

The processes, measures, and tools presented in this work represent a starting point for the application of information theory concepts to the analysis of polling data; there is ample room to further develop this capability. Recommendations for further work include:

- further investigation of the information measures and how they correspond to degree of association;

- investigation of how response aggregation affects information measures, and how aggregation may be optimized to increase the chance of identifying sub-groups in the population;

- investigation of error calculation, for example, to determine confidence intervals for the information measures based on sampling error; and

- development of tools to automate both the analysis and interpretation of the results to further improve the capability, potentially making comparisons between large numbers of questions feasible.

References

1. Kendall M., Stuart A "The Advanced Theory of Statistics" Vol. 2, 4th Edition, 1979, pp. 566-615.

2. Lewis B.N. "On the Analysis of Interaction in Multi-Dimensional Contingency Tables" *Journal of the Royal Statistical Society. Series A*, Vol. 125, No. 1 (1962) pp. 88-117.

3. Goodman L.A., Kruskal W.H., "Measures of Association for Cross Classifications" *Journal of the American Statistical Association* Vol. 49, No. 268 (1954) pp. 732-764.

4. Renyi A. "On Measures of Dependence" *Acta Math. Acad. Sci. Hung.*, Vol. 10 (1959) pp. 441-451.

5. McGill W.J. "Multivariate Information Transmission" *Psychometrika* Vol 19, No.2 (1954) pp. 97-116.

6. Bell C.B. "Mutual Information and Maximal Correlation as Measures of Dependence" *The Annals of Mathematical Statistics*, Vol. 33, No. 2 (1962) pp. 587-595.

7. Holloway J., Leith J., Woodbury, M.A. "Application of information theory and discriminant function analysis to weather forecasting and forecast verification" *Technical Report No. 1, Meteorological Statistics Project*, Institute for Cooperative Research, University of Pennsylvania, (1995).

8. Dom B.E. "An information-theoretic external cluster-validity measure" *Research Report*, IBM, RJ 10219, 2001.

9. Shannon C.E. "A Mathematical Theory of Communication" *Bell Sys. Tech. Journal*, Vol 27 (1948) pp. 379-423, 623-656.

10. Yeung R.W. "A First Course in Information Theory", Springer, 2002.

11. Vincent E.; Eles P.; Turnbull A.; Smith P.; Chapman B.; Connell D.; Woodliffe E."Kandahar Province Survey – Wave 4 Summary – February 2008" Internal CEFCOM document. DRDC CORA publication in preparation.

Annex A Some Instructive Examples

A.1 Example 1: Calculating The Mutual Information Between Two Questions

Consider two questions asked in a poll of 480 individuals. Suppose question one (Q1) has three possible answers (I, II, III), and question two (Q2) has four possible answers (A, B, C, D). Suppose Table 3 represent the contingency table that summarizes the results of the poll.

Table 3: Contingency table for example

		Q1			
		I	**II**	**III**	**Subtotal:**
Q2	**A**	80	20	20	120
	B	20	80	20	120
	C	20	20	80	120
	D	40	40	40	120
Subtotal:		160	160	160	

By examining Table 3, one can immediately see that though the responses to each question are equally distributed among the population, there is a strong association between the answers. Those who answered A to Q2 were more likely to answer I to Q1, those who answered B likely answered II, and those who answered C likely answered III. Respondents who answered D to Q2, were equally likely to give any of the answers to questions Q1.

One can see directly from Table 3 that given an individual respondent's answer to Q1, there is a 50% chance of guessing their answer to Q2 (compared to a 25% chance if the questions were unassociated). Conversely, given a respondent's answer to Q2, there is a 66.7% chance of guessing their answer to Q1 if their answer was A, B or C, and a 33.3% chance if their answer was D. On average, there is a 58.3% chance of guessing the answer to Q1 given Q2. If the answers were unassociated, there would be a 33.3% chance of guessing the answer to Q1.

Using the data from Table 3, the IE for each question can be calculated from Equation (1):

$$H(Q1) = -3(160/480)\log_2(160/480) = 1.6\ bits \tag{A.1}$$

$$H(Q2) = -4(120/480)\log_2(120/480) = 2\ bits \tag{A.2}$$

It is instructive to consider Q2 for a moment. There are four possible answers, so it would take two bits to encode them all (00, 01, 10, and 11). From Table 3, it is clear that all four possibilities

are equally likely. Thus, the frequency distribution gives no information regarding what might be expected if one answer were drawn randomly from the dataset, and so the information obtained when that answer is revealed is maximal. Therefore, it is expected that the IE of Q2 is 2 bits, as indeed it is (Equation (A.2)).

Applying Equations (2) and (3), and using the results from Equations (A.1) and (A.2), the MI for the two questions can be calculated:

$$H(Q1,Q2) = -6(20/480)\log_2(20/480) - 3(40/480)\log_2(40/480) - 3(80/480)\log_2(80/480) = 3.335\ bits \quad \text{(A.3)}$$

$$I(Q1;Q2) = 1.585 + 2.000 - 3.335 = 0.25\ bits \quad \text{(A.4)}$$

Thus, Q1 and Q2 share 0.25 bits of information.

Figure 8 summarizes the results of Equations (A.1)-(A.4) by way of an information diagram. The area of each circle is proportional to the IE for each question, and the area of the overlap is proportional to the MI. The numbers in the non-overlapping part of the circle represent the information remaining once the answer to the other question is revealed. As a summary of Figure 8, it can be stated that Q1 contains 0.25 bits of the 2 bits of information contained in Q2 (i.e. 12.5% of the bits). Similarly, Q2 contains 0.25 bits of the 1.59 bits of information contained in Q1 (i.e. 15.8% of the bits).[16] The meaning of these numbers is discussed in Annex B. For now, it is sufficient to say that calculating the MI for all pairs of questions highlights those questions which are *most* associated.

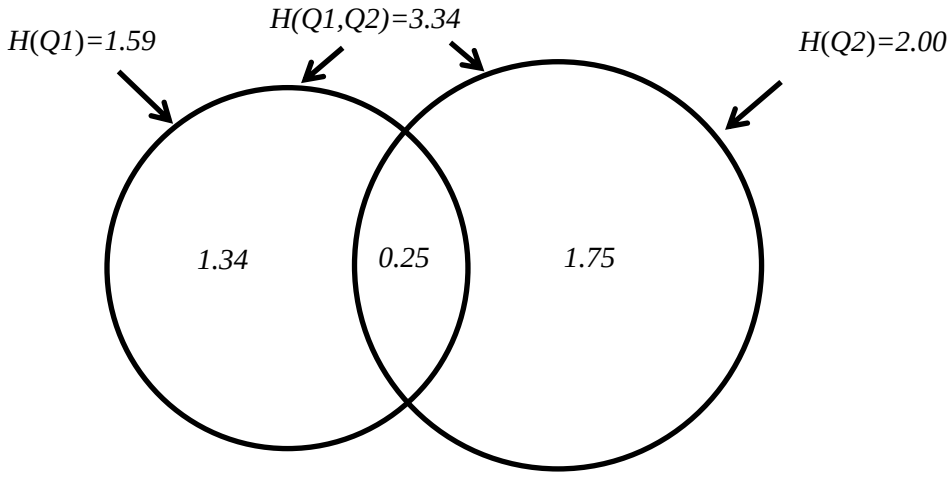


Figure 8: Information diagram derived from hypothetical poll data in Table 3.

[16] It is a general result that for two questions, where Q1 has 3 possible answers, and Q2 has 4 answers, the maximum possible MI is 1.19 which is 75% of the maximal IE of Q1, and 59% of the maximal IE of Q2.

A.2 Example 2: Mutual Information Between Three Questions Can be Positive or Negative

Consider an example involving three questions: A, B and C. Consider also that the following information measures have been calculated: H(A)=1; H(B)=1; H(C)=1; I(A;B)=0.5; I(A,C)=0.5; I(B,C)=0.5; and I(A,B,C)=0.25.[17] The resulting information diagram is shown in Figure 9. The following are interpretations of the diagram:

1. Question A shares 0.5 bits of information with question B, and 0.5 bits with question C. Together, questions B and C contain only 0.75 bits of A's information (i.e. *less* than the sum of the two pair-wise MIs).

2. If the answer to question B is revealed, then question C still holds 0.25 bits of question A's information.

3. If the result of questions B and C are revealed, then 0.25 bits of question A still remain unknown.

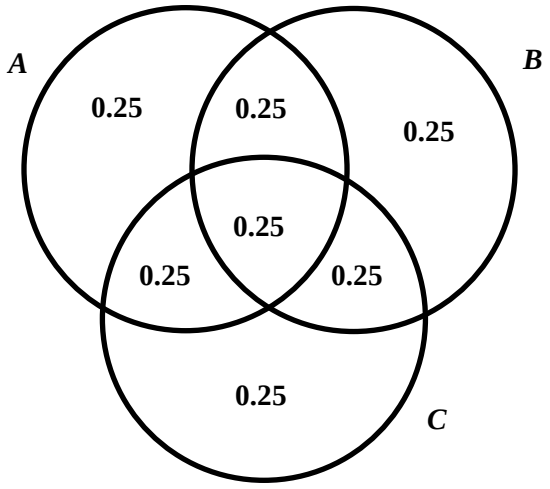


Figure 9: An information diagram for 3 questions with a positive MI. The areas are not to scale.

Interestingly, the MI between more than two questions can be positive or negative. In this example, it was positive. The following example shows what happens when it is negative.

[17] Assuming that questions A, B, and C have two possible answers each (call them a/b, I/II, and i/ii respectively), the data that would produce these information measures and the resulting information diagram might have the following contingency table:

	I		II	
	i	ii	i	ii
a	390	55	55	0
b	55	0	0	445

Consider the same information measures as in the previous example except that the MI is negative: H(A)=1; H(B)=1; H(C)=1; I(A;B)=0; I(A,C)=0; I(B,C)=0; I(A,B,C)=-0.25.[18] The resulting information diagram is shown in Figure 10. The following are interpretations of the diagram:

4. Question A shares no information with question B, and no information with question C. However, together questions B and C contain 0.25 bits of A's information (*more* than the sum of the two pair-wise MIs).

5. If the answer to question B is revealed, then question C now holds 0.25 bits of question A's (more than it did before B was revealed).

6. If the result of questions B and C are revealed, then 0.75 bits of question A are still unknown.

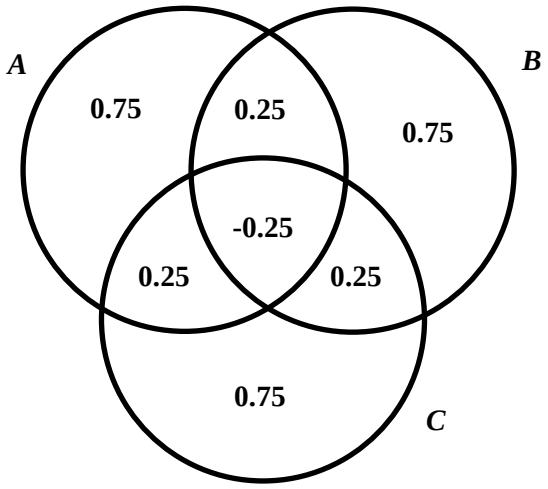


Figure 10: A sample information diagram for 3 questions where the MI is negative. The areas are not to scale.

Generally speaking, for three questions, a negative MI means that knowledge of the answer to two of the questions together tells you more than knowing the answers to each question individually. Knowing B or C individually may not tell you anything about A, but knowing them both tells you a lot. A positive MI suggests that some of the information contained in the three questions is redundant. Knowing the answer to B tells you something about A, but subsequently finding out the answer to C doesn't tell you much more because B also tells you about C.

[18] Assuming that questions A, B, and C have two possible answers each (call them a/b, I/II, and i/ii respectively), the data that would produce these information measures and the resulting information diagram might have the following contingency table:

	I		II	
	i	ii	i	ii
a	196	54	54	196
b	54	196	196	54

Annex B Interpreting Information Measures

Though calculating information measures is straightforward enough, interpreting the results is less trivial. What does it mean that two questions share 0.25 bits of information, or that one question contains 12.5% of the information in the other? Does this suggest that the two questions are associated, or not? Some insight into interpreting information measures is provided in this section.

B.1 Interpreting Bits as Units of Information

Consider a binary string of $n>1$ bits which encode 2^n possible values with equal probability. A random guess of what that string is would have a $1/(2^n)$ chance of being correct. If, however, one of the bits is disclosed, then only $n-1$ bits remain unknown, and a random guess would have a $1/(2^{n-1})$ chance of being correct. Thus knowledge of one bit doubles the chance of guessing correctly.[19]

These arguments suggest that revealing m bits of information increases the uncertainty in the guess by a factor of 2^m. Though this type of argument only applies to integer values of m, fractional values may also be reasonably expected to follow such a relation.[20] That being said, too much should not be read into this interpretation because it was obtained by considering only the case when all possible values of an n-bit sequence are equally likely. The important result to retain from this discussion is that the degree of association between two questions may be expected to scale exponentially with the number of bits of MI.

B.2 Interpreting Mutual Information in Terms of Degree of Association

Another way to examine the link between bits of information and degree of association is to construct an example where the degree of association is intuitively defined, and see what the information measures are. The following is such an approach.

Consider two poll questions Q1 and Q2. Assume that for α percent of respondents, their answers to the two questions are completely *associated* (their answer to Q1 automatically determines their answer to Q2), while the responses of the remaining $(1-\alpha)$ percent of the population are completely *unassociated* (their answer to Q1 has no bearing on their answer to Q2). For reference, call the two subpopulations the associated and unassociated groups. Further assume

[19] As an example, consider the string of 4 bits: *1010*. If someone were to try to guess the bit sequence with no prior knowledge, they would have a $1/2^4=1/16=6.25\%$ chance of guessing correctly. If, however, it was revealed that the first bit is a *1*, then they would have a $1/2^3=1/8=12.5\%$ chance of guessing correctly. Revealing that the first two bits are *10* increases the chance of guessing correctly to $1/2^2=1/4=25\%$.

[20] One can apply this simplified interpretation to the example in Annex A.1 where 0.25 bits of Q2's 2 bits were contained in Q1. Thus, knowledge of Q1 may be expected to increase the chance of guessing Q2 correctly by a factor of $2^{0.25}=1.19$ (i.e. by 19% more than without Q1). The value calculated in Annex A.1 was 25%.

that all possible answers to Q1 are equally likely for both groups, and that all possible answers for Q2 are equally likely for the unassociated group (whereas the associated group's response is determined by their answer to Q1). The contingency table for the case where both questions have 2 possible answers is shown in Table 4. From this table, the MI can easily be determined using Equations (1), (2) and (3).

Table 4: Percentage of responses to two questions.

		Q1		
		I	**II**	
Q2	**A**	$\alpha/2+(1-\alpha)/4$	$(1-\alpha)/4$	50%
	B	$(1-\alpha)/4$	$\alpha/2+(1-\alpha)/4$	50%
		50%	50%	

Figure 11 shows the MI between the two binary questions as a function of the fraction of the population in the associated group. From the figure it is evident that the MI increases non-linearly with increasing size of the correlated group. When the fraction of the population in the associated group is small (less than ~30%), the MI remains fairly low (i.e. the slope is shallow). Even small values of *I(Q1, Q2)* can indicate significant associations within the population. For example, MI of 5%, 10% or 20% of the maximum value, corresponds to 26%, 37% and 51% of the population being in the associated group respectively.

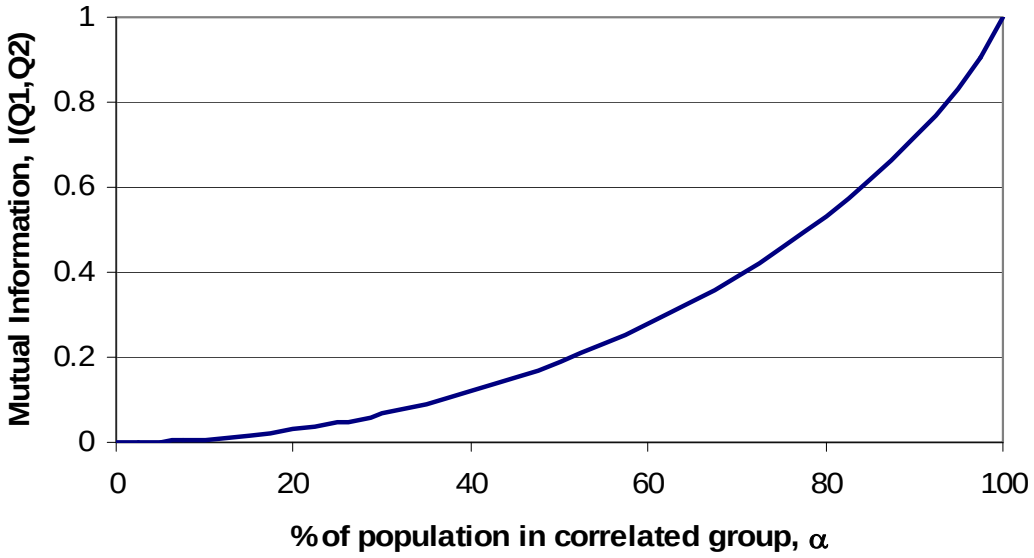


Figure 11: MI (in bits) for two binary questions as a function of the size of the population in the associated group.

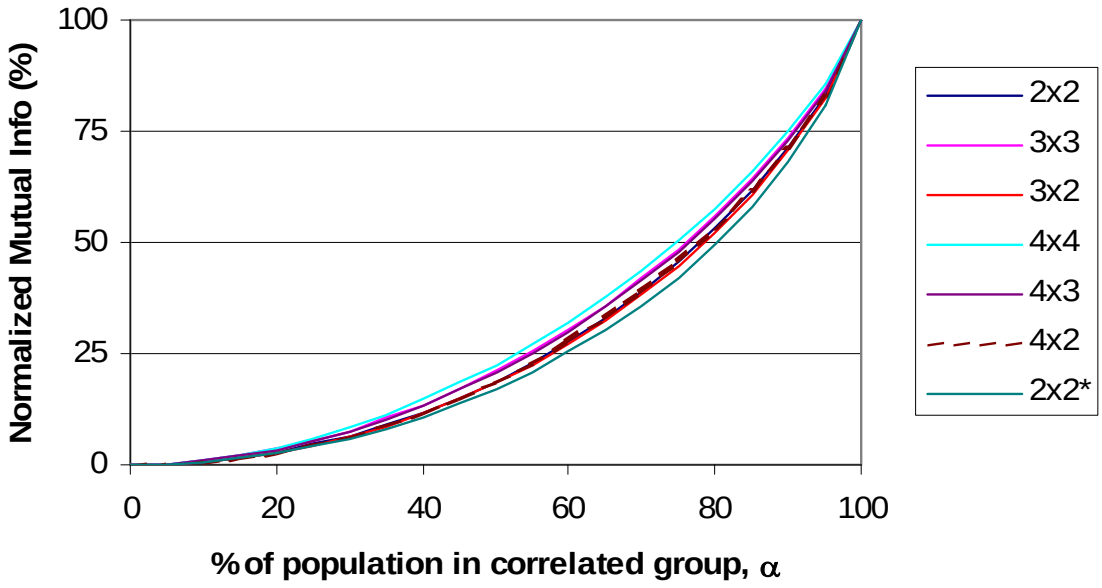


Figure 12: MI versus percent of the population in the associated group (as in Figure 11) for questions with varying number of possible answers. The MI is shown as a percent of the maximum possible value for that pair of questions. The 2×2 line is the case where the answers to Q1 were skewed 80:20 in favour of one answer for both the associated and unassociated groups.*

One might expect that this result would depend on the number of possible answers to the two questions (indeed there is no reason to think that the result should be the same). However, when the exercise is repeated for other pairs of questions, each having 2, 3 or 4 possible answers, the results are surprising. When the MI is written as a percentage of the maximum value for each pair of questions, the resulting lines on a graph of MI versus population in the associated group are very similar, as shown in Figure 12. This seems to hold even when the answers to the questions are not equally likely. Redoing the analysis for two binary questions with one answer to Q1 being four times as likely as the other (for both the associated and unassociated groups), the result (labeled 2x2* in Figure 12) still follows a similar relation.

These results suggest that MI can be compared between pairs of questions as long they are compared as percentages of the maximum possible value. This holds even if the questions have differing numbers of possible answers, or differing degrees of skewness towards one answer. The other important result is that seemingly small values of the normalized IO can be indicative of significant levels of association.

This page intentionally left blank.

List of symbols/abbreviations/acronyms/initialisms

CORA	Centre for Operational Research and Analysis
CEFCOM	Canadian Expeditionary Forces Command
DND	Department of National Defence
DRDC	Defence Research & Development Canada
IE	Information Entropy
IO	Information Overlap
ISAF	International Security Assistance Force
MI	Mutual Information
OR	Operational Research
QoI	Question of Interest
Q1	Question One
Q2	Question Two
R&D	Research & Development

This page intentionally left blank.

Distribution list

Document No.: DRDC CORA TM 2008-022

LIST PART 1: Internal Distribution

	DRDC-CORA
1 CD copy	DG DRDC CORA
1 CD copy	CSci CORA
1 CD copy	Section Head, Land and Operational Command OR
1 CD copy	Section Head, Pers & Social Sciences ORA
1 email	All CORA Team Leaders
4x Hard copies + 2x CD copies	Authors
1 CD copy	DRDKIM 3
	DRDC-Toronto
1 CD copy + email	Section Head Adversarial Intent, DRDC Toronto Carrol McCann 1133 Sheppard Ave W, PO Box 2000 Toronto ON, M3M 3B9 Canada Mccann.ca@forces.gc.ca
1 CD copy + email	Thinking, Risk and Intelligence Group Leader, DRDC Toronto Dave Mandel DRDC-Toronto 1133 Sheppard Ave. West, P.O. Box 2000 Toronto ON, M3M 3B9 Canada david.mandel@drdc-rddc.gc.ca mandel.d@forces.gc.ca
9x CD Copies 4x Hard Copies + emails	TOTAL LIST PART 1

LIST PART 2: External Distribution

1 CD copy	ADM(S&T) (for distribution)
1 CD copy	Director S&T IC (Paul Comeau)
1 CD copy	CISTI
1 CD copy	Library and Archives Canada
1 CD copy	Document Exchange Manager DSTO Research Library Defence Science & Technology Organisation PO Box 44 Pyrmont NSW 2009 AUSTRALIA
1 CD copy	Dr. Neville J Curtis Research Leader Land Operations Research 75 Labs Land Operations Division PO Box 1500 Edinburgh SA 5111 AUSTRALIA
1 CD copy	Michael Gillman (for dist'n and library) Chief Technologist Land Battlespace Systems Dstl Integrated Systems Room 31, Bldg A3, Fort Halstead Sevenoaks, Kent, UK, TN14 7BP
1 CD copy	Dr. Jason Field Land Battlespace Systems Dstl Integrated Systems Fort Halstead Sevenoaks, Kent, UK, TN147BP
1 CD copy	Director, US AMSAA ATTN: AMSRD-AMS-S 392 Hopkins Road APG, MD 21005-5071 USA
1 CD copy	Mr. Patrick O'Neill Chief, Combat Support Analysis Division USAMSAA (ATTN: AMSRD-AMS-S) 392 Hopkins Road APG, MD 21005-5071 USA

1 CD copy	Dr. James T. Treharne OCA Division Center for Army Analysis 6001 Goethals Road Fort Belvoir, VA 22060-5230
1 CD copy	Mr. Robert Barrett Chief, International Activities Center for Army Analysis 6001 Goethals Road Fort Belvoir, VA 22060-5230
1 CD copy	Mr. John Hughes HQ, TRADOC Analysis Center (TRAC) Programs & Resources Directorate (PRD) 255 Sedgwick Avenue Fort Leavenworth, Kansas 66027-2345
1 CD copy	Ms. Belinda Smeenk TNO Defence, Security and Safety Information and Operations P.O. Box 96864, 2509 JG The Hague, The Netherlands
1 CD copy	Mr. Bob Barbier TNO Defence, Security and Safety Information and Operations P.O. Box 96864, 2509 JG The Hague, The Netherlands
15 CD copies	TOTAL LIST PART 2
24x CD Copies 4x Hard Copies + emails	**TOTAL COPIES REQUIRED**

This page intentionally left blank.

DOCUMENT CONTROL DATA

(Security classification of title, body of abstract and indexing annotation must be entered when the overall document is classified)

1. ORIGINATOR (The name and address of the organization preparing the document. Organizations for whom the document was prepared, e.g. Centre sponsoring a contractor's report, or tasking agency, are entered in section 8.)

Defence R&D Canada - Centre for Operational Research and Analysis
101 Colonel By Drive
Ottawa, Ontario K1A 0K2

2. SECURITY CLASSIFICATION (Overall security classification of the document including special warning terms if applicable.)

UNCLASSIFIED

3. TITLE (The complete document title as indicated on the title page. Its classification should be indicated by the appropriate abbreviation (S, C or U) in parentheses after the title.)

Information Theory Based Measures of Association Applied to Opinion Polls:

4. AUTHORS (last name, followed by initials – ranks, titles, etc. not to be used)

Eles PT; Vincent E

5. DATE OF PUBLICATION (Month and year of publication of document.)

July 2008

6a. NO. OF PAGES (Total containing information, including Annexes, Appendices, etc.)

54

6b. NO. OF REFS (Total cited in document.)

13

7. DESCRIPTIVE NOTES (The category of the document, e.g. technical report, technical note or memorandum. If appropriate, enter the type of report, e.g. interim, progress, summary, annual or final. Give the inclusive dates when a specific reporting period is covered.)

Technical Memorandum

8. SPONSORING ACTIVITY (The name of the department project office or laboratory sponsoring the research and development – include address.)

9a. PROJECT OR GRANT NO. (If appropriate, the applicable research and development project or grant number under which the document was written. Please specify whether project or grant.)

9b. CONTRACT NO. (If appropriate, the applicable number under which the document was written.)

10a. ORIGINATOR'S DOCUMENT NUMBER (The official document number by which the document is identified by the originating activity. This number must be unique to this document.)

DRDC CORA TM 2008-022

10b. OTHER DOCUMENT NO(s). (Any other numbers which may be assigned this document either by the originator or by the sponsor.)

11. DOCUMENT AVAILABILITY (Any limitations on further dissemination of the document, other than those imposed by security classification.)

Unlimited

12. DOCUMENT ANNOUNCEMENT (Any limitation to the bibliographic announcement of this document. This will normally correspond to the Document Availability (11). However, where further distribution (beyond the audience specified in (11) is possible, a wider announcement audience may be selected.))

13. ABSTRACT (A brief and factual summary of the document. It may also appear elsewhere in the body of the document itself. It is highly desirable that the abstract of classified documents be unclassified. Each paragraph of the abstract shall begin with an indication of the security classification of the information in the paragraph (unless the document itself is unclassified) represented as (S), (C), (R), or (U). It is not necessary to include here abstracts in both official languages unless the text is bilingual.)

A methodology is described for analyzing opinion poll data and, in particular, for measuring associations between poll questions. The methodology is based on information theory and relies on calculating the information content of each question and the information shared between questions. In this context, shared information is defined as the amount of information revealed about an individual's response to one poll question once his/her response to another question is known. Information diagrams are introduced to facilitate interpretation of results. The paper focuses on defining a process for identifying the most significant associations between questions from polls that have a large number of questions. The key results of this work are: the definition of information theory based measures of association, the introduction of a step-by-step process for identifying relevant associations from large datasets, insights into the interpretation of calculated measures of association, a summary of lessons learned from practical application of the technique, and a demonstration of the technique with a real-world example taken from an opinion poll of the Afghan population in Kandahar Province commissioned by the Canadian Forces in February 2008.

Il s'agit de la description d'une méthodologie lors de l'analyse des données d'un sondage d'opinion, notamment pour mesurer les associations entre les questions du sondage. La méthodologie se base sur la théorie de l'information et repose sur le calcul de la quantité d'information de chaque question ainsi que sur l'information échangée entre les questions. Dans ce contexte-ci, l'information échangée se définit par la quantité d'information tirée de la réponse d'un individu à une question du sondage dès que sa réponse à une autre question est connue. La présentation d'information à l'aide de schémas est adoptée afin de faciliter l'interprétation des résultats. L'orientation de ce présent document porte sur l'élaboration d'un processus servant à la mesure des associations les plus pertinentes entre les questions de sondages qui contiennent un grand nombre de questions. Les résultats clés de cet ouvrage sont les suivants : la définition des mesures d'association basées sur la théorie de l'information; la mise en place d'un processus systématique pour relever des associations pertinentes provenant de grands ensembles de données; l'aperçu de l'interprétation des mesures calculées d'association; un résumé de leçons apprises à partir de l'application de la technique; la démonstration de la technique à l'aide d'un exemple concret provenant d'un sondage d'opinion effectué auprès de la population afghane de la province de Kandahar, réalisé à la demande des Forces canadiennes en février 2008.

14. KEYWORDS, DESCRIPTORS or IDENTIFIERS (Technically meaningful terms or short phrases that characterize a document and could be helpful in cataloguing the document. They should be selected so that no security classification is required. Identifiers, such as equipment model designation, trade name, military project code name, geographic location may also be included. If possible keywords should be selected from a published thesaurus, e.g. Thesaurus of Engineering and Scientific Terms (TEST) and that thesaurus identified. If it is not possible to select indexing terms which are Unclassified, the classification of each should be indicated as with the title.)

Poll; Survey; Analysis; Association; Measure of Association; Information Theory; Mutual Information

DRDC CORA

www.drdc-rddc.gc.ca