PNWD-3550

Methodology, Metrics and Measures for Testing and Evaluation of Intelligence Analysis Tools

Frank L. Greitzer

Glass Box

March, 2005

Pacific Northwest Division
Battelle Memorial Institute

THIS PAGE INTENTIONALLY BLANK

Contents

List of Figures

List of Tables

THIS PAGE INTENTIONALLY BLANK

Methodology, Metrics and Measures for Testing and Evaluation of Intelligence Analysis Tools

1. Introduction

The intelligence analysis (IA) professional is confronted each day with high demands for rapid, yet accurate assessments that require discovery and marshalling of evidence, integration and synthesis of data from disparate sources, interpreting and evaluating data and information that are constantly changing, and making recommendations or predictions in the face of inconsistent and incomplete data. Recognizing the difficulty of the IA task, stakeholders and the research community have been seeking technology-based solutions to reduce the analyst's workload and improve the throughput and quality of IA products. Research conducted by and for the intelligence community (IC), such as the Advanced Research and Development Activity's (ARDA) Novel Intelligence from Massive Data (NIMD) program, aims to develop tools for analysts that enhance such activities as information collection, hypothesis generation and tracking, integration of information from large data sets, and analysis/assessment of evidence bearing on alternative hypotheses. It is expected that such research conducted by leading scientists from academic and commercial R&D communities will yield products that, once deployed in operating IA facilities, will produce measurable performance improvements. A challenge for the research community is to develop useful and valid metrics and measures that may be used to assess the impact of software tools and products intended to improve IA performance.

Recent workshops and conferences supporting the IC have highlighted the need to characterize the difficulty or complexity of intelligence analysis (IA) tasks in order to facilitate assessments of the impact or effectiveness of IA tools that are being considered for introduction into the IC. Some fundamental issues are: (a) how to employ rigorous methodologies in evaluating tools, given a host of problems such as controlling for task difficulty, effects of time or learning, and small-sample size limitations; (b) how to measure the difficulty and complexity of IA tasks; and (c) how to develop more rigorous (summative), performance-based measures of human performance beyond the more traditional reliance on formative assessments (e.g., subjective ratings). This critical challenge must be addressed to ensure that tools and techniques introduced into the IA process are effective.

2. Background

An active area of research is the design and development of tools to improve IA. Evaluating the effectiveness of tools proposed for introduction into the IC requires the use of realistic IA tasks and an appropriate research methodology that controls for task difficulty. Typical evaluations, if they occur at all, do not employ proper experimental (or quasi-experimental) methodology and address only formative levels of evaluation (e.g., user ratings). Thus, the evaluation methodology must be at least sound and rigorous (defensible) if not "scientifically valid." It is difficult to conduct true experiments in our domain, but we should avoid logical and methodological pitfalls that would call the conclusions of such studies into question (e.g., due to factors that may "confound" the results). Quasi-experimental research methods may apply. One important ingredient of a sound methodology is to demonstrate some control over the independent variables—specifically, the nature/difficulty of the tasks used in the study.

Thus, greater control of task difficulty and performance based metrics are needed to assure more scientific, rigorous, and objective decisions. There is a need to characterize the difficulty or complexity of IA tasks. Some fundamental issues are: (a) how to employ rigorous methodologies in evaluating tools, given a host of problems such as controlling for task difficulty, effects of time or learning, small-sample size limitations; (b) how to measure the difficulty/complexity of IA tasks in order to establish valid experimental/quasi-experimental designs aimed to support evaluation of tools; and (c) development of more rigorous (summative), performance-based measures of human performance during the conduct of IA tasks, beyond the more traditional reliance on formative assessments (e.g., subjective ratings).

3. Objective

Three objectives of this research, and the accompanying report, are to:

- Examine possible research methodologies and designs that may be applied effectively and expediently to support tool evaluation and to identify issues and pitfalls in flawed designs that should be avoided.
- Discuss concepts and describe factors that should be considered in developing a useful set of task difficulty dimensions to support evaluation and testing of IA tools.
- Begin to address issues relating to behavioral, task-based performance measures that that are focused on specific areas of performance that a prospective tool is designed to impact.

Each of these three sets of issues and challenges are summarized in the major sections of this report.

4. Research Methodology

4.1 Overview of the Research Methodology Problem

Concerns about research methodology arise because of the constraints of testing proposed tools in real-world settings. It is difficult or impossible to design and carry out a controlled experiment with rigorous experimental and control groups. Therefore, there are risks to evaluation validity, such as the following:

- Are selected tasks representative of IA in general?
- Do we know enough about how to take task difficulty into account or to control this factor?
- Are the tasks sufficiently understood to enable evaluation of the quality of the product?
- Are there enough participants for the study?
- Do the participants represent the broader population of intelligence analysts?
- Are there a sufficient number of tasks available for study?
- In the design of the study, how can we account for or control learning effects (i.e., the fact that participants will gain experience with the tool being tested and therefore improve naturally as the study continues)?
- In the design of the study, how can we account for or control order effects? That is, we need to make sure that the manipulations or interventions introduced in the study do not all occur in the same order for all participants—this is related to the learning effect issue, above.

Challenges for designing valid and rigorous studies to assess the effectiveness of IA tools center on finding practical solutions to these questions. The problem with small sample sizes (due to budget limitations and practical limitations) applies both to the number of tasks selected for study as well as the number of analysts who are able to participate in the study. The effects of time/learning effects are particularly challenging given the small sample sizes that are available.

4.2 Scope of Discussion on Methodology, Metrics and Measures

There are three main types of evaluation approaches or paradigms for testing the effectiveness and acceptability of software tools: heuristic evaluation, observational/ethnographic, and user testing. Heuristic observations are performed by human factors/human-computer interface experts based on established or agreed-upon criteria such as described by Nielsen (1992; 1993). Observational studies involve recording of user activities while working with the product; recordings are made both manually and using audio/video recording equipment. Ethnographic studies are of this type but they are distinct in that the researcher becomes “embedded” in the environment and the culture of the group being observed. A notable ethnographic study of intelligence analysts was reported by Johnston (2003a). User testing is a preferred method of evaluating software tools because it examines user performance on tasks of interest in controlled settings, with objective performance data such as times to complete tasks or errors made, and with subjective evaluations by users in the form of questionnaires and interviews.

The focus of this report is on user testing. Further, the focus is specifically on assessing the impact of tools or products based on their effectiveness in meeting their intended design requirements and criteria for success, as identified by stakeholders and developers. Thus, although the assessment of usability features such as ease of use, user interface design issues, etc., is important to the ultimate success of a product, this area of human factors has provided a well-established methodology and practices that need not be described here, except insofar as they are affected by specific constraints imposed by user testing in the intelligence community. The focus here is on user testing.

The context of such testing is in realistic settings such as in actual operating intelligence analysis working groups or in similar environments. One such environment is Battelle's Colonial Place Operations (BCPO) facility, which is the home of the NIMD Glass Box analysis environment. Battelle Glass Box analysts have been operating within this test and evaluation environment for over two years. Another environment that is being used for testing and experimentation is at NIST facilities in Gaithersburg, MD, where experienced, reserve intelligence officers participate as part-time subjects in usability studies. Both of these test environments enable researchers to control the nature of tasks, time available to perform tasks, the tools available, amount of collaboration possible, and so forth. The Glass Box instrumentation, the tools being studied, and, in some cases, observations by evaluators and questionnaire/interviews provide data to support the usability studies. These environments enable the researchers to have a fair amount of control over many variables that affect performance, but by no means can they control all such variables as might be expected in laboratory experiments. A paper by Johnston (2003b) provides an appreciation of the kinds of variables that may affect intelligence analysis performance, including systemic variables (organizational, political), systematic variables (operations, information, reporting), idiosyncratic variables (psychological/cognitive limitations and biases, education, training, readiness), and communicative variables (formal and informal communications, technology for network access and collaboration).

A more formal test and evaluation setting will be available at a future time through the Research, Development and Engineering Center (RDEC). The RDEC is an operating IA facility that provides large numbers of analysts who may act as subjects in controlled experiments designed to test IA tools. The RDEC provides perhaps the only test environment in a realistic/operational IA setting in which experiments may be performed involving groups with sufficient sample sizes to conduct statistical tests with sufficient power to detect meaningful experimental effects. This enables experiments to be performed with sufficient controls, number/types of experimental/control groups, and numbers of subjects to overcome many of the methodological pitfalls that were described in Section 4.1. However, tools to support IA must be pre-tested and vetted sufficiently to establish their worthiness before they move on to the RDEC test bed. The scope of the present discussion is therefore largely focused on the less formal testing that is being done within the NIMD program at BCPO and NIST, or perhaps at individual NIMD contractor's facilities.

4.3 Mitigating Risks to Experimental Validity

To help deal with some of the issues in studies that are constrained to provide less than "true experimental control" of variables, a number of "quasi-experimental" research designs have been discussed (e.g., Campbell and Stanley, 1963). Some simple examples with varying degrees of scientific rigor are:

- The one-shot case study—a single group of participants is studied only once, subsequent to (or during) the "treatment" presumed to cause change (e.g., use of the IA tool). This methodology has a total absence of control and has almost no scientific value. Fortunately, the NIMD plans for evaluation do not follow this tenuous design.
- The one-group Pretest-Posttest Design—This design is diagrammed as: O1→ X → O2 where O1 is the pretest, O2 is the posttest, and X is the experimental intervention. In this design, time between observations is a threat to internal validity because there is no control or accounting for events occurring between the pretest and posttest. This is close to the design for evaluation that has been in practice, and that is planned for future NIMD tool evaluations.
- The posttest-only control group design—This is diagrammed as:
  R X → O2
  R X → O1,
  where *R* indicates random assignment to the experimental and control groups. This is a valid experimental design. However, it requires a sample of subjects that can be divided into two groups (experimental and control groups), which is not generally possible for the NIMD evaluations. It is a preferable experimental design that should be used if/when NIMD tools are introduced for evaluation into the RDEC.

Thus, for NIMD tool evaluations, we are constrained by budget limitations and practical limits on the number of subjects to use a *within-subjects pretest-posttest* design. The performance of each analyst in the study should be compared with himself or herself. Examples of viable manipulations are:

Example 1:

1. *Pretest*: Analyst performs analysis without tool (as part of baseline period, for example)
2. *Tool Introduction and Training*: Analyst receives training and has time to become familiar, if not proficient, with the tool.
3. *Posttest*: Analyst performs analysis with the tool.

An alternative to the Pretest activity above is for the analyst to perform the analysis with the NIMD tool, but without a particular function enabled. This allows the effectiveness of the function to be assessed.

Example 2:

1. *Pretest*: Analyst performs analysis without tool (as part of baseline period, for example)
2. *Tool Introduction and Training*: Analyst receives training and has time to become familiar, if not proficient, with the tool.
3. *Posttest 1*: Analyst performs analysis with the tool
4. *Posttest 2*: Analyst performs analysis with the tool, but without a specific function that is being tested.

In example 2, it would be prudent to randomly order steps 3 and 4; that is, flip a coin to decide whether an analyst uses the complete functionality of the tool in step 3 and reduced functionality in step 4, versus using the reduced functionality in step 3 and complete functionality in step 4. This randomization helps to combat biasing results due to order effects.

The preceding examples and discussion do not address a significant risk to validity in this experimental design: confounding of results due to the uncontrolled effects of task difficulty and/or to uncontrolled effects of individual (analyst) differences. The problem with task difficulty is that it is impossible to give an analyst the identical task to perform under the pre- and posttest conditions, since once having performed the task the analyst and the problem will not be the same. The only solution for a within-subjects design is to be able to assign tasks to pre- and posttest conditions that are generally "equivalent" in difficulty.[1] This challenge of characterizing, and therefore controlling, task difficulty is addressed in the next section.

Analyst individual differences are less of a problem for the design of the research than they are a problem for interpreting the results (e.g., the generality of the conclusions). If it is possible to use analysts with different levels of experience on a problem, it will be easier to compare the utility of a tool for analysts with different "readiness" levels. However, managing analyst experience as an *experimental* variable would require a larger sample of analysts to employ in the studies than is currently available. This question should be addressed when larger samples of experimental subjects are available, as in the RDEC environment.

5. Task Difficulty[2]

Why is there a concern about understanding how to characterize or measure task difficulty? The issue is fundamental to experimental methodology. Task difficulty metrics are needed so that we can have some confidence that tasks (problems) chosen to be worked on during the evaluation are comparable. A scientifically valid evaluation can be done only if we are able to "control" task difficulty as we study the impact of proposed tools. The challenge derives from the fact that it is impossible to use the same task for both the experimental and control conditions, particularly if the experimental and control conditions are both applied to each subject (a within-subjects design). It is obviously not possible to "erase" the analyst's memory and start fresh on a task if it has already been worked on in a prior experimental condition. This issue is not as serious if the experiment uses a "between-subjects" design, but task difficulty still needs to be understood and controlled if the experiment aims to apply to a range of IA problems that vary in type or difficulty.

[1] If a *between-groups* experimental design is used in which one group of analysts works on a task with the tool being tested (experimental group) and the other group works on the same task without the tool (control group), the confounding effects of task difficulty are eliminated. This requires more subjects than would be available on a practical basis. If this were possible, then care should be taken to randomly assign analysts to experimental and control groups.

[2] Material in this section has been submitted for publication (Greitzer, 2005).

5.1 Informal Conceptualizations

There are many informal characterizations of what makes IA hard, including the oft-cited "information overload" problem, but the problem is much more complex. The information overload problem is generally interpreted in terms of "too much data." However, it is equally common for analysts to struggle with "too little data." Moreover, the *quantity* of data, per se, would not seem to underlie the problem so much as the problems inherent in the data, such as consistency, reliability, heterogeneity. Thus, for example, it is argued that a massive data set that tends to be consistent and homogeneous in its content or interpretation would not pose as difficult a problem as a much smaller data set that lacks consistency and homogeneity. While heterogeneity is often described in terms of incompatible data formats, other possible manifestations of heterogeneity are even more challenging, such as the need to tie together multiple "threads" of evidence from disparate domains (financial, political, military, law enforcement, etc.) to produce a more complete picture that we call "situational awareness." A careful analysis and study of task difficulty dimensions will lead to a better understanding of these factors and their interdependencies.

5.1.1 An Initial Set

At a recent "Friends of the Intelligence Community" (FOIC) workshop, Bonnie Wilkinson (2004) presented some views on the dimensions of difficulty for IA tasks that provide an excellent summary of how the IC (informally) views task difficulty and associated performance challenges faced by IA professionals. Wilkinson described the following dimensions of difficulty: predicting the future, human behavior, low observability, lack of physical/hard data, high data ambiguity, low confidence in sources, lack of specificity, multiple data times, multiple subjects, too many variables, many organizations, and insufficient time. In an unpublished document, I examined these in a bit more detail (Greitzer, 2004), and the main points are repeated here.

Predicting the Future is an excellent characterization of one dimension of task difficulty. I suggested that the underlying dimension might be described in terms of **Characterization versus Prediction**, where characterization focuses on developing biographical profiles, company/country capability or science/technology profiles and the like; while prediction focuses on "what-if" analyses about hypothetical actions. (As an aside, it is generally considered that prediction is a more difficult type of task than characterization).

Human Behavior refers to the difficulty of reporting on the thoughts, motives, inclinations, or possible actions of individuals. I suggested that an underlying dimension might be described as **Sociological Complexity**, where the focus of the analysis might be an individual, at one end of the dimension, and a group, a State, or a region at the other end of the dimension.

The next several dimensions offered by Wilkinson are related to various aspects of the data, which I suggest might be combined into a general dimension termed **Data Uncertainty**: *Low Observability*, *Lack of Physical/Hard Data*, *Data Ambiguity*, *Low Confidence in Sources*, *Data Specificity*, and *Multiple Data Times*. I have interpreted these to mean, for example, that behavior or characteristics being studied are not easily observed or that they are difficult to interpret—which could arise from a lack of data, from ambiguous, deceptive, or unreliable data, or because the data are dynamic (changing over time). This dimension is arguably the most difficult to characterize and to operationally define; it is possible that further study would lead us to separate this aggregated dimension into two or more dimensions. To the extent that it is difficult to interpret or discriminate among these aspects of the data in operational/practical terms, we will retain this simplified "uncertainty" dimension for expediency.

The next three difficulty dimensions suggested by Wilkinson are *Multiple Subjects*, *Too Many Variables*, and *Too Many Organizations*. These characteristics seem to relate to a dimension reflecting the extent to which the analysis topic is narrowly focused versus broad and open-ended. For this reason I suggest the use of the term **Breadth of Topic** to encompass these factors.

Finally, the last dimension offered by Wilkinson is *Time*. To be sure, the amount of time available to conduct the analysis is a significant determinant of the difficulty in carrying out the task. This has been observed in earlier experimental research (Patterson, Roth, and Woods, 2001) and cognitive task analyses (e.g., Hutchins, Pirolli and Card, in press). Note that the time factor seems different from all the others mentioned above because it represents a variable that can be *manipulated directly and independently* in an experimental situation (i.e., one can control the time pressure by setting the deadline for the product). Despite this distinction from the other factors, the time factor (perhaps more appropriately called **Time Pressure**) is included among the list of difficulty dimensions. It is a valid experimental factor that is available for experimental manipulation (and it is being used currently in experiments conducted within the NIMD program).

Another observation provided by Bonnie Wilkinson is that IA is difficult because intelligence assessments can change the future, and because there is no opportunity for immediate feedback on predictions about actions that haven't yet occurred. These are valid examples of what makes IA difficult and stressful, but beyond using the Prediction dimension that has already been described, there does not seem to be any further need to define an additional task difficulty dimension in terms of lack of feedback. However, the feedback issue is certainly applicable to *performance* metrics – reflecting the extent to which the IA product was accurate or correct. Performance metrics are a necessary and important topic in their own right, but beyond the scope of this paper.

5.1.2 Some Additional Factors

Data Availability. In the spring of 2004, a number of IC professionals and researchers began an idea exchange on task difficulty concepts as part of an ARDA Metrics Challenge workshop. John Bodnar (2004) provided a preliminary assessment of his perspective on task difficulty: "The degree of difficulty in assessing any WMD program (or indeed any problem) is related mainly to the data available." He suggested that one way to assess task difficulty is to compare the amount of data that is potentially "out there" on the topic with the actual amount of data that is realistically available (i.e., possible to obtain or perhaps already obtained). Bodnar's basic idea and approach to characterization of the data set suggests a **Data Availability** dimension.

Problem Structure. In preparation for the ARDA Metrics Challenge study, further collaboration occurred in the spring of 2004 between Jean Scholtz and Emile Morse (NIST), Tom Hewett (Drexel University), and the author (PNNL). This produced a questionnaire that was used for studying task difficulty in problems used for NIMD research (Glass Box analysis taskings) and for research on question answering methods and tools being conducted by ARDA's AQUAINT (Advanced Question Answering for Intelligence) program. Most of the proposed set of dimensions described above (and in Greitzer, 2004) were translated into a number of Likert scale questions. An additional factor was added: **Problem Structure** (extent to which the problem is highly structured with a clearly defined objective, compared to the case in which the problem is ill-structured and requires the analyst to impose a structure). In addition, some supporting items were included in the questionnaire to collect information about the IA assignment, the product requested, and some analyst demographic variables. A description of the questionnaire and some preliminary results is provided in Hewett and Scholtz (2004).

Data Synthesis. Another dimension that seems to be a factor in IA task difficulty is the need to synthesize multiple sources of information, also referred to as data fusion. We will adopt this as a proposed **Data Synthesis** dimension. As Hutchins et al. (in press) observe, data synthesis is particularly problematic when multiple sources of disparate types of data are involved, when different pieces of data have varying degrees of validity and reliability, and when different types of domain expertise are needed to analyze each type of data. One of these task difficulty influences (data validity/reliability) is already addressed in the proposed Data Uncertainty dimension; another (analyst expertise) may be best represented by factors associated with analyst variables rather than task difficulty dimensions (and thus it may be considered outside the scope of the present discussion).

5.1.3 Problem Complexity

A concept missing from the above discussion is the notion of problem complexity. Indeed, this notion has not received much attention in discussions about task difficulty among IC researchers—even though it is fundamental to understanding the IA process, developing tools to support it, and defining metrics for task difficulty and performance effectiveness. The task difficulty concept relates to the mental processes involved in problem solving, which, despite a long history of study in psychology, still eludes our rigorous understanding. While we have considered the dimension of Problem Structure to distinguish between well-defined problems and ill-specified problems, we have not focused sufficiently on the mental activity that makes up the analysis process itself. As Heuer (1999, p. 31) observes: "Intelligence analysts should be self-conscious about their reasoning process. They should think about *how* they make judgments and reach conclusions, not just about the judgments and conclusions themselves." Most psychological research on problem-solving has been focused on well-defined problems: those for which we know a solution exists, and that we will recognize the solution when we find it (for example, we know when we solve a puzzle or prove a theorem). Ill-specified problems lack such tests because there are no criteria for "the correct answer" in these problems.[3] Real-world problems, including most IA tasks, are largely ill-specified. Nevertheless, we can gain some insight into such activities by considering what has been learned about problem solving with well-defined problems. Psychological research shows that successful problem solving, particularly for well-defined problems, is characterized by two principles: it must be hierarchical, and it must be goal-directed (e.g., Newell and Simon, 1972). "Hierarchical" means that complex problems must be decomposed into sub-problems until each sub-problem becomes simple enough to be solved—today this is referred to as "decomposition" (e.g., Heuer, 1999). "Goal directed" means, for example, that the process is guided by heuristic principles that concentrate the search on promising regions of the problem space to avoid getting bogged down pursuing blind alleys. For the purposes of defining task difficulty metrics, we can distill notions from the problem-solving research literature and, as taught by Frank Hughes at the Joint Military Intelligence College, from philosophers and thinkers in the legal field, to describe a dimension that reflects the complexity of the analytical problem. In modern and IA relevant terms, we are concerned about how many possible hypotheses must be considered in carrying out the IA task, and how much evidence must be gathered to "pull the threads" in pursuing answers and resolving questions about the status of these hypotheses. We must be concerned about the total number of such threads to follow, and the level of reasoning that is required to reach a conclusion. These are the ingredients of a dimension that reflects the complexity of the analytical problem. Frank Hughes observes that the *analysis* activity has received insufficient attention:

> For many years, the intelligence collection process has best been described as the task of trying to collect *everything* with the hope of finding *something*. This accounts in part for the sheer volume of information being gathered by the many agencies in the Intelligence Community. But throwing massive amounts of information at an intelligence analysis problem will not, by itself, solve this problem. What is so frequently left out of the equation is the process by which the information is analyzed….
>
> Any intelligence analysis task involves three major ingredients that must be generated or discovered by an analyst: hypotheses (possible explanations, predictions, or conclusions), evidence, and arguments linking evidence and hypotheses. (Hughes and Schum, 2003).

Acknowledging, then, that a **Problem-Complexity** metric would be useful, what sort of computational approach might apply? One possible approach could be based on the application of inference networks in the context of evidence marshalling and analysis (as provided in the teachings of Frank Hughes and David Schum, for example: Schum, 1999; Wigmore, 1937; Anderson and Twining, 1999; Tillers and Schum, 1988). Inference networks may be represented graphically in diagrams using nodes and links between nodes that represent propositions. Therefore, one possible measure of complexity could be based on the number of nodes or perhaps on the pattern of links between the nodes of the network (Tillers, 2004). An illustration is shown in Figure 5-1.

[3] This is reminiscent of an age-old problem stated in the dialogues of Plato over 2000 years ago: "And how will you enquire, Socrates, into that which you do not know? What will you put forth as the subject of enquiry? And if you find what you want, how will you ever know that this is the thing which you did not know?" (Plato)

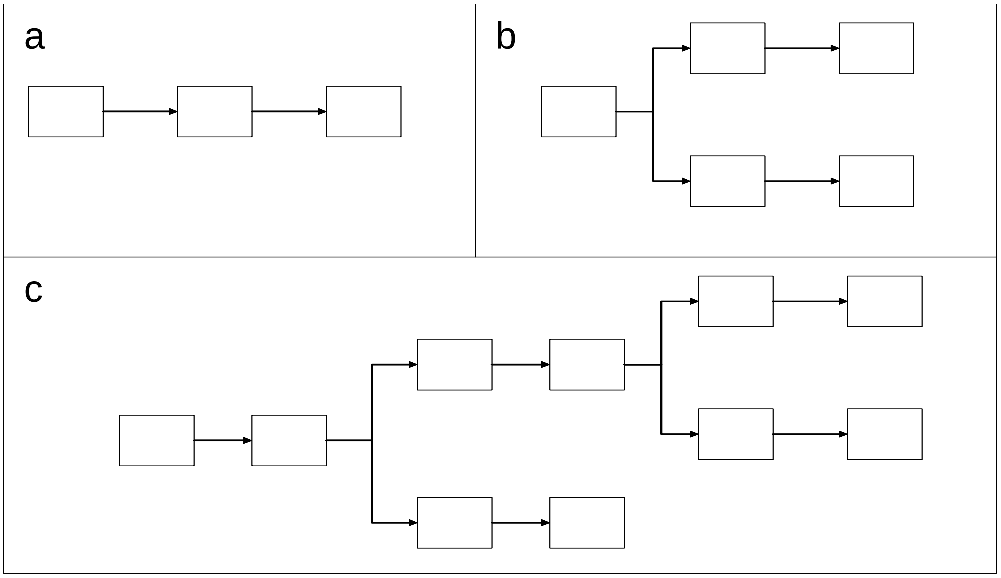


Figure 5-1. Graphical Representations of Three Examples of Inference Networks. Problem or inference complexity might be characterized by the number of links and/or nodes. (a) 3 nodes; (b) 5 nodes; (c) 10 nodes. Other, more complex, non-hierarchical inference networks may also be considered.

Of course, use of such a measure (or possibly other measures of complexity for inference networks) requires that the solution has already been produced. This is acceptable for after-the-fact measures, but less useful when attempting to select tasks that are comparable in difficulty to control experimental variables. For such cases, it would be advisable to have expert analysts solve the problems first so such measures can be obtained before the tasks are used in an experiment.

There are a number of other possible factors that have been implicated in the issue of "what makes IA hard" that have not been explicitly mentioned above, but that are implicit or related by association to those dimensions that have already been described. High on this list is what is commonly referred to as "high cognitive workload" or "information overload." It is suggested that high workload can be understood as a result of introducing one or more of the task difficulty dimensions already described. For example, dealing with an overwhelming stream of information *produces* high cognitive workload through its imposition of time pressure along with other dimensions such as data uncertainty, data synthesis, and problem complexity. Information overload, similarly can be described in terms of one or more dimensions such as time pressure, data availability, and data synthesis.

5.2 Working Set of Task Difficulty Dimensions

The dimensions and concepts discussed in the previous section are summarized in Table 5-1, which provides a recommended set of Task Difficulty factors. The table is divided into two sections: Better-defined factors (i.e., they have operational definitions and have been applied in recent research, e.g., supporting the NIMD and Metrics Challenge programs); and factors that are less well-defined (they have not been applied using operational definitions, and they require considerably more research to develop appropriate metrics and means to measure them). The last three factors shown are members of this second category of dimensions that need a considerable amount of further study to develop measurable metrics and/or to further refine the concept.

As a preliminary assessment of the adequacy of at least some task difficulty dimensions, a questionnaire incorporating some of the factors discussed here was administered to working analysts (Hewett and Scholtz, 2004). The questionnaire included five of the factors from the first section of Table 5-1 and six additional factors, plus an overall task difficulty factor. They obtained a correlation of 0.85 between the average difficulty *ranks* of each task over the eleven factors with the task's average overall difficulty rating. While this result is encouraging, there were a limited number of analysts and, importantly, the set of IA tasks that made up the survey comprised problems that were similar with respect to several of the proposed difficulty dimensions (e.g., Characterization vs. Prediction, Sociological Complexity, Breadth of Topic, Problem Structure; and Data Synthesis).

Table 5-1. Recommended Set of IA Task Difficulty Factors

Factor	**Description**
Factors That Have Been Defined at Least Operationally	
Characterization vs. Prediction	Does the task require a description of current capabilities (situation assessment) or does it ask for a prediction about future capabilities or actions?
Sociological Complexity	Does the task focus on the behavior of an individual, an organization, a group, an industry, a State, or a region?
Time Pressure	Number of hours/days available to the analyst to conduct the task.
Breadth of Topic	How narrow/focused versus broad/open-ended is the topic? Does it deal with many variables or a few?
Data Availability	Relative amount of data that is realistically/practically accessible to the analyst compared to the total amount of data that is considered "potentially" available.
Problem Structure	The extent to which the problem is highly structured with a clearly defined objective, compared to the case in which the problem is ill-structured and requires the analyst to impose a structure
Factors That Are Less Well-Defined (Requiring Further R&D)	
Data Uncertainty	A general (ill-defined as yet) construct that reflects the heterogeneity, consistency, number and reliability of sources, and confidence in the data.
Data Synthesis	The extent to which data must be integrated and analyzed from multiple sources of disparate types of data (including source types and data formats challenge data ingest capabilities).
Problem Complexity	A measure of complexity of the reasoning process used in obtaining the solution. Research on measures of complexity of inference networks may be applicable.

5.3 Speculation about Further Research on Task Difficulty

At this early stage of research, we do not have a deep understanding of the factors or dimensions that underlie IA task difficulty, but the concepts described here and listed in

Table 5-1 represent a start. One area for research is to conduct additional studies with a greater number of participants and a broader range of problems, with a goal of further refining the initial set, to determine if any factors have been omitted and to assess the extent to which the existing factors are independent. The extent to which these factors are *predictive* and independent is not known: Additional research is needed to apply statistical analyses to this and related questions. Some questions to address through questionnaire and performance studies are:

- As has been suggested, the study by Hewett and Scholtz (2004) should be replicated in a follow-up study with a more diverse set of IA tasks. Also, the factors addressed in the questionnaire should be "tightened up" based on arguments that I have made above. For example, Hewett and Scholtz included a question on analyst experience; while this is a relevant "demographic" or *analyst* variable, it is not a *task* variable and does not belong in a list of task difficulty dimensions[4]. A question to address in a follow-up questionnaire study would then be: Looking at the nine possible difficulty factors in Table 1, to what extent do these factors "predict" an overall task difficulty judgment for IA tasks?
- Another question that applies to analysis of questionnaire data is: Would a multiple regression analysis reveal dependencies in the proposed dimensions? Would it reveal stronger correlations (regressions) for some of these factors on overall task difficulty than for other factors? (Which ones are the strongest determinants of difficulty?)
- Would it be possible to extract/derive sufficient meaning from task descriptions to enable them to be "sorted" or "valued" along any of the proposed dimensions? Dimensions that would most likely be amenable to such analysis are: Characterization vs. Prediction, Sociological Complexity, Time Pressure, Breadth of Topic, and Problem Structure. It is possible that key words and semantic/ontological relationships could be used to identify these factors. It is noted that the other factors depend to a large extent on the nature and availability of the data or on the complexity of the solution and possible alternatives—information that would not be known without actually conducting the analyses.

Beyond the use of survey-based studies, it would be desirable to conduct a study and an analysis of *behavioral* data to see if one can find correlates of the proposed task difficulty variables in data collected during IA activities. For example, data collected using the Glass Box instrumentation has been used to examine behavioral correlates, such as an analysis of dwell times to support inferences about the analyst's interest (Sanquist, Greitzer, Slavich, Littlefield, Littlefield, and Cowley, 2004). The analysis of behavioral data may serve to validate, disambiguate, or further refine the proposed dimensions into a more useful set that can be used to guide further research. Some speculative research questions about such behavioral correlates of proposed task difficulty dimensions are:

[4] Experience certainly affects performance—but this does not mean that it should be incorporated into difficulty metrics. A more familiar example may clarify the distinction: Readability measures, such as the Flesch Reading Ease or Flesch Kincaid Grade Level (e.g., Flesch, 1994), are based entirely on the content of the material (number of words per sentence and number of syllables per word) and are meant to be independent of an individual's reading ability.

- *On data uncertainty*: Would it be useful to collect data from (Glass Box) analysts on their level of confidence in evidence that they collect? Current Glass Box functionality enables analysts to easily indicate "relevance" of material, and a confidence measure could be implemented in the same way. Potential advantages of this behavioral measure are that it may shed light on the extent of uncertainty in the data – i.e., the Data Uncertainty dimension.

- *On data synthesis*: As described in Table 5-1, this factor generally refers to the multiple data formats and types that may need to be digested and analyzed. It is likely that expert analysts can provide a reasonable assessment of the extent to which this factor plays a role in a given IA task merely by responding to a questionnaire item, similar to those that we created for the Metrics Challenge project (Hewett and Scholtz, 2004). The research community has had an opportunity to observe analysts struggling with the synthesis and interpretation of open source data or data that have been made available through artificial sources, which is a formidable problem in itself; the data synthesis problem is compounded in classified environments in which data are collected from even more diverse sources, with varying and incompatible formats. Data synthesis presents a significant challenge for both open source and all-source analysts, and a difficulty metric would be useful in evaluating the impact of data ingest tools.[5]

- *On problem complexity*: As described in the discussion of problem solving and evidential reasoning, it is well-established through behavioral research and common sense that one of the most effective ways to deal with complexity is to "divide and conquer" – in other words, decomposition of the problem into simpler parts. It is almost a "given" that analysts do this, whether or not it is done consciously. The question is: Can such strategies be observed in (actually, inferred from) the data collected through such means as the Glass Box instrumentation? On one level, it is possible to see evidence of task decomposition by looking at sub-tasks that analysts are free to create for themselves while planning and conducting the analysis. Similarly, certain organizational artifacts may be evident in the file system structure that analysts set up and use on their computers. On a more challenging (and indirect) level of analysis, can such decompositions be inferred from queries entered in the Internet browser? In any of these cases, such decompositions could be used to estimate the complexity of the analysis task—not as direct a measure as would be obtained by analyzing an inference network, but perhaps more expedient since analysts do not explicitly create such networks to support their reasoning process.

[5] It has been observed that existing visualization and analysis tools—designed to support the collection and synthesis of such data—do not deal adequately with the ingest of such data. There is a critical need for such visualization and analysis tools to provide a more usable front-end that analysts can use, with only a limited amount of training, to support ingest of such data (e.g., Badalamente and Greitzer, 2005).

6. Performance Measures

A third challenge in evaluating the impact of IA tools concerns the need for performance measures. Performance measures are measured quantities used to compare performance with and without tools. Performance measures are needed to answer questions such as: Does tool *X* improve the throughput of analytic tasks of type *Y*? Does it yield more efficient or higher quality output for certain types of tasks, or for certain "phases" of analysis? In considering possible impacts of tools and technologies, it is important to consider not only the effects on the collection, analysis, and production processes, but also on the *vetting* process (e.g., will a proposed tool/technology make it easier or quicker for other IA professionals and clients to assess and interpret the IA product?).

Performance measures are usually interpreted in terms of usability. There are several sources of established guidelines for usability testing (e.g., Nielsen, 1993). Commonly used criteria include efficiency, learnability, and memorability. Usability measures address the experience of users—whether or not they found the tool useful, easy to learn, easy to use, and so forth. Often, users are asked to provide this sort of feedback using qualitative measures obtained through verbal ("out loud") protocols and/or post-hoc comments (via questionnaires, interviews, ratings). Additional usability criteria address what might be called the utility of the tool—how effective was the tool in supporting the user's needs? Quantitative measures that assess utility include efficiency in completing the task (time, accuracy, completeness). These will be most useful in comparing the utility of alternative tools or assessing the utility of a given tool versus baseline performance without the tool.

6.1 Measures and Metrics

In Section 5 we discussed the need to characterize task difficulty and offered some conceptualizations of metrics that could be used. In this section we discuss measures. Before going further, we should be sure that there is a common understanding of the difference between metrics and measures in this context. We have been using the term "metrics" to indicate a standard of measurement. Consider the concept of distance: in mathematics, the Euclidean metric is but one of a number of ways to conceptualize distance. In our daily lives, we often use the concept of time to characterize distance—as in the example: "The hotel is only a 10 minute drive from the airport." Metrics are chosen for their relevance in characterizing the quantity or quality of interest. Units of measure are associated with specific methods of determining these quantities or qualities—for example, miles represent a unit of measure for distance.

6.2 Some Qualitative Measures

Measures of effectiveness come in several varieties, but all are aimed at assessing the impact of the tool. User satisfaction is a necessary, but not sufficient measure. Overall quality of the output is a useful measure—the AQUAINT Metrics Challenge project used a creative ranking method to measure product quality that takes advantage of opinions from multiple judges to help improve the objectivity of the measure.

We can measure the quality of the output of a process, a tool, or an entire effort. For IA products, the customer can assess the quality of a report; if reports are produced using different methods or tools, then these reports can be assessed individually or they can be ranked to compare quality. NIST used such a method by asking analysts to rate the quality of reports generated by different analysts working on the same problem, using various tools that were being evaluated (Hewett and Scholtz, 2004).

When NIMD products are being evaluated within the Glass Box environment, users are able (and should be encouraged) to provide qualitative ratings at any level they wish. The Glass Box annotation function allows the analyst to make a quick annotation on the quality of any aspect of a tool that is deemed noteworthy. An advantage of using this feature (over the alternative of waiting until the task or day is finished and filling out a general-purpose questionnaire at that time) is that the analyst who waits to make such judgments is more likely to forget about such fine details altogether.

Table 6-1 lists some typical qualitative measures.

Table 6-1. Typical Qualitative Measures of Effectiveness

Measure	**Example**
Quality of Product • Points or grades assigned to products by experts (or users) • Products are ranked by experts (or users)	• Experts judge quality of a report on a 1-10 scale • User ranks quality of hypotheses delivered by a hypothesis generation tool
Confidence • User confidence in findings	• User rates confidence in result (e.g., relevance rating for documents offered/returned by tool)
Cognitive workload • Difficulty assessment ratings • Cognitive workload ratings (NASA Task Load Index, TLX; Hart and Staveland, 1988)	• User rates difficulty of performing a task using a tool • User ranks the task on set of criteria established for the NASA TLX
*Standard usability measures** • Efficiency • Learnability • User control • Consistency • Error prevention • Feedback	• These may be addressed qualitatively using questionnaire items. Some may also be addressed quantitatively… see Section 6.3.

*See Nielsen (1992) for a more complete list of measures used in heuristic evaluations.

6.3 Quantitative Performance Measures

Usability guidelines that suggest criteria such as efficiency, learnability, memorability, preventing errors, etc. (e.g., Nielsen, 1992; 1993) can be applied in a general way, such as indicated in Table 6-1 above, to assess the overall experience of users. More objective, performance-based measures represent an important, and certainly a more challenging, means of assessing the impact of tools on performance. To do this effectively, a close collaboration is needed between the tool developers and cognitive scientists/human factors researchers. For while general usability guidelines suffice for the general-purpose usability testing that is prevalent in the field of product design and application development, the level of feedback that is gained may not be sufficient to meet the specific goal of the IC to deploy effective tools that will have an impact and that will be used.

A more detailed evaluation is necessary to allow IC stakeholders to answer the question, "Should the investment be made to deploy this tool in the intelligence community?" For regardless of the monetary investment that went in to the development of a prospective tool, the cost of training, maintaining, and sustaining the tool are substantial future costs that should be avoided if it is determined that the tool does not meet the needs of its users. The only effective way to assess this question is to conduct more specialized evaluations that are based on explicit performance criteria and requirements. These should be obtained from the stakeholders who funded the development and from the developers, who are in the best position to identify specific functions and features that they expect to positively impact IA performance. To stimulate the specification of these detailed criteria, one should ask:

- What aspects/phases of the IA process are most affected by the application of the tool?
- How will the tool affect this performance? These effects should be manifested in measures such as time to perform an activity, accuracy of the result, completeness of the result, etc.
- What data can the tool provide that are not available through any other means? There is the concept of measures obtained "outside" the tool (such as measures provided by examining Glass Box data) versus measures that are best provided by the tool, given that the tool captures and logs such data. If we're trying to assess the effectiveness of a tool in enhancing an analyst's cognitive activities, then in a real sense we must rely on these same tools to provide relevant data—for such data may not exist anywhere else.[6]

6.3.1 Examples of Quantitative Measures

Table 6-2 shows some examples of possible performance measures that may be obtained if the tool developer makes plans in advance of the evaluation studies to record such data. In some cases, the Glass Box instrumentation may also be able to obtain such data—the likely

[6] For example, suppose that a tool helps the analyst determine relationships among three events that are part of a large collection of data. Analysis activities that the tool helps the analyst perform, such as manipulating representations, examining elements of events, comparing facts among different events, etc., are expressly or indirectly observable by the tool itself, but most likely not amenable to monitoring or recording by any other software (however, the Glass Box logging function provides a convenient means of storing system data provided by tools that are integrated with the Glass Box).

source of the data is indicated in columns two and three. Some measures must be determined by a human, as shown in column four. In several cases, the notion of "target criteria items" is used to represent items that are identified as success criteria for the products being evaluated. For example, a product designed to generate/track hypotheses should be evaluated in part on the basis of the number of "good" hypotheses that are generated compared with a baseline, and in particular compared with what experts consider to be the most relevant and important hypotheses. Similarly, products that support evidence marshalling should be evaluated in terms of the number of appropriate documents, URLs, etc. that they yield in comparison with baseline data and expert judgments. Other variations on such measures should also include the number of target items missed and the number of useful/relevant items produced that were not considered *a priori* (this reflects the ability of the tool to facilitate *novel* intelligence). Other measurement concepts relate to temporal factors that reflect efficiency and effort.

Table 6-2. Examples of Quantitative Performance Measures

Examples of Measures	**Tool Log**	**Glass Box**	**Human/ Manual**
Quality—measures of Accuracy			
Comparison with "expert" solution			√
Percent agreement between system and analyst			√
Product correctness compared to established criteria			√
Amount of evidence used in analysis	√		
Number of target criteria items* considered by system	√		
Number of target criteria items missed by system			√
Number of new/novel items produced**			
Does analyst *choose* to use tool when freely available?	√	√	√
Temporal—Measures of Efficiency			
Time spent solving problem	√		√
Time spent with tool compared with other applications			√
Workload—Measures of Effort			
Number of queries made	√	√	
Number of links examined	√	√	
Depth of links examined	√	√	
Number of documents read (accessed)		√	
Number of times each document was accessed		√	
Number of steps needed to perform a function	√		
Rate of growth of the draft report document		√	

* Target criteria items, as described in the text, include variables, hypotheses, documents found, etc. that experts identify as relevant and important data that should be produced—i.e., success criteria.

**Novel items, as described in the text, are data or information that was not identified *a priori* as target criteria items.

Most of the examples in Table 6-2 reflect relatively high-level performance measures. Exceptions are measures of effort such as depth of links examined, number of steps needed to perform a function, and rate of growth of a report document. These measures are more difficult to obtain, but may be worth the effort because of their greater potential for illuminating the extent to which an IA tool is effective in augmenting the analyst's cognitive process. Therefore, at the risk of providing more information or detail than some readers would welcome, consider another example and a description of a more detailed elaboration of performance criteria that could be used to evaluate IA tools.

Hughes and Schum (2003) have carefully described the construction of an argument, which is a chain of reasoning that connects evidence to hypotheses of interest in the analysis. Figure 6-1, adapted from Hughes and Schum (2003), shows but one of many chains of evidence in an inference network that would represent the thought process behind an IA product. As described by Hughes and Schum, links may be characterized in terms of uncertainty about the credibility of the evidence. Reasoning from one link to another is justified by generalizations that provide rationale for such reasoning, and evidence used in the argument may be directly relevant or indirectly relevant (ancillary—i.e., not directly relevant but that can be inferred). Hughes and Schum observe that "generalizations and ancillary evidence supply the 'glue' that holds our arguments together."

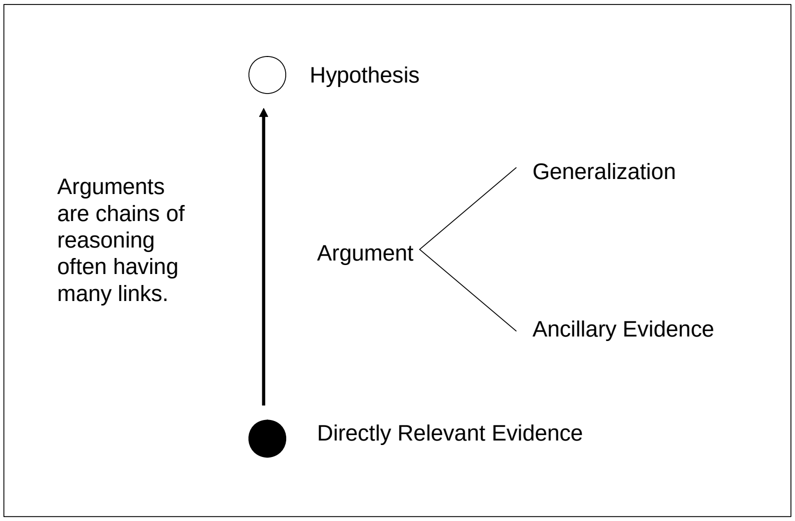

Hypothesis
Arguments are chains of reasoning often having many links.
Generalization
Argument
Ancillary Evidence
Directly Relevant Evidence

Figure 6-1. Argument Represented as a Chain of Reasoning. (After Hughes and Schum, 2003).

Consider an IA tool that is designed to help the analyst identify hypotheses and marshal evidence that will ultimately form the basis of the analytic product.[7] A detailed analysis of data collected during the IA task should be able to identify hypotheses, directly relevant evidence, and ancillary evidence that are used in the argument. (Generalizations may not be directly observable in the data, as they may result from tacit knowledge.) For the observable data, consider possible "outcomes" that may be extracted from the data to

[7] The arguments described here may be applied to any basic "elements" of an IA product that can be identified by "experts" as important, necessary, critical, etc. Hypotheses and evidence are used here for illustrative purposes.

assess the performance of the human-machine system. Outcomes may be either successful discovery/identification of important concepts or failures to find or reach conclusions about such concepts. Some outcomes may be attributed primarily to the analyst, some to the tool, and some to their integration or human-computer interface. One purpose of attempting to conduct this more detailed analysis is to distinguish among the effects of the system, the human, and the joint human-computer system.

A possible analysis approach is inspired by early work of Kantor (1980) for evaluating library science applications, which was recently recast to apply to the evaluation of question-answering systems for intelligence analysis.[8] Kantor specifies a chain of behavior (failure types) that exhibits a natural, logical order, and that associates specific "atoms" of information with each step. Some elements of the chain represent failures of system functions and others may be the human's responsibility—the key is that each function is dependent upon the successful performance of preceding steps in the chain. As an example using our context, consider the possible ways that the human-computer system might fail to identify a specific hypothesis, ℋ:

Potential Failures in the Chain of Behavior	**Possible System Limitation**	**Possible User Limitation**
A. The system did not generate ℋ for the analyst to consider	*Hypothesis Generation*	
B. Not A, but the analyst did not examine ℋ		*Attention, Information overload*
C. Not A or B, but the system did not provide evidence ℰ to support ℋ	*Evidence Marshalling*	
D. Not A, B or C, but the analyst did not examine ℰ	*User Interface**	*Attention, Information overload*
E. Not A, B, C, or D, but the system did not properly associate ℰ and ℋ to validate ℋ	*Hypothesis Tracking*	
F. Not A, B, C, D, or E, but the analyst did not accept ℰ as evidence for ℋ		*Judgment*
G. None of the above, but the analyst did not interpret ℋ properly		*Interpretation*

*Ideally, either the system or the user should be identified as responsible for a shortcoming shown in a row. Here, it does not appear feasible to distinguish system- from user-limitations without further study.

By applying this structured approach to identified steps in a process, it may be possible to specify performance measures and success criteria that can be used to identify needs for system improvements to correct its deficiencies or to accommodate user limitations. Steps in the IA process have been described by several researchers based on cognitive task analyses, but it is likely that it would be necessary to add more detail about the cognitive process and how the IA tool is used. This evaluation research approach deserves further consideration. For more detailed discussions of cognitive models, cognitive task analyses, and the coupling between human and computer components of a system, see D'Amico et al. (2004), Elm (2004), Badalamente and Greitzer (2005), and Woods (2005).

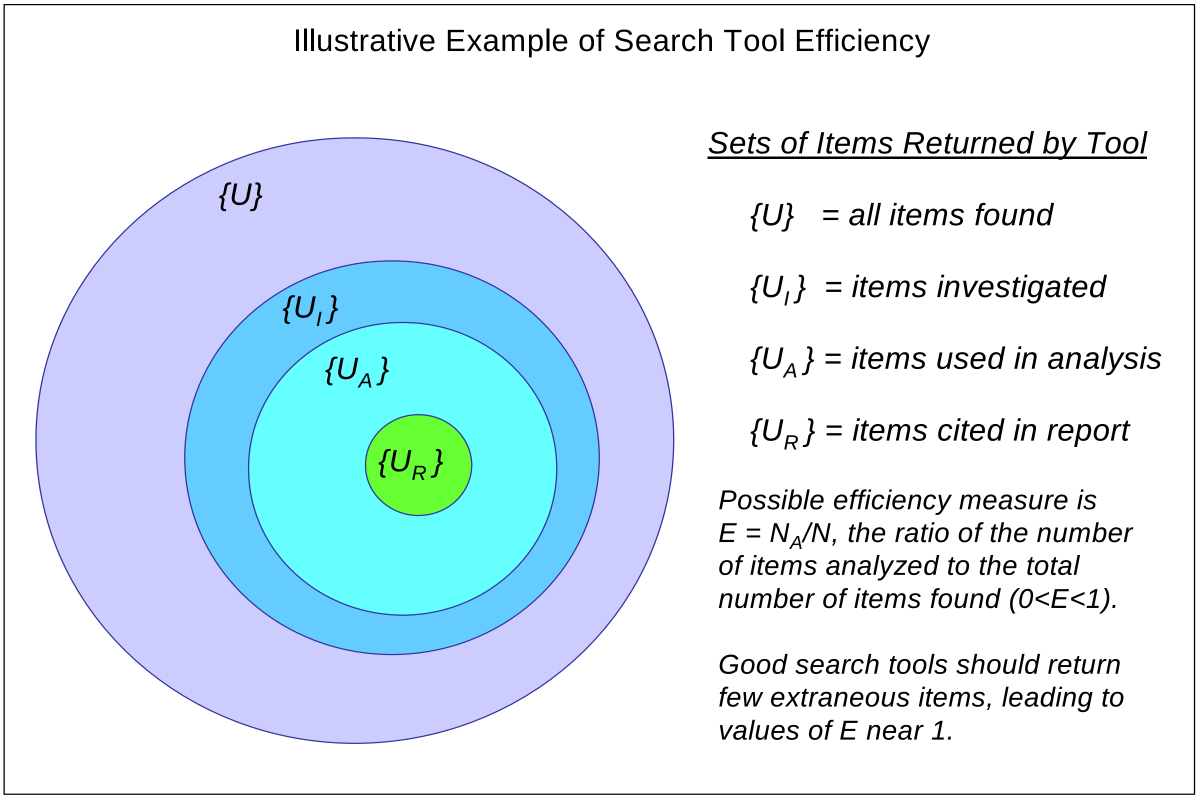
[8] Material prepared by Paul Kantor for the AQUAINT Program Winter 2005 Symposium, Palm Springs, CA, February 2005.

6.3.2 Measures of Effectiveness

Measures of effectiveness (MOE) are derived from the performance measures. In some cases, a MOE is virtually the same as a performance measure—e.g., a grade/rating/score given to a product by an expert evaluator. In many cases, performance measures are used or combined to produce MOEs. MOEs should directly reflect a product's success criteria—i.e., if a product is designed to improve the search/evidence marshalling process, then one or more MOEs should be developed that reflect this capability. An illustrative example is provided in the box below and in Figure 6-2.

Illustrative Example: Impact of a Search Tool

Consider the situation in which we assess the impact of a tool that is designed to improve evidence marshalling by enhancing search capabilities. One measure of performance is the number of documents or URL "hits" that the tool provides to the analyst for further study. Underlying this is a collection of increasingly specific sets of items. The total collection ("universe") of items returned by the search tool is designated $\{U\}$. Some of these items appear to be relevant and are investigated further (set $\{U_I\}$, a subset of $\{U\}$); some of those are found to be important and are used in the analysis (set $\{U_A\}$, a subset of $\{U_I\}$); some of those are even cited in the final report (set $\{U_R\}$, a subset of $\{U_I\}$). There is a collection of items that are not relevant and never investigated (the set $\{U\}$-$\{U_I\}$). One possible derived efficiency measure of interest is the proportion of items that are actually used in the analysis—e.g., the number of items in set $\{U_A\}$ divided by the total number of items returned in set $\{U\}$, which we can denote $E = \mathrm{N_A/N}$ where $\mathrm{N_A}$ is the size of set $\{U_A\}$ and N is the size of set $\{U\}$. This measure can be obtained for the tool being evaluated (E_T) as well as for a "standard" or baseline tool (E_S) such as Google. A MOE might then be the ratio of the respective measures, E_T/E_S. A success criterion might be to realize a 50% improvement in this ratio, or $E_T/E_S = 1.5$.



Illustrative Example of Search Tool Efficiency

Sets of Items Returned by Tool

$\{U\}$ = all items found

$\{U_I\}$ = items investigated

$\{U_A\}$ = items used in analysis

$\{U_R\}$ = items cited in report

Possible efficiency measure is $E = N_A/N$, the ratio of the number of items analyzed to the total number of items found ($0<E<1$).

Good search tools should return few extraneous items, leading to values of E near 1.

Figure 6-2. Illustration of One Form of an Efficiency Measure Calculation.

7. Summary and Recommendations

7.1 Summary

There are some very interesting—and difficult—issues that need to be addressed in order to conduct a technically-sound evaluation of the impact or effectiveness of a tool. Three major issues and challenges concern: (1) research methodology/experimental design issues aimed at achieving the most objective, scientifically valid conclusions based on the research; (2) metrics for task difficulty/complexity that may be used to manipulate or control this important task variable; and (3) performance measures for the effectiveness of the tool. These three areas are intimately inter-related: If our judgment/method/approach/measures are faulty in any one of these three areas, the overall value of the evaluation is at risk.

This report has developed in more detail some of the initial thoughts that were outlined in a draft report (Greitzer, 2004) focused on evaluation/research methodology and task difficulty metrics within the context of evaluation challenges for NIMD products. The present report builds on this earlier work by discussing the need to design robust and valid experimental or quasi-experimental research to support evaluation of new tools and technologies; by recommending some task difficulty metrics as an initial step in the complex process toward defining rigorous research methods and approaches to assessing the impact of IA tools; and by discussing the need for and concepts for performance measures to be used in tool evaluations.

7.2 Recommendations

Recommendations on evaluation methodology, task difficulty metrics, and performance measures follow from the considerations and discussion in the three major sections of this report:

- For evaluation of IA tools, use of a "one-shot case study" is not recommended because it does not provide any control over variables that could confound the results. Between-group experiments provide the best "experimental controls" of factors that threaten the validity of the evaluation, but it is recognized that the number of subjects required for such experiments is not typically available.

- A within-subjects pretest-posttest design is recommended for NIMD evaluations because it does not require as many subjects as between-groups experiments. The purpose of the pretest is to provide baseline performance data. The challenge to validity is to be able to pick experimental analysis tasks that are relatively equivalent in difficulty.

- There is a need to better understand task difficulty. A research priority should be to develop task difficulty metrics that are useful and practical and that may be used to support evaluation studies.

- To develop meaningful performance measures, tool developers should identify success criteria that indicate when the tool has met its objectives and performance requirements. Performance measures should be closely tied to success criteria.

- In addition to subjective data that are relatively easily obtained through ratings and questionnaire methods, objective, behavioral performance measures should be identified. Objective measures may be based on temporal factors or on the extent to which the human-computer system is successful in meeting specific performance objectives.

- Performance measures should not be limited to data that are captured through traditional methods such as usability testing:

    - One example of "out of the box thinking" about performance measures is to consider the impact of a tool on the vetting process as well as the analysis process (time or efficiency benefits are possible throughout the "life cycle" of an intelligence product).
    - Research should focus on development of performance measures based on criteria other than subjective ratings or high level/global ratings of overall quality. Examples of more detailed performance measures were provided, and should be studied further.

- Measures of effectiveness may be defined as functions of the more basic performance measures. Examples of some new ways of conceptualizing measures of effectiveness were discussed that may provide additional insights about the locus of the effect of a tool or method, or the locus of the deficiencies in the human-computer system that need improvement.

Developing a useful, predictive set of IA task difficulty metrics and associated measures is important to the IC community, researchers, and stakeholders because such measures are needed to assess the impact of new methods and tools that are being considered for introduction into the field. It is hoped that the present paper has been successful in laying out the overall research issues and challenges as well as laying some groundwork for development of operationally defined difficulty metrics and performance measures to support the intelligence community.

8. References

Anderson, T., and W Twining. 1999. *Analysis of Evidence: How to Do Things with Facts Based on Wigmore's Science of Judicial Proof.* Evanston, IL: Northwestern University Press.

Badalamente, R. V., and FL Greitzer. 2005. Top ten needs for intelligence analysis tool development. *First International Conference on Intelligence Analysis Methods and Tools.* McLean, VA, May 2-5, 2005.

Bodnar, J. 2004. What's a 'Difficult' Intelligence Problem? (Unpublished Manuscript).

Campbell, D. T., and JC Stanley. 1963. Experimental and Quasi-Experimental Designs for Research. Chicago: Rand McNally & Company.

D'Amico, A., D Tesone, K Whitley, B O'Brien, M Smith, and E Roth. 2004. *Understanding the Cyber Defender: A Cognitive Task Analysis of Information Assurance Analysts*. Interim Report—Volume 1. Report No. CSA-CTA-1-1. Secure Decisions. December, 2004. [FOUO]

Elm, WC. 2004. *Beauty is more than skin deep: Finding the decision support beneath the visualization.* Technical Report, Cognitive Systems Engineering Center. National Imagery and Mapping Agency. Reston, VA.

Flesch, R. F. 1994. The art of readable writing. New York: Wiley.

Greitzer, F. L. 2004. Preliminary thoughts on difficulty or complexity of intelligence analysis tasks and methodological implications for glass box studies. (Unpublished Manuscript).

Greitzer, F. L. 2005. Toward the development of cognitive task difficulty metrics to support intelligence analysis research. (Manuscript submitted to the *4th IEEE International Conference on Cognitive Informatics*, August 2005).

Hart, S. G., and LE Staveland. 1988. Development of a multi-dimensional workload rating scale: Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human Mental Workload*. Amsterdam, The Netherlands: Elsevier.

Hewett, T. and J Scholtz. 2004. Towards a Metric for Task Difficulty. Paper presented at the NIMD PI Meeting, November 2004, Orlando, FL.

Hughes, F. J., and DA Schum. 2003. Preparing for the future of intelligence analysis: Discovery—Proof—Choice. Unpublished Manuscript, Joint Military Intelligence College.

Hutchins, S. G., PL Pirolli, and SK Card. (in press). What Makes Intelligence Analysis Difficult? A Cognitive Task Analysis of Intelligence Analysts. In Robert Hoffman (Ed.), *Expertise Out of Context*, Hillsdale, NJ: Lawrence Erlbaum.

Johnston, R. 2003 (a). Integrating methodologists into teams of substantive experts. *CIA Studies in Intelligence*, *47(1)*, Available on the Internet: http://www.cia.gov/csi/studies/vol47no1/article06.html.

Johnston, R. 2003 (b). Developing a taxonomy of intelligence analysis variables. *CIA Studies in Intelligence*, *47(3)*, http://www.cia.gov/csi/studies/vol47no3/article05.html.

Kantor, P. B. 1980. Availability analysis. *Journal of the American Society for Information Science*. *27(6)*, 311-319. Reprinted in *Key Papers in Information Science*. White Plains, NY: Knowledge Industry Publications, Inc., 1980, pp. 368-376.

Nielsen, J. 1992. Finding usability problems through heuristic evaluation. In *Proceedings of CHI'92*, 373-380.

Nielsen, J. 1993. *Usability Engineering*. San Francisco: Morgan Kaufmann.

Newell, A., and HA Simon. 1972. *Human Problem Solving*. Englewood Cliffs, N.J.: Prentice-Hall.

Patterson, E. S., EM Roth, and DD Woods. 2001. Predicting Vulnerabilities in Computer-Supported Inferential Analysis Under Data Overload. *Cognition, Technology & Work, 3*, 224-237.

Plato. *The Dialogues of Plato (Meno dialogue)*. Translated into English by B. Jowett. Oxford: Clarendon Press, 1875. http://www.classicallibrary.org/plato/dialogues/10_Meno.htm

Sanquist, T. F., FL Greitzer, A Slavich, R Littlefield, J Littlefield, and P Cowley. 2004. Cognitive Tasks in Intelligence Analysis: Use of Event Dwell Time to Characterize Component Activities. Human Factors and Ergonomics Society 48th Annual Meeting. New Orleans, LA: 20-24 September 2004.

Schum, D. 1999. Marshaling thoughts and evidence during fact investigation. *South Texas Law Review. Vol. 40(2)*, 401-454.

Tillers, P. 2004. Picturing Inference. An Essay in Honor of Professor Lothar Philipps. http://tillers.net/pictures/picturing.html

Tillers, P., and D Schum. 1988. Charting new territory in judicial proof: Beyond Wigmore. *Cardozo Law Review, Vol. 9 (3)*, 907-966.

Wigmore, J. H. 1937. *The Science of Judicial Proof: As Given by Logic, Psychology, and General Experience and Illustrated in Judicial Trials*. 3rd ed. Boston: Little, Brown & Co.

Wilkinson, B. 2004. What makes intelligence analysis hard? Presentation at the Friends of the Intelligence Community meeting, Gaithersburg, MD.

Woods, DD. 2005. Supporting cognitive work: How to achieve high levels of coordination and resilience in joint cognitive systems. To appear in *Naturalistic Decision Making 7*. Amsterdam, The Netherlands, June 15, 2005.