

## Toward the Development of Cognitive Task Difficulty Metrics to Support Intelligence Analysis Research

Frank L. Greitzer  
Battelle—Pacific Northwest Division  
[frank.greitzer@pnl.gov](mailto:frank.greitzer@pnl.gov)

### Abstract

*Intelligence analysis is a cognitively complex task that is the subject of considerable research aimed at developing methods and tools to aid the analysis process. To support such research, it is necessary to characterize the difficulty or complexity of intelligence analysis tasks in order to facilitate assessments of the impact or effectiveness of tools that are being considered for deployment. A number of informal accounts of "What makes intelligence analysis hard" are available, but there has been no attempt to establish a more rigorous characterization with well-defined difficulty factors or dimensions. This paper takes an initial step in this direction by describing a set of proposed difficulty metrics based on cognitive principles.*

### 1. Introduction

Intelligence analysis (IA) professionals are confronted each day with high demands for rapid, yet accurate assessments that require discovery and marshaling 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. This research aims to produce tools that will yield measurable performance improvements when deployed in operating IA facilities. 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 these conditions are applied in a within-subjects design, which requires use of different tasks. Thus, rigorous metrics are required to characterize task difficulty.

The purpose of this paper is to examine concepts that have been associated with the notion of IA task difficulty in an attempt to produce an initial set of variables or constructs that will move the R&D community closer to developing metrics.

### 2. Background

In its most simple terms, this paper relates to why some tasks are more difficult to perform than others. It does not, however, focus on what differentiates expert from novice performance. There is a large body of judgment/decision making and cognitive science research that demonstrates areas where experts and novices behave in similar ways (e.g., demonstrating biases and limitations in decision making) and areas where experts excel over novices (e.g., strategies, automated processes). Reviewing this research in the context of describing the role of task characteristics in expert performance, Shanteau [1] observed that expertise is domain specific: task characteristics determine whether or not an expert will behave competently. Characteristics that make tasks hard include dynamic data, lack of predictability, decisions about people versus things, and lack of feedback.

Intelligence analysis is among the more difficult problem domains because it is associated with decisions about human intentions and actions, largely unpredictable, with little or no feedback available, and with dynamic and unreliable data.

## 2.1. Initial set of task difficulty dimensions

During the past two years, the IA research community held discussions aimed at a more rigorous understanding of IA task difficulty. At a “Friends of the Intelligence Community” workshop held in January 2004, Bonnie Wilkinson [2] 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. A series of discussions and more detailed analyses followed that produced an initial characterization [3] [4] [5] that includes:

- *Characterization versus Prediction.* Does the task require a description of current capabilities or does it ask the analyst to forecast future capabilities or actions? 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.
- *Sociological Complexity.* Is the focus of the analysis on an individual, group, State, or region? Shanteau [1] cited the distinction between “decisions about things” versus “decisions about behavior” and Wilkinson [2] referred a human behavior factor. Greitzer [3] [4] suggested the more general dimension of sociological complexity to reflect the nature of the social network.
- *Data Uncertainty.* Are the data difficult to observe or interpret? Greitzer [3] [4] suggested this dimension to encompass several factors identified by Wilkinson [2] such as low observability, lack of physical/hard data, data ambiguity, low confidence in sources. Data uncertainty could also arise from a lack of data, from ambiguous, deceptive, or unreliable data, or because the data are of multiple types, different levels of specificity, or dynamic/changing over time.
- *Breadth of Topic.* Wilkinson [2] used descriptions like multiple subjects, many variables, and many organizations to describe this idea; Hewett and Scholtz [6] described this factor

in terms of the extent to which the analysis topic is narrowly focused versus open-ended.

- *Time Pressure.* The amount of time available to conduct the analysis influences the difficulty in carrying out the task, as has been observed in experimental research [8] and in cognitive task analyses (e.g., [9] [10]). Time pressure seems different from the other dimensions because it represents a variable that can be *manipulated directly and independently* (i.e., one can control the time pressure by setting the deadline for the product). Thus, it may be argued whether the time variable *per se* is a true task difficulty dimension, as opposed to a possible experimental variable to be studied.
- *Data Availability.* As pointed out by John Bodnar [7] (also described in [5]): “The degree of difficulty in assessing 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).
- *Problem Structure.* To what extent is the problem 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 [6]?
- *Data Synthesis.* To what extent does the analyst need to synthesize multiple sources of information (also called data fusion)? 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 levels of domain expertise are needed to analyze each type of data [10].

## 2.2. Consideration of additional factors

One possible task difficulty factor that has been suggested is the lack of feedback [1] [2]. 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. This is in part due to the nature of predictive tasks (which has already been described as a task difficulty dimension), and in part due to lack of feedback (which is not necessarily intrinsic to the task but can be manipulated as an experimental variable). Lack of feedback, then, may be more appropriately called a task variable.

Information overload is another factor implicated in the issue of “what makes IA hard.” Information overload is often attributed to “too much data.” The *quantity* of data, per se, may not underlie the problem so much as the problems inherent in the data, such as consistency, reliability, and heterogeneity. Thus, for example, a massive data set that tends to be consistent and homogeneous may not pose as difficult a problem as a much smaller data set that lacks consistency and homogeneity. Thus, information overload may be reflected in terms of one or more dimensions such as time pressure, data availability, and data synthesis. Similarly, high workload may result from 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.

As a preliminary assessment of the adequacy of an initial set of task difficulty dimensions as of the spring 2004 timeframe, a questionnaire incorporating possible task difficulty factors was administered to working analysts [6]. The questionnaire included five of the initial set of task difficulty dimensions in the previous section (Characterization vs. Prediction, Sociological Complexity, Breadth of Topic, Problem Structure; and Data Synthesis) and six additional variables, plus an overall task difficulty factor. This study used a limited number of analysts (eight) 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. Nevertheless, a statistically significant 0.85 correlation was obtained between the average difficulty *ranks* of each task over the eleven factors with the task’s average overall difficulty rating. This result does not imply that the set of dimensions and variables studied would provide a reliable or complete characterization of task difficulty.

Indeed, there is at least one potential factor that appears to have been overlooked in our initial discussions—problem complexity. The complexity of the analysis task has not received much attention in discussions about task difficulty among IC researchers. This concept is fundamental to understanding the IA process, developing tools to support it, and defining metrics for task difficulty and performance effectiveness. This important factor is discussed in the next section.

### 3. Problem complexity

Problem complexity relates to the mental processes involved in problem solving, which, despite a long history of study in psychology, largely 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. Heuer [11] 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.” (p. 31) Similarly, Hughes and Schum [12] claimed “What is so frequently left out of the equation is the process by which the information is analyzed.” For the purposes of defining task difficulty metrics in terms of the complexity of the analytical problem, we can distill notions from psychological research on problem-solving and, as taught by Frank Hughes at the Joint Military Intelligence College, from philosophers and thinkers in the legal field.

Most psychological research on problem-solving has focused on well-defined problems: those for which we know a solution exists, and for which 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.<sup>1</sup> 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 [14]. “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” [11]. “Goal directed” means, for example, that the process is guided by

---

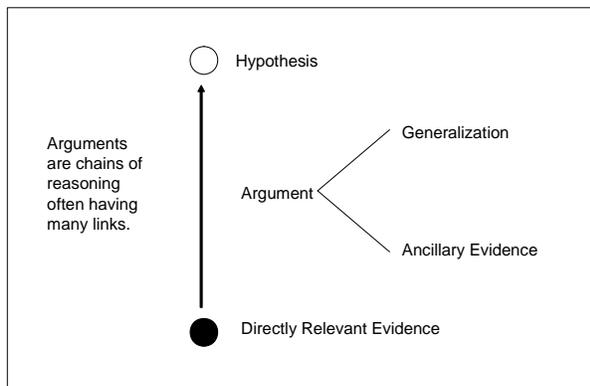
<sup>1</sup> Plato [13] described this age-old problem over 2000 years ago in his *Meno Dialogue*: “And how will you enquire, Socrates, into that which you do not know? What will you put forth as the subject of enquiry? And if you find what you want, how will you ever know that this is the thing which you did not know?”

heuristic principles that concentrate the search on promising regions of the problem space to avoid getting bogged down pursuing blind alleys.

Based on this perspective, relevant problem-complexity parameters include 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. The total number of such threads to follow and the level of reasoning that is required to reach a conclusion are also relevant parameters for problem complexity.

Hughes and Schum [12] observed that “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.” They carefully describe the construction of an argument, which is a chain of reasoning that connects evidence to hypotheses of interest in the analysis. Figure 1, adapted from Hughes and Schum, shows but one of many chains of evidence in an inference network that would represent the thought process behind an IA product. 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.”

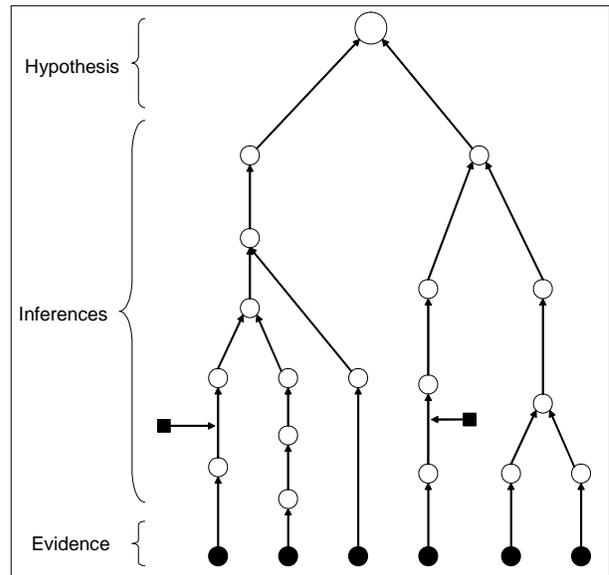
The structured aggregation of many chains of reasoning that make up the test of IA hypotheses may



**Figure 1. Argument represented as a chain of reasoning.**

be described by inference networks that represent evidence marshaling and analysis [15] [16] [17] [18] [19]. A graphical representation of an inference network is shown in Figure 2, where nodes represent evidence and inferences and links between nodes represent propositions. Directly observed evidence is shown as filled circles, auxiliary evidence is shown as filled squares, and inferences based on the evidence are shown as open circles. The illustration shows three chains of reasoning that weigh on the hypothesis, shown at the top of the diagram as a larger circle.

What sort of computational approach to defining a Problem-Complexity metric might apply? 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 [20]. Of course, use of metrics based on inference networks requires that the solution (network) 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).



**Figure 2. Graphical illustration of an inference network.**

#### 4. Discussion

Research is needed to clarify distinctions or dependencies among the task difficulty dimensions and to determine if any factors have been omitted. Several suggestions for research are discussed.

#### 4.1. Refinements in questionnaire studies

The study by Hewett and Scholtz [6] should be replicated with a more diverse set of IA tasks. Their analysis included some factors that should be reconsidered in deriving task difficulty metrics, particularly analyst experience. Experience is a relevant “demographic” or analyst variable that certainly affects performance—but this does not mean that it should be incorporated into difficulty metrics. It is not a task variable as such.<sup>2</sup> The follow-up questionnaire study should include the problem complexity dimension and the focus of the analysis should be to determine the extent to which these factors “predict” an overall task difficulty judgment for IA tasks. Another question (perhaps addressed using a multiple regression analysis) is whether dependencies in the proposed dimensions would be revealed in stronger correlations (regressions) for some of these factors on overall task difficulty than for other factors. Finally, the analysis can make use of analyst experience data by examining possible differences in the strength of such relationships exhibited in ratings of experienced versus inexperienced analysts.

#### 4.2. Multidimensional scaling studies

Expert analysts should be asked to rate similarities between pairs of IA task descriptions or to sort task descriptions into groups based on subjective similarity. Several multidimensional scaling techniques are available (e.g., [22]) to examine or derive dimensions from such data. Dimensions that would most likely be amenable to such analyses 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.

#### 4.3. Behavioral studies

Behavioral data should be collected and analyzed to reveal possible correlates of the proposed task difficulty variables in observed IA activities. This may serve to validate, disambiguate, or further refine the proposed dimensions into a more useful set that can be used to guide further research. An important development that supports behavioral data collection

---

<sup>2</sup> A more familiar example may clarify the distinction: Readability measures, such as the Flesch Reading Ease or Flesch Kincaid Grade Level (e.g., [21]), 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.

for IA tasks is the NIMD program’s Glass Box instrumentation [23]. For example, to support inferences about the analyst’s interest in material gathered during the analysis process, “dwell times” exhibited in Glass Box data have been analyzed [24].

Glass Box data may also be useful in gathering information on task complexity by examining behavioral data or artifacts recorded by the Glass Box instrumentation. For example, one of the most effective ways to deal with complexity is to “divide and conquer” (decompose the problem into simpler parts). Can such strategies be inferred from the Glass Box data? 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, 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? 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 typically create such networks to support their reasoning process.

Finally, to improve our understanding of the data uncertainty dimension, we may ask analysts to provide confidence ratings on evidence that they collect. The Glass Box functionality currently enables analysts to easily indicate “relevance” of material; a confidence measure could be implemented in the same way.

### 5. Conclusions

A useful, predictive set of IA task difficulty metrics and associated measures are needed to assess the impact of new methods and tools that are being considered for introduction into the field. Task difficulty is part of a broader context of evaluation research—here applied to intelligence analysis—that includes fundamental questions about research methods, task difficulty dimensions, and performance measures [4].

This paper has described task difficulty dimensions that should be taken into account in the design of evaluation studies for IA tools and methods. 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. While continued use of the traditional questionnaire approach will be useful for refining and validating the proposed dimensions, this paper recommends a complementary approach that incorporates behavioral measures to address more cognitive factors and correlates of the task difficulty dimension.

## 6. References

- [1] J. Shanteau, "Competence in experts: The role of task characteristics." *Organizational Behavior and Human Decision Processes*, 1992, 53, 252-266.
- [2] B. Wilkinson, "What makes intelligence analysis hard?" Presentation at the Friends of the Intelligence Community meeting, Gaithersburg, MD, January, 2004.
- [3] F. L. Greitzer, "Preliminary Thoughts on Difficulty or Complexity of Intelligence Analysis Tasks and Methodological Implications for Glass Box Studies (Interventions)." Unpublished Report, Battelle—Pacific Northwest Division. February 2004.
- [4] F. L. Greitzer, "Methodology, Metrics and Measures for Testing and Evaluation of Intelligence Analysis Tools." Technical Report PNWD-3550. Richland, WA: Battelle—Pacific Northwest Division, March 2005.  
URL: <http://www.pnl.gov/cogInformatics>
- [5] F. L. Greitzer and K. Allwein, "Metrics and measures of intelligence analysis task difficulty." Panel Session, 2005 *International Conference on Intelligence Analysis Methods and Tools*. Vienna, VA, May 2-5, 2005.
- [6] T. Hewett and J. Scholtz., "Towards a metric for task difficulty." Paper presented at the NIMD PI Meeting, Orlando, FL, November 2004.
- [7] J. Bodnar, "What's a 'Difficult' Intelligence Problem?" Unpublished paper, April, 2004.
- [8] E. S. Patterson, E. M. Roth, and D.O. Woods. "Predicting vulnerabilities in computer-supported inferential analysis under data overload." *Cognition, Technology & Work*, 2001, 3, 224-237.
- [9] S. G. Hutchins, P. Pirolli, and S. K. Card, "A new perspective on use of the critical decision method with intelligence analysts." Paper presented as part of a Panel on Designing Support for Intelligence Analysts. In *Proceedings of the 48<sup>th</sup> Human Factors and Ergonomics Society Annual Meeting*, New Orleans, LA, September 21-24, 2004.
- [10] S. G. Hutchins, P. Pirolli, and S.K. Card (in press). What makes intelligence analysis difficult? A cognitive task analysis of intelligence analysts. In R. Hoffman (Ed.), *Expertise out of Context*, Hillsdale, NJ: Lawrence Erlbaum.
- [11] R. J. Heuer, Jr., *Psychology of Intelligence Analysis*, CIA Center for the Study of Intelligence, Washington, D.C., 1999.
- [12] F. J. Hughes and D.A. Schum. "Preparing for the future of intelligence analysis: Discovery—Proof—Choice." Washington, D.C., Joint Military Intelligence College, January, 2003.
- [13] Plato. *The Dialogues of Plato (Meno dialogue)*. Translated into English by B. Jowett. Oxford: Clarendon Press, 1875.
- [14] A. Newell and H.A. Simon, *Human Problem Solving*. Englewood Cliffs, N.J.: Prentice-Hall, 1972.
- [15] D. Schum, "Marshaling thoughts and evidence during fact investigation." *South Texas Law Review*. 1999, 40(2), 401-454.
- [16] J. J. Wigmore, *The Science of Judicial Proof: As Given by Logic, Psychology, and General Experience and Illustrated in Judicial Trials*. 3rd ed. Boston: Little, Brown & Co., 1937.
- [17] P. Tillers, and D. Schum, "Charting new territory in judicial proof: Beyond Wigmore." *Cardozo Law Review*, 1988, 9 (3), 907-966.
- [18] T. Anderson and W. Twining. *Analysis of Evidence: How to Do Things with Facts Based on Wigmore's Science of Judicial Proof*. Evanston, IL: Northwestern University Press, 1999.
- [19] E. Waltz, *Knowledge Management in the Intelligence Enterprise*, Boston, MA; Artech House, 2003.
- [20] P. Tillers, "Picturing Inference. An Essay in Honor of Professor Lothar Philipps," accessed online on May 4, 2005 at <http://tillers.net/pictures/picturing.html>.
- [21] R. F. Flesch, *The art of readable writing*. New York: Wiley, 1994.
- [22] J. B. Kruskal and M. Wish, *Multi-dimensional Scaling (Vol 11, Quantitative Applications in the Social Sciences)*. Thousand Oaks, CA: Sage Publications, 1978.
- [23] P. Cowley, L. Nowell, and J. Scholtz, "Glass Box: An instrumented infrastructure for supporting human interaction with information." *Hawaii International Conference on System Sciences, Island of Hawaii*, January 3-6, 2005.
- [24] T. F. Sanquist, F. L. Greitzer, A. Slavich., R. Littlefield, J. Littlefield, and P. Cowley, "Cognitive Tasks in Intelligence Analysis: Use of Event Dwell Time to Characterize Component Activities." *Proceedings, Human Factors and Ergonomics Society 48th Annual Meeting*. New Orleans, LA, September 21-24, 2004.