ConMap:

Investigating New Computer-Based Approaches to Assessing Conceptual Knowledge Structure in Physics

A Dissertation Presented

by

Ian D. Beatty

Submitted to the Graduate School of the
University of Massachusetts Amherst in partial fulfillment
of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2000

Physics

CONMAP:
INVESTIGATING NEW COMPUTER-BASED APPROACHES TO ASSESSING CONCEPTUAL KNOWLEDGE STRUCTURE IN PHYSICS

A Dissertation Presented

by

IAN D. BEATTY

Approved as to style and content by:

William J. Gerace, Chair

Robert A. Guyer, Member

Robert J. Dufresne, Member

Allan Feldman, Member

John F. Donoghue, Department Head
Physics and Astronomy

Dedicated to

Lee David Beatty

1925 — 1998

Loyal, brave, and humble when it mattered most.

Acknowledgements

I am deeply grateful for the assistance of my committee during the researching and writing of this dissertation. Bob Guyer had an apparently infinite amount of time available to discuss any aspect of the work, despite his considerable other obligations. Without Bill Gerace's conviction that the research in question did have merit and that I was capable of carrying it out, I would have quit many times over.

Abstract

CONMAP:
INVESTIGATING NEW COMPUTER-BASED APPROACHES
TO ASSESSING CONCEPTUAL KNOWLEDGE STRUCTURE
IN PHYSICS

MAY 2000

IAN D. BEATTY, B.S., UNIVERSITY OF MASSACHUSETTS AMHERST

Ph.D., UNIVERSITY OF MASSACHUSETTS AMHERST

Directed by: Professor William J. Gerace

There is a growing consensus among educational researchers that traditional problem-based assessments are not effective tools for diagnosing a student's knowledge state and for guiding pedagogical intervention, and that new tools grounded in the results of cognitive science research are needed. The *ConMap* ("Conceptual Mapping") project, described in this dissertation, proposed and investigated some novel methods for assessing the conceptual knowledge structure of physics students.

A set of brief computer-administered tasks for eliciting students' conceptual associations was designed. The basic approach of the tasks was to elicit spontaneous term associations from subjects by presenting them with a prompt term, or problem, or topic area, and having them type a set of response terms. Each response was recorded along with the time spent thinking of and typing it.

Several studies were conducted in which data was collected on introductory physics students' performance on the tasks. A detailed statistical description of the data was compiled. Phenomenological characterization of the data (description and statistical summary of observed patterns) provided insight into the way students respond to the tasks, and discovered some notable features to guide modeling efforts. Possible correlations were investigated, some among different aspects of the ConMap data, others between aspects of the data and students' in-course exam scores. Several correlations were found which suggest that the ConMap tasks can successfully reveal information about students' knowledge structuring and level of expertise. Similarity was observed between data from one of the tasks and results from a traditional concept map task.

Two rudimentary quantitative models for the temporal aspects of student performance on one of the tasks were constructed, one based on random probability distributions and the other on a detailed deterministic representation of conceptual knowledge structure. Both models were reasonably successful at approximating the statistical behavior of a typical student's data.

Table of Contents

List of Figures

List of Tables

1. Introduction

Physics education needs new assessment methods. Physics education research, informed by cognitive science, must devise them. To accomplish this, better models of physics student's learning, knowing, and knowledge application are required.

This dissertation describes and discusses an ongoing research effort called the *ConMap* ("Conceptual Mapping") project. The fundamental aim of the project is to investigate the utility of a particular set of proposed assessment tools — brief, computer-administered tasks for eliciting spontaneous conceptual associations — for probing the quality and extent of a physics student's conceptual knowledge structure (CKS) in an introductory physics domain.

Section 1.1 presents the shortcomings of traditional, problem-based physics assessments and notes the need for a suitable cognitive model of physics knowing and learning in the construction of a better assessment approach. Section 1.2 outlines the context of existing research in which the ConMap study has been pursued. Section 1.3 identifies the goals of the ConMap project and describes the organization of the remainder of this dissertation.

1.1. See the Need

1.1.1. The Inadequacy of Traditional Assessments

Traditionally, educators attempt to measure a student's physics knowledge by presenting the student with a set of problems to solve and grading the student's solutions. In situations where the time required for detailed grading is prohibitive, machine-graded multiple-choice problems are typically given. A student who can solve a sufficient fraction of the problems is considered to have "learned" the material adequately.

Appropriately-chosen problems can provide a reasonable measure of a student's ability to "do" physics, if doing physics is defined as being able to solve problems of the sort assigned. Many physics instructors, however, assert a more ambitious goal: they want to help students develop a conceptual understanding of the topic material, qualitative reasoning abilities, transfer of knowledge to different contexts, and the general capabilities of "thinking like a physicist". Traditional problem-based exams fail to test for the success of these objectives. According to Robert J. Mislevy of the Educational Testing Service (Mislevy 1993),

> Educational measurement faces today a crisis that would appear to threaten its very foundations. The essential problem is that the view of human abilities

> implicit in standard test theory… is incompatible with the view rapidly emerging from cognitive and educational psychology. Learners increase their competence not only by simply accumulating new facts and skills, but by reconfiguring their knowledge structures, by automating procedures and chunking information to reduce memory loads, and by developing strategies and models that tell them when and how facts and skills are relevant. The types of observations and the patterns in data that reflect the ways that students think, perform, and learn cannot be accommodated by traditional models and methods.

Furthermore, traditional problem-based exams provide little information about *why* a particular student failed on the particular problems he or she did not get right, and even less about what specific pedagogic interventions the instructor should employ to help the student resolve their difficulties. There are many reasons why a student might fail to solve a problem correctly. Among them:

- Failure to interpret the problem situation as the assessor intended;
- Insufficient or incorrect physical intuition to understand what is happening in the problem situation;
- Ignorance of the necessary principle;
- Failure to recognize the correct principle;
- Conceptual mistake during application of the correct principle;
- Cognitive overload: general confusion and failure to keep track of enough information and lines of thought;
- Algebraic error during calculations;
- Numerical error or calculator keypress error;
- Error determining units or powers of ten;
- Failure to answer the precise question being asked.

If a student's written solutions are hand-graded, it might be possible to identify the point at which the student went awry. Even then, the *reason* for the mistake — the underlying misconception, missing piece of information or insight, etc. — can only be guessed at. This is complicated by the fact that students frequently make "careless errors" caused by the failure to apply knowledge they have, rather than by missing knowledge.

Kikumi K. Tatsuoka of the Educational Testing Service, a pioneer in statistical pattern recognition and classification approaches to the interpretation of standard problem-based test results, admits:

> The problem of cognitive diagnosis has an additional difficulty [beyond statistical decision theory], because cognitive processes that should be extracted as feature variables are not observable, and knowledge states as classification categories cannot be obtained directly from observations.

In other words, a student's "knowledge state" cannot be obtained directly from his or her exam performance.

In short, traditional problem-based exams have limited *evaluative* value and even more limited *diagnostic* value. They might be capable of providing a crude measure of student competence, narrowly defined, but are insufficient for diagnosing specific weaknesses in a student's knowledge and are therefore insufficient to guide an instructor in remediation.

1.1.2. Alternative Assessments

The field of physics education research (PER) has made significant advances in physics pedagogy, providing a fundamentally new perspective on the role of the instructor in the learning process and developing many new and demonstrably efficacious teaching tools (Larkin 1979; Mestre and Touger 1989; Mestre 1991; Gerace 1992; Redish 1994; Gerace, Leonard et al. 1997; McDermott and Redish 1999). The progress made on the development of new assessment methodologies has been less dramatic, perhaps because the "assessment problem" is more difficult than the "instruction problem". Nevertheless, some progress has been made.

If one wishes to assess a student's "knowledge state", rather than merely summarize the parts of the assessment the student did and didn't succeed at, it is obvious that one needs a model of what a knowledge state is and how it is probed by the assessment. Thus, the development of assessment methodologies must be founded on the results of cognitive science research into physics knowledge, learning, and task performance. In particular, an effective diagnostic assessment must describe a student with reference to some suitably detailed model of physics knowing, learning, and application.

Physics education research has provided general qualitative descriptions of knowledge structuring in physics, and cognitive science has furnished detailed models of limited aspects of learning and knowledge application. Despite significant recent progress, however, no sufficiently specific model of knowledge structuring and accessing exists now which can serve as a basis for detailed diagnostic assessment of conceptual understanding.

1.2. Understand the Need

This section provides a brief outline of research that has been conducted into the nature of physics knowledge, the development of student models, and attempts at developing new assessment approaches. The purpose of the review is to set the context for the ConMap study, not to provide a comprehensive bibliography. It is a difficulty of the field that relevant research is scattered among disparate disciplines like physics, cognitive science, psychology, education, computer science, and neurobiology, so the references contained in this review should be taken as a representative sampling rather than a comprehensive bibliography. Further references can be found within the referred-to papers.

1.2.1. Research on Knowledge Structure

It has been said that “knowledge representation is one of the thorniest issues in cognitive science” (Anderson 1993). Much of the thorniness is concerned with the details of knowledge representation at its lower, more primitive levels: the microscopic “constituents” of knowledge. At higher levels, some general perspectives have emerged on how physics domain knowledge should be represented.

Cognitive scientists make a distinction between two fundamental types of knowledge: *declarative* and *procedural* (Anderson 1993). In essence, declarative knowledge is explicit knowledge of facts, which can be stated or reported; procedural knowledge is tacit knowledge of how to perform operations, which can be demonstrated but not stated. Knowing that force is equal to the product of mass and acceleration is declarative knowledge; knowing how to draw a free-body diagram is procedural, although the individual might also know several declarative facts about how free-body diagrams ought to be drawn. Frequently, knowledge is acquired first as declarative elements which are consciously consulted to carry out operations. With practice and repetition, those operations become automated as procedural elements, and the declarative elements are no longer required. With time, declarative elements may atrophy.

Physics knowledge involves both declarative and procedural components interacting. Most research to date on the nature of physics knowledge structure has investigated declarative knowledge, probably because it is easier to probe. Studies on physics experts’ and novices’ problem-solving behavior suggest that at least within the domain of physics, declarative knowledge can be represented as divided into four general, approximate categories (Larkin 1979; Chi, Feltovich et al. 1981; Gerace 1992; Mestre, Dufresne et al. 1993; Gerace, Leonard et al. 1997): *conceptual* knowledge, *operational and procedural* knowledge, *problem-state* knowledge, and *strategic* knowledge. (The category of “operational and procedural" knowledge refers to declarative knowledge about physics operations and procedures, as distinct from automated, non-declarative "procedural knowledge". The choice of terminology is unfortunate, especially since many operational skills have both declarative and procedural components.) Figure 1.1 depicts a representation of an expert's knowledge store, omitting strategic knowledge.

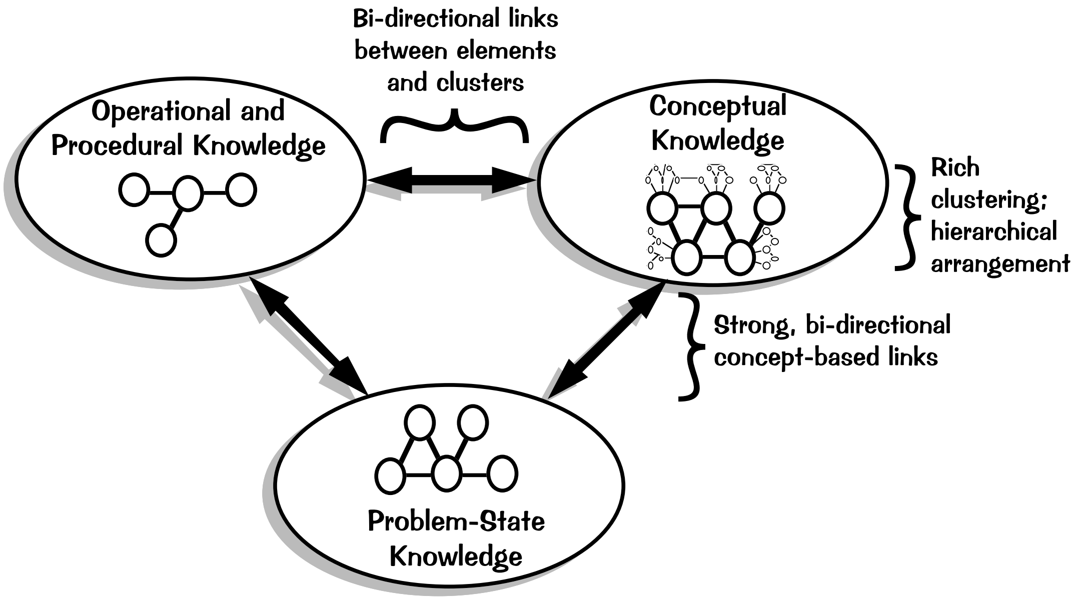

Bi-directional links between elements and clusters
Operational and Procedural Knowledge
Conceptual Knowledge
Rich clustering; hierarchical arrangement
Strong, bi-directional concept-based links
Problem-State Knowledge

Figure 1.1: Graphic depiction of a physics expert's declarative knowledge store, showing conceptual, operational/procedural, and problem-state knowledge (Gerace, Leonard et al. 1997).

Characteristics of the physics knowledge store of importance to mastery of the domain are brought to light by comparisons between experts and novices thinking and problem-solving. Table 1.1 lists the main differences between experts' and novices' knowledge characteristics based on such studies.

Several studies have revealed that experts and novices are distinguished not just by the content of their knowledge stores, but by the organization (Zajchowski and Martin 1993): one needs contextually-appropriate *access* to, not just possession of, knowledge (Redish 1994); it is the *structure of interconnections* between knowledge elements which allows such access (Mestre and Touger 1989); and expert's knowledge is structured around key *principles* (Hardiman, Dufresne et al. 1989).

These findings suggest that for purposes of assessing students' degree of expertise with respect to a physics topic, a model is required that describes students' declarative knowledge state in terms of the knowledge elements present and especially the structure of interconnections between those elements.

Expert	Novice
Store of domain-specific knowledge	Sparse knowledge set
Knowledge richly interconnected	Knowledge mostly disconnected and amorphous
Knowledge hierarchically structured	Knowledge stored chronologically
Integrated multiple representations	Poorly formed and unrelated representations
Good recall	Poor recall

Table 1.1: Summary of the main differences between experts' and novices' declarative knowledge characteristics, from expert-novice problem-solving studies (Gerace, Leonard et al. 1997).

1.2.2. Research on Cognitive Modeling

A variety of knowledge models have been constructed by the cognitive science community, for such domains as physics, computer programming, chess, land navigation, maintenance of aircraft hydraulic systems, and electronic circuit design (Chipman, Nichols et al. 1995). These models tend to fall into two distinct categories: network models of declarative knowledge and rule-based models of procedural knowledge.

Semantic networks are the classic network model of knowledge (Bara 1995), originally developed for the analysis of natural language. A semantic network consists of a set of *nodes* interconnected by labeled, oriented *links*. The nodes typically represent concepts. The links represent relationships between the nodes, with a label describing the nature of the relationship. The links are oriented so that asymmetric relationships may be described. Figure 1.2 displays an example of a small semantic network.

The primary difficulty with semantic networks as knowledge models is that they are tremendously general: without specific rules to limit the labels allowed on links, almost any degree of conceptual complexity can be buried within a creative choice of label. Consider, for example, the number and subtlety of concepts subsumed in the simple declarative statement "Newton's laws describe force" from Figure 1.2. Nevertheless, semantic networks do capture the general idea that declarative knowledge, or at least the conceptual part of it, gets its meaning from the interrelationships between conceptual elements.

Artificial neural networks (ANNs) form a class of models superficially similar to semantic networks, but with a very different origin and intent (McClelland and Rumelhart 1986; Hertz, Krogh et al. 1991; Watkin and Rau 1993). Inspired by the neurobiological functioning of the brain, ANNs were developed as an alternative computational paradigm to the traditional one where a single, complex processor executes one instruction at a time from a very large sequence of such instructions (a *program*). In the ANN paradigm, a large number of simple, identical processing elements (*nodes*, analogous to neurons) are interconnected by *links* (analogous to synapses) of varying strengths. The dynamics of the system is determined by the strengths of the links, and can be very complex.

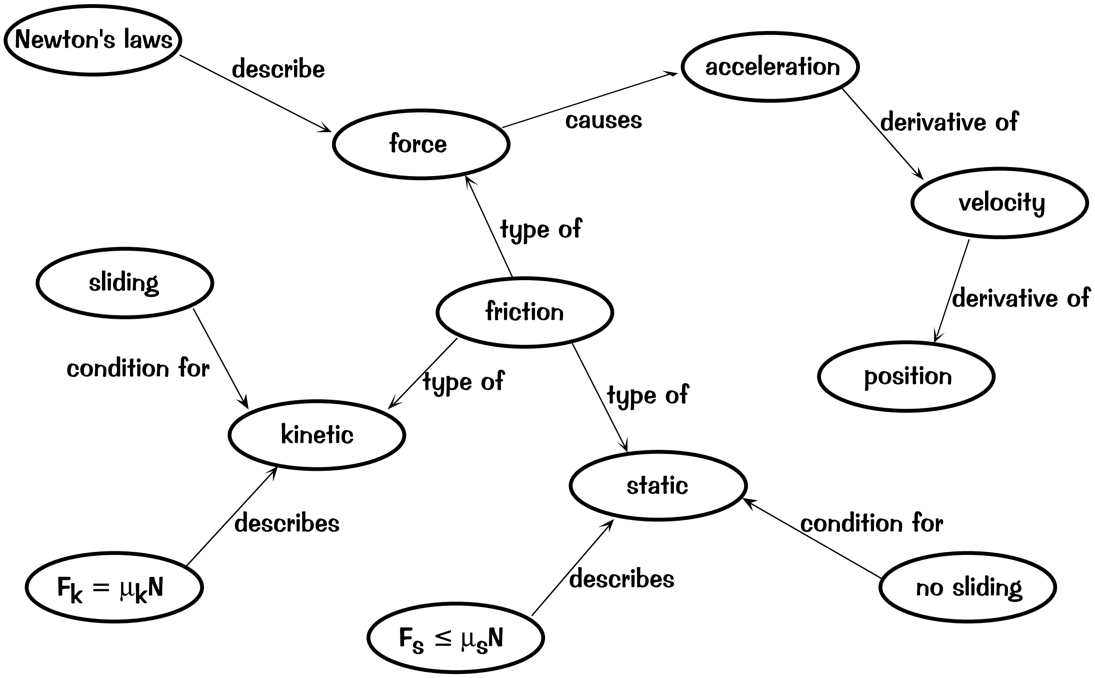


Figure 1.2: Example of a (partial) semantic network for concepts related to "friction".

Like a semantic network, an ANN consists of nodes interconnected by directed links. Unlike a semantic network, no explicit meaning is assigned to the nodes or links of an ANN; any meaning to the network's components and behavior must be assigned by some observer, and is not part of the model. The most significant difference between the two, however, is that semantic networks are static descriptions and ANNs have *dynamics*. That is, they have a *state* which *evolves* in time. The state of the network can be described by the subset of nodes which are in an "active" state (analogous to a neuron firing). The dynamics of each node and the set of links and link strengths determine which nodes become active or cease being active as time progresses.

Most research on applications of ANNs to cognition has concentrated on modeling of very specific, low-level neural and cognitive processes like associative recall (Hopfield 1982; Hopfield 1984), visual processing (Bialek and Zee 1988; Kohonen 1993), short-term memory (Lisman and Idiart 1995), and behaviorist stimulus-response training (Donahoe, Burgos et al. 1993). Some attempts to construct models of larger, multiple-function systems have been made, but they more resemble electronic circuit design than cognitive modeling (Trehub 1991) and still describe a low level of neural detail. A stumbling block in the application of ANN models to practical physics education research is that ANN dynamics does not itself describe observable behavior, but needs to be connected to observable behavior by additional model constructs.

In contrast to the network-based knowledge models are the rule-based models, most of which are termed *production systems*. These models describe procedural knowledge as a large set of simple if-then rules that act in concert to describe decision-making and action. The most complete and well-developed such model (Chipman, Nichols et al. 1995) is probably John R. Anderson's ACT-R system (Anderson 1993; Corbett, Anderson et al. 1995). ACT-R is a complete, runnable model implemented on a computer, capable of modeling computer programming, game-playing, and some other skilled pursuits for which rule sets have been developed. It has been used with some success as the basis for intelligent tutoring systems.

Although ACT-R's focus is procedural knowledge, it contains an ANN-style component to model declarative knowledge. The model's primary practical limitation is the amount of effort required to apply the model to a new domain like problem-solving in introductory physics, because the set of rules for a domain must be carefully crafted and tested. In contrast, John H. Holland's *Classifier System* models (Holland, Holyoak et al. 1986; Holland 1990) incorporate a dynamic for evolving the rule set according to external reinforcement, so that the rule set may be "trained" through trial and error. Classifier systems therefore attempt to model learning, whereas production systems like ACT-R model knowledge application.

Taking a more ad-hoc approach, Martin and VanLehn (Martin and VanLehn 1995) have attempted to develop a computerized physics assessment tool named OLAE which models students by inferring which of a large library of physics knowledge "rules" they know. The library is static and predesigned, and contains incorrect as well as correct "knowledge". Unlike production systems and classifier systems, the model is purely descriptive and has no dynamics.

1.2.3. Research on Assessment

A major realm of assessment research, called *psychometric test theory*, aims to extract more complete, detailed, and reliable information from traditional problem-based exams (Mislevy 1993; Corter 1995; Tatsuoka 1995). This approach applies sophisticated statistical analysis to the selection of test problems and to the investigation of patterns in a population of students' results. It has some implications for cognitive science because analysis of a population's test results can reveal the atomic *psychometric attributes* which serve as the cognitive elements that individual students either do or do not know. Psychometric attributes are purely empirical entities, revealed via statistics and not based on any *a priori* cognitive perspective; as such, they may serve to guide efforts in cognitive model-building for a domain. However, because its intent is to devise tests which discriminate between students for the purposes of evaluation and ranking, the theory focuses on differences between individuals rather than on their commonalties. Information about the commonalties would be most useful for the construction of cognitive models.

Psychometric testing is essentially nothing more than traditional exam preparation and interpretation stretched to its limits. Although the statistical techniques developed are likely to be useful for other kinds of assessment, nothing qualitatively new is offered.

The alternative to traditional exams with psychometric analysis is loosely termed cognitive test theory, because it builds on the results and models of cognitive science (Chipman, Nichols et al. 1995). Assessment methods within this category generally posit a knowledge model, and attempt to elicit or probe information about a student's knowledge structure within the framework defined by that model. If a network-based model like a semantic network is assumed, assessment instruments and analyses are developed to infer a student's set of concepts and inter-concept connections. If a rule-based model like a production system is assumed, assessment instruments and analyses attempt to identify, perhaps probabilistically, the subset of rules within the model that the student "knows".

For assessment of declarative knowledge, assuming a network model of knowledge structure, approaches can be categorized as either *direct* or *inferred*. The classic direct approach is the *concept map* task, in which a student is asked to draw a nodes-and-links representation of his or her understanding of a domain topic area (Novak and Gowin 1984). The resulting map is taken as a description, perhaps partial, of the student's declarative knowledge structure for the subject. Many variants of the task have been developed and investigated (Ruiz-Primo and Shavelson 1996). Sometimes subjects are asked to draw the entire map without assistance, with labeled or unlabeled links; sometimes they are given a set of terms to arrange into a map; sometimes a partial map is given and they are asked to fill in the remainder; sometimes a complete map without link labels is presented and subjects are asked to label all links. Scoring systems also vary widely, with credit given for the number of nodes, the number of links, the number of nodes or links deemed "relevant", the degree of similarity to a reference map, or some combination of these possibilities.

In general, research has found concept map assessments can be generally valid, in the sense that scores appropriately derived from students' concept maps tend to correlate with other, trusted indicators of student domain mastery (Young 1993; Ruiz-Primo and Shavelson 1996; Rice, Ryan et al. 1998). Instructors and researchers employing concept map assessments have also found that such assessments tend to be tedious and time-consuming to administer and to score and analyze, rendering them poorly suited for mass adoption by educators (Regis, Albertazzi et al. 1996; Ruiz-Primo and Shavelson 1996). Some researchers have implemented concept map assessments by computer and automated the scoring procedures (Ju 1989), but so far no widely adopted assessment tools have resulted, perhaps because of doubts about the reliability of the scoring protocols chosen.

Regardless of whether concept map tasks can be developed into useful assessment methodologies to supplement traditional exams, they have shown great promise as pedagogic tools (Cliburn 1990) and research instruments (Nakhleh 1994; Trowbridge and Wandersee 1996).

Whether or not students are capable of drawing a concept map that accurately describes their actual knowledge structure is open to significant doubt. One likely reason why the answer might be negative is that drawing a concept map is a time-consuming and attention-intensive activity, and a student is unlikely to be able to draw a map of any completeness for more than a very small set of concepts. In an attempt to probe students' domain knowledge more thoroughly, and to capture information about the relative strengths of inter-concept links as well as the presence or absence of such links, inferred approaches to declarative knowledge assessment have been developed. Inferred approaches typically follow a three-step paradigm (Goldsmith, Johnson et al. 1991):

1. Elicit raw data on knowledge structure via some kind of association-probing task;

2. Re-represent that data to reveal the underlying organization imposed by the student;

3. Evaluate this representation according to the assessment's objectives.

Association-probing tasks that have been used for step 1 include:

- Sorting tasks, such as partitioning a set of items into categories and sub-categories (Konold and Bates 1982);

- word association, in which subjects reply to a prompt term with whatever term comes to mind (Johnson 1964; Cooke, Durso et al. 1986);
- term proximities in free prose responses, in which subjects write essay-style responses to questions, and the average distances between various pairs of relevant words are calculated (Miyamoto, Suga et al. 1990); and
- item relatedness judgments, in which subjects are presented with pairs of items (typically words or terms) and asked to rate the relatedness of each pair on a numerical scale (Goldsmith, Johnson et al. 1991; Gonzalvo, Cañas et al. 1994).

Typically, the data from step 1 consists of a matrix of numbers which represent the relatedness of every pair of items in the task. For an item relatedness judgment task, these numbers might be the rating given to each pair by the subject; for other tasks, some calculation might be necessary to arrive at a matrix representation. Two general re-representation techniques are common for step 2: scaling procedures and network-construction algorithms. A scaling procedure interprets the matrix numbers as measures of distance (perhaps the inverse of distance), conceives of the items as occupying a point in a multidimensional space, and analyzes the distribution of the points in that space (Cooke, Durso et al. 1986; ter Braak 1995). *Cluster analysis* is a scaling procedure which groups the items into clusters, and then into clusters of clusters, and so on based on their relative distances in the space. *Multidimensional scaling* is a different scaling procedure which attempts to find the minimum dimensionality space necessary to adequately embed the items, preserving their relative proximities, and then enable attribution of meaning to each of the dimensions (Goldsmith, Johnson et al. 1991; Gonzalvo, Cañas et al. 1994; Johnson, Goldsmith et al. 1995). Scaling procedures are sometimes applied to the aggregated data of an entire population rather than of individuals (Mashhadi and Woolnough 1996), although its value for individual assessment is then lost.

Rather than representing items as points embedded in a multidimensional space, network-construction algorithms attempt to represent items as a network of nodes, assigning link strengths consistent with the proximity matrix values. The most common algorithm is the Pathfinder algorithm (Cooke, Durso et al. 1986; Goldsmith, Johnson et al. 1991; Gonzalvo, Cañas et al. 1994; Johnson, Goldsmith et al. 1995), although others have been used (Britton and Tidwell 1995). Network representations tend to be easier to interpret than scaling representations, and they have the advantage that they reveal but do not impose a hierarchical organization of the items. Cluster analysis, on the other hand, imposes such an organization regardless of the data.

The methods used for step 3 of the inferred approach, evaluation of the re-represented data, depends on the specific goals of the assessment. Qualitative judgments by domain experts may be used. It is common to compare the generated network or scaling procedure result to a reference network or embedding determined from a panel of experts, and define various quantitative measures of the degree of similarity (Goldsmith, Johnson et al. 1991; Gonzalvo, Cañas et al. 1994; Johnson, Goldsmith et al. 1995). If a network-construction algorithm has been employed, the problem of evaluation is now similar to that for explicitly-drawn concept maps, and techniques developed for that problem may be employed.

Overall, investigations into the validity of inferred approaches to declarative knowledge structure assessment have been generally positive: measures comparing the similarity of students' derived structures (networks or scaling procedure results) to experts' referent structures correlate significantly, but not completely, with more traditional measures of domain mastery (Goldsmith, Johnson et al. 1991; Gonzalvo, Cañas et al. 1994; Johnson, Goldsmith et al. 1995). Much of the research concerns itself with what comparison measures, or combination of comparison measures, produces the cleanest correlation. Some studies suggest that network-construction algorithms and scaling procedures are sensitive to somewhat different and complementary aspects of knowledge organization, and that a combination of the two approaches is more reliable than either alone.

Researchers who adopt a rule-based model of knowledge employ very different assessment methods. Since rule-based models describe tacit procedural knowledge rather than introspectively accessible declarative knowledge, they must observe the task performance of subjects engaged in activity, and infer from observations the rule set responsible for the performance (Martin and VanLehn 1995; Mislevy 1995). Typically, a "rule space" is defined, representing all possible knowledge states that a subject might have with respect to a specified set of rules — that is, every possible combination of knowing some rules and not knowing others. Bayesian inference methods are then applied to assign, based on the subject's observed behavior, a probability to each point in the rules space describing the likelihood that that point describes the subject's state of knowledge. In practice, such calculations are numerically very intensive, and much of the research surrounding such assessment methods aims to find algorithms for reducing the complexity of the calculations to manageable levels.

When applied to relatively simple models of skill application in relatively limited domains, Bayesian inference assessment methods have proven to be successful enough to incorporate into intelligent tutoring systems. The immediate limitation which prevents their utility for more ambitious rule-based models of larger domains, such as physics problem solving, seems to be the numerical complexity of the calculations required for the Bayesian inference. Much of the research surrounding such assessment methods aims to find algorithms for reducing the complexity of the calculations to manageable levels, and as affordable computers become ever more powerful, this limitation should recede.

1.3. Meet the Need

It has been argued that physics education needs new assessment methods, and that new assessments must be based on better models of physics learning and expertise than are currently available. It is also true that the development of better models depends upon data derived from appropriate assessment techniques. This is the same "chicken and egg" problem as confronts physics research: theory both depends upon and guides experiment, as depicted in Figure 1.3.

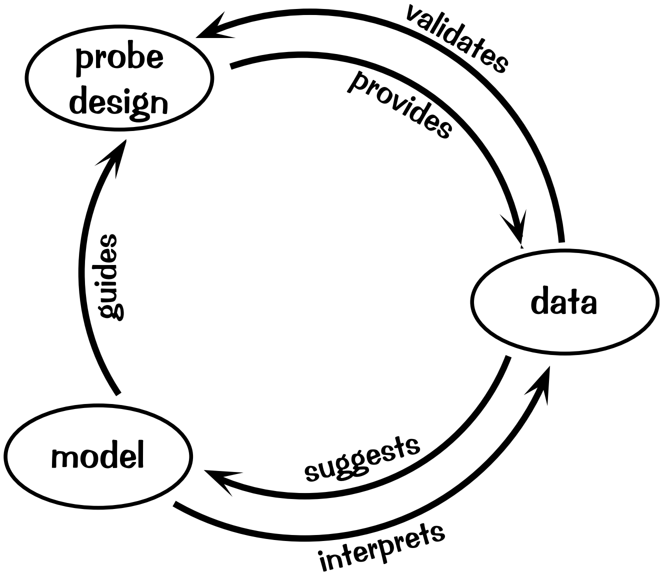


Figure 1.3: Interrelationships between the design of an experimental probe, the data gathered and the model constructed.

In mature research fields, focusing on one stage at a time — designing a better experimental probe, for example, or revising a model — is possible and generally recommended. In a newly emerging field, however, wherein researchers are still struggling to determine what the measurable and modelable quantities might be and how they ought to be represented, the three stages cannot be separated so cleanly. One's model of the system being studied, as preliminary and vague as it might be, guides the design of an experimental probe, and analysis of the data provided by that probe continuously modifies the model; at the same time, a viable model provides an interpretation to the data, and interpretable data validates the probe design used.

With this in mind, the following interrelated goals were set for the ConMap project:

1. Devise and test possible strategies for probing physics students' knowledge structures (probe design);

2. Search for potentially meaningful patterns and correlations in data provided by the probes (data analysis); and

3. Develop quantitative models of knowledge structure and access consistent with the gathered data (modeling).

Chapter 2 describes the design of the ConMap tasks (probes) devised, and the design of the research studies which investigated the application of those probes to physics students. Chapter 3 presents extensive analysis of the data from those studies. Chapter 4 describes some preliminary attempts at constructing quantitative models for a physics student's performance on the tasks. Finally, Chapter 5 summarizes and discusses the findings of the study.

Two important characteristics of the research should be borne in mind. First, the ConMap project represents an initial, exploratory foray into conceptual knowledge assessment and modeling in physics. As such, trying a variety of approaches and possibilities was given higher priority than performing a thorough, well-controlled test of a few hypotheses. Subsequent research to follow up any interesting findings has always been intended.

Second, the project was directed towards quantitative measures and models. Although some qualitative analysis is inevitable when dealing with concepts and similar difficult-to-quantify entities, it was a goal of the project to bring as quantitative and physics-like as possible an approach to physics education research. A great deal of potentially fascinating qualitative analysis of the data was therefore eschewed in favor of the search for revealing quantitative measures and correlations.

2. ConMap Design

Section 2.1 of this chapter presents the design of the ConMap tasks, the experimental "probes" being investigated for their utility for assessing physics students' knowledge structures. Section 2.2 describes the ConMap studies, in which the designed tasks were presented to physics students and task results were gathered. Section 2.3 presents some reflections on the task and study designs, based on the administrators' experiences conducting the studies (not on analysis of the resulting data).

2.1. ConMap Tasks

2.1.1. Design Objectives

As stated in Chapter 1, the intent of the ConMap project is to investigate the utility of a particular set of proposed assessment tools — brief, computer-administered tasks for eliciting spontaneous conceptual associations — for probing the quality and extent of a physics student's conceptual knowledge structure (CKS) in an introductory physics domain.

Ultimately, an ideal assessment task would provide a complete "map" of a student's knowledge structure, indicating connections that ought to be present for expert-like knowledge but were not, and connections that existed but should not (misconceptions). With such a student-specific map, an instructor could design specific pedagogic activities to benefit that student. More modestly and immediately, one might hope to develop assessment measures which characterize general qualities of a student's knowledge structure: how extensive it is, whether it is richly or sparsely interconnected, whether the organizational scheme is haphazard or systematic, and so on. The ConMap tasks were designed with both the ultimate and immediate goals in mind.

Why elicit conceptual associations? As discussed in Subsection 1.2.1, it is believed that a major component of domain expertise in physics is the possession of a richly interconnected, hierarchically organized network of associated concepts. A primary goal of physics instruction is to facilitate a learner's development of such a conceptual "understanding". ConMap tasks were therefore designed to elicit, as directly as possible, information on a subject's possession of concepts and inter-concept associations.

Why elicit spontaneous conceptual associations? It was desired that the tasks probe conceptual associations that were "readily accessible" to the subject, in the belief that such associations represent the automated knowledge inherent in the subject's knowledge structure. Therefore, task designs in which a subject was asked to reflect upon his or her conceptual associations were avoided, in favor tasks eliciting spontaneous associations. An additional consideration was that tasks involving a subject's considered judgments on his or her associations can confound measurement of the actual knowledge structure with measurement of the subject's introspective abilities and inventiveness. (Hand-drawn concept maps, which do not address spontaneous associations, were included in the study for comparison purposes.)

Why use computer-administered tasks? ConMap is not primarily intended to be abstract cognitive science research, but rather a step towards the development of new, practical assessment methods which are grounded in cognitive science findings. It is hoped that the tasks developed or their offspring may ultimately have practical value to educators by providing meaningful and timely diagnosis of students. If the tasks are not easy to administer and evaluate, they will not be useful.

Why use brief tasks? The first practical use of such tasks would likely be for ongoing diagnostic assessment during a course of instruction, rather than for occasional comprehensive evaluations. The tasks must therefore be relatively brief and non-demanding. This rules out more comprehensive, detailed, and thorough tasks of the types often employed for cognitive science research.

2.1.2. Terms as Concepts

In order to probe the conceptual portion of declarative knowledge, most of the ConMap tasks attempt to elicit subject's associations between *terms*. The focus is on terms rather than on equations, propositions, or other kinds of entities because terms seem to be the closest accessible approximation to "conceptual building blocks". This study is not concerned with the underlying cognitive nature of such building blocks, or with the neurological details of their representation, storage, and retrieval.

It has proven difficult to rigorously define *term*. When instructing subjects, a term was loosely defined to be one or perhaps two or three words describing one concept, idea, or thing. Some examples of terms drawn from introductory mechanics are:

- kinematics
- Newton's first law
- velocity
- pulley

- second
- impulse-momentum theorem
- problem-solving

Statements like “energy is conserved in an elastic collision” were not considered to be terms, but rather propositions involving multiple terms and their relationship. “Conservation of energy”, on the other hand, would be accepted as a term, since it serves as a name for a physics concept. In practice, the line between single-concept terms and compound statements of relationship is not well-defined, and subjects frequently wandered dismayingly far over it.

2.1.3. The Specific Tasks

The ConMap tasks were developed with the intention of probing the set of terms and inter-term associations that constitute the conceptual portion of a subject’s declarative knowledge store for a domain. Each task was intended to elicit a somewhat different aspect of that knowledge store. The following subsubsections describe the tasks that were developed and investigated.

2.1.3.1. Free Term Entry (FTE)

For the *Free Term Entry* (FTE) task, subjects are given a general topic area like “introductory mechanics” or “the material covered in Physics 151”. They are asked to think of terms that they associate with this topic area, spontaneously and without strategy, and to type these terms into a dialog box (shown in Figure 2.1) as the terms come to mind. When each term is completed, the subject presses the “return” key on the keyboard (equivalent to clicking on the “Enter” button in the dialog box), and the typed term disappears, leaving the typing box empty and ready for a new term.

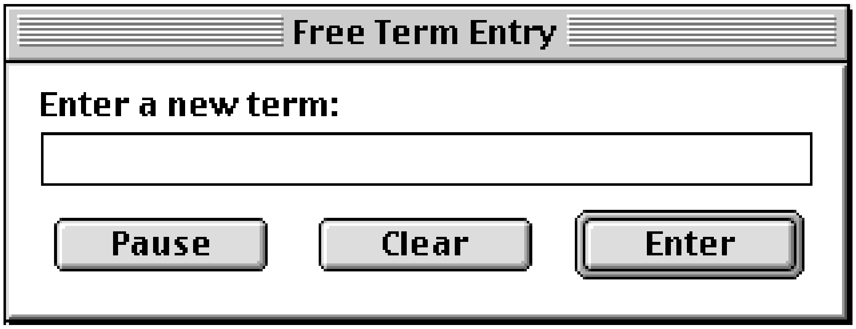


Figure 2.1: Term entry dialog box for the FTE task.

The data gathered consist of the list of terms in the order they were entered, together with the time at which typing began for each term (the moment at which the first character was typed into an empty field), and the time at which each term was entered (the moment at which the return key or "Enter" button was pressed). The task runs for a specified total duration, typically 20 to 45 minutes, before terminating.

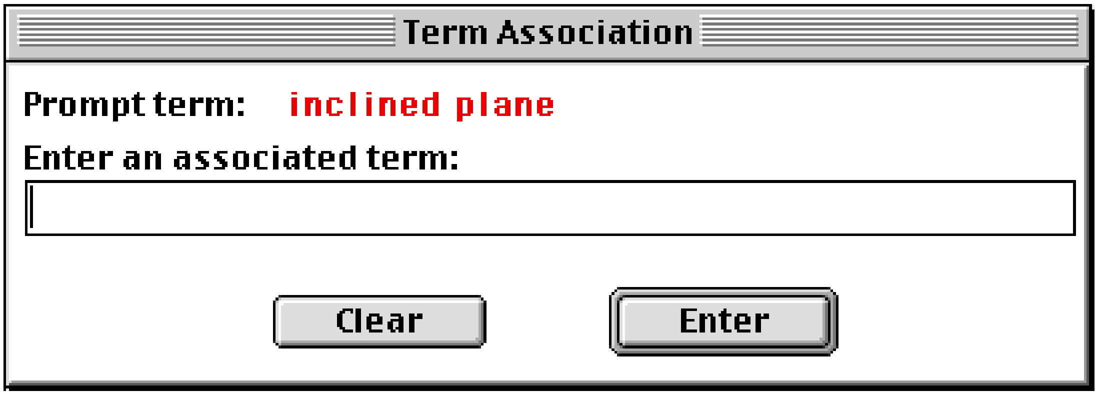
This task was intended to explore the space of terms constituting a subject's active vocabulary of concept-describing terms for the topic area, without influencing the responses by providing terms through external prompting. The result was conceptualized as a kind of "random walk" through the space of a subject's active vocabulary. It was hoped that the duration of pauses between term entries, and the grouping of term entries into clusters separated by longer fallow periods, might reveal some information about what terms a subject associates closely. Since the list of terms and times comprising a FTE data set forms a one-dimensional series, and the space of conceptual knowledge elements and their interconnections requires two dimensions to represent (for example, as a matrix of connection strengths), it was clear from the beginning that the FTE task can never provide a complete probe of a subject's conceptual interconnections. Nevertheless, it was a first attempt at exploring the space. In addition, it was hoped that overall statistical patterns in a subject's FTE data might reveal global aspects of his or her knowledge and cognition, perhaps serving as bulk measures in much the way that thermodynamic quantities like temperature and pressure characterize global statistical aspects of a collection of microscopic particles.

2.1.3.2. Term-Prompted Term Entry (TPTE)

For the *Term-Prompted Term Entry* (TPTE) task, subjects are given a prompt term. They are asked to think of terms they consider related to this prompt term, spontaneously and without strategy, and to type these terms into a dialog box (shown in Figure 2.2) as the terms come to mind. The prompt term stays visible throughout, and typed terms disappear from view as they are entered. Data gathered is the same list of terms, term start times, and term entering times as for the FTE task. The process repeats for several different prompt terms.

Several schemes for terminating a subject's entering of response terms have been considered. In initial trials, a subject's entering of response terms was terminated the first time ten seconds of inactivity was detected while the term-entry field was empty, on the assumption that this indicated the subject was having difficulty thinking of another relevant term to enter. The task would also be terminated after the tenth response term entry. For later trials, the task was terminated by the same criteria, except that the task would not terminate until at least three terms had been entered.

The TPTE task was intended to explore subjects' conceptual associations in a more focused and directed way than the FTE task allows, eliciting the strongest associations a subject has with a particular prompt term. A mode of operation envisioned (but not implemented in any studies to date) was to first give a subject the FTE task on a topic, and then use the set of responses gathered as TPTE prompt terms to fill out a web of connections between those terms. It was hoped that such a procedure might allow computer-based elicitation and construction of a "concept map" representation of a subject's knowledge of a topic, in a manner more spontaneous and therefore presumably more genuine than occurs for traditional hand-drawn concept map tasks.



Prompt term: inclined plane

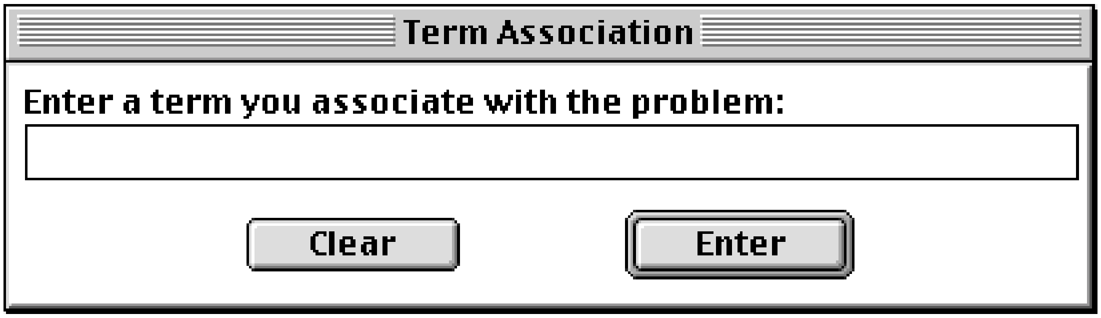
Figure 2.2: Term entry dialog for the TPTE task.

2.1.3.3. Problem-Prompted Term Entry (PPTE)

The *Problem-Prompted Term Entry* (PPTE) task is identical to the TPTE task, except that instead of being prompted with a term, subjects are directed to read the description of a problem or problem situation. Subjects then enter terms they associate with the problem in a dialog box (see Figure 2.3). The process is repeated for several prompt problems. In all studies to date, prompt problems have been provided on paper in a ring binder, and the computer program implementing the task has instructed students when to turn the page and read a new problem.

The data gathered is identical to that gathered in for the TPTE task.

The PPTE task was intended to explore the relationship between problem solving and conceptual associations. By intent, most ConMap tasks target conceptual knowledge structure and ignore other skills and knowledge types relevant to problem-solving; the PPTE task is an exception in that it addresses the interface between conceptual knowledge and problem-state knowledge.



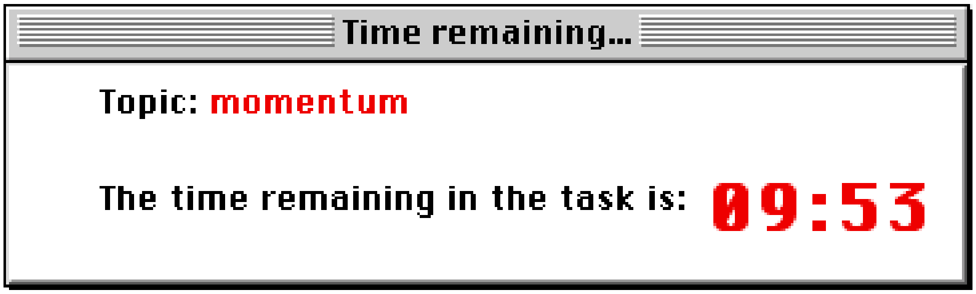
Enter a term you associate with the problem:

Figure 2.3: Term entry dialog for the PPTE task.

2.1.3.4. Hand-Drawn Concept Map (HDCM)

Strictly speaking, the *Hand-Drawn Concept Map* (HDCM) task is not considered to be a ConMap task, but rather a more traditional task which has been studied elsewhere (cf. Subsection 1.2.3) and which was included in the ConMap project for comparison purposes. Since variants of the HDCM task have been extensively researched, it was hoped that a comparison of HDCM data with data from other ConMap tasks for the same subjects and topic matter might aid in the interpretation of ConMap data.

In the variant of the HDCM task implemented here, subjects are instructed to draw by hand a concept map consisting of nodes that contain terms and links interconnecting those nodes. A prompt term is provided as the central node, defining the topic to be mapped, and the subject must provide and choose all other terms. Links are not to be labeled, and the structure need not be hierarchical. The task runs for a specified time, typically eight or ten minutes, before terminating. The computer program presents task instructions and the topic to be mapped (the prompt term) and counts down the allowed time with a visible display (see Figure 2.4), but the map is drawn in pencil on 11 inch by 17 inch paper.



Topic: momentum

The time remaining in the task is: 09:53

Figure 2.4: Prompt term and timer dialog for HDCM task.

2.1.3.5. Term-Prompted Statement Entry (TPSE)

For the *Term-Prompted Statement Entry* (TPSE) task, subjects are given a prompt term as in the TPTE task, but instead of entering isolated terms for responses, they are asked to enter statements — complete sentences — which state important "things" (facts, ideas, relationships) about or involving the prompt term. Subjects are given a specific amount of time per prompt (four minutes in the studies conducted). This task was incorporated in only one of the studies performed to date, with five prompt terms distributed among two sessions. It was implemented as a paper-and-pencil task with a human administrator keeping time, with no computer involvement whatsoever; a computer implementation may eventually be developed, although subjects' typing skills may then be an issue. For each prompt term, subjects were given a sheet of paper with the prompt written at the top and nine bullets down the left side of the page. Subjects were instructed to write one statement next to each bullet, until nine had been written or until the session administrator terminated the task.

The objective of this task was to provide some insight into the meanings of the associations subjects make between terms. It was envisioned as supporting the TPTE task by providing a more thorough understanding of some of the term associations revealed by TPTE data and aiding in the interpretation of that data.

2.1.3.6. Term Proximity Judgment (TPJ)

An additional task which was considered but not implemented in any of the studies to date is the *Term Proximity Judgment* (TPJ) task, in which subjects are presented with every possible pairing of terms from a predefined list, one pairing at a time. For each pairing, subjects are asked to assign a "relatedness" rating from a numerical scale (e.g. integers from 1 to 9 inclusive). Tasks of this type have been employed extensively in other studies of alternative assessments (Goldsmith, Johnson et al. 1991; Gonzalvo, Cañas et al. 1994).

The TPJ task does not meet the criteria set for ConMap tasks: it requires an extensive amount of time to administer for any reasonable number of terms, it provides a set of externally selected terms to the subjects instead of allowing subjects to choose their own terms, and it asks subjects to reflectively judge term relatedness rather than eliciting spontaneous term associations. Nevertheless, it could provide valuable data for research into cognition and the utility of ConMap tasks; its primary virtue is its capability of providing a complete set of connection data for all possible pairings of a term set.

This virtue imposes a significant drawback: for a set of N terms, it requires order N^2 queries of the subject to elicit complete data. If some independent basis can be found for ruling out a large fraction of the pairings as uninteresting, the usefulness of the task would be greatly increased. No such basis currently exists in the context of the ConMap studies, so the TPJ task has not included in any studies yet conducted.

2.2. ConMap Studies

As part of the ongoing ConMap project, several studies have been conducted, with varying population sizes, duration, task inclusion, and degree of planning. Many of the "studies" were not intended to provide rigorous data for full analysis, but rather to furnish preliminary data and experience as an aid to the design of more reliable studies. Only two contained a large enough population of subjects for serious analysis, and they provided most of the data for Chapter 3.

The *preliminary study* was actually no more than a collection of test cases. It consisted of the loose and informal presentation of various tasks, predominantly the FTE, to various individuals of various backgrounds and levels of expertise, in a nonsystematic way, under inconsistently-controlled conditions. The objective was to test and debug tasks.

The *Physics 119 Fall 1997* (p119f97) study drafted all 8 students from Physics 119/597T (introductory mechanics for prospective high-school physics teachers, taught by Profs. William Gerace and Robert Dufresne). The study consisted of one FTE task on "energy" given near the end of the course. As part of the course itself, a HDCM task was given to students by the course instructors.

The *Physics 152 Fall 1997* (p152f97) study selected 18 subjects for pay from a pool of volunteers enrolled in Physics 152 (thermodynamics, electricity and magnetism for engineers, taught by Prof. Jose Mestre). The study consisted of one FTE task on the entire course domain, given at the end of the course.

The *Physics 151/2 Summer 1998* (p15Xs98) study recruited five subjects for pay from the students enrolled in the summer sessions of both Physics 151 (introductory mechanics for engineers) and Physics 152; one of the five recruits did not complete the study. The study consisted of two sessions, each a battery of multiple tasks. The first session was given between the end of Physics 151 and the beginning of Physics 152, mostly on p151 material, with a "pre-test" FTE task on p152 material. The second session was given at end of Physics 152, on p152 material. An interview with subjects was conducted and recorded after each session.

The *Physics 151 Spring 1999* (p151s99) study selected sixteen subjects for pay from a pool of volunteers taking Physics 151 (taught by Prof. Jose Mestre). The study consisted of ten weekly sessions during the semester, each of fifteen minutes duration, except for the last which lasted 1.5 hours. A variety of tasks and prompts and topics was presented throughout, with significant repetition designed in.

The *Physics 172 Spring 1999* (p172s99) study selected five subjects for pay from volunteers taking Physics 172 (thermodynamics, waves, fluids, and other miscellaneous topics, taught by Prof. Ross Hicks) who had taken Physics 171 (introductory mechanics for physics majors, taught by Prof. Gerace) the previous semester. The study consisted of two sessions. The first included various tasks and prompts given during the first few sessions of the p151s99 study. The second was identical to the final session of the p151s99 study, except for the inclusion of some additional TPTE prompt terms. The intention of the study was to provide some comparison data for the p151s99 study. Differences were expected because of (among other factors) the fact that the Physics 171 course subjects had taken was highly focused on conceptual structuring.

The following subsections detail the three studies from which the data analyzed in Chapter 3 were drawn.

2.2.1. Physics 152 Fall 1997 (p152f97) Study

The purpose of this study was straightforward: to obtain FTE task data from a relatively large sample of subjects from a typical physics course, and see what the data might reveal.

Volunteers were sought from the body of students taking Physics 152 during the Fall of 1997, taught by Prof. Jose Mestre. The course was aimed at engineering and physical science majors, and covered thermodynamics for approximately three weeks, and then spent the rest of the semester on electricity and magnetism. Volunteers were solicited, and asked to fill out an application form, near the end of the semester; pay was offered. Most of the subjects knew the author as their discussion section teaching assistant for the course and had a friendly relationship with him. Twenty of the volunteers were selected and scheduled, based on their grades on the three course exams that had been given and on their availability during the time slots planned. The selection was made in an attempt to get a study population with a relatively uniform distribution of course exam grades, from the "D" level to the "A" level. Of the twenty, eighteen showed up for their scheduled session.

Sessions were conducted with four or five subjects simultaneously, in the same room, each with a separate computer. All sessions occurred during the same day, a Saturday after the end of classes and before the beginning of final exams. The computers were arranged to prevent any subject from easily or accidentally viewing another subject's display. Subjects were first given a brief written questionnaire designed to extract some basic profile information. They were then given verbal instructions for the task and allowed to ask questions.

After that, they were verbally given the topic area for the FTE task, and instructed to commence. The task was halted after approximately 45 or 50 minutes, but the administrator waited until none of the present subjects had entered a term for several seconds, so as not to interrupt any subject in the middle of entering a cluster of terms.

2.2.2. Physics 151 Spring 1999 (p151s99) Study

The overall purpose of this study was to examine in more detail what kinds of information the different ConMap tasks can reveal about subjects, to what extent that information is consistent with traditional assessments and drawn concept maps, and whether the tasks can detect the signature of evolving knowledge as a course progresses.

After the first course exam, volunteers were solicited from the body of students taking Physics 151 in the Spring of 1999, taught by Prof. Jose Mestre. The course was aimed at engineering and physical science majors, and covered introductory mechanics. Volunteers were offered financial compensation for their participation. Subjects were chosen from the resulting pool of volunteers, with the following objectives:

1. Have the subject population's grades on the first course exam span the range from slightly below course average and up, with a reasonably uniform distribution, on the assumption that low-performing students are less likely to take the course and the study seriously;

2. Attain as even a distribution of men and women as possible, to allow investigation of correlations with gender;

3. Omit subjects with poor English skills (determined during telephone contact), since most of the study tasks are verbal and require reading and writing facility.

A total of sixteen subjects were selected and scheduled, all of whom completed the study. One of the subjects was not given one of the tasks during the final session, due to a logistical error.

Ten weekly sessions were held, one every class week from March first through the end of the semester. The first nine sessions lasted approximately fifteen minutes each, suitable for one or two tasks. The last session lasted approximately 1.5 hours.

Session A (starting 3/02) was intended to familiarize subjects with the tasks. The TPTE included a majority of non-physics terms. After the two tasks, subjects were given instructions for the HDCM task but did not actually draw a map, to save time during the next session.

- *Task 1:* TPTE (food, travel, democracy, tree, acceleration, vector)
- *Task 2:* PPTE

Session B (starting 3/09) was intended to get some basic TPTE responses, mostly on kinematics. A HDCM was given to serve as a basis of comparison with an end-of-course HDCM using the same topic area, and to compare with TPTE data.

- *Task 1:* TPTE (displacement, force, free-fall, energy, acceleration, graph)
- *Task 2:* HDCM (force)

Session C (starting 3/23) presented TPTE and PPTE tasks focused on forces (previously covered in the course) and work and energy (being covered at that time). One PPTE prompt problem was presented as a "problem situation" without a question, to investigate how the presence or absence of a question impacts subjects' responses.

- *Task 1:* PPTE
- *Task 2:* TPTE (energy, force, inclined plane, equilibrium, reaction force, work, normal force)

Session D (starting 3/30, the week of Exam 2) was a follow-up to Session C, using many of the same prompts, to see how course coverage of the material and studying for the exam impacted task results. The "problem situation" from Session C was presented with a question, and another problem from Session C was presented as a situation without a question.

- *Task 1:* PPTE
- *Task 2:* TPTE (conservative, inclinded plane [sic], equilibrium, reaction force, work, spring, normal force)

Session E (starting 4/06) included as PPTE prompts two problems given on the second course exam, to compare PPTE responses to exam performance. TPTE problems were primarily drawn from momentum ideas, which the course was beginning its treatment of.

- *Task 1:* PPTE
- *Task 2:* TPTE (center of mass, power, momentum, range, collision, potential, impulse)

Session F (starting 4/13) was intended as a first trial of the TPSE task.

- *Task 1:* TPSE (acceleration, force, energy)

Session G (starting 4/20) was primarily intended to present a PPTE task with diagrams rather than with problems or problem situations.

- *Task 1:* PPTE
- *Task 2:* HDCM (energy)

Session H (starting 4/27, the week of Exam 3) revisited the momentum TPTE prompts used in Session E, to look for evolution in responses due to course coverage of the topic. A HDCM on "momentum" was presented to allow another comparison of TPTE to HDCM data.

- *Task 1:* TPTE (center of mass, power, momentum, range, collision, potential, impulse)
- *Task 2:* HDCM (momentum)

Session I (starting 5/04) presented problems from the third course exam as PPTE prompts, to allow comparison of exam performance with PPTE data. The TPSE task was tried again.

- *Task 1:* PPTE
- *Task 2:* TPSE (momentum, friction)

Session J (starting 5/11, the final week of classes) was a 1.5-hour session designed to cover many tasks and extensively repeat many previously-used prompts. The time-intensive FTE task was presented for the first and only time in the study. Subjects were also given a group interview on their perceptions of the tasks, and a questionnaire to elicit profile information. Because of last-minute scheduling difficulties, only ten of the sixteen subjects were given task 5, and one of the sixteen subjects was never given task 4.

- *Task 1:* FTE ("the material covered in Physics 151")

- *Interlude 1*: Group Interview
- *Task 2:* TPTE (inclined plane, conservative, rotation, vector, displacement, energy, force, graph, spring, free-fall, friction, velocity)
- *Interlude 2:* Profile Questionnaire
- *Task 3:* PPTE
- *Task 4:* HDCM (force)
- *Task 5:* HDCM (momentum) [only for some subjects]

Most of the student contact required by this study — contacting and scheduling subjects and administering sessions — was done by Dan Miller, another graduate student.

2.2.3. Physics 172 Spring 1999 (p172s99) Study

The purpose of this study was to give a subset of the tasks and prompts from the p151s99 to subjects expected to have more highly structured knowledge, in order to look for a signature of that structuring in the data.

Volunteers were solicited from the subset of students taking Physics 172 during Spring 1999 that had taken Physics 171 the previous semester. Physics 171 and 172 are the first two semesters of the introductory physics sequence for physics majors; 171 covers mechanics, and 172 covers thermodynamics, waves, fluids, and miscellaneous other topics. Physics 171 had been taught by Prof. William Gerace, and 172 was taught by Prof. Ross Hicks. No financial compensation was offered. Prof. Gerace chose five subjects from the pool of volunteers, attempting to get a reasonable distribution of ability levels based on his recollection of each student's overall performance in 171.

This particular population of students was targeted because in the Physics 171 course, Prof. Gerace strongly and explicitly emphasized the structuring of conceptual knowledge to students. It was hoped that this might leave an observable signature in subjects' ConMap task data. The study was conducted during the middle and end of the subsequent physics course due to constraints on the scheduling of the study, not for any specific research purpose.

Two sessions were held. The first was of about one-half hour duration, within a week of 4/22/99. The second was held during the final week of classes and lasted for approximately 1.5 hours.

Session A (starting 4/22) presented TPTE, PPTE, and HDCM tasks to subjects, using a subset of the prompts given in the p151s99 study.

- *Task 1:* TPTE (food, democracy, “big ideas” of mechanics, acceleration, force, inclined plane, energy, equilibrium, graph, momentum, collision, work, friction)

- *Task 2:* PPTE

- *Task 3:* HDCM (force)

Session B (starting 5/13) was nearly identical to Session J of the p151s99 study. Some additional TPTE terms drawn from Physics 172 course material were added for contrast, and the HDCM prompt was changed so that the two HDCM tasks given in the p172s99 study used different prompts.

- *Task 1:* FTE ("the material covered in Physics 171 last fall")

- *Interlude 1:* Group Interview

- *Task 2:* TPTE (inclined plane, conservative, rotation, vector, displacement, energy, force, graph, spring, free-fall, friction, velocity, wave, gravity, sound, light)

- *Interlude 2:* Profile Questionnaire

- *Task 3:* PPTE

- *Task 4:* HDCM (momentum)

Student contact for this study was also handled by Dan Miller.

2.3. Reflections on the Administration of Tasks

This section describes some of the difficulties encountered during administration of the tasks. Weaknesses of the task and study designs that were revealed during data analysis will be discussed in Chapter 3.

When administering term-entry tasks (FTE, TPTE, and PPTE), it was occasionally necessary to remind subjects to restrict themselves to physics terms. Subjects were sometimes inclined to include terms from chemistry, biology, or everyday experience. A small number of non-serious “joke” terms were entered. Some subjects included terms descriptive of the course and instructor as a whole rather than of the subject matter, especially during FTE tasks. Clearer, more explicit instructions with examples and counterexamples might be useful in this regard, but a design decision was made not to provide subjects with any examples even vaguely related to the domain being investigated, for fear of influencing their responses.

Despite reminders during several sessions, some subjects in the p151s99 study demonstrated a distressing tendency to include multi-term phrases and statements in term-entry tasks. Two examples are "depends on distribution of mass" (TPTE response to "center of mass") and "object in motion tends to stay in motion" (TPTE response to "momentum"). Note that for the second example, "Newton's first law" would communicate essentially the same meaning, but would be accepted as a single term (representing one principle); this second example might therefore be considered borderline legitimate. Reminding subjects more frequently to restrict themselves to isolated terms might help curb this tendency. It would be possible to restrict the length of response terms by having the computer program refuse to accept more than a set number of characters (perhaps 20), which might help to prevent the most flagrant offenses, but this option was rejected for fear that subjects might be confused and distracted from the task if the program did not act as they expected or caused them to wonder whether a term they thought of would fit.

A similar but less frequent problem occurred for the TPSE task: some subjects included occasional compound, multiple-clause, multiple-idea sentences as a single "statement". One example from the p151s99 study is "Friction does work on a moving object, but points in the opposite direction of motion"; another is "Kinetic energy is the energy of a moving object, when it is stationary, kinetic energy is zero [sic]." Subjects seemed able to avoid this when specifically reminded to.

A loophole was discovered in the termination scheme for TPTE and PPTE tasks. Normally, the task terminates if the subject is idle for ten seconds and a specified minimum number of terms (typically three) has been entered. In order to avoid interrupting a subject during the entry of a term, the ten-second cutoff acts only if the term entry box is empty, not if some text has been typed but not yet entered via the return key. Some subjects apparently discovered that they could type some text, think for any amount of time, and then delete the typed text and enter a term without having the task terminate. There is strong evidence that some subjects deliberately abused this loophole. Subject p151s99-14 had six thinking times greater than the ten-second cutoff in each of three different tasks. Several other subjects had as many as five, six or seven such illegal times for some tasks. Thinking times of 79.98 seconds and 155.54 seconds were seen. Fortunately, subjects seemed cooperative when they were informed of the loophole and asked to avoid exploiting it; deliberate abuse appeared to cease, and only a few instances, probably accidental, occurred thereafter.

When subjects are given a PPTE task, it is difficult to control how long they spend considering the problem before beginning to enter response terms. Since subjects' reading speeds and the complexity of the prompt problems varied significantly, it was difficult for an observer to estimate how contemplative subjects might be during the reading phase. It was not intended that subjects pre-think their responses at all, but some time was clearly necessary to "digest" the problem. This difficulty may be unavoidable when using complex prompts for spontaneous association tasks.

A related PPTE task complication arises when subjects wish to pause in their entry of response terms and remind themselves of some aspect of the prompt problem by looking back at it. Such a desire seems reasonable, given that subjects have been instructed to keep their responses relevant to the prompt problem and that they are unlikely to keep every detail of the problem in mind after one reading. For the TPTE task, the prompting term is kept visible and prominent directly above the term-entry box, and subjects are expected to glance at it frequently to re-focus themselves. For the PPTE task, however, a subject's re-reading of the problem can introduce a significant and difficult-to interpret thinking time into the data, or even cause the task to terminate. This problem might also be unavoidable when using a complex prompt for a spontaneous but constrained association task.

When carrying out the HDCM task, some subjects demonstrated a misunderstanding about how maps were supposed to be drawn. Some of the maps drawn had more than one node containing the same term, presumably either through forgetfulness or as a convenience for the mapmaker. Some maps had branching links which connected more than two nodes together. Subjects had read brief written instructions on how to draw a proper map, and had been shown an example map for a non-physics topic. More explicit instructions and training are apparently necessary.

Another difficulty which occurred with the HDCM task, and to a lesser extent with term entry tasks, was that subjects sometimes entered incomplete terms whose meaning was only clear from the context in which the fragment appeared. For example, a concept map might have "kinetic" and "static" as nodes connected to a node for "friction", and might have other nodes for "kinetic" and "potential", connected to "energy". Technically, this is a case of duplicate nodes, since two nodes both contain the term "kinetic". However, from their context, the two nodes clearly refer to different concepts: "kinetic friction" and "kinetic energy". Most such term fragments could be completed with ease by a domain expert during data analysis, but would pose a significant problem for computerized analysis procedures and for integrated ConMap systems which might, for example, use terms harvested from one task as prompts in another.

Better instruction and training of subjects might reduce the incidence of the problem, but would probably not eliminate it entirely, since subjects may not realize their term fragments are ambiguous.

For the long-duration FTE and HDCM tasks, subjects were intended to concentrate on the task until time expired, even if that required them to search their minds for minutes at a time to think of additional terms or map elaborations. Sometimes, however, subjects appeared to relax and cease working on such a task before time had run out, as if they had decided they had nothing more to add. This is perhaps not a very serious problem, since a subject inclined to make that decision might not have entered much more had they remained on task. The more general issue of subject attention and distraction plagues all studies in which subjects are required to concentrate for extended periods of time, and may be unavoidable.

For the most part, the problems noted during administration of the tasks were not major and did not appear to impact the data seriously enough to prevent the preliminary analysis intended. The one exception is subjects' exploitation of the TPTE/PPTE task termination loophole, which introduced a nontrivial number of spurious data points into the timing data. Most of the problems should be addressable through improved instruction and training procedures in future studies, and the remainder are probably unavoidable consequences of the natures of the tasks themselves.

3. ConMap Analysis

Data gathered during the ConMap studies has been subject to three different kinds of “analysis”: phenomenological description, hypothesis testing, and modeling.

A phenomenological description of the data is useful as a starting point for further analysis and for evaluation and modification of the tasks, and for the design of future studies. A major analytical goal within the ConMap project was to characterize the data returned by the various tasks, and to draw attention to features of the data which may be of significance and which may lend themselves to a model-based interpretation. The phenomenological analysis conducted focuses primarily on the patterns across subjects and sessions with the study populations, rather than differences between them, since the samples are small enough and the data noisy enough to prevent reliable identification of significant discrepancies between individuals. Some attempts were made, however, to find aspects of the data which distinguish subjects from each other and which might reveal natural variables for describing or categorizing subjects.

With data from some of the tasks, several hypotheses were investigated statistically. These hypotheses typically predicted a correlation between different aspects of the ConMap data, or between some aspect of the ConMap data and an external measure like course exam performance. The general purpose of such hypothesis testing is *validation* of the ConMap tasks as subject probes: demonstrating that patterns in the data captured by ConMap tasks are not mere statistical curiosities, but do in fact reveal something meaningful about subjects with some degree of reliability.

Evidence for validation can take two forms: *external* and *internal*. External validation uses an established measure like performance on an already-validated assessment as a standard of comparison. Validation of concept map and Pathfinder-based assessment methodologies, for example, generally compare measures of domain mastery extracted from the methodology with traditional exam scores. Internal validation uses the internal consistency of a set of data as evidence for its objective validity by demonstrating that relevant patterns are reliably robust and reproducible.

External validation is the stronger form. Unfortunately, it was not of much use with ConMap. The only available data from the ConMap studies that could be considered “external” to the ConMap tasks was subjects’ course grades. Of these, homework grades reflect effort more than knowledge state, and in-class quiz grades were collaborative and therefore corrupt as measures of individual knowledge, leaving only exam grades for comparison. But the original motivation of the ConMap product was the belief that traditional, problem-based exams serve as poor indicators of knowledge structure; exam grades are therefore not expected to correlate more than weakly with interesting measures from the ConMap data. While some attempts were made to compare ConMap measures against course exam performance, strong and compelling results were not expected.

It should be possible to construct exams or other instruments to probe knowledge content and structure more effectively than traditional exams; for research purposes, these instruments would not need to be constrained by standard course requirements for practicality (amount of student or evaluator time required, for example). Such instruments could in principle be used to validate ConMap-based assessments, although none were designed into the ConMap studies. Future studies should rectify this shortcoming.

The remainder of this chapter presents analysis of data from the ConMap studies. Each of the chapter's sections addresses one type of ConMap task. Data from the task are described and summarized, and phenomenological analysis of the data is presented. With data from the FTE, TPTE, and PPTE tasks, the in-depth investigation of some specific hypotheses is discussed. For all tasks, suggestions for follow-up studies are made, including recommendations for design changes to rectify inadequacies discovered in the present study's design.

3.1. Free Term Entry (FTE) Data Analysis

As described in Subsubsection 2.1.3.1, for a Free Term Entry (FTE) task, subjects are given a target domain like "introductory mechanics" or "the material covered in Physics 152", and asked to type into a dialog box terms that they associate with the domain, one term at a time, pressing the return key after each term. Each term disappears when they press the return key. Subjects are asked to enter the terms in the order they think of them, as close as possible to the time they think of them, with minimal disruption of their train of thought.

Section 3.1 analyzes the data from the FTE component of the p151s99, p152f97, and p172s99 studies. Subsection 3.1.1 takes a phenomenological approach, describing observable statistical features of the data. Subsection 3.1.1 addresses the specific question of whether the amount of time subjects spend thinking before entering a term is correlated with how related that term is to immediately previous terms. Subsection 3.1.3 investigates whether a correlation exists between subjects' in-course exam scores and the frequency with which their FTE response terms are apparently unrelated to immediately previous terms. Subsection 3.1.4 summarizes the findings.

3.1.1. Phenomenological Description of Data

3.1.1.1. Raw Data

The raw data captured for each subject on each FTE task is a list of the terms entered, in the order entered. Along with each term, the time at which the first letter of the term was typed (*start time*), and the time at which the return key was pressed to complete the term (*enter time*), are recorded. Times are determined by the system clock of the computer presenting the task, and recorded to one sixtieth of a second. For later analysis, the start time of the first entered term was subtracted from these times, defining the "t = 0" point.

As an immediate data processing step, a *typing time* and *thinking time* are calculated for each term. The typing time is the difference between the term's enter and start times, indicating how long the subject spent typing the term. The thinking time is the difference between the term's start time and the previous term's enter time, indicating how much time passed between the two terms while the subject was not typing. The thinking time for the first entered term was defined to be zero.

Such data was collected and analyzed for each of three studies: p151s99 (16 subjects), p172s99 (5 subjects), and p152f97 (18 subjects). Data from one of the p152f97 subjects was discarded because the subject clearly misunderstood the task instructions and carried out the task in a way which made the data meaningless.

3.1.1.2. Times as Random Variables

Consider a timeline to be a series of start times $\{t_1, t_2, \ldots, t_N\}$ for the N term entries. Define the series of time differences $\{\Delta t_1, \Delta t_2, \ldots, \Delta t_N\}$ by $\Delta t_n = t_n - t_{n-1}$ and $t_0 = 0$, so that the time difference for a term is equal to the term's thinking time summed with the previous term's typing time. Define a term's *index* to be 1 if it was the first term entered in a subject's FTE response set, 2 if it was the second entered, etc. Figure 3.1, Figure 3.2, and Figure 3.3 show start time difference vs. term index, thinking time vs. term index, and typing time vs. term index respectively for the data set of an example subject (p151s99 study, subject 01 on task J1_FTE).

The start time differences and thinking times appear randomly distributed inside an envelope that increases with term index. The typing times do not tend to increase significantly with term index. For all three plots, short times appear more common than longer ones.

Disregarding for now the systematic trend of increasing times with term index, the sets of start time differences, thinking times, and typing times can each be analyzed as a set of uncorrelated values drawn from a random distribution, and the nature of those distributions can be explored. For the same example data

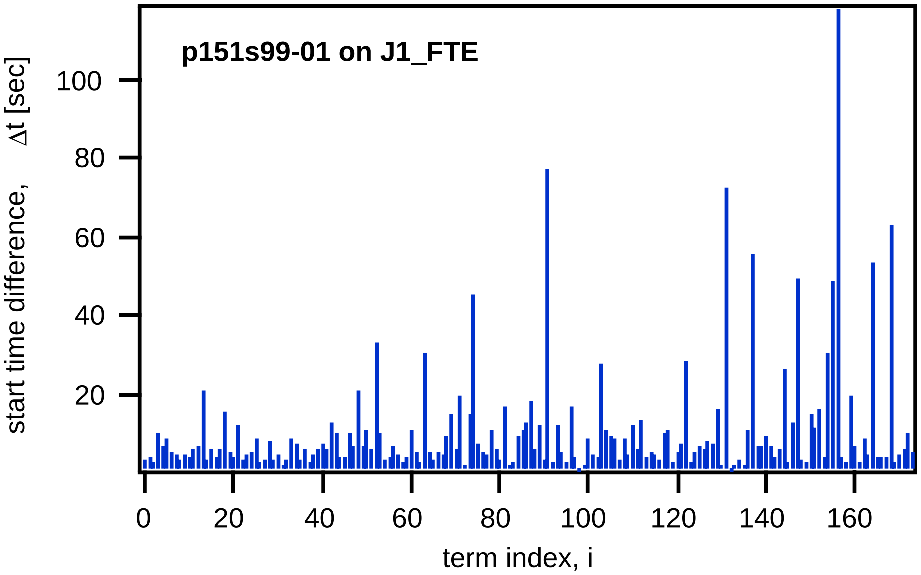


Figure 3.1: Start time differences vs. term index for study p151s99, subject 01 on task J1_FTE.

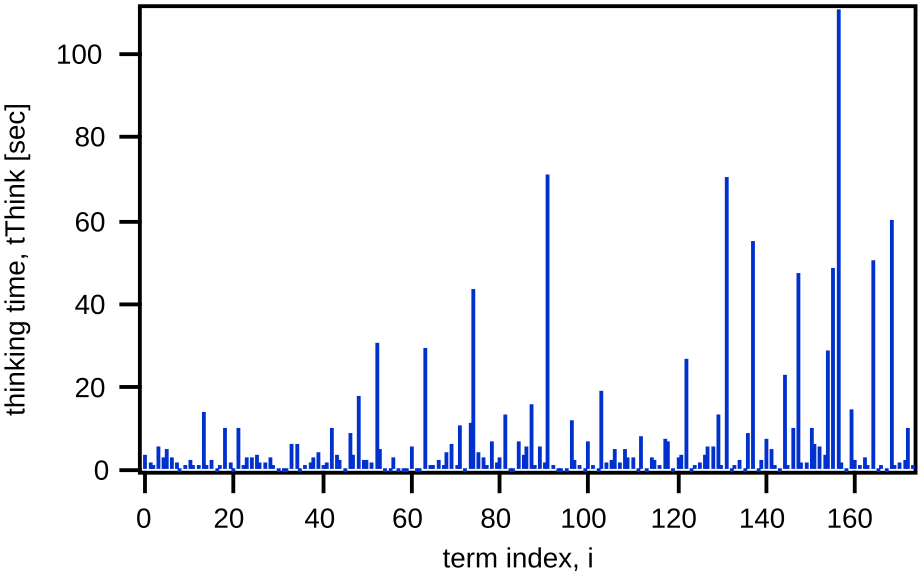


Figure 3.2: Thinking time vs. term index for study p151s99, subject 01 on task J1_FTE.

set as above, Figure 3.4, Figure 3.5, and Figure 3.6 show histograms of the natural logarithms of the start time differences, thinking times, and typing times. The natural logarithm of the times has been used rather than the times themselves because short times are far more common than long times, and a linear scale that included the longest times would lose detail for the short times. Although the set

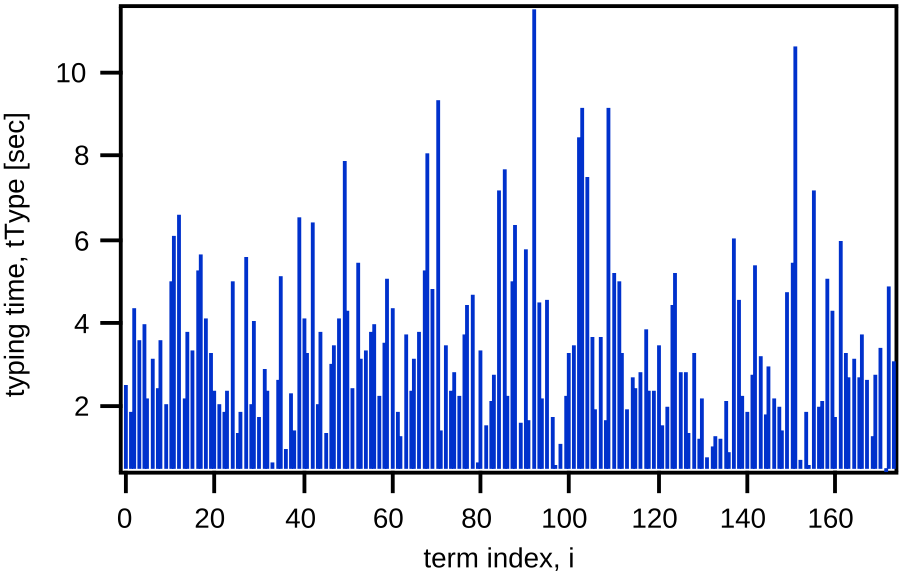


Figure 3.3: Typing time vs. term index for study p151s99, subject 01 on task J1_FTE.

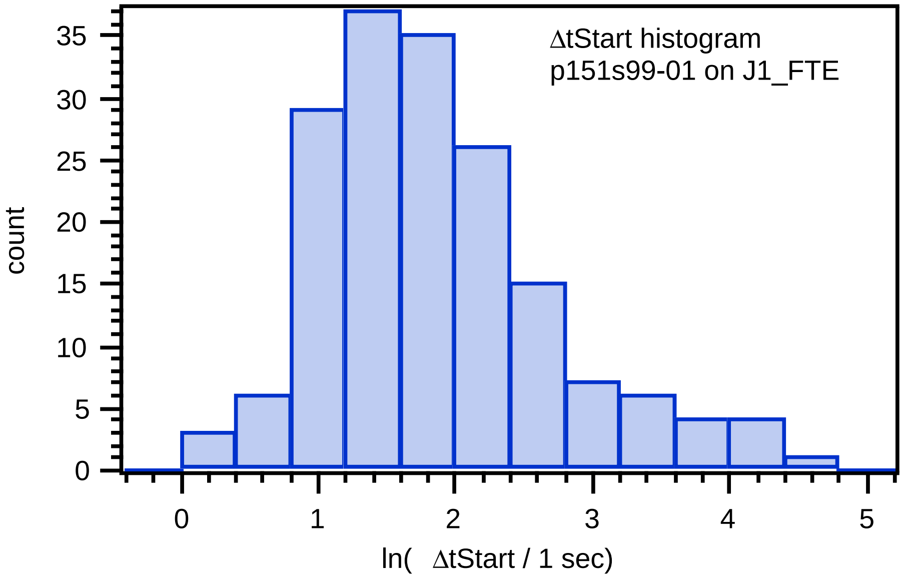


Figure 3.4: Histogram of the natural logarithms of the start time differences for study p151s99, subject 01 on task J1_FTE.

of typing times does not include as wide a range of times as does the set of thinking times, the same logarithmic scale was used for consistency and ease of comparison.

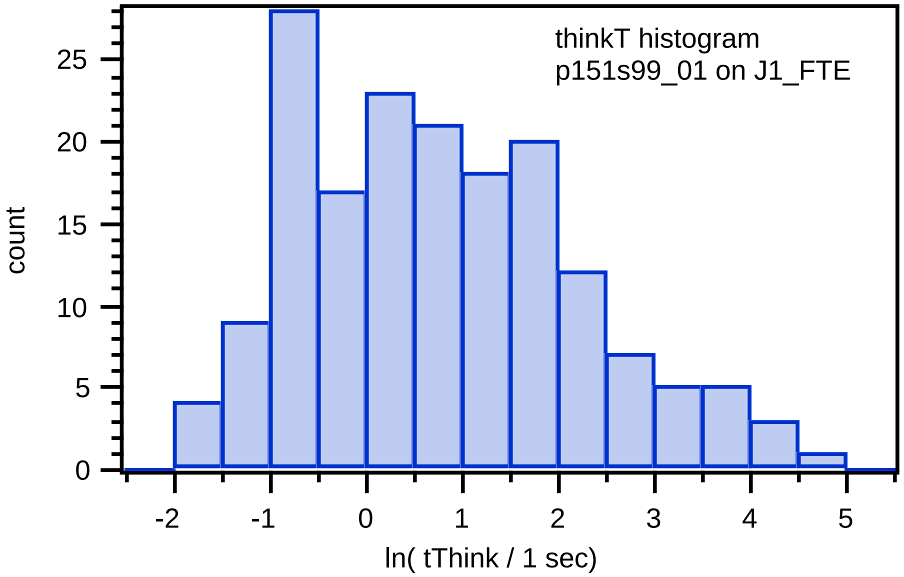


Figure 3.5: Histogram of the natural logarithms of the thinking times for study p151s99, subject 01 on task J1_FTE.

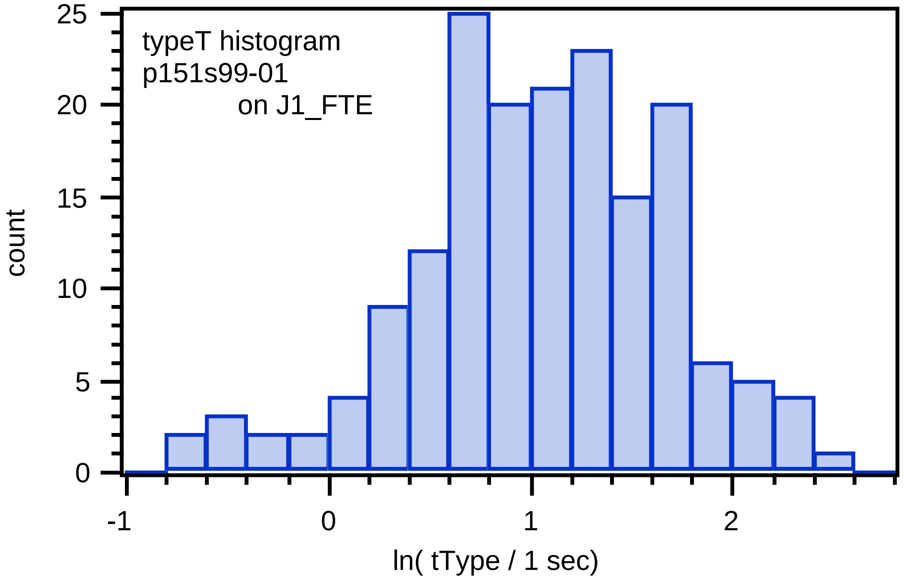


Figure 3.6: Histogram of the natural logarithms of the typing times for study p151s99, subject 01 on task J1_FTE.

The distributions of the logarithms of start time differences and thinking times look generally normal, with a noticeable skew to the right. The thinking time histogram has a pronounced spike to the left of its peak. The logarithmic typing time distribution, on the other hand, has a slight tail to the left. The fact that all three histograms are at least crudely normal indicates that the distributions are approximately log-normal, and justifies the decision to look at the distributions of the logarithms rather than of the times themselves. (A random variable obeys a *log-normal* distribution if its logarithm obeys a normal, i.e. Gaussian, distribution.)

Because we expect thinking times and typing times to be the fundamental, approximately independent quantities indicative of subjects' mental machinations during a FTE task, and because start time differences are dependent quantities calculable from thinking and start times, the following analysis will focus on thinking and typing times and not on start time differences.

3.1.1.3. Thinking Time Distribution

Thinking time are interesting because they might plausibly provide information about the cognitive process underlying a subject's responses to the FTE task. At the very least, a long thinking time probably indicates significant cognitive processing of some kind. Characterization of thinking time statistics is therefore of interest for characterizing individual subjects and for guiding theoretical modeling efforts of cognitive structure and processing.

To the extent that the thinking and typing times in a subject's FTE response set are approximated by a log-normal distribution, the response set can be characterized by the parameters necessary to fit such a distribution to the time sets. These parameters might serve as useful overall measures of a subject's performance on the task. Residual differences between the actual subject distributions and the best-fit curves might be illuminating, if divergences between individual subjects' patterns and the log-normal model can be given a cognitive interpretation.

Fitting a normal distribution to the logarithms of a set of times produces the same best-fit parameters as fitting a log-normal distribution to the times themselves, and is computationally and conceptually easier. Also, rather than fit a distribution's probability density function (PDF) to a histogram of data, it is advantageous to fit the distribution's cumulative distribution function (CDF) to a quantile plot of the data, so as to avoid the arbitrariness introduced by choosing histogram bins. A quantile plot for a set of times is constructed by sorting the set into increasing order and assigning to each time an ordinate equal to the fraction of times in the set less than or equal to that time. A time whose quantile value is 0.5 is therefore the median of the set. If a random distribution's PDF (properly normalized for total number of points and bin width) should fit a measurement set's histogram, then its CDF should fit the corresponding quantile plot. For the histogram of thinking times shown above in Figure 3.5, the corresponding quantile plot is presented in Figure 3.7.

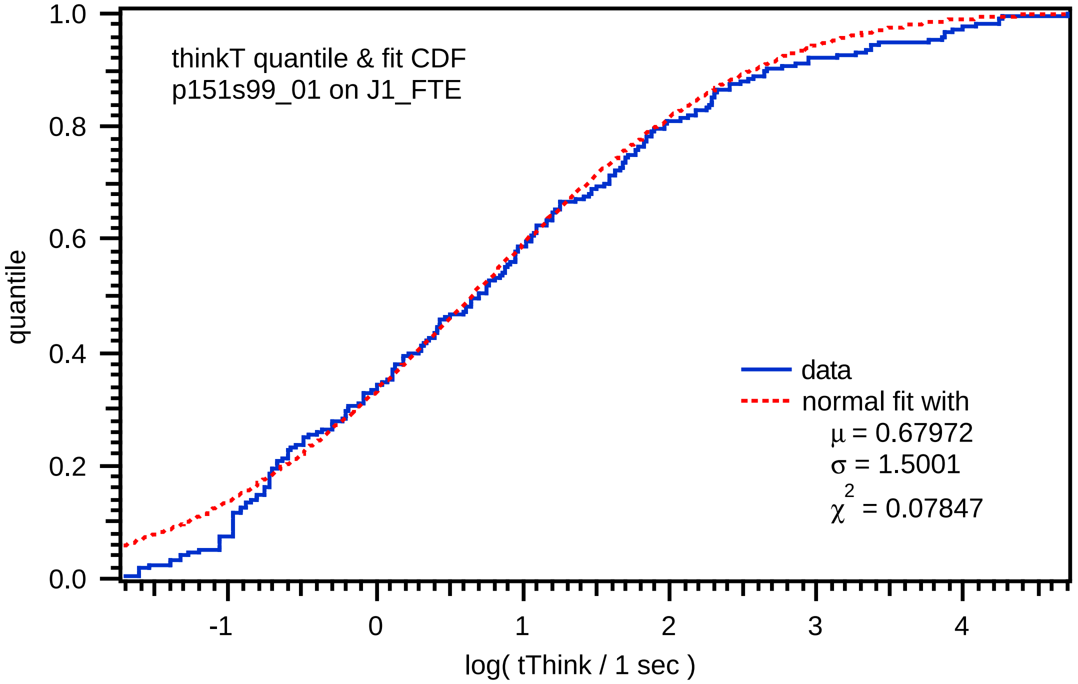


Figure 3.7: Quantile plot of thinking time logarithms for study p151s99, subject 01 on task J1_FTE.

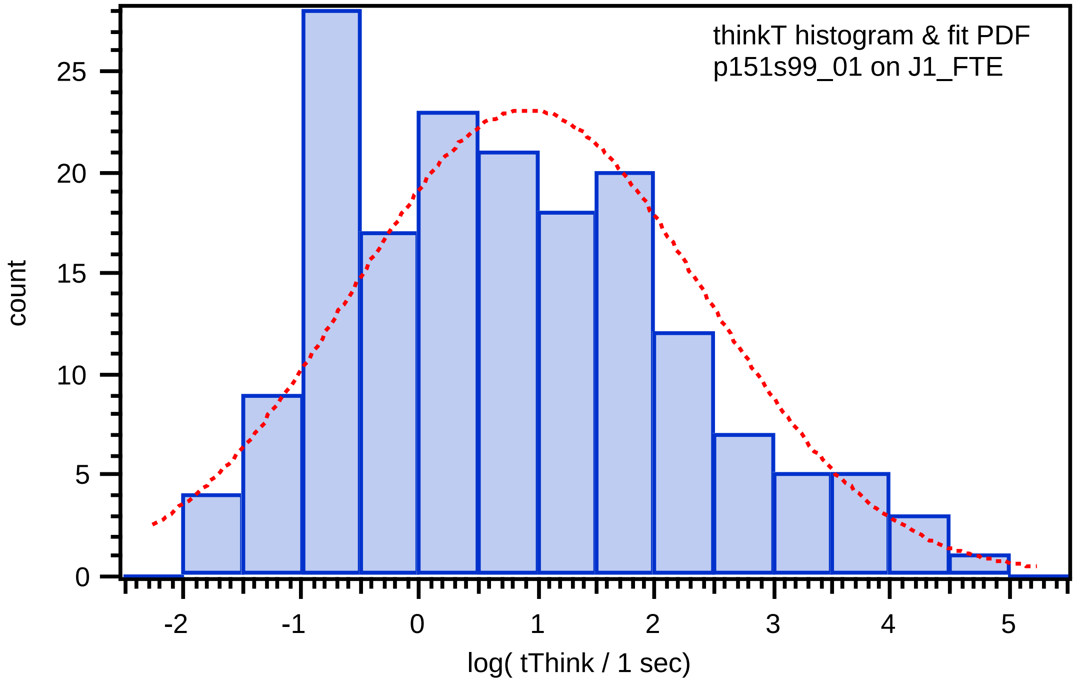


Figure 3.8: Histogram of logarithms of thinking times for subject p151s99-01 on the J1-FTE task, with best-fit curve for normal (Gaussian) probability density function, normalized for total number of counts.

The dashed line represents the best-fit curve for the normal distribution's cumulative distribution function

$$P_{\mathrm{normal}}(t;\mu,\sigma) \equiv \int_{-\infty}^{t} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[\frac{-1}{2}\left(\frac{t'-\mu}{\sigma}\right)^2\right] dt' = \frac{1}{2}\left[1+\mathrm{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right] \qquad \text{(Eq. 3.1),}$$

where μ and σ are fit parameters. The best-fit parameter values can be used to generate a PDF for comparison with the data histogram, as shown in Figure 3.8.

p151s99	μ	σ	χ*χ
-01	0.68	1.50	0.078
-02	1.21	1.82	0.100
-03	1.63	2.12	0.043
-04	1.62	1.85	0.050
-05	1.83	1.95	0.055
-06	0.90	1.61	0.033
-07	1.42	1.65	0.088
-08	1.50	1.97	0.061
-09	1.67	1.77	0.082
-10	0.85	1.38	0.160
-11	1.38	2.11	0.096
-12	1.21	1.51	0.100
-13	0.98	2.15	0.067
-14	1.39	1.20	0.038
-15	1.12	1.60	0.040
-16	1.21	1.73	0.113
mean	1.29	1.75	0.075
std. dev.	0.33	0.28	0.034

p152f97	μ	σ	χ*χ
-01	2.10	2.23	0.132
-02	2.34	1.66	0.140
-03	1.64	3.05	0.130
-04	2.48	1.53	0.127
-11	2.03	1.81	0.028
-12	2.03	2.40	0.120
-13	2.52	1.40	0.016
-14	1.30	1.78	0.238
-15	1.84	1.78	0.184
-21	2.06	2.12	0.097
-22	2.21	1.95	0.116
-24	1.64	2.22	0.185
-25	2.14	2.07	0.121
-31	2.76	2.09	0.099
-32	*	*	*
-33	1.74	1.94	0.056
-34	1.07	1.57	0.242
-35	0.23	1.60	0.024
mean	1.89	1.95	0.121
std. dev.	0.61	0.40	0.067

p172s99	μ	σ	χ*χ
-01	1.71	1.74	0.053
-02	1.06	2.03	0.234
-03	1.90	1.96	0.038
-04	2.18	2.26	0.045
-05	1.50	2.13	0.062
mean	1.67	2.02	0.087
std. dev.	0.42	0.19	0.083

Table 3.1: Best-fit parameters when a normal distribution is fit to the set of logarithms of thinking times, for the FTE task, of subjects in the p151s99, p152f97, and p172s99 studies. (Subject p152f97-32 misinterpreted the task instructions in a way that made that his/her data worthless.)

For each FTE task in each study, this two-parameter fit was performed on each subject's set of thinking time logarithms using a chi-squared algorithm. Table 3.1 shows the resulting fit parameters for the three studies. χ^2 values for the fits are indicated to provide a relative sense of fit quality.

Of all 38 analyzed subjects in the three studies, the subject with the largest χ^2 value for the thinking time fit is p152f97-34, with a χ^2 value of 0.242. According to the χ^2 measure, this is the thinking time data set for which the log-normal fit is the poorest. The quantile plot with CDF fit is shown in Figure 3.9, while Figure 3.10 displays the corresponding histogram of thinking time logarithms and PDF curve for the fit parameters generated by the CDF fit.

Both Figure 3.8 and Figure 3.10 show a marked spike on the left side of the distribution. Examination of the corresponding histograms for all 38 subjects in the three studies shows that most of the histograms display a pronounced spike on the left edge of a broad peak, and very few of the histograms do not have at least a rudimentary bump there. Given the low number of counts in many data sets and the general noise level of the data, all the data sets might plausibly obey a distribution with a strong spike superimposed on a broader peak.

To further investigate the random variable distribution which might model thinking time measurements, the data from the 38 subjects was aggregated into one large set. Such aggregation loses subject-specific details of the individual data sets, but might help reveal more general patterns common to the sets but obscured by statistical noise.

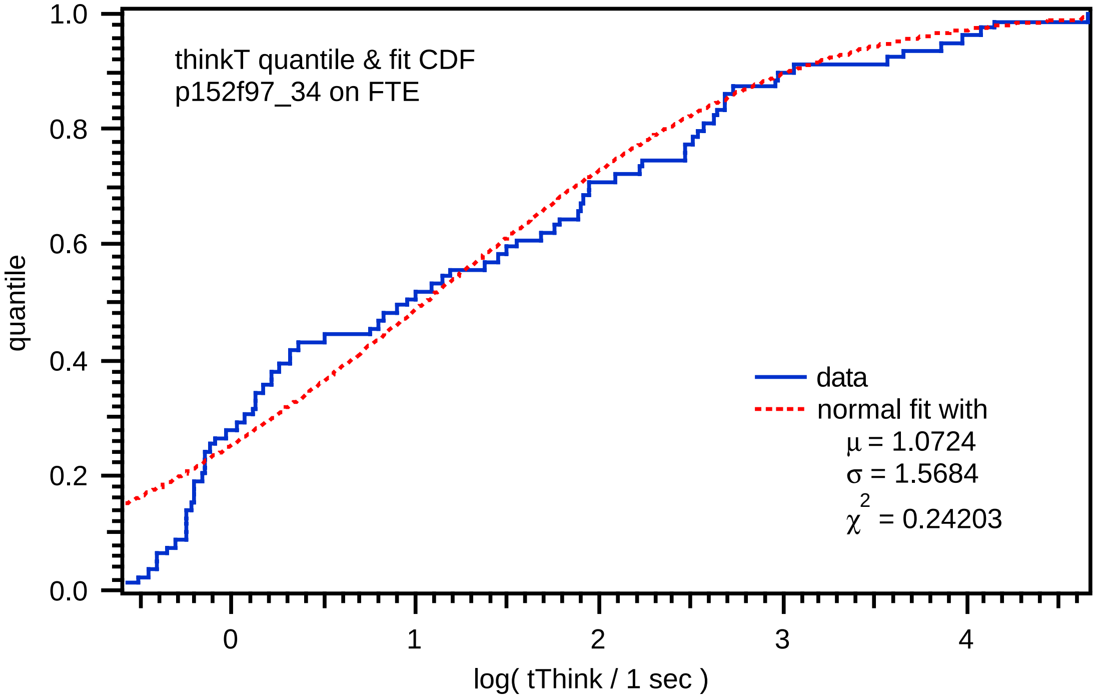


Figure 3.9: Quantile plot of logarithms of thinking times for subject p152f97-14 on FTE task, with best-fit curve for normal (Gaussian) cumulative distribution function.

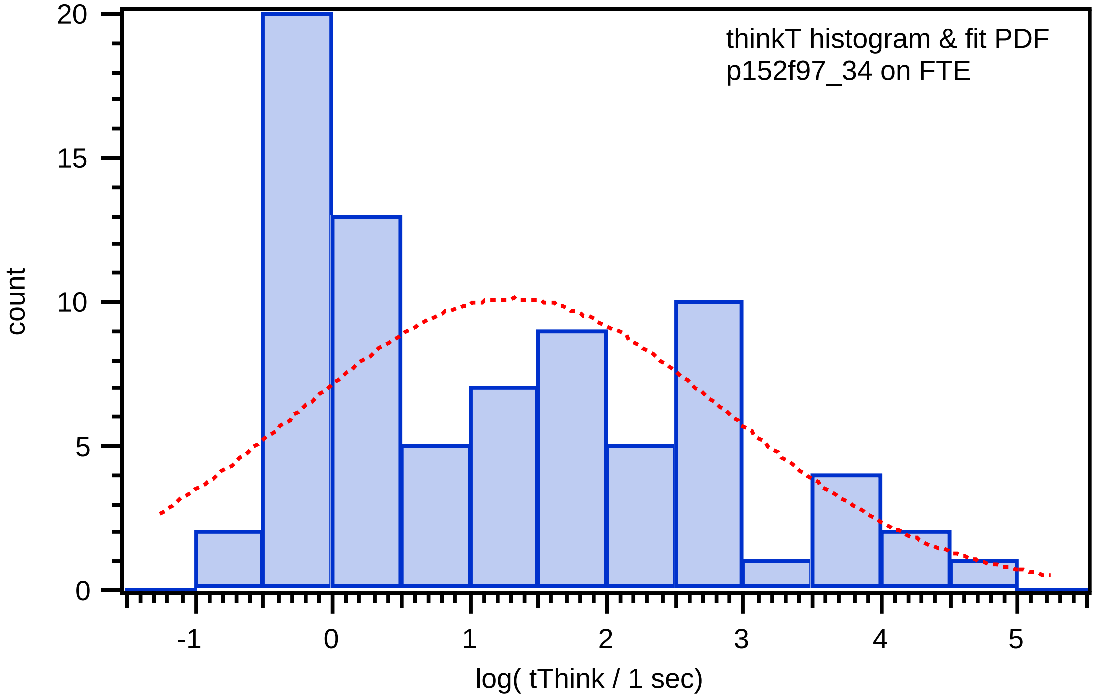


Figure 3.10: Histogram of logarithms of thinking times for subject p152f97-14 on FTE task, with best-fit curve for normal (Gaussian) probability density function, normalized for total number of counts.

To aggregate the data, it was necessary to assume that all the data sets are roughly described by a normal distribution of thinking time logarithms. The thinking time logarithms from each set could then be “standardized”, i.e. scaled so that the best-fit normal distribution to the set is the “standard” normal distribution with a mean of zero and a standard deviation of unity. If the best-fit parameters of a normal distribution to a subject’s data set are μ and σ, and τ_i is the logarithm of the i^{th} thinking time, then the variable transformation which standardizes that subject’s thinking time logarithms $\{\tau_i\}$ is $x_i \equiv (\tau_i - \mu)/\sigma$. Once each subject’s data was standardized, all 38 sets studies were aggregated into a larger data set.

Figure 3.11 shows a quantile plot for the resulting aggregate data. Two fits to the data are included: a CDF fit for a normal distribution, and a CDF fit for a sum of two normal distributions (“double normal” distribution), which will be discussed below. The best-fit values for the normal distribution are close to $\mu = 0$ and $\sigma = 1$; this is expected because all individual subject data sets were standardized to those values before aggregating.

Figure 3.12 displays a histogram of the aggregated values, with PDF curves shown for the two fits obtained from the quantile plot. Here the data is clearly seen to two separate peaks. This is not surprising given the spike-plus-peak shape seen in the individual data sets. The broad main peak of the various sets aggregate to a well-defined peak, with statistical noise smoothed out; the leading spikes from the various sets do not all occur at the same time value, even when standardized, so they aggregate to a second peak of similar size and smaller width rather than to a narrow, tall spike.

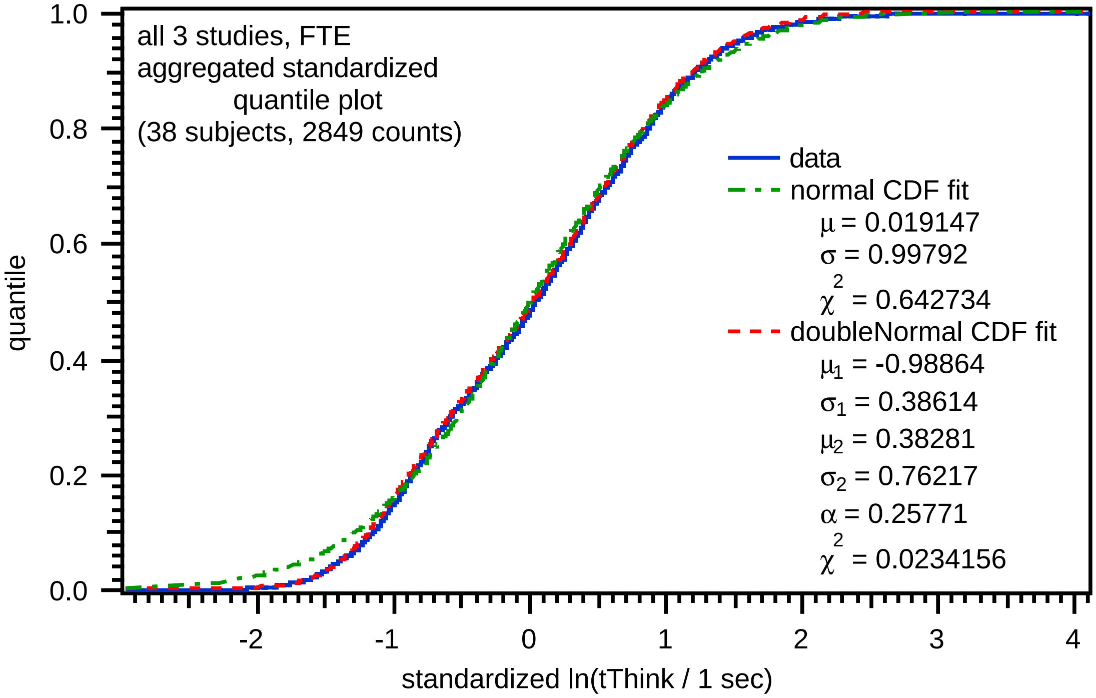


Figure 3.11: Quantile plot for logarithms of thinking times, standardized by subject to a normal distribution, aggregated across all subjects in the p151s99, p172s99, and p151f97 studies. CDF curves for the best-fit normal (dot-dashed line) and double-normal (dashed line) distributions are indicated. The double-normal fit lies close enough to the data that it is difficult to distinguish.

Since the histogram displays two peaks, a normal distribution is clearly only a very approximate description of the data. A two-peaked distribution would be more appropriate, motivating the application of a “double-normal” distribution, defined to be the normalized sum of two normal distributions. The PDF is

$$p_{\mathrm{DN}} \equiv \frac{\alpha}{\sigma_1\sqrt{2\pi}}\exp\left(-\frac{(x-\mu_1)^2}{2\sigma_1^2}\right)+\frac{1-\alpha}{\sigma_2\sqrt{2\pi}}\exp\left(-\frac{(x-\mu_2)^2}{2\sigma_2^2}\right) \qquad \text{(Eq. 3.2),}$$

where μ_1, σ_1, μ_2, σ_2, and α are fit parameters. The CDF for this distribution is

$$P_{\mathrm{DN}}(x)=\frac{1}{2}\left[1+\alpha\,\mathrm{erf}\left(\frac{x-\mu_1}{\sigma_1\sqrt{2}}\right)+(1-\alpha)\mathrm{erf}\left(\frac{x-\mu_2}{\sigma_2\sqrt{2}}\right)\right]. \qquad \text{(Eq. 3.3).}$$

The double-normal CDF was fit to the aggregate quantile plot with an iterative χ^2 procedure; the resulting best-fit parameters are indicated in Figure 3.11, and the corresponding CDF and PDF curves are shown on the quantile plot and histogram, respectively.

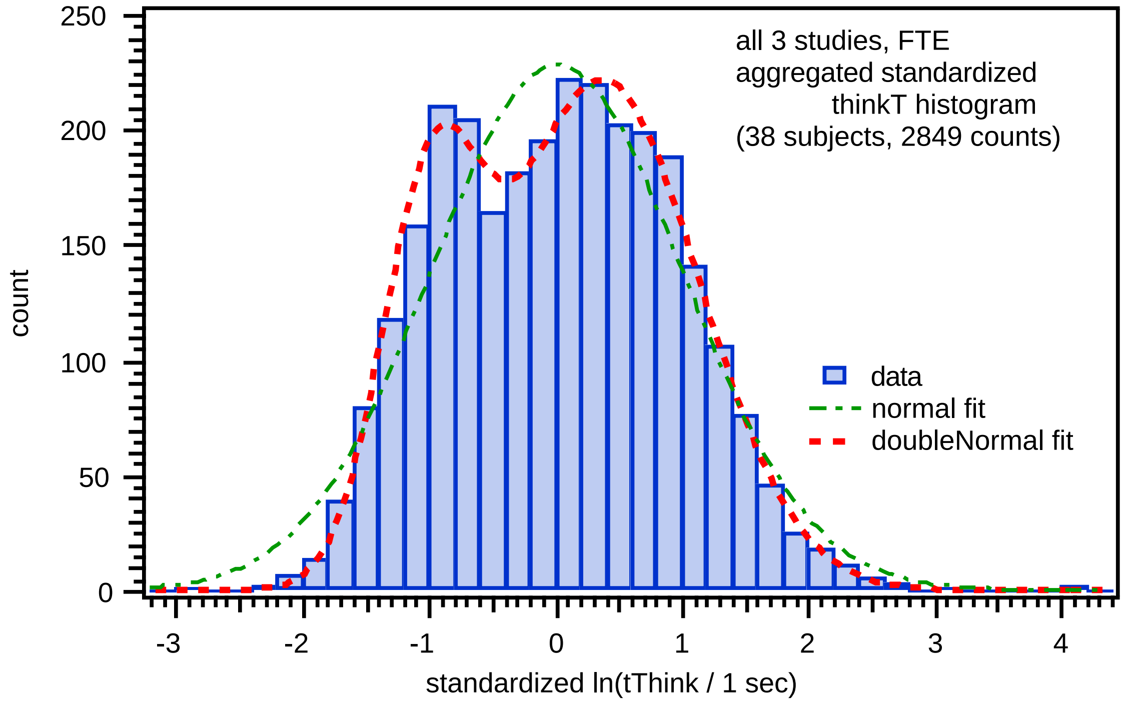


Figure 3.12: Histogram for logarithms of thinking times, standardized by subject to a normal distribution, aggregated across all subjects in the p151s99, p172s99, and p151f97 studies. PDF curves for the normal and double-normal distributions fit to the previous quantile plot are indicated.

On both the quantile plot and histogram, the double-normal distribution can be seen to fit significantly better than the normal distribution, especially along the leading edge of the data. A comparison of the χ^2 values for the two fits, indicated on the quantile plot, supports this observation.

For a comparison between study populations, subject data sets within each of the studies can be aggregated and fit via the same procedure as used above. Figure 3.13 and Figure 3.14 show a quantile plot and histogram for aggregated p151s99 data, with fits; Figure 3.15 and Figure 3.16 show the same for p152f97 data; and Figure 3.17 and Figure 3.18 show the same for p172s99 data.

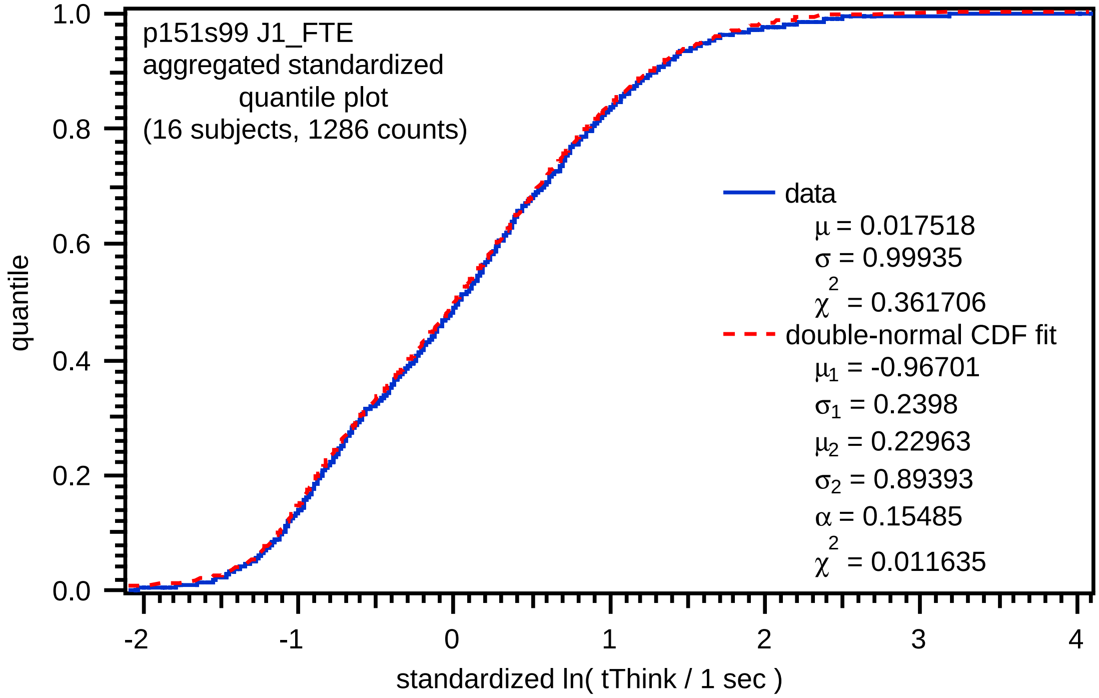


Figure 3.13: Quantile plot with best-fit double-normal CDF for aggregated p151s99 data.

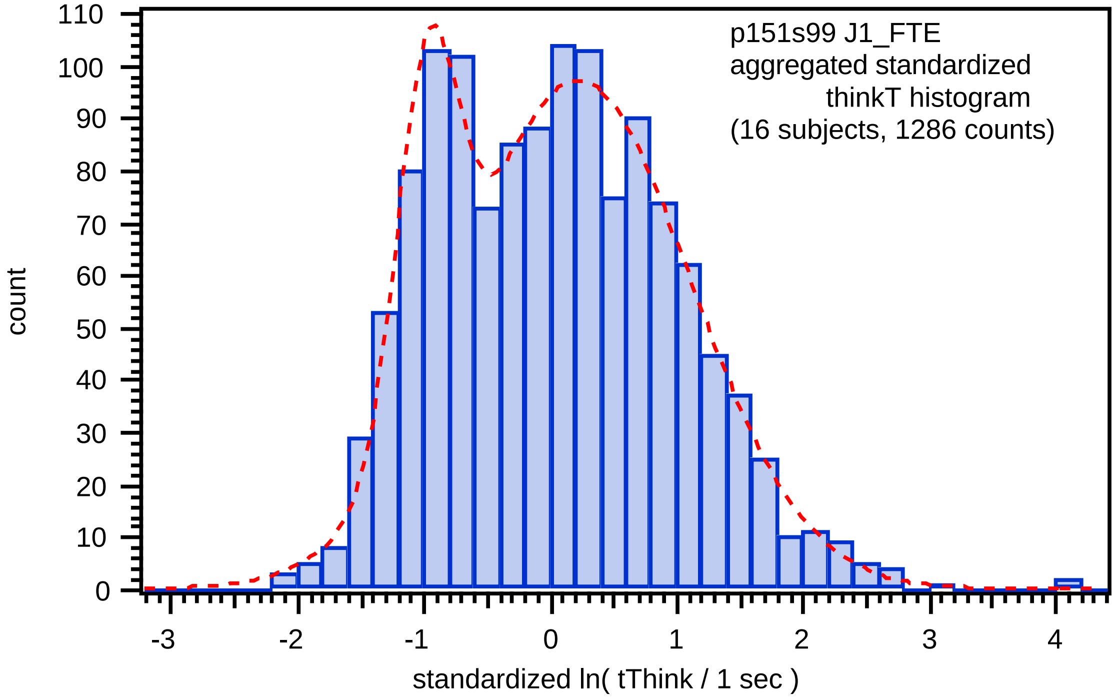


Figure 3.14: Histogram with best-fit double-normal PDF for aggregated p151s99 data.

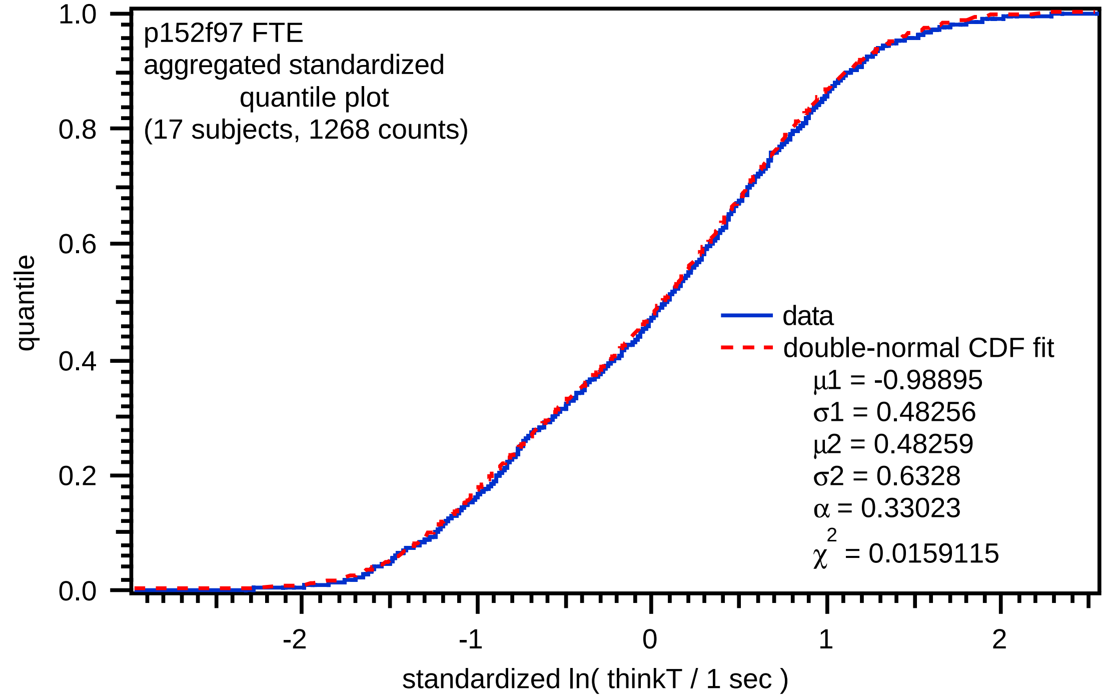


Figure 3.15: Quantile plot with best-fit double-normal CDF for aggregated p152f97 data.

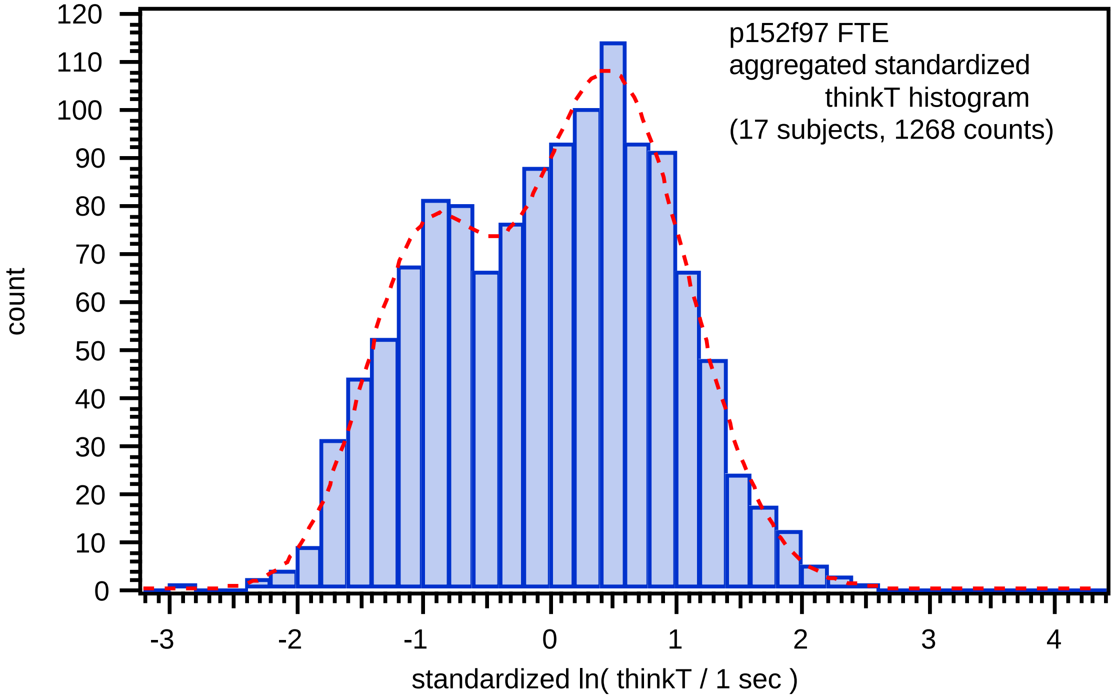


Figure 3.16: Histogram with best-fit double-normal PDF for aggregated p152f97 data.

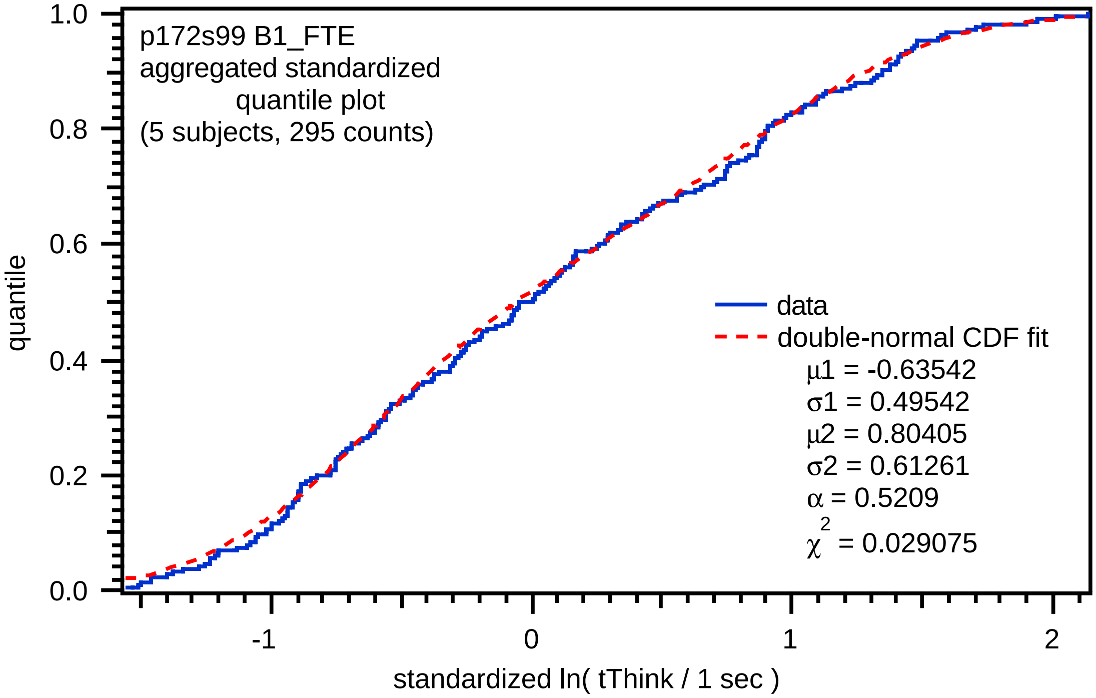


Figure 3.17: Quantile plot with best-fit double-normal CDF for aggregated 172s99 data.

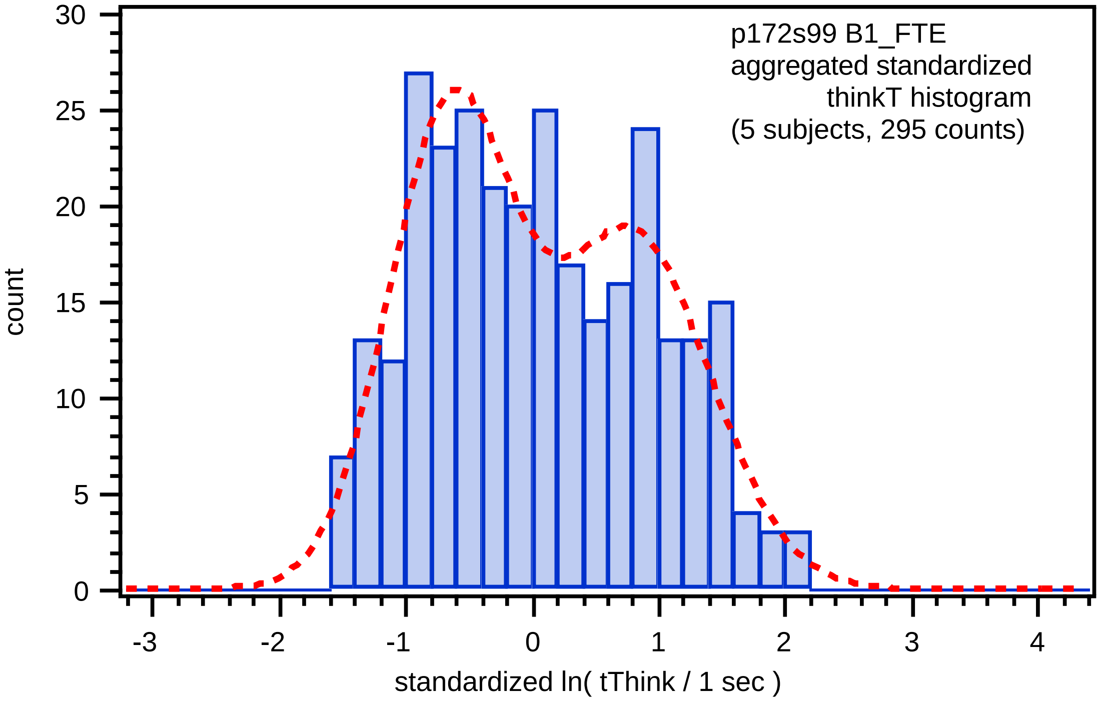


Figure 3.18: Histogram with best-fit double-normal PDF for aggregated 172s99 data.

Both the p151s99 and p152s97 aggregated sets are well fit by the double-normal distribution, although for somewhat different best-fit parameter values. The fit to the p172s99 data is less satisfactory. This could be attributable to poorer statistics because of the significantly smaller sample size (5 subjects for 295 counts, vs. 16 subjects for 1286 counts in the p151s99 study and 17 subjects for 1268 counts in the p152f97 study). It could also be a consequence of the fact that the p172s99 FTE task was administered at the end of the semester *subsequent to* the course whose domain was the topic of the FTE, whereas the p151s99 and p152f97 FTE tasks were administered directly at the end of the relevant course. Histograms for individual subjects in the p172s99 study (not shown here) are quite noisy but display some tendency towards a leading spike, suggesting that the underlying distribution is similar to that of the other studies, but obscured by noise.

It would be desirable to return to the subject-specific, non-standardized data sets and fit them with the double-normal distribution, obtaining best-fit parameters which might prove more useful and revealing as subject characterizers than the parameters for the normal fit listed in Table 3.1. Unfortunately, the five-parameter double-normal fit is quite sensitive to noise in smaller data sets, and frequently produces a fit different from what the human eye would consider most appropriate. Figure 3.19 shows eight arbitrarily-chosen sample cases from the p151s99 study to demonstrate the range of fits which can result.

For subject 01 and especially for subject 06, the spike and main peak are not distinct enough, and the two peaks of the double-normal distribution merge together into a single peak. For subjects 02, 04, and 05, the fit captures two similar, roughly equal peaks. For subjects 03 and 07, the fit captures and perhaps exaggerates the leading spike and the broad main peak. For subject 08, the histogram has two narrow spikes and a broad central plateau, and the fit included the plateau with the first peak rather than the second, producing a trailing rather than a leading spike.

Overall, the double-normal distribution seems to do an adequate job of fitting the data, but the noise level is sufficiently high so that the fit often fails to display the spike-plus-broader-peak distribution that seems suggested by the overall pattern of the data. Therefore, fitting individual subject data with a double-normal distribution is not fruitful for extracting fit parameters to use as subject-specific characterization measures. The fact that a double-normal distribution does seem to describe the overall pattern of the data, however, may be useful in guiding the construction of theoretical models of underlying cognitive processes.

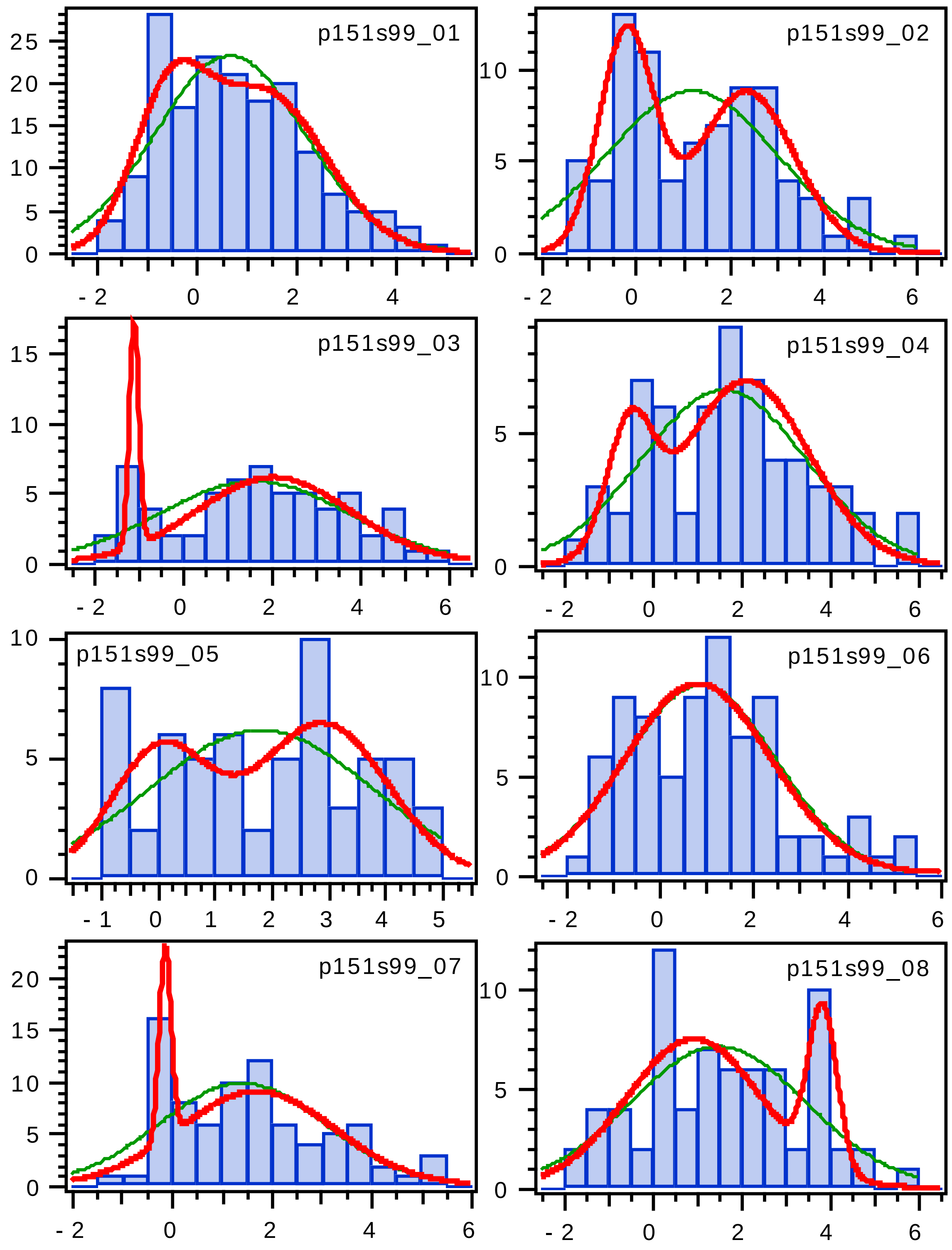


Figure 3.19: Histograms of thinking time logarithms for subjects p151s99-01 through -08, with PDF for best-fit double-normal (heavy dash) and best-fit normal (light dot-dash) distributions.

As an alternative, single-subject data sets which show an identifiable spike and peak could be crudely characterized by the locations (thinking-time values) of the spike and peak, and perhaps the fraction of events ascribable to the spike. The thinking times describing the spike and peak locations might be interpretable as characteristic cognitive times of the subject. Further analysis in this direction is warranted.

3.1.1.4. Typing Times

Although typing times might shed light on how the mechanism of the FTE task impacts the interpretation of thinking times and other data, we do not expect them to be particularly revealing of subjects' domain knowledge and cognitive processes, so they have not been analyzed as thoroughly as have thinking times.

Because histograms of the logarithms of each subjects' typing times, like thinking times, look approximately normally distributed, a quantile plot of typing time logarithms was constructed for each subject, and fit with a normal distribution CDF. The resulting best-fit parameters are listed in Table 3.2.

These parameters might be useful in describing subjects' typing facility, which could have relevance to the cognitive interpretation of FTE task results. For example, some subjects (notably skilled typists) might be able to "parallel process" and think about the next term to enter while typing a response term. Some of the expert subjects who tried the FTE task in preliminary studies observed that by the time they had finished typing a term, they had frequently already though of the next term to enter. In such situations the "thinking time" does not accurately describe the time required to think of the next term, but rather the physical time required to type the first letter of the term after pressing the enter key.

If a subject accomplishes a significant amount of thinking while typing, one might expect to see a tendency for shorter thinking times to follow longer typing times. To look for such a tendency, a scatterplot was constructed for each subject's FTE data, in which each thinking time was plotted against the previous term's typing time. Figure 3.20 shows an example of such a plot.

If thinking while typing caused thinking times to be slightly shorter when following long typing times, then such a scatterplot should reveal the correlation as an excess of points in the top-left and bottom-right quadrants of the plot, and a scarcity in the bottom-left and top-right. Plots equivalent to that of Figure 3.20 were constructed and examined for each of the subjects in the p151s99, p172s99, and p152f97 studies, and none displayed such a correlation. In fact, a few of the scatterplots showed a very slight tendency towards a positive correlation, with a deficit of points in the top left and bottom right corners, although the trend was weak enough to be of questionable significance. Such a positive correlation might result if longer typing times correlate with terms that are more difficult to type (perhaps due to unfamiliar spelling or approximate mathematical notation), with a corresponding cognitive load that prevents thinking of the subsequent term. It might also indicate general fatigue, so that slower typing accompanies the longer pauses which occur late in the task. Another possibility is that subjects tend to enter longer terms later in the task as, having exhausted their vocabulary of shorter, simpler terms, they enter longer and more esoteric (perhaps compound) terms.

p151s99	μ	σ	$\chi*\chi$
-01	1.07	0.59	0.328
-02	1.23	0.62	0.071
-03	1.20	0.53	0.015
-04	1.23	0.65	0.078
-05	1.42	0.61	0.041
-06	1.85	0.54	0.110
-07	1.19	0.54	0.114
-08	1.33	0.64	0.047
-09	1.05	0.63	0.035
-10	1.86	0.77	0.034
-11	1.48	0.51	0.031
-12	1.43	0.66	0.020
-13	0.98	0.62	0.012
-14	1.28	0.64	0.052
-15	1.41	0.75	0.044
-16	1.56	0.57	0.477
mean	1.35	0.62	0.094
std. dev.	0.25	0.07	0.127

p152f97	μ	σ	$\chi*\chi$
-01	0.95	0.67	0.029
-02	1.58	0.59	0.041
-03	1.49	0.79	0.026
-04	1.99	0.52	0.021
-11	2.54	1.09	0.166
-12	1.58	0.50	0.049
-13	2.29	0.99	0.102
-14	1.25	0.63	0.038
-15	1.45	0.72	0.028
-21	1.20	0.85	0.030
-22	1.46	0.74	0.040
-24	1.24	0.60	0.061
-25	1.77	0.45	0.053
-31	1.56	0.80	0.050
-32	*	*	*
-33	1.57	0.66	0.050
-34	2.22	0.77	0.035
-35	1.44	0.53	0.023
mean	1.62	0.70	0.050
std. dev.	0.42	0.17	0.036

p172s99	μ	σ	$\chi*\chi$
-01	1.05	0.63	0.044
-02	1.68	0.57	0.052
-03	1.55	0.60	0.059
-04	1.35	0.74	0.069
-05	1.55	0.56	0.053
mean	1.43	0.62	0.056
std. dev.	0.24	0.07	0.009

Table 3.2: Best-fit parameters when a normal distribution is fit to the set of logarithms of thinking times, for the FTE task, of subjects in the p151s99, p152f97, and p172s99 studies. (Subject p152f97-32 misinterpreted the task instructions in a way that made that his/her data worthless.)

Further analysis of the existing data might suggest one or another of these possibilities, but the issue of how the mechanical aspects of term entry interact with the cognitive aspects is general and important enough to warrant a carefully designed study of its own. Such a study might have subjects perform an FTE verbally, with timing data extracted after the fact from an audio recording, or in writing, with timing data extracted from a videotape. Comparisons with a typed, computer-mediated FTE as used in this study might clarify the impact of typing on task performance. Independent measures of subjects' facility at typing would be another useful data source.

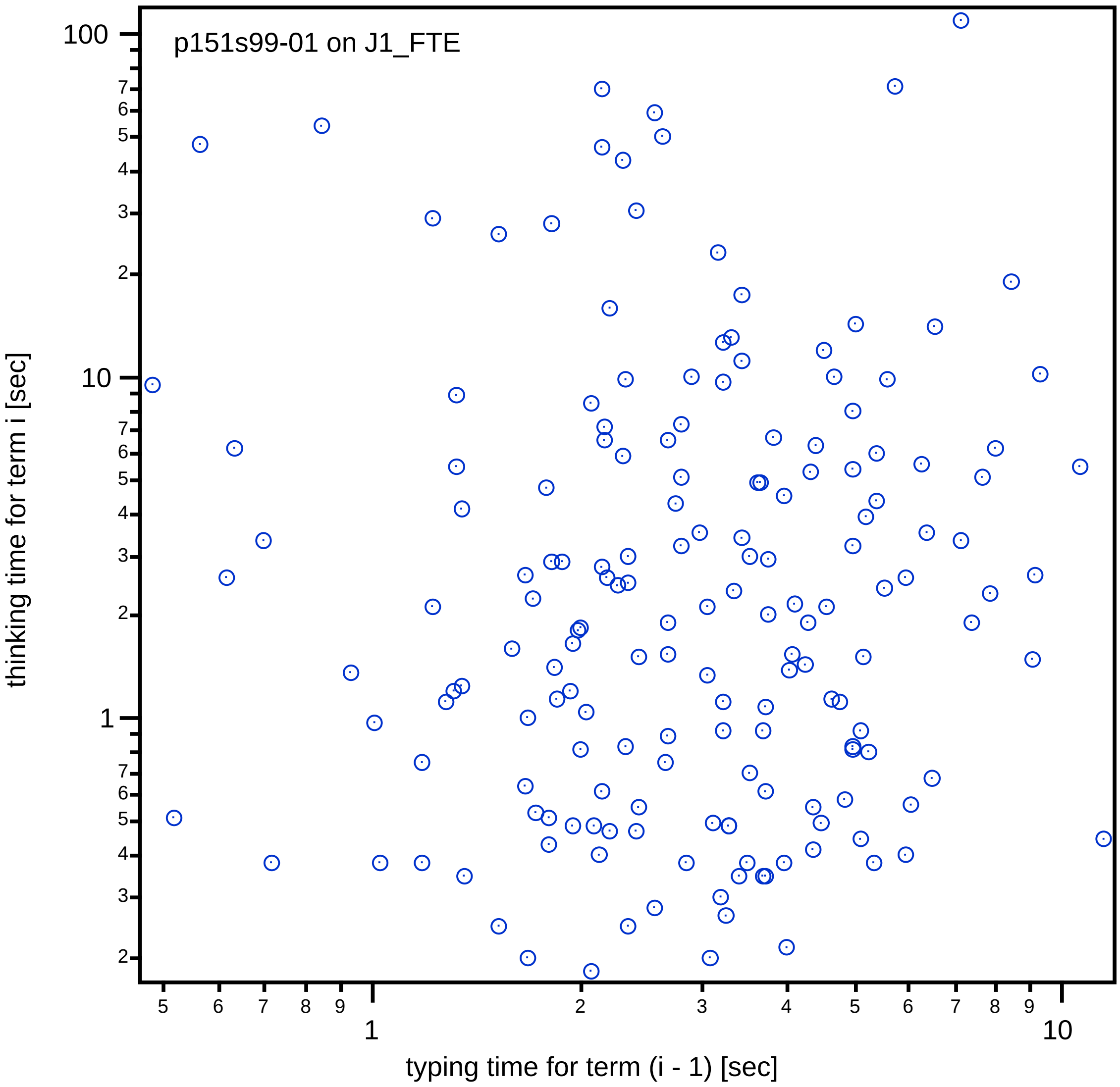


Figure 3.20: Successor scatterplot for thinking time against previous typing time, for subject p151s99-01 on task J1_FTE.

3.1.1.5. Temporal Correlations

The previous sections have examined FTE start time differences, thinking times, and typing times as if they were uncorrelated numbers drawn from a random distribution. Such a description is incomplete: the times form a well-ordered sequence from the beginning to the end of the task, and both overall trends and correlations between neighboring values are possible.

A *timeline* displays a FTE response set's start times along a one-dimensional timeline. The timeline for subject 01 of the p151s99 study, task J1_FTE, is displayed in Figure 3.21. Two global patterns appear:

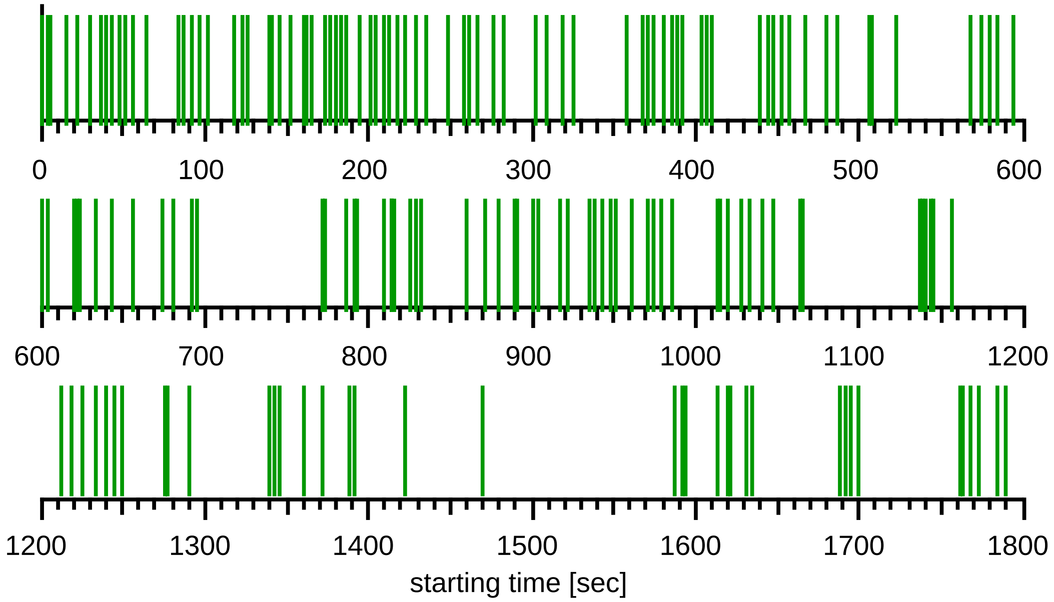


Figure 3.21: Timeline of FTE entries for study p151s99, subject 01 on task J1_FTE.

1. As the task progresses, term entries become increasingly sparse;

2. Term entries appear to occur in clusters, especially later in the task.

The same general patterns are observable in timelines for the other subjects.

Both patterns make intuitive sense: term entries become increasingly sparse later in the task because subjects have already entered most of their readily accessible terms, and have to think hard to recall additional terms; and term entries tend to occur in clusters because when a subject thinks of a term, it often suggests other connected terms. Both apparent patterns will be investigated below.

Decreasing Term Entry Rate

To make more rigorous the qualitative observation that term entries in the timeline become sparser as the task progresses, rate of term entry can be plotted vs. elapsed time for each subject's data set. A "moving window" technique is used to calculate the locally-averaged rate of term entry at any given time during the task: for a time of t and a window of width w, the average term entry rate is defined to be the number of start times in the interval $(t - w/2, t + w/2)$ divided by w. The larger the window, the smoother the resulting function for term entry rate; the smaller the window, the more sensitive the function is to local variations in term entry rate.

For subject p151s99-01, Figure 3.22 shows a plot of average term entry rate vs. elapsed task time for a window of width $w = 100$ seconds; Figure 3.23 shows the same for a window of width $w = 500$ seconds. To avoid edge effects, the term entry rate has not been calculated closer than $w/2$ to the beginning or end of the task.

Both plots show clearly that the average rate of term entry decreases as the task progresses. Both also show that the average rate fluctuates significantly throughout the task. Note that the larger window width is greater than 1/4 of the entire task duration, so very significant averaging occurs, and yet fluctuations are still quite evident. This is caused by the local density fluctuations — the "clustering" — seen in the data.

Not all subject data sets show such a clear overall decrease in term entry rate. Figure 3.24 displays the same plot as Figure 3.23, but for subject p152f97-02. The overall trend is not so simple.

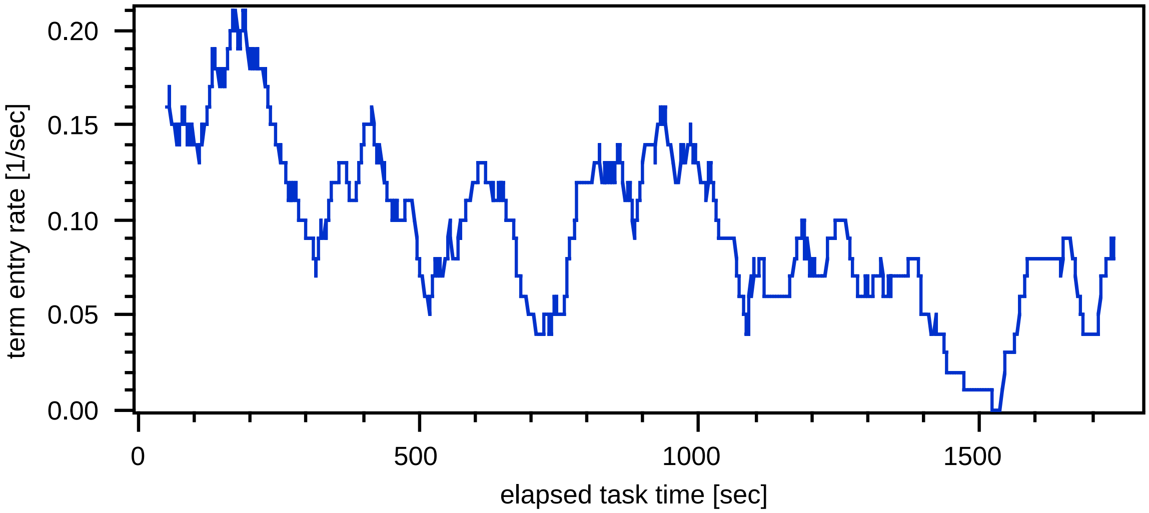


Figure 3.22: Term entry rate vs. elapsed task time for subject p151s99-01 on task J1_FTE, with window width 100 sec.

Another representation of term entry rate vs. elapsed task time, which has the advantage of not requiring an arbitrary window width choice, is cumulative number of terms entered (as a fraction of the total number of terms for the task) vs. elapsed task time. This is analogous to examining a distribution's CDF rather than a histogram corresponding to its PDF. In this representation a constant rate of term entry from task beginning to end would result in a straight line of positive slope, and a term entry rate that decreases with time would result in a curve that begins from the origin with large slope and then approaches a horizontal line as the elapsed time increases (like the classic "charging capacitor" curve). Figure 3.25 shows such a plot for the data set represented in Figure 3.22 and Figure 3.23, and Figure 3.26 shows one for the data set in Figure 3.24.

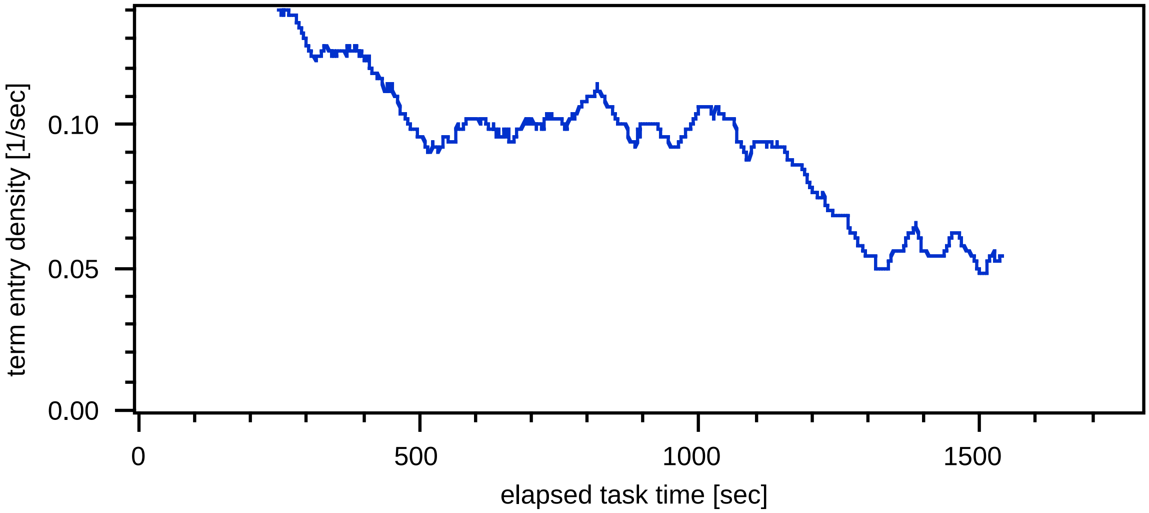


Figure 3.23: Term entry rate vs. elapsed task time for subject p151s99-01 on task J1_FTE, with window width 500 sec.

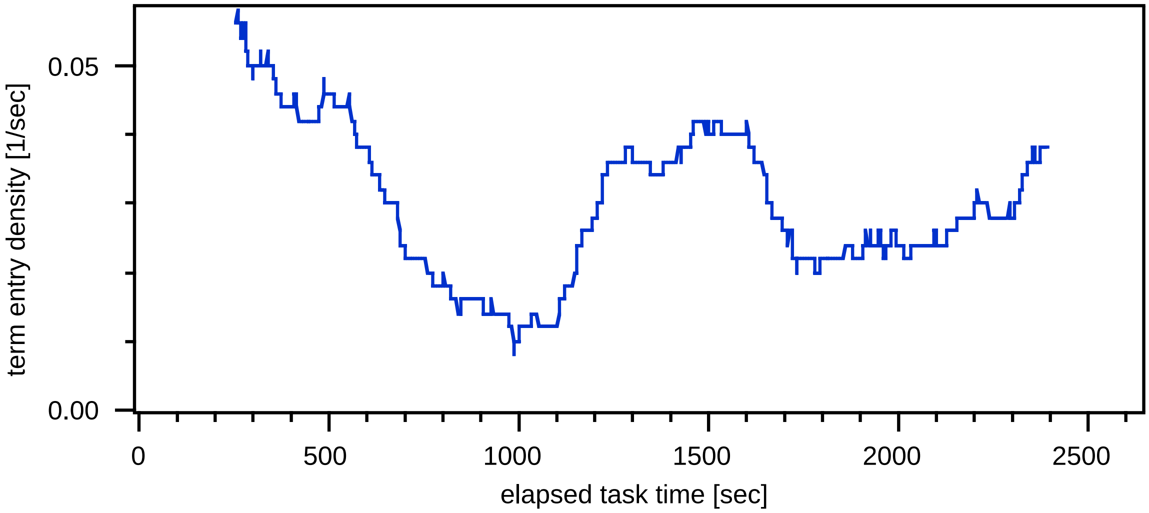


Figure 3.24: Term entry rate vs. elapsed task time for subject p152f97-02 on FTE task, with window width of 500 sec.

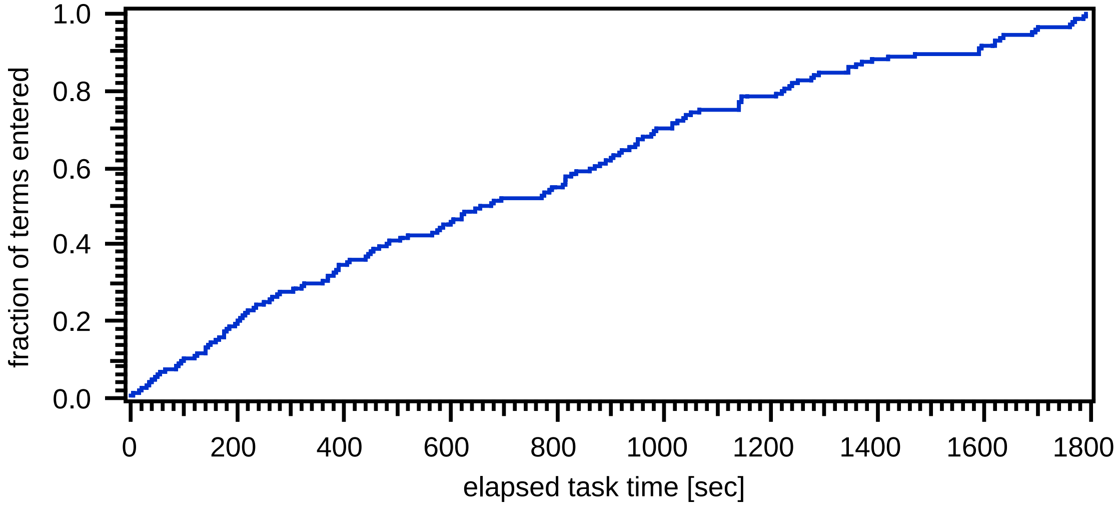


Figure 3.25: Cumulative number of terms entered vs. elapsed task time for subject p151s99-01 on task J1_FTE.

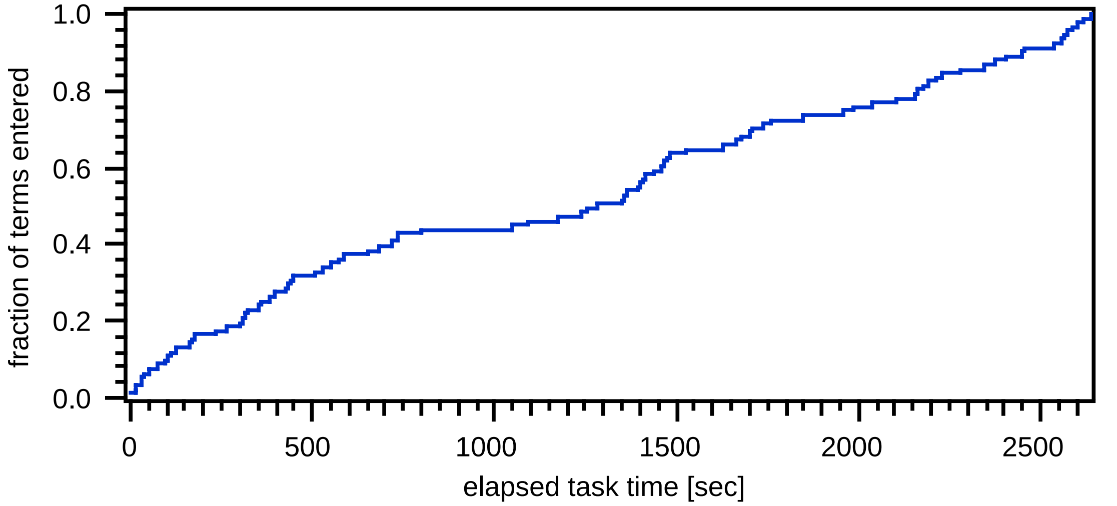


Figure 3.26: Cumulative number of terms entered vs. elapsed task time for subject p152f97-02 on FTE task.

These representations of the data clarify the subjective impression of increasing term sparseness conveyed by the timelines, and could be used to quantify the trend. For example, one could fit a simple function like a second-degree polynomial or $\left(1-e^{-\alpha t}\right)$ saturation curve to the cumulative term count vs. elapsed time plots, and use the fit parameters to characterize the changing density. This might be useful to constrain and test theoretical models of the task dynamics. In the absence of a model, however, there seems to be little reason to pursue the issue.

Clustering

Although the timelines appear to show clustering, it is not obvious that this clustering isn't an illusion of statistical fluctuations. Here, we use the term *clustering* to mean a tendency for terms to come in runs separated by short times that differs statistically from what would be expected for uncorrelated random variables. Any series of numbers drawn from a random distribution weighted towards small values (like a log-normal distribution) will be punctuated by occasional large values, which, when interpreted as gap lengths in a timeline, would give the appearance of "clustering" of the intervening shorter values. Such clustering is no more meaningful than the various runs of consecutive "heads" that occur during repeated tosses of a biased coin. Clustering of term entries in a timeline, if statistically significant, indicates that the time differences in the series are not uncorrelated, but that short time differences tend to come close together. In other words, the data displays statistical behavior not modelable by a sequence of uncorrelated values drawn from a random distribution.

To determine whether the apparent clustering does in fact describe a significant feature of the data, the series of time differences must be checked for correlations. A simple way to do this is to construct a *successor correlation plot*, a scatterplot of Δt_n vs. Δt_{n-1} for all terms in a FTE response set. If time differences are in fact uncorrelated, the points on the plot should show no correlation. If, on the other hand, short time differences tend to come in clusters, the data points should fall along a diagonal line with positive slope. Figure 3.27 displays successor correlation plots for subjects in the p152f97 study.

Figure 3.27 and similar plots for all other subjects in the p151s99, p172-s99, and p152-f97 studies show no obvious correlation, which suggests that the apparent clustering visible in the timelines can be described as statistical fluctuations in an uncorrelated random variable. Note that this does not imply that the thinking times in a FTE data set are truly random in origin, or that response terms are in fact unrelated to each other, or that the apparent clusters are without meaning; it merely means that the set of thinking times has the statistical properties of a sequence of uncorrelated random variables. It remains quite possible that a detailed cognitive model of the processes elicited by the task could explain the observed data without invoking random distributions.

This "no correlation" result does not even imply that there is no statistical difference between the observed data and the pattern expected for an uncorrelated random variable, merely that this test is not sensitive to any difference that might exist. Other tests might be worth investigating.

Perhaps the most significant conclusion to be drawn from this result is that it appears unrealistic to seek a purely statistical criterion by which to identify "clusters" in the response list for use in further analysis. If we wish to define clusters, for example to test the hypothesis that clusters contain sequences of terms which are related in the topic domain, an external rule must be imposed. An example of such a rule might be "a cluster is defined as a sequence of terms separated by thinking times of less than τ_c, preceded and followed by thinking times greater than τ_c,", which depends on a choice of τ_c.

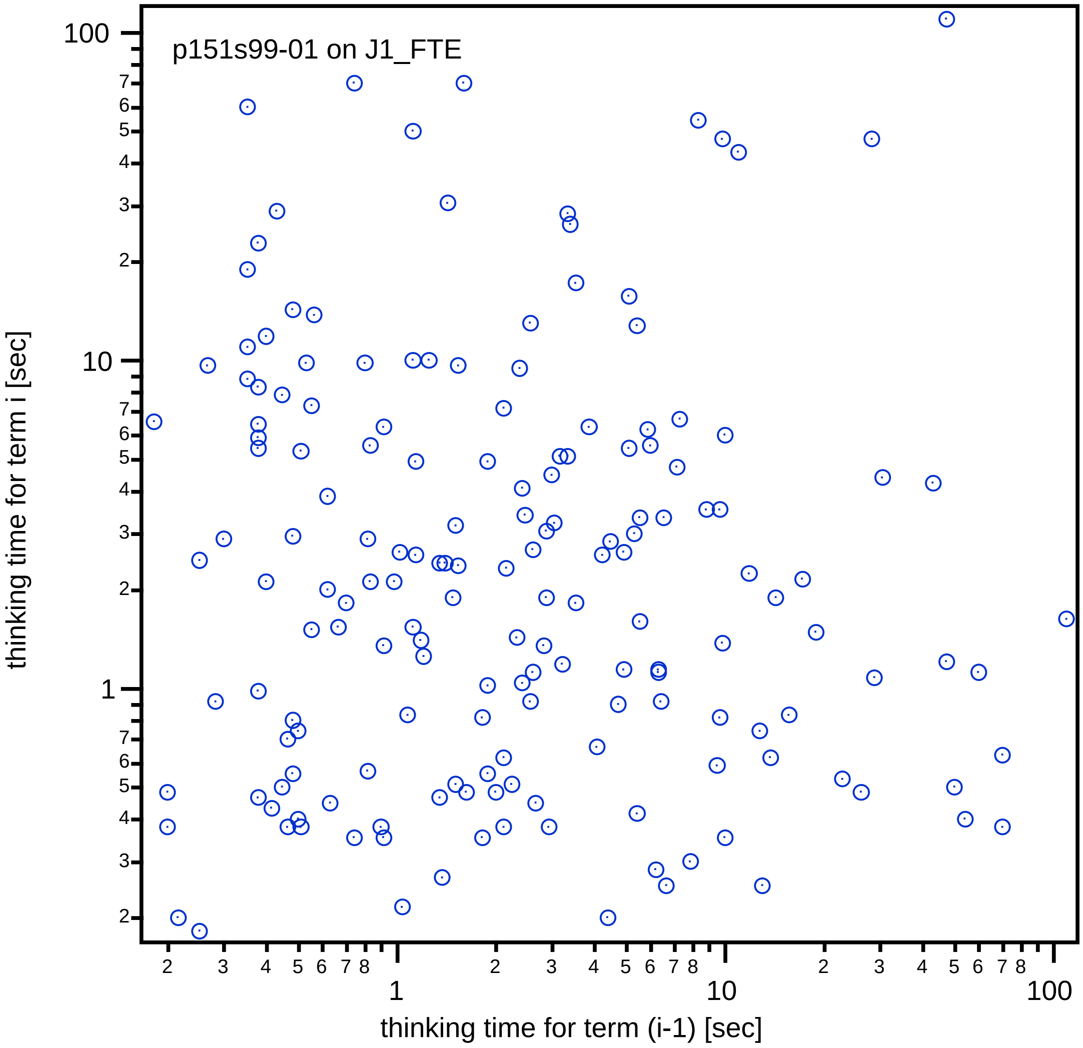


Figure 3.27: Successor correlation plot for thinking times, for subject p151s99-01 on task J1_FTE.

3.1.2. FTE Thinking Times vs. Term Relatedness

In Subsection 3.1.1, analysis of FTE data focused on timing information and ignored the meanings of the terms entered by subjects. This section will attempt to relate thinking times to one aspect of term meanings, the degree to which adjacent terms in the FTE response list are related or unrelated to each other.

3.1.2.1. Introduction

According to the introspective testimony of experts who served as FTE subjects in preliminary studies, term entry events can be crudely classified into two types: those for which the term to be entered was immediately suggested by terms immediately preceding it, and those for which the subject had to search his/her memory for some period of time to think of the term. According to these expert subjects, the immediately-suggested terms were generally closely related in meaning to one of the prior few terms, while the terms thought of after a period of mental searching were usually related only distantly to prior terms. This suggests the following hypothesis: let the term *jump* refer to a term which is "relatively unrelated" to any one of the previous *n* terms in a FTE response list; then, in FTE response data, longer thinking times should be statistically more likely to occur for jumps, while short thinking times should be more likely to occur for non-jumps.

To make this hypothesis testable, "relatively unrelated" must be defined, and a value for *n* must be chosen. For this initial investigation, *n* was arbitrarily set at three, which seemed reasonable based on perusal of the response lists and interviews with expert subjects. Each term in each response list analyzed was compared to the three previous responses in the list by one domain expert (the author), and subjectively declared to be relatively related to one of the three (non-jump) or relatively unrelated to all three (jump).

A neighborhood of multiple preceding terms was used, rather than the one immediately preceding term, because interviews and perusal of term lists suggests that subjects often enter a term and then enter a sequence of multiple terms that come to mind approximately simultaneously. That is, a subject enters term A, and quickly thinks of terms B and C which are related to A but not necessarily related to each other; the subject then enters B and C. C is therefore a jump if one only considers it relative to B, but not if A is also part of the context. It is the testimony of some expert subjects that sometimes when they think of a term they perceive a "fork" in the mental path, with two possible "threads" of related terms that they could follow. In such a case they often try to follow one thread while it is productive, and then return to the fork and pick up the other thread. It seems intuitively reasonable that if the first thread is longer than about three terms, remembering and returning to the other thread is likely to introduce a significant thinking time, thus warranting classification as a jump.

Making explicit the criteria used to judge whether any pair of terms is related or not has proven to be quite troublesome. Experts seem to possess an intuitive notion of whether terms are related, but have difficulty explicitly identifying their criteria. In addition, experts seem to use contextual information in their judgments, inferring what the subject was thinking while he/she entered a series of terms, and deciding whether a term is a jump in that context.

The following list attempts to specify some of the criteria used to decide whether a pair of terms was "relatively related":

- They were both within a sufficiently small topic area (e.g. collisions, graphs, angular momentum);
- They were analogous elements of a set or list (e.g. kinds of forces, units of measure, key principles of mechanics);
- One was a subclass or special case of the other (e.g. "force" and "spring [force]", "motion" and "rotation");
- One was a situation or problem type in which the other figures significantly (e.g. "falling objects" and "gravity", "collision" and "impulse");
- They were mathematically related (e.g. "work" and "impulse", "velocity" and "position");
- One was a feature of the other (e.g. "slope" and "graph", "force" and "free-body diagram").

This is not a complete list, but it illustrates the kinds of relationships considered.

Note that a very important question has been ignored so far: related *to whom*? The original hypothesis, based on experts' introspection on their own experience while performing a FTE, was that long thinking times correlate with terms unrelated to immediately preceding terms *according to their own knowledge structure*. When an expert analyzing the data examines a subject's list of responses and identifies terms as jumps or non-jumps, however, the judgment of relatedness is made according to the expert's understanding of the domain, not the subject's. So, even if the hypothesis is completely correct and thinking times correlate perfectly with jumps, analysis by an expert would not show a perfect correlation unless the expert and subject were in complete agreement about what terms are and are not strongly related.

We assume, however, that an expert with experience teaching the domain material can make judgments based on a structure that is reasonably close to what earnest students, or at least the more apt ones, possess. With that assumption, the operational hypothesis to test is that long thinking times will correlate noisily but significantly with jumps as perceived by an expert.

In fact, if the original hypothesis is correct and thinking times reveal what are and are not jumps to the subject, then this task could provide a mechanism for comparing parts of a subject's conceptual knowledge structure to an expert's. If a subject entered a term after a short thinking time but the term appears to be a jump to an expert, then perhaps the subject has attached importance to a link which ought not to be so important; this might indicate a misconception. The converse case seems less informative: if a subject enters a term that an expert considers related but does so after a long thinking time, it is not clear whether the subject does not in fact associate the term with its predecessors very strongly, or whether he/she considered several other terms and rejected them (perhaps because they were entered earlier in the task), or whether he/she was simply distracted for a span of time.

But first, a correlation between thinking times and term relatedness must be established.

3.1.2.2. Distributions of Jump and Non-Jump Thinking Times

For each of the 16 subjects in the p151s99 study, an expert in introductory mechanics with experience teaching the subject (the author) reviewed the list of response terms for the task J1_FTE, and classified each term as a *jump* or *non-jump* as explained above. The set of thinking times for the subject's task performance was then divided into a subset containing thinking times for jumps and a subset containing thinking times for non-jumps. Figure 3.28 shows histograms of these two subsets for subject p151s99-01, superimposed on the same axes. For comparison, Figure 3.29 shows the two distributions as stacked histograms, revealing the histogram for the set of all thinking times. In keeping with Subsection 3.1.1, the natural logarithms of the thinking times have been used rather than the times themselves.

Figure 3.30 displays one of the noisier of such histogram comparisons, for a subject whose data contains relatively few terms. While some of the data sets are too noisy to identify a clear peak for both histograms, for all but one of the 16 subjects, the mean and median of the jump distribution is clearly larger than the mean and median of the non-jump distribution. The one exception is subject p151s99-14, whose data set contains atypically few points, resulting in atypically sparse, noisy histograms with similar means and medians.

The general pattern is clear: for any given subject, the thinking times associated with jumps are generally larger than the thinking times associated with non-jumps, but the two distributions overlap significantly. There are typically more non-jumps than jumps, although the ratio varies by subject. For

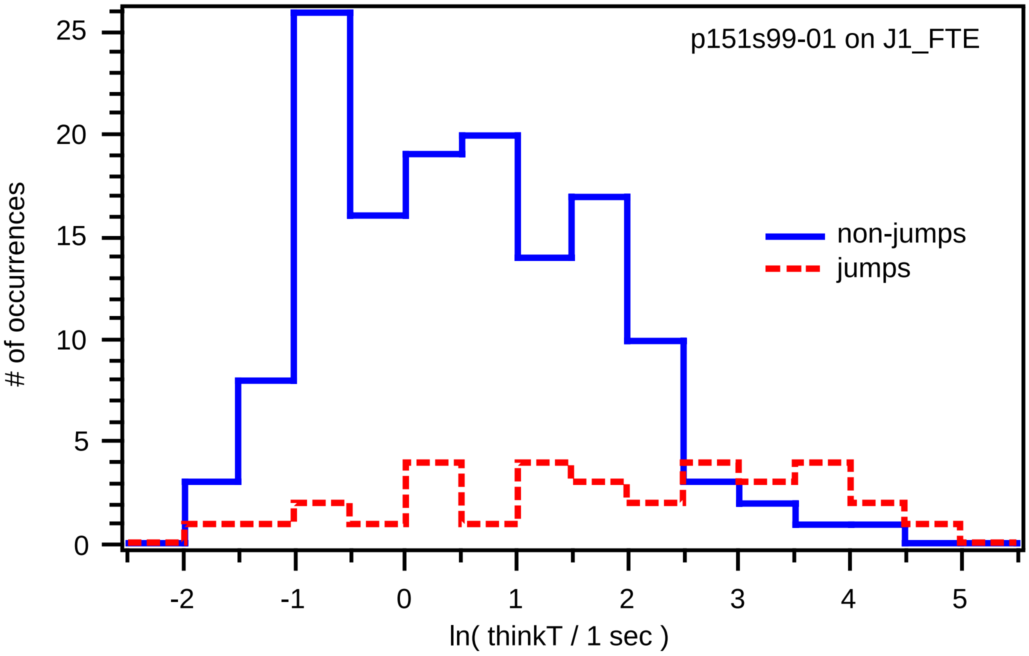


Figure 3.28: Comparison of histograms of logarithms of thinking times for jumps and for non-jumps, for subject p151s99-01 on task J1_FTE.

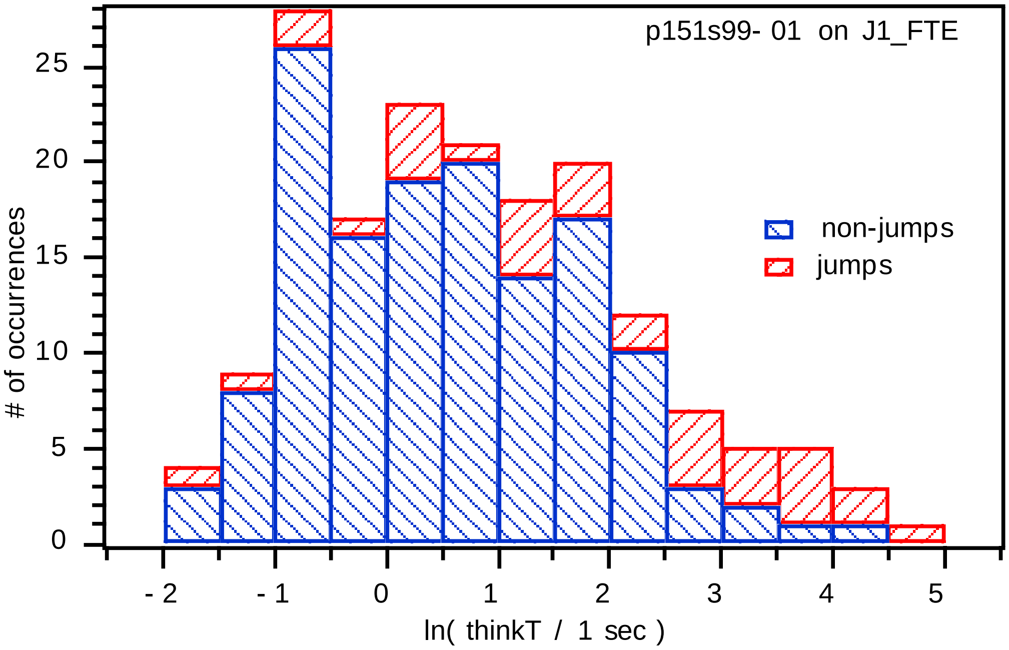


Figure 3.29: Stacked histograms of logarithms of thinking times for jumps and for non-jumps, for subject p151s99-01 on task J1_FTE, revealing histogram of all thinking times.

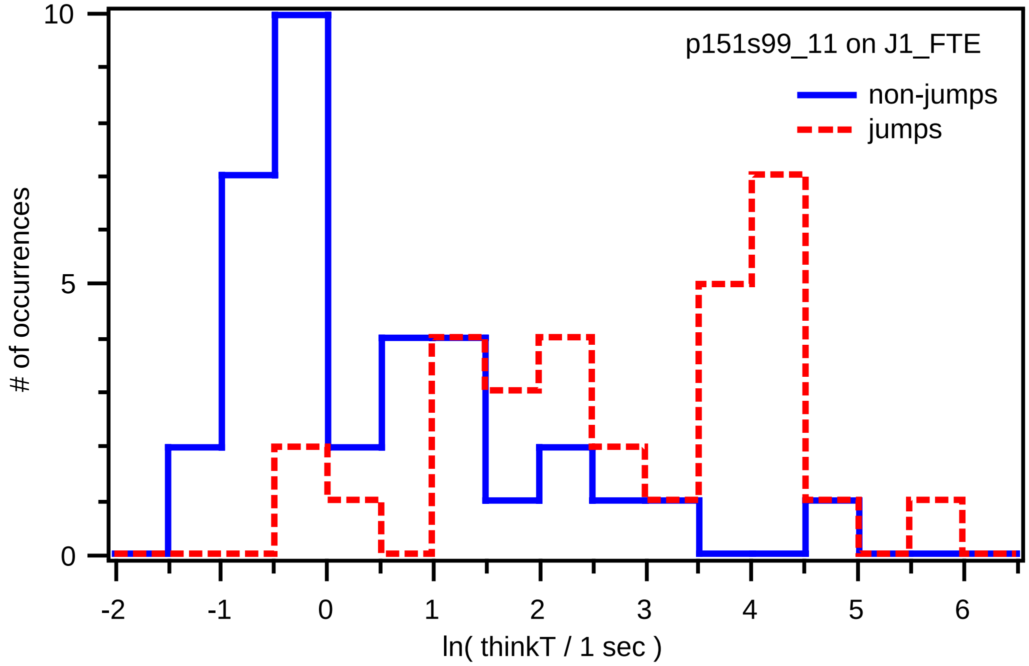


Figure 3.30: Same as Figure 3.28, for subject p151s99-11.

each subject, the fraction of terms which were identified as jumps (*jump rate*) was calculated; the resulting set of 16 jump rates had a mean of 0.36, standard deviation of 0.10, minimum of 0.19, maximum of 0.54, and median of 0.35.

3.1.2.3. Predicting Jumps by Thinking Time Threshold

As previously mentioned, it could be useful to identify terms with long thinking times which an expert classified as a non-jump, and terms with short thinking times which an expert classified as a jump. One way to attempt this is to define a threshold time, and predict that at all terms whose thinking times are greater than the threshold time will be jumps, and all terms whose thinking times are less than the threshold time will be non-jumps. Terms for which the predicted categorization differs from the expert-assigned categorization can then be identified for possible cognitive or pedagogic interpretation.

For each subject, define the *success rate* of the prediction to be the fraction of terms for which the predicted and expert-assigned categorizations agree. To produce a set of predicted categorizations for a subject's FTE data, it was necessary to specify a threshold time. Three possible methods were considered:

1. Choose a threshold time which equals the thinking time at which the thinking-time distributions for jumps and non-jumps cross, in a plot like Figure 3.28.

2. Choose a threshold time which produces the same jump rate as an expert's categorizations;

3. Choose a threshold time which maximizes the success rate of the resulting predictions.

Methods 1 and 3 are in fact equivalent if there are enough data points so that the data's discreteness is not an issue. This can be understood by looking at Figure 3.28 or Figure 3.29 and considering a vertical line drawn at a horizontal coordinate where the thinking time is equal to the chosen threshold time. The total number of events to the left of that line due to both distributions is the predicted number of non-jumps, while the total number to the right is the predicted number of jumps. Moving that line to the right (i.e. increasing the threshold time) increases the number of predicted jumps. Every time the line passes a thinking time corresponding to a term entry while moving to the right, the predicted classification of that term changes from incorrect to correct if the term is part of the non-jump histogram, increasing the success rate; if the term is part of the jump histogram, the success rate is decreased. Assuming the distribution for jumps peaks farther to the right than the distribution for non-jumps, the maximal success rate must therefore occur at the point at which the two distribution curves (approximated by histograms) cross.

With discrete data rather than idealized continuous distributions, multiple crossing points are possible, in which case the success rate has multiple local maxima; the largest should be chosen. There may exist multiple maxima of equal height, in which case a rule must be defined to resolve the ambiguity.

Figure 3.31 and Figure 3.32 show plots of success rate vs. threshold time for the two example subjects of Figure 3.28 and Figure 3.30. The effect of discreteness for small data sets is clearly visible: the first subject entered 174 terms, and the second entered 67.

Table 3.3 shows optimal threshold times and the corresponding maximized success rates for each subject as determined by method 3, calculated numerically from the data rather than from histograms to avoid binning effects. For a given subject, if the maximum success rate value occurred for multiple values of the threshold time, the reported threshold time value is the logarithmic mean of those values.

When interpreting the success rates, consider that if the threshold-time prediction and the expert assignment are perfectly correlated, the success rate will be 1; if they are completely uncorrelated, it will have a statistical expectation value of $f_p f_j + \left(1 - f_p\right)\left(1 - f_j\right)$, where f_p is the jump rate according to the threshold-time prediction, and f_j is the jump rate according to the expert's categorization judgments. The table includes columns for the jump rate according to the expert judgments, the jump rate according to the threshold prediction, and the success rate expected if the predictions and judgments were uncorrelated.

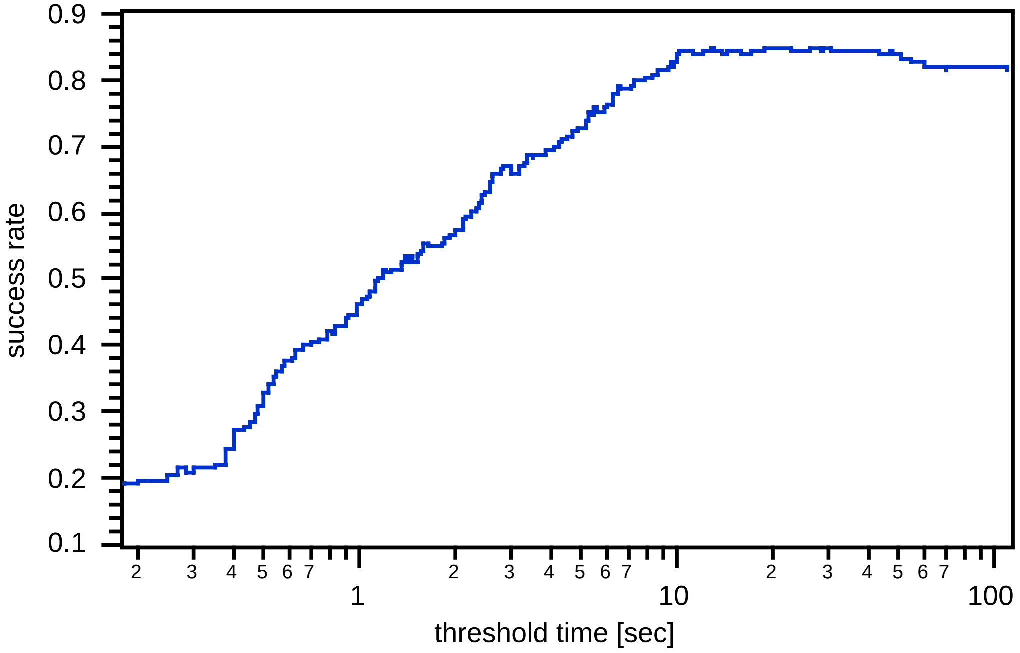


Figure 3.31: Success rate vs. threshold time for jump/non-jump prediction, for subject p151s99-01 on task J1_FTE (174 terms entered).

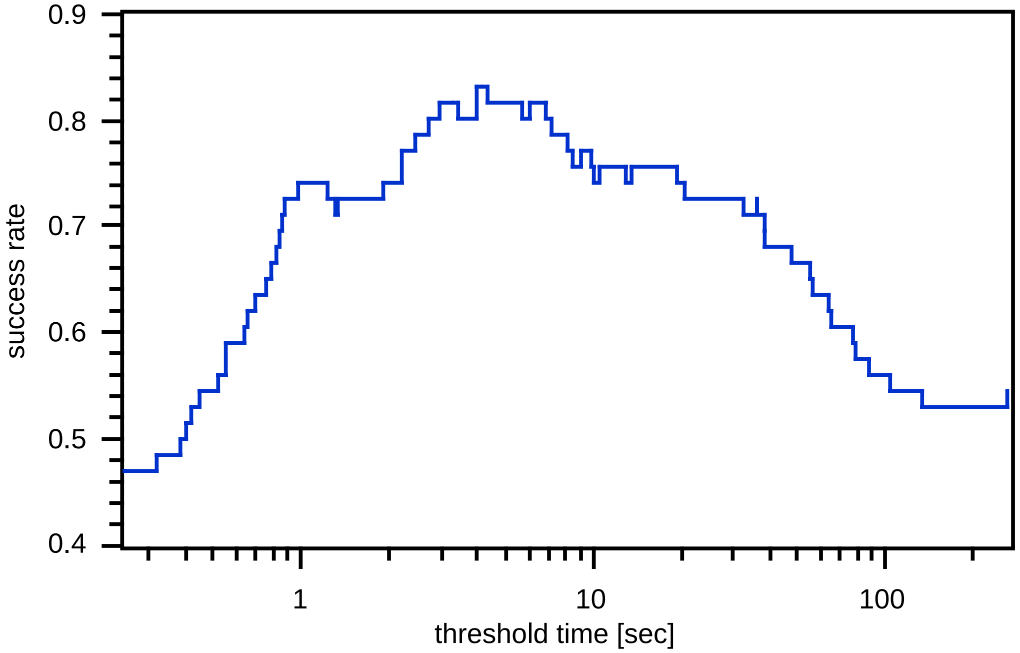


Figure 3.32: Same as Figure 3.31, for subject p151s99-11 (67 terms entered).

		cutoff	max	judged	pred.	uncor.	error
	cutoff	time	success	jump	jump	success	rate
subject	time	ln()	rate	rate	rate	rate	ratio
p151s99-01	20.67	3.03	0.85	0.19	0.08	0.76	0.63
-02	14.03	2.64	0.79	0.30	0.20	0.62	0.56
-03	7.00	1.95	0.79	0.40	0.44	0.51	0.43
-04	18.74	2.93	0.72	0.42	0.23	0.54	0.61
-05	3.86	1.35	0.70	0.54	0.60	0.51	0.61
-06	7.00	1.95	0.87	0.27	0.26	0.61	0.33
-07	10.53	2.35	0.79	0.43	0.27	0.53	0.45
-08	2.54	0.93	0.79	0.46	0.63	0.49	0.42
-09	13.22	2.58	0.77	0.33	0.33	0.56	0.52
-10	19.12	2.95	0.86	0.24	0.08	0.71	0.49
-11	4.07	1.40	0.83	0.46	0.48	0.50	0.33
-12	11.99	2.48	0.94	0.25	0.22	0.64	0.17
-13	14.13	2.65	0.84	0.32	0.24	0.59	0.40
-14	22.94	3.13	0.77	0.28	0.05	0.70	0.77
-15	3.82	1.34	0.70	0.46	0.48	0.50	0.60
-16	8.39	2.13	0.79	0.37	0.32	0.55	0.45

Table 3.3: Selected threshold times and corresponding success rates for comparison of predicted and expert-judged “jump” vs. “non-jump” term categorization, for p151s99 study, task J1_FTE; with comparison to success rate expected if prediction and expert judgment are uncorrelated (see text).

Define the *error rate* of a prediction to be the success rate subtracted from one; that is, the fraction of terms that were mispredicted. The final column shows the ratio of the error rate of the prediction to the error rate expected for uncorrelated predictions; values less than one indicate a smaller error rate (better prediction), while values greater than one indicate a higher error rate (poorer prediction). The average of that ratio across subjects is 0.49, indicating that the threshold-time prediction method employed in this section produces about half the errors that would be obtained by a random coin-toss with bias equal to the number in the “predicted jump rate” column.

Whether the listed success rates are considered adequate depends on the use one intends for the resulting predictions. For an ideal case where a subject’s distribution of thinking times fell into two distinct peaks, and where an expert judged most of the terms comprising the first peak to be non-jumps and most in the second peak to be jumps, identifying the few jumps in the first peak and the few non-jumps in the second peak would likely be of value for pedagogic and research purposes. For such a case, the threshold method described above would suffice. But for a case like that displayed in Figure 3.28, the threshold method seriously overpredicts the number of non-jumps. If the threshold is selected by methods 1 or 3, almost all terms are predicted to be non-jumps. As a result, the majority of jumps are mispredicted as non-jumps.

As discussed at the end of Subsubsection 3.1.2.1 above, jumps with the timing signature of non-jumps are likely to be of more cognitive and pedagogic interest than the converse case. The threshold method tends to overpredict such events, reducing their usefulness. Threshold-determination method 2, requiring the jump rate to be the same for predictions and expert judgments, would force the threshold line left of the histograms' crossing point on a case like that of Figure 3.28, reducing the number of falsely predicted non-jumps at the expense of very sharply increasing the number of falsely predicted jumps. This might be of benefit to a cognitive or pedagogic analysis. Without a specific analysis in mind, further discussion is not fruitful.

3.1.2.4. Incorporating Elapsed Task Time in Jump Predictions

Figure 3.33 shows a plot of thinking time vs. start time for an example subject. Each data point represents one term-entry event, and the horizontal axis indicates the start time of the event (the time elapsed in the task when the term was entered). Data point markers indicate whether each term was classified as a non-jump (cross) or jump (circle) by the expert judge.

Examining such plots for all subjects in the p151s99 study reveals some general trends:

1. Thinking times are scattered within an envelope that increases as the task progresses (i.e. as start time increases), in agreement with the discussion on decreasing term entry density in Subsection 3.1.1.5.

2. The density of jumps relative to non-jumps is higher in the later part of the task than in the earlier part.

3. Overall, jumps have larger thinking times than non-jumps, in agreement with the previous section's findings.

In this representation, the threshold time prediction method of the previous section corresponds to drawing a horizontal line through the plot, and predicting that all points above the line correspond to jumps and all points below the line correspond to non-jumps. The fact that no such line cleanly divides the jump points from the non-jump points is consistent with the fact that the two histograms of Figure 3.28 overlap.

It is likely that a non-horizontal line, or even some kind of parameterized curve, might be more successful at partitioning the jumps from the non-jumps. This is equivalent to modifying the threshold-time prediction method to use a threshold that varies with elapsed task time (start time). Although success rates for such a method have not been calculated, examining graphs like Figure 3.33

for all subjects in the p151s99 study suggests that for some subjects it would be significantly more successful, while for others (including the example subject shown above) the improvement would be minor. Again, whether such methods are useful depends on the purpose one has for the results.

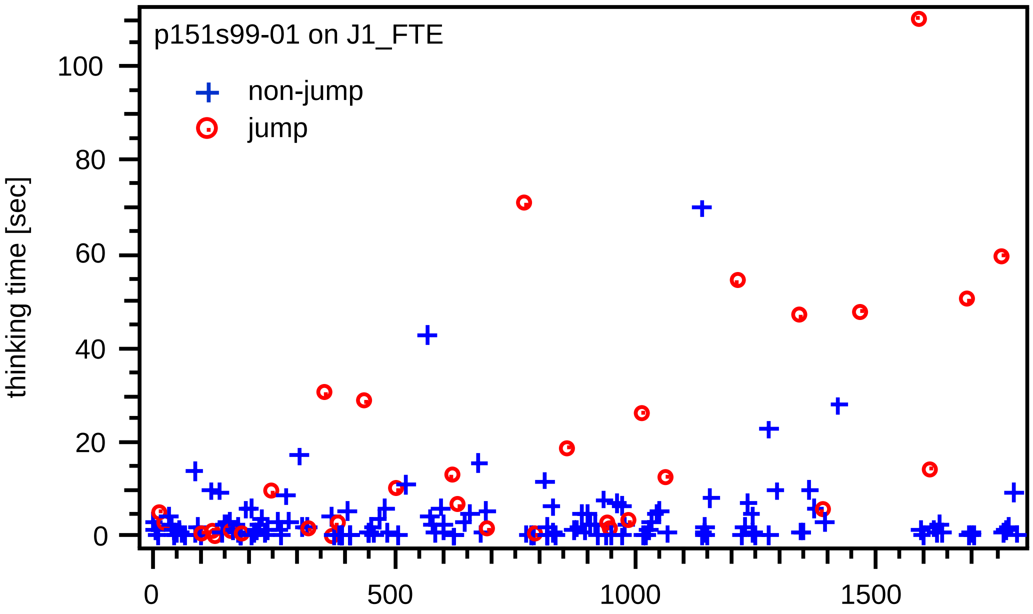


Figure 3.33: Term thinking time vs. term start time (relative to beginning of task) for subject p151s99-01 on task J1_FTE. Symbols indicate expert's classification of terms (see text).

3.1.2.5. Suggestions for Further Research

The line of inquiry discussed here in Section 3.1.1 could be pursued in several ways. One would be to reduce the noise introduced by the expert's judgment of which terms should be categorized as jumps or non-jumps. A simple improvement would be to have a panel of experts make the judgments, rather than one expert. Explicitly identifying criteria for the experts to apply should aid consistency of judgment.

Going a step further in this direction, a "reference proximity matrix" could be constructed, with each cell containing a numerical value representing the proximity or "relatedness" of a pair of terms. Constructing such a matrix would be a laborious task, perhaps achieved by subjecting several experts to a "term proximity judgment" (TPJ) task (cf. Section 2.1.3.6). Once the matrix exists, a rule could be defined which categorizes terms as jumps or non-jumps according to a numerical criterion based on that term's proximity values to a specified number of preceding terms. As a coincidental benefit, comparing a matrix-based categorization to an expert's subjective judgment might shed light on the relationship of considered, TPJ-style judgments to subjective judgments in the context of a FTE response list.

The original hypothesis with which we started was that short thinking times correspond to successions of related terms, while long thinking times correspond to jumps to unrelated terms, *from the subject's point of view.* As described then, comparing subject thinking times to expert categorizations confounds this correlation with the correlation between subjects' and experts' sense of term relatedness. This confounding could be eliminated by giving subjects another task which extracts their sense of what terms are related, perhaps a TPJ-style task. Note that the time required for a TPJ task could be reduced significantly by only filling in the elements of the matrix corresponding to the term pairs needed for categorizing terms in the subject's FTE response list; this reduces the order N^2 general TPJ problem to an order N problem. For term lists of 100 terms or more, this is a crucial improvement.

3.1.2.6. Conclusions

The analysis described in Section 3.1.1 found that there exists evidence to connect FTE thinking times with the conceptual relatedness of the corresponding terms. The correlation is noisy, however, and confounded because an expert's judgment of what terms subjects ought to relate was substituted for the subjects' own conceptual associations. Nevertheless, the line of inquiry is promising, and some follow-up research has been suggested which would remove the confounding problem and better answer the question of how the timing information in FTE tasks relates to subjects' conceptual knowledge structure.

3.1.3. FTE Jump Rates vs. Exam Scores

The previous subsection (3.1.1) sought evidence of conceptual structure in a subject's FTE responses by correlating thinking times with the presence or absence of "jumps" to locally unrelated terms in the sequence of term responses. This section examines whether the frequency with which such jumps occur in a subject's response list correlates with his/her mastery of the domain material, as measured by course exam scores. Although exam scores are not likely to capture subjects' level of expert-like structuring in their conceptual knowledge store (a fact central to the motivation of this entire body of research), no other measure independent of the experimental ConMap data was available, so exam scores were employed.

3.1.3.1. The Idea

By hypothesis, the less well-structured a subject's relevant domain knowledge is, the more frequently terms in their response sequence will appear to be unrelated to immediately preceding terms. We expect that a subject with richly structured knowledge could easily "walk" that structure, with each term they enter suggesting other related terms. In contrast, we expect the response pattern of a subject with amorphous, poorly-structured knowledge to appear much more random, with more frequent leaps from one subdomain to another.

As before, define a term-entry event in an FTE response list to be a "jump" if it does not appear to be "reasonably related" to one of the few immediately-preceding response terms, according to a domain expert's judgment. For the analysis of this section, the jump/non-jump identifications of the previous section's analysis were used. Define a subject's *jump rate* on a FTE task to be the number of such jumps divided by the total number of response terms (not including the first). A jump rate of 0 would indicate that every term is related to one of the previous few terms; a jump rate of 1 would indicate that every term was unrelated to all of the previous few.

3.1.3.2. Analysis

The p152f97, p151s99, and p172s99 studies each presented subjects with one FTE task. For each subject in each of these studies, a jump rate for their FTE response list was calculated. The degree of correlation between subjects' jump rates and their exam scores was checked for each of the exams in the relevant course. For each exam, a scatterplot of FTE jump rate vs. exam score was constructed, with each data point representing one subject. A line was fit to the plot, and "Pearson's *r*-value", a measure of the significance of the correlation, was extracted from the fitting procedure. Figure 3.34 shows an example of such a plot.

The p152f97 study was conducted at the end of the Physics 152 course, on the material from that course (thermodynamics, electricity and magnetism); subjects' FTE jump rates from that study were compared with their exam performance on each of the exams during Physics 152. The FTE task of p151s99 study was conducted at the end of the Physics 151 course, on material from that course (mechanics), and subject's jump rates were compared with their exam performance on each of their Physics 151 exams. The FTE task of the p172f99 study was conducted at the end of Physics 172, on material from Physics 171 (mechanics), which the subjects had taken the previous semester; subjects' FTE jump rates were compared with their exam performance on each of the exams during Physics 171.

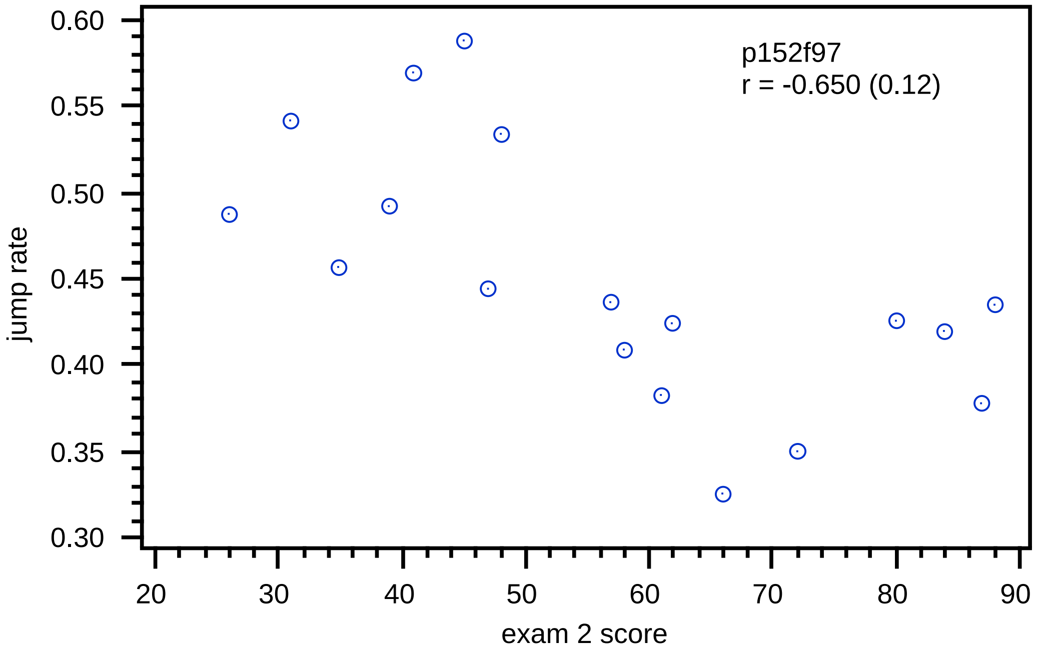


Figure 3.34: Subject FTE jump rate vs. Exam 2 score for p152f97 study.

During the identification of "jumps", we noticed a general trend: more jumps seemed to occur in the second half of the FTE response sequence than in the first. This agrees with subjects' testimony when interviewed: toward the end of an FTE task, subjects generally experience a greater sense of "hunting around" in their memories for terms that they haven't yet entered, whereas in the beginning they enter terms almost continuously.

This suggests that the first half of the FTE response sequence may be a better indicator of structure than the later part. We therefore repeated the comparison of *r*-values with exam scores, using only the first half of each subject's FTE response sequence to calculate a jump rate.

For the p152f97 study, Table 3.4 shows *r*-values for the correlation between subjects' jump rates and their various exam scores. Table 3.5 shows results for the same calculation for the p151s99 study's J1_FTE task. Table 3.6 shows the same for the p172s99 study's B1_FTE task.

	Exam 1	Exam 2	Exam 3	Final
all responses	-0.489	-0.650	-0.059	-0.665
first 1/2 responses	-0.244	-0.586	-0.587	-0.506

Table 3.4: Pearson's *r*-value for correlation between subject jump rates and exam scores in p152f97 FTE task. $|r| > 0.12$ for statistical significance with an 18-point sample.

	Exam 1	Exam 2	Exam 3	Final
all responses	-0.067	-0.44	-0.053	-0.091
first 1/2 responses	-0.36	-0.59	-0.32	-0.20

Table 3.5: Pearson's *r*-value for correlation between subject jump rates and exam scores in p151s99 J1_FTE task. $|r| > 0.13$ for statistical significance with a 16-point sample.

	Exam 1	Exam 2	Final
all responses	0.679	0.677	0.648
first 1/2 responses	0.827	0.804	0.730

Table 3.6: Pearson's *r*-value for correlation between subject jump rates and exam scores in p172s99 B1_FTE task. $|r| > 0.25$ for statistical significance with a 5-point sample.

	mean jump rate	mean jump rate
	(all responses)	(first half)
p151s99	0.319 (0.125)	0.263 (0.122)
p172s99	0.419 (0.097)	0.351 (0.091)
P152f97	0.450 (0.073)	0.408 (0.077)

Table 3.7: Mean jump rates for p151s99, p172s99, and p152f97 populations on their respective FTE tasks. Numbers in parentheses indicate the corresponding standard deviations.

To compare the mean jump rates of the different study populations, all subjects' FTE jump rates were averaged within each study population, for both the whole and first-half data sets. Table 3.7 displays the results, with parentheses indicating the corresponding standard deviations.

3.1.3.3. Discussion

For the p152f97 data, all calculated coefficients of correlation are above the threshold for statistical significance, indicating an unquestionable (though perhaps noisy) correlation. Three of the four *r*-values are slightly lower for the half-data sets than for the full data sets, and the fourth is nearly identical, weakly suggesting that jump rate during the later part of the FTE is a slightly better predictor of ability than jump rate during the earlier part. This is contrary to our hypothesis about first-half jump rates.

For the p151s99 data, calculated *r*-values are above the statistical significance threshold for one of the four exams when calculated from entire response sequences, but exceed the threshold for all four exams when calculated from first-half response sequences. Ignoring the second half of the response sequences significantly increases the *r*-value for each exam comparison, suggesting that jump rate during the first part of the FTE task is a better predictor of ability than jump rate during the second part.

For the p172s99 data, calculated *r*-values are well above the statistical significance threshold, even though that threshold is higher than for the other two studies since this study only included five subjects. Unfortunately, the correlation is positive, whereas it is negative for the other two studies. If taken seriously, this would suggest that students who do better on exams have *higher* jump rates.

Looking at the plots for the p172s99 data, it appears that the strong positive correlation is entirely attributable to two outlying subjects: one with very weak exam scores and a very low jump rate, and another with very strong exam scores and a very high jump rate. The other three subjects are all clustered in the middle on both axes. This suggests that the strong positive correlation may be an artifact of two atypical subjects. The subject with top exam scores and high jump rate was in fact a high-school student at the time, taking Physics 171 and 172 for high-school credit; this makes him a rather atypical subject.

Overall, the evidence suggests that jump rate correlates (albeit noisily) with exam performance. The evidence is less clear, but still suggestive, that the early part of the FTE response sequence correlates more strongly. When considering the discrepancy between the p151s99 and p152f97 results — specifically, the fact that for one the first-half data sets correlated more strongly with exam scores than did the entire data sets, while for the other it correlated more weakly — it is important to note that task conditions for the two groups were different in many ways. For one thing, p151f97 subjects were given approximately 45 minutes for the task, while p151s99 subjects were given 30 minutes. For another, the FTE task was the only task in the only session p152f97 subjects underwent, whereas it was presented to the p151s99 subjects during their tenth session of tasks.

Regarding the comparison between the mean jump rates for p151s99 and p172s99 groups: one might expect the p172s99 group to show more of a signature for "expertise", since the subjects took a more advanced version of the introductory mechanics course, and since only students who survived the course well enough to continue in Physics 172 were in the sample. On the other hand, an entire semester had passed between their completion of the relevant course and their participation in the ConMap study, which might have a significant effect on term recall and knowledge structure. In addition, all five subjects had taken the sequel course Physics 172 during the intervening semester, which might have impacted their knowledge structure and recall, perhaps "diluting" it with the addition of new links.

3.1.3.4. Suggestions for Future Research

The correlation between jump rate and exam performance seems strong enough to warrant further study. Improving the procedure for identifying jumps, as discussed in Section 3.1.2.5, would be of benefit.

A better indicator of conceptual domain expertise than course exam scores is crucial. The study should be repeated, presenting subjects with a "problem-solving task" along with a FTE task. The new task should require subjects to solve carefully-crafted problems designed to test conceptual understanding of the domain material. This should remove a tremendous amount of noise and confounding from the correlation being studied and allow a more reasonable assessment of the hypothesis that FTE jump rate correlates with expertise.

3.1.4. Summary of FTE Findings

To summarize the findings of Section 3.1: the timing information contained in a subject's FTE response list data can be separated into a set of *thinking times* which describe the approximate amount of time the subject spent thinking about each term, and a set of *typing times* which describe the approximate amount of time the subject spent typing each term. For an entire response list, the set of thinking times approximately follows a log-normal distribution, although for most subjects there is a narrow, tall spike superimposed on the leading edge of the generally Gaussian peak when the distribution is viewed on a logarithmic scale. The typing times do not display this peak. When individual subjects' sets of thinking time logarithms were rescaled to a common mean and width and then aggregated together, the resulting aggregate set displayed a clear two-peaked shape which was fit well by a linear combination of two Gaussian peaks. Individual data sets were in general too noisy to fit well with this five-parameter curve, however.

In checking the hypothesis that at least some subjects thought about their next terms while typing a term, it was found that there was no correlation between a term entry event's thinking time and the previous event's typing time. If significant thinking occurred during typing, one might expect to see an inverse correlation. A few subjects showed a slight tendency for longer thinking times to follow longer typing times, a phenomenon which has various plausible explanations.

Subjects' response lists showed a general pattern of decreasing density, meaning that the number of terms entered per unit time, suitably averaged, decreased throughout the duration of the task. Density fluctuations were significant, however, so the decrease was not smooth even when averaging over a moving window 1/4 the duration of the entire task. All subjects' highest rate of term entry occurred near the beginning of the task, but not all had even approximately monotonic rate decreases throughout.

The "density fluctuations" in term entry rate give the appearance that term entry events tend to occur in clusters, but a check for a correlation between the thinking time for an event and the previous event's thinking time was negative, suggesting that the apparent clustering is statistically equivalent to the random fluctuations observable if thinking times were independently drawn from an appropriate random distribution. This is not a statement about any lack of meaning conveyed by the clustering, but merely an observation about the statistical properties of the data.

Section 3.1.1 investigated the hypothesis that longer thinking times in a subject's FTE response list are associated with *jumps* in the list of terms, where a jump is defined to be the entering of a term which is not strongly related to one of the previous few terms entered. Without a mechanism to determine how strongly a subject associates term pairs, it was necessary to compare thinking times to an expert's assessment of how related pairs of terms should be for student subjects. Each subjects' set of thinking times was partitioned into two subsets, one for the events classified as jumps and one for the non-jumps. For each subject, the two sets were distinctly different, with the distribution of jumps having its peak at a larger value of thinking time. The two sets overlapped significantly, however.

A simple threshold rule was defined which predicted each term entry event to be a jump or non-jump based on only its thinking time. With optimal choice of thinking time, such predictions were moderately successful for most subjects, making on average one-half the number of erroneous predictions as would have been made by a coin-toss biased to the same ratio of jumps to non-jumps. In particular, the rule underpredicts jumps, which may be detrimental to likely cognitive and pedagogic uses of this kind of analysis.

Elaborations of the threshold rule that employ information about how thinking times correlate with elapsed task time might improve predictions somewhat, but it is clear thinking time does not correlate strongly enough with the expert's categorizations for any rule to predict significantly greater success.

Section 3.1.3 investigated the hypothesis that subjects' jump rates — the fraction of term entry events categorized by an expert as jumps — on FTE tasks should correlate inversely with their level of expertise in the domain material and therefore with their performance on associated course exams. Calculations of Pearson's *r*-value for subject jump rates against exam score, by study population and exam, show a significant correlation as expected. There is some weak evidence that jump rates for the first half of subjects' response lists correlate more strongly, at least for the p151s99 study.

Overall, analysis of data from FTE tasks indicates that this task does in some ways probe subjects' knowledge structure, and therefore further study is merited.

3.2. Hand-Drawn Concept Map (HDCM) Data Analysis

Unlike most of the ConMap tasks, the Hand-Drawn Concept Map (HDCM) task was not computer-administered. Instead, subjects were provided with a large (11 inches by 17 inches) piece of blank paper and a pencil, and instructed to draw a concept map elaborating a given prompt term. A clock display on the computer counted down the remaining time in the task. The data collected consisted of each subject's drawn map.

This section presents a brief phenomenological description of maps drawn during the p151s99 study. The p172s99 study also employed the HDCM task, but since it only included five subjects, data from that study is not discussed here. In the p151s99 study, HDCM tasks were given during Sessions B, G, H, and J. The prompt terms given were "force", "energy", "momentum", and "force" respectively, with time limits of 8, 10, 10, and 10 minutes.

Subsection 3.2.1 defines and presents some quantitative measures of the data. Subsection 3.2.2 examines these quantities for possible correlations with subjects' course exam performance. Since the primary purpose that the HDCM task was intended to serve in the study was as a basis for comparison with the TPTE task, no further analysis of HDCM data is discussed in this section; subsection 3.3.6 compares TPTE data with HDCM data. Subsection 3.2.3 summarizes the findings of the section.

3.2.1. Data Quantification

In order to allow some quantitative analysis, some measures of subjects' HDCM task maps were tabulated. For each map, the *node count* (the number of nodes included by the subject, excluding the node containing the prompt term) was determined, as was the *link count* (the number of inter-node links). The ratio of link count to node count was computed. Each node was assigned a *level* indicating how far removed the node was from the prompt node: nodes directly linked to the prompt node were defined to be at level one; nodes directly linked to a level-one node but not the prompt node were defined to be level two; and so on for as many levels as was necessary to describe the entire map. For each map, the *level counts* — the number of nodes assigned each level number — were tabulated. Table 3.8 displays these quantities for each subjects' map from the H2_HDCM task. Similar tables were constructed for the other three HDCM tasks of the study (not shown here).

subject	#nodes	#links	ratio	level counts
p151s99-01	21	34	1.62	{9, 10, 2}
p151s99-02	18	29	1.61	{9, 8, 1}
p151s99-03	16	22	1.38	{5, 6, 4, 1}
p151s99-04	23	35	1.52	{5, 10, 5, 1, 1, 1}
p151s99-05	16	19	1.19	{7, 6, 3}
p151s99-06	31	37	1.19	{4, 9, 6, 4, 5, 3}
p151s99-07	40	73	1.83	{6, 14, 18, 2}
p151s99-08	37	56	1.51	{8, 11, 10, 8}
p151s99-09	29	52	1.79	{9, 13, 7}
p151s99-10	9	10	1.11	{6, 3}
p151s99-11	20	23	1.15	{4, 10, 5, 1}
p151s99-12	32	42	1.31	{12, 10, 5, 5}
p151s99-13	24	45	1.88	{10, 12, 1, 1}
p151s99-14	25	39	1.56	{4, 8, 6, 6, 1}
p151s99-15	25	33	1.32	{6, 11, 7, 1}
p151s99-16	29	47	1.62	{5, 11, 8, 5}

Table 3.8: HDCM Statistics for subjects' maps from task H2_HDCM. See text for column definitions.

A few subjects misunderstood the task instructions and drew maps with invalid constructs. Two kinds of invalid construct were encountered: *duplicate nodes* and *branching links*. In order to analyze these maps and generate the quantitative data required, an "equivalent" valid map construct was created to replace each invalid construct, and analysis proceeded with the valid constructs.

A duplicate node occurred when the subject put more than one node containing the same term on a map. To create an equivalent valid construct, this was corrected by treating all duplicate nodes as if they were topologically one node. Thus, all duplicate versions would have the same level, determined by the level of the one nearest to the prompt term node.

Branching links occurred when a subject drew a line that had branches or intersections, so that it connected more than two nodes together. Determining an equivalent valid construct required a subjective judgment to be made about the subject's intentions when drawing the branching link, which were not always obvious. For example, if a link from node A forked to connect to nodes B and C, should that be replaced by three valid links connecting all three pairs of nodes, or only links from A to each of B and C? The decisions made during analysis in such cases were based on the lengths and shapes of the drawn lines and the angles at the fork, as well as on the interpretation that seemed likely given the meanings of the node terms.

3.2.2. Correlations with Exam Performance

To investigate whether gross numerical measures of subjects' drawn maps correlate with their course exam performance, scatterplots were constructed to determine whether the node counts of subjects' maps, the link counts, or the ratio of links to nodes correlated with subjects' exam scores. Figure 3.35 shows each subject's node count for task H2_HDCM, plotted against the sum of the subject's raw exam scores over the semester. Figure 3.36 shows link counts against exam score. Figure 3.37 shows the ratio of links to nodes against exam score. All plots display the best-fit line, with the associated coefficient of correlation *r*.

The plots shown for H2_HDCM are typical of all four HDCM tasks in the study. Table 3.9 displays the coefficients of correlation resulting from linear fits to these plots and equivalent plots for the other three HDCM tasks. For B1_HDCM, G2_HDCM, and H2_HDCM, the statistical significance threshold is 0.130 for sixteen data points. For J4_HDCM, the statistical significance threshold is 0.136 for 15 data points, since one subject missed that task due to logistical problems during administration of the session.

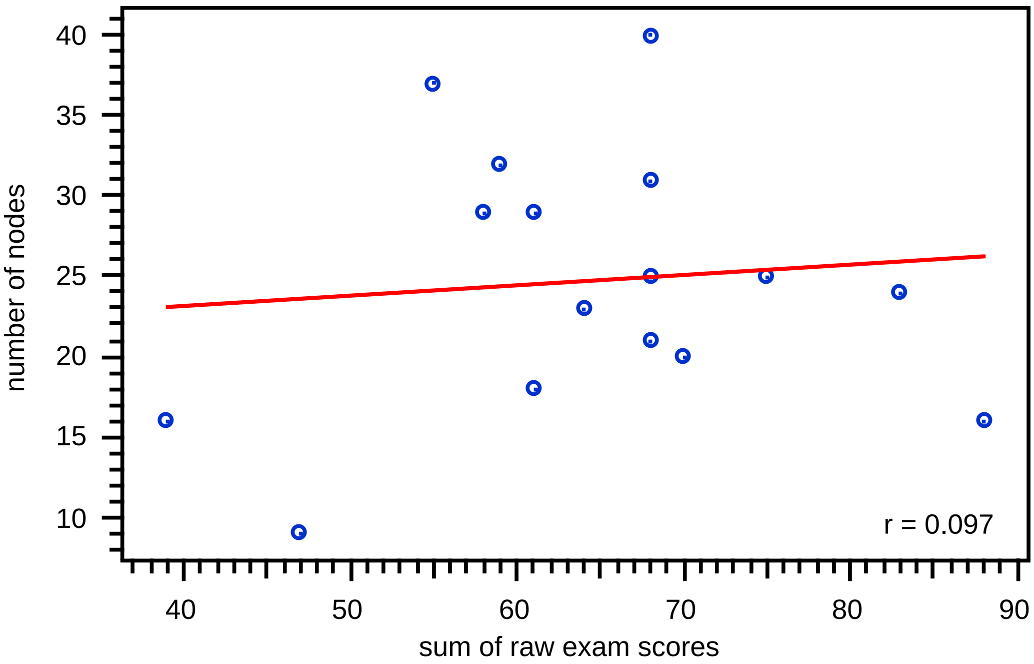


Figure 3.35: HDCM node counts vs. subject's course exam performance, for task H2_HDCM of p151s99 study. The best-fit line and coefficient of correlation are indicated.

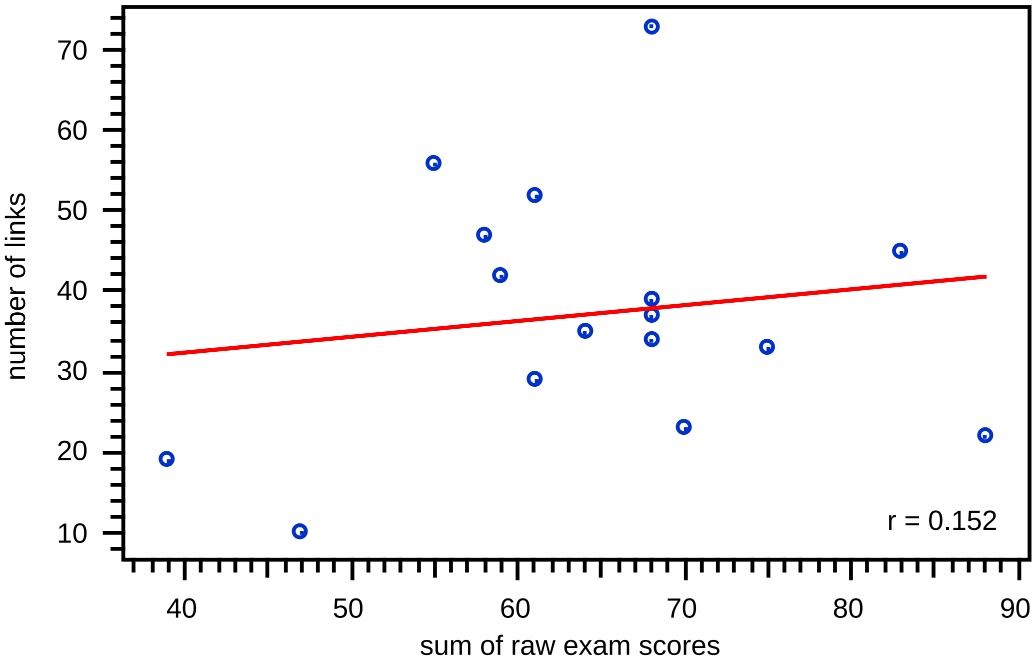


Figure 3.36: HDCM link counts vs. subject's course exam performance, for task H2_HDCM of p151s99 study. The best-fit line and coefficient of correlation are indicated.

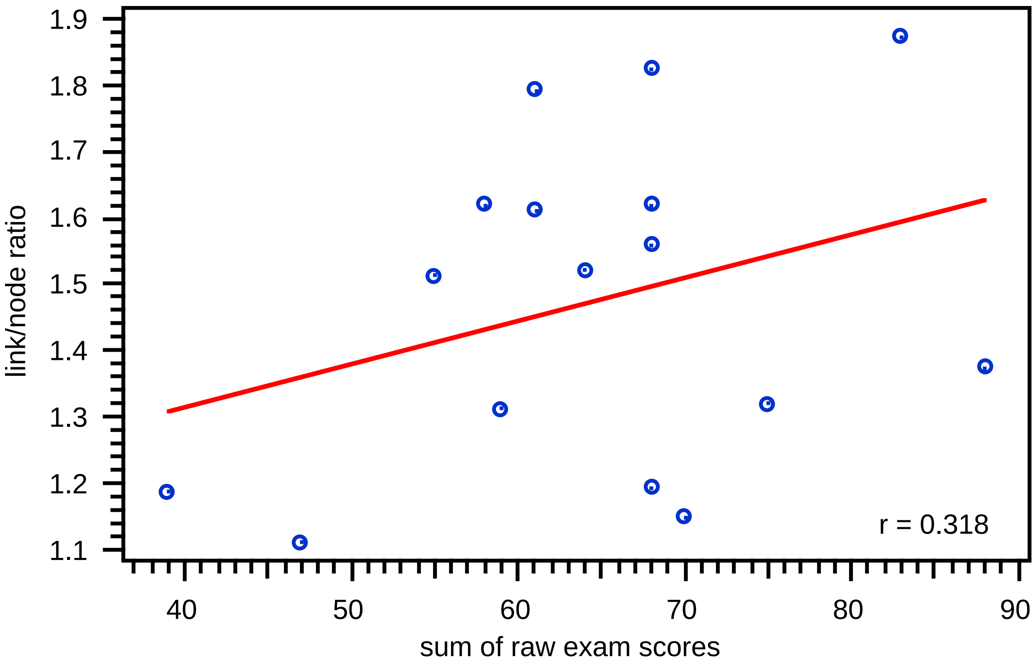


Figure 3.37: Ratio of link count to node count vs. subject's course exam performance, for task H2_HDCM of p151s99 study. The best-fit line and coefficient of correlation are indicated.

Task	coefficient of correlation (*r*)		
	#nodes vs. exam score	#links vs. exam score	ratio vs. exam score
B1_HDCM	0.444	0.217	-0.057
G2_HDCM	0.169	0.181	0.176
H2_HDCM	0.097	0.152	0.318
J4_HDCM	0.116	0.140	0.095

Table 3.9: Coefficients of correlation for plots of node count vs. exam score, link count vs. exam score, and link/node ratio vs. exam score for all four HDCM tasks in the p151s99 study.

According to the table, evidence in favor of a correlation between any of the three map-characterizing quantities and exam performance exists, but is weak. Two of the four *r*-values for node count against exam score are above the statistical significance threshold, one by a large margin, and all four are positive; this suggests subjects with better exam performance tend to include more nodes in their concept maps. All four of the *r*-values for link count against exam score are above the statistical significance threshold, and all positive, suggesting that subjects with better exam performance do tend to include more links in their concept maps — not surprising if they include more nodes. Two of the four *r*-values for link to node ratio are above the significance threshold, one by a relatively large margin, although one of the two below-threshold values is negative; this suggests weakly that subjects with better exam performance might include more links per node in their maps.

As discussed elsewhere, exam performance is not expected to be a strong, clean indicator of the kind of conceptual expertise and knowledge structuring that concept maps and other ConMap tasks were designed to probe. The fact that a strong correlation was not found here is therefore not surprising. One possible explanation for the evidence found in support of a weak positive correlation is that more earnest, dedicated students are likely to do better on course exams, and also likely to put more effort into drawing and elaborating concept maps for the study.

Further study along the lines in this subsection should use a more appropriate standard of comparison than course exam scores. Conceptually difficult, quantitatively easy problems chosen specifically for the study might be more suitable. Another improvement might be to count only "relevant" and "correct" nodes and links, as judged by a domain expert; it is possible that less able subjects pad their concept maps with relatively worthless nodes and links.

3.2.3. Summary of HDCM Findings

This section introduced some quantitative measures derived from subjects' hand-drawn concept maps, for use here and in comparisons with TPTE data. The *level* of a node was defined as it's topological distance form the map's prompt node. Total node counts, link counts, link to node ratios, and node counts by level were tabulated. Mechanisms for dealing with invalidly drawn maps were discussed.

Weak evidence was presented suggesting that node and link counts and link to node ratios might correlate with subject's exam scores.

Further analysis of HDCM data can be found in subsection 3.3.6, where the content of TPTE data and HDCM data is compared.

3.3. Term-Prompted Term Entry (TPTE) Data Analysis

As described in Subsection 2.1.3.2, the Term-Prompted Term Entry (TPTE) task presents subjects with a series of prompt terms. For each prompt term, they are required to type response terms that they associate with the prompt term. Their entering of response terms is cut off after they have entered ten terms, or the first time they pause without typing for ten seconds (unless they have not yet entered a specified minimum term count, usually three). They are then presented with the next prompt term. As with the FTE, the raw data collected consists of the list of response terms entered, along with the times at which the subject started and finished typing each.

Section 3.1 presents the analysis performed to date on the TPTE data collected. Since the p151s99 study was the only study conducted which employed the TPTE task and which included enough subjects for reasonable analysis, only that study is discussed here. Subsection 3.3.1 explains the "term mapping" mechanism used to facilitate TPTE analysis. Subsection 3.3.2 analyzes the statistics of *response term counts*, the number of terms a subject entered in response to a prompt term. Subsection 3.3.3 presents term frequencies, the number of subjects entering a specific response term to a particular prompt and session.

Subsection 3.3.4 defines a measure called *similarity* to quantify the degree to which a subject's responses are typical or atypical of the general study population, and investigates whether similarity values are characteristic of individual subjects or prompts, and whether they vary systematically by session over the study. Subsection 3.3.5 introduces a rubric for scoring TPTE response lists according to domain experts' opinions of how relevant and insightful the response terms are, and looks for evidence that the resulting scores correlate with subjects' degree of domain mastery. Subsection 3.3.6 compares the data from

TPTE and HDCM tasks to see whether the two types of task reveal similar information about subjects. Finally, Subsection 3.3.7 summarizes the section's findings about the TPTE task and makes some suggestions for further research.

3.3.1. TPTE Term Mapping for Standardization

Much of the analysis done on TPTE data was based on the terms subjects entered as responses, not merely on the timing information associated with term entry. Since subjects were free to enter their own idiosyncratic choices of spelling, phraseology, and even concept representation, a method was necessary for comparing terms based on their meaning rather than on the character strings themselves. To that end, an *ad hoc* "term mapping" methodology was developed.

For each TPTE prompt term, the union of all subjects' response list terms was formed, resulting in a list of all terms entered by any subject in the study. Each of these *raw terms* was mapped to a *standardized term* to which it was considered equivalent, and which would be used for further analysis. For many of the terms, the raw and standardized versions were identical. For misspelled terms, the standardized term was a corrected version of the raw term. For sets of raw terms with the same meaning but different word choice or ordering or tense, one form was chosen as the standardized form, and all of the equivalent raw terms were associated with that one standardized term ("many-to-one mapping").

Because some subjects typed in algebraic representations of formulas (e.g. "mv") and others typed verbalized versions (e.g. "mass times velocity" or "mass * velocity"), all such mathematical expressions were considered equivalent and mapped to one standardized form. In cases where the left-hand side of an equation was clearly implied by the prompt term, forms with and without it were considered equivalent (e.g. "f = ma" vs. "mass * acceleration" when the prompt term was "force"). Formulas and their names (e.g. "f = ma" and "Newton's second law") were *not* considered equivalent, since this difference of representation could be significant and cognitively revealing. Other clearly implied words were supplied in the standardized form (e.g. "conservation of" was mapped to "conservation of energy" if the prompt term was "energy").

A few raw terms were considered to be lists of multiple distinct terms (e.g. "inelastic/elastic", "friction or frictionless") and were mapped to more than one standardized term. As a result, a few response lists were treated as having 11 responses, even though the task specified an upper limit of 10.

The mapping scheme has been characterized as *ad hoc* because there were many cases where two raw terms did not have exactly the same meaning but were quite close, necessitating a subjective judgment as to whether they should be mapped to the same standardized term. No explicit criteria were identified as to how this decision should be made in each case; a domain expert (the author) made a subjective judgment on the matter. A more sophisticated analysis scheme would probably include some kind of hierarchical categorization scheme, in which terms could be identified as equivalent or not at whatever degree of detail the analysis in question demanded.

3.3.2. Response Counts

In the p151s99 study, 16 subjects were presented with 52 TPTE prompts distributed among seven of the ten study sessions. For the 52 prompts, 28 unique prompt terms were employed, a few once, most twice, and a few three times. With 16 subjects and 52 response lists per subject, the total number of TPTE response lists collected throughout the study was 832.

The maximum number of responses allowed per prompt term was ten. The minimum was generally three; for the six prompt terms of Session A, however, the minimum was one. For all subjects, TPTE tasks, and prompts in the study, the mean number of responses per list was 6.68, with a standard deviation of 2.62; the median was seven. Figure 3.38 shows a normalized histogram of the response counts.

Figure 3.39 displays each subject's mean number of responses per prompt, averaged over all prompts in all TPTE tasks in the study. Error bars indicate the associated standard deviation (not standard error). It is clear that some subjects were regularly more prolific than others. For example, subject p151s99-15 usually entered nine or ten terms, while subject p151s99-10 typically entered between three and seven. This raises the question of whether subjects' mean response counts might correlate with their general domain competency. To investigate this, a scatterplot of subject mean response count vs. course exam score was constructed, and is shown in Figure 3.40.

According to the plot and coefficient of correlation (r-value) calculation, there is a statistically significant but extremely noisy correlation. The correlation would vanish if the two subjects with the lowest exam scores were omitted, raising doubt about the real "significance" of the correlation. Further study with more subjects and cleaner data is required to resolve the issue.

For each prompt term in each task, an average response count across subjects was calculated. The resulting set of 52 averages had a mean of 6.7, of course. More interestingly, the set of 52 averages had a standard deviation of 1.2, indicating that there was some characteristic variation by prompt term and session, but that this variation was not strong. By comparison, the standard deviation of the subjects' mean response counts (as displayed in Figure 3.39) was 1.5. It appears that response term counts vary more by subject than by prompt.

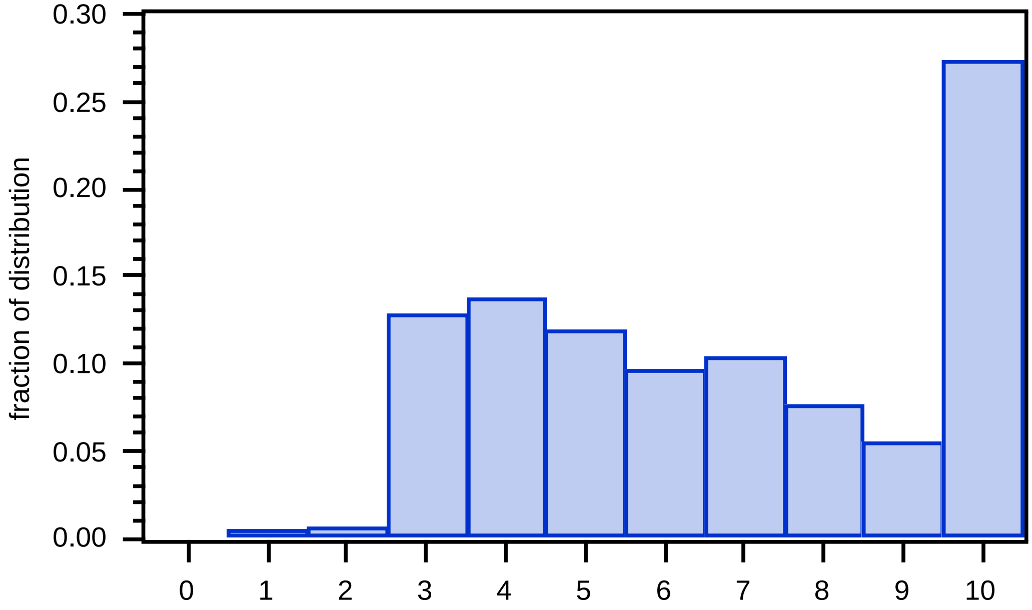


Figure 3.38: Histogram of response counts for all TPTE tasks/prompts in the p151s99 study. Total number of counts was 832.

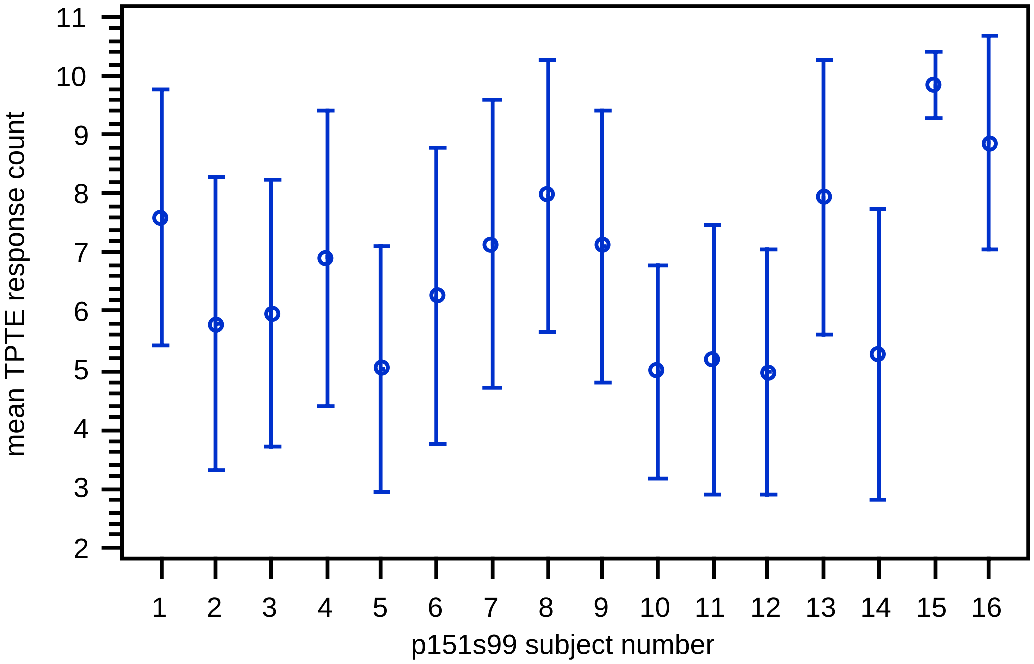


Figure 3.39: Mean TPTE response list term counts by subject (p151s99 study). Error bars indicate the standard deviation of the set of term counts.

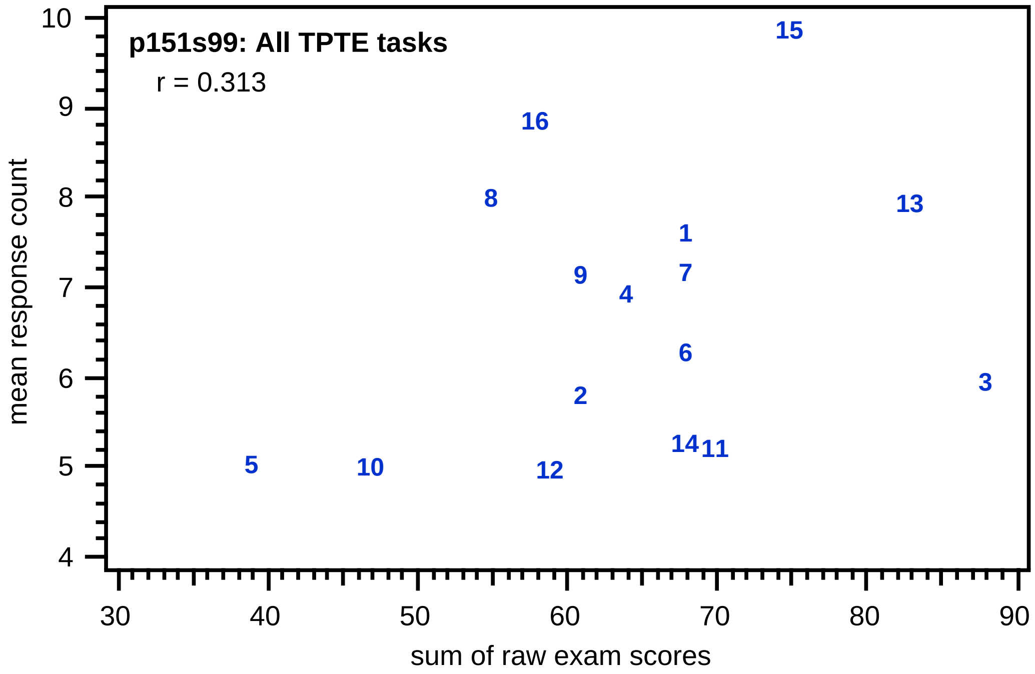


Figure 3.40: Mean number of response terms for all TPTE prompts vs. total of course exam scores, by subject for p151s99 study. Numbers for data points indicate subject. Coefficient of correlation *r* = 0.313, where 0.13 is the significance threshold for 16 data points.

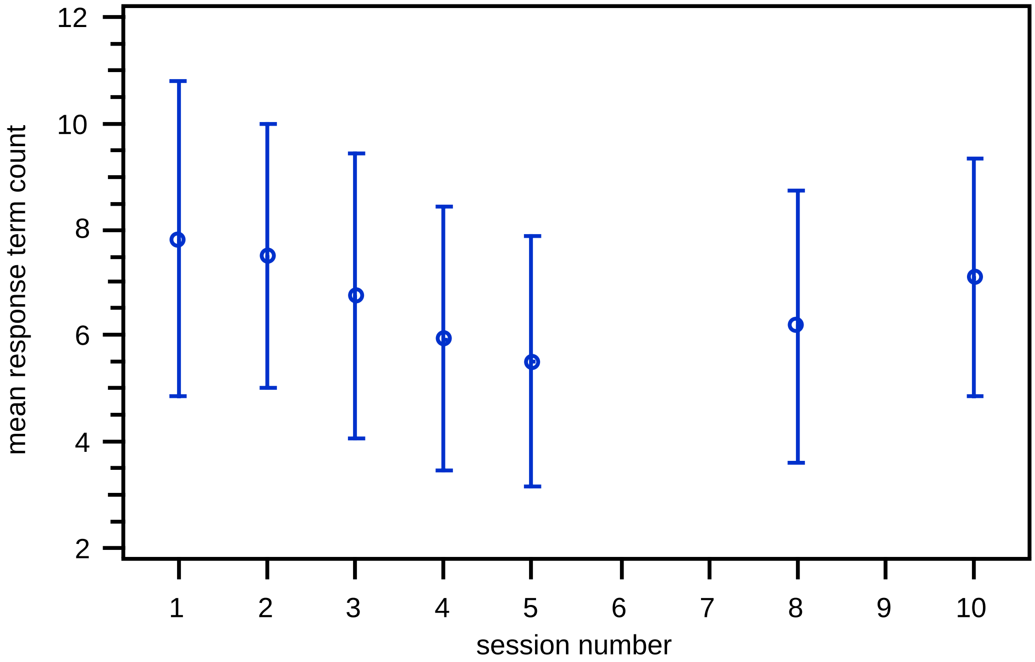


Figure 3.41: Mean number of response terms by session, averaged across subjects and prompts. Error bars indicate standard deviations. (Session 1 = "Session A", etc.)

Figure 3.41 shows mean response term counts by session, averaged across subjects and prompts. A suggestive trend is evident: mean response counts appear to drop monotonically and significantly — by about two terms per session — through the first five sessions. Response counts rise again for the last two TPTE tasks, which occur in sessions H (8) and J (10). (Sessions F (6), G (7), and I (9) did not include a TPTE task.) It is difficult to posit an explanation for the trend, because many factors varied from session to session, such as:

- subjects became more familiar with the task and study;
- subjects were exposed to additional course material;
- the prompt terms employed varied in a way that partially tracked the course syllabus, but later sessions were more likely to include "review" terms from earlier topics;
- prompts in later sessions were more likely to be repeats that subjects had seen in earlier sessions, because prompt terms were usually repeated two or three times.

No comparisons of different termination criteria have yet been conducted. Analysis of whether response counts correlate with the specific prompt term used, independent of session, has yet to be done.

Overall, there is evidence that the number of response terms subjects typically enter to a TPTE prompt might be an informative quantity for cognitive and pedagogic purposes, and further study is warranted. Alternative task-termination criteria should be investigated, as the set of criteria chosen will directly influence the number of responses subjects can enter and might strongly impact the usefulness and noisiness of the response-count measure.

3.3.3. Response Term Frequencies

3.3.3.1. Tabulation of Frequencies

For a particular session/task and prompt term, a list of all response terms entered by all subjects in the study could be compiled, and the number of subjects entering each term in the list could be tabulated and examined. This was done for the eight prompt terms "force", "energy", "inclined plane", "potential", "momentum", "power", "impulse", and "spring" in each of the sessions in which they appeared. Before compiling the list and tabulating the frequencies, all terms were standardized as described in Subsection 3.3.1. Table 3.10 shows the resulting frequencies of occurrence for responses to the prompt “force”, for the three sessions in which it was presented.

The pattern observed for the prompt “force” was common to all eight of the prompts whose frequency tables were inspected: of typically 40 or so unique responses (after standardization), around half were idiosyncratic to one subject, perhaps a third to a sixth of the terms were entered by at least a quarter of the subjects, and five or fewer were entered by at least half of the subjects. For the prompt “force”, the most popular four responses were responsible for 1/3 of the counts in Sessions B and J, and almost that large a fraction in Session C. This pattern is also typical of the other prompts.

Some prompts elicited more agreement among subjects (higher frequencies for a core set of terms) than others. “Force” and “potential” produced higher frequencies, while “impulse” and “power” had few common terms and many idiosyncratic to one subject. For the most part, the small set of most popular responses remained approximately the same session to session for each prompt, although their frequencies might vary slightly.

Some interesting differences between sessions were evident in the data, although their interpretation is not obvious. In Table 3.10, it can be seen that the term “F = m a” (or an equivalent term which was mapped to that) was entered by eight of the sixteen subjects during Session B; by four during Session C; and by none at all during Session J. Meanwhile, the term “Newton’s second law” was entered by four, five, and seven subjects respectively in those sessions. In the table of response frequencies for the prompt “energy” (not shown), the response “work-energy theorem” rose from a frequency of one subject in Session B to three in Session C to nine in Session J. Meanwhile, “work” was the most frequently mentioned response in Sessions B and C, included by thirteen subjects each time, but it was only entered by four subjects in Session J.

3.3.3.2. Identifying Subjects Responsible for Frequency Counts

When a response term has approximately the same frequency of appearance in all sessions in which it was used, it is not obvious from the frequency data alone whether the same subset of subjects entered that term each time. For example, according to Table 3.10, ten subjects entered “acceleration” as a response to the prompt “force” during Session B, eleven entered it during Session C, and eleven during Session J. Was there a set of five subjects who did *not* enter “acceleration” during all three sessions, or (at the other extreme) were there six who did not mention it during Session B, five others who did not enter it during Session C, and five still different ones who didn’t enter it during Session J? The first case indicates that subjects are quite consistent across sessions; the second that subjects are quite variable, and frequency counts describe only the likelihood that any given subject will enter a particular response for a particular prompt.

B1_TPTE	#	C2_TPTE	#	J2_TPTE	#
gravitational	12	acceleration	11	gravitational	12
normal	12	friction	11	acceleration	11
friction	11	gravitational	10	normal	11
acceleration	10	mass	9	friction	10
newton	9	normal	9	mass	10
F = ma	8	vector	8	newton	8
vector	8	newton	7	net	7
mass	6	applied	5	Newton's second law	7
contact	5	net	5	spring	7
direction	4	Newton's second law	5	tension	6
Newton's second law	4	tension	5	weight	6
equal and opposite	3	direction	4	conservative	5
free body diagram	3	F = ma	4	applied	4
magnitude	3	weight	4	free body diagram	4
net	3	work	4	vector	4
Newton's laws	3	distance	3	action at a distance	2
weight	3	contact	2	contact	2
applied	2	dot product	2	external	2
distance	2	magnetic	2	internal	2
exerted	2	magnitude	2	Newton's laws	2
force	2	Newton's laws	2	nonconservative	2
push	2	strong	2	angular acceleration	1
tension	2	weak	2	coefficient of friction	1
9.8 m/s^2	1	action at a distance	1	direction	1
9.8 N/kg	1	add	1	displacement	1
action at a distance	1	air resistance	1	distance	1
air resistance	1	angle	1	electromagnetic	1
components	1	centripetal force	1	force	1
electromagnetic	1	components	1	G	1
F	1	displacement	1	kinetic friction	1
kg m/s^2	1	electromagnetic	1	magnitude	1
magnetic	1	F=W/d	1	moment of inertia	1
move	1	free body diagram	1	push	1
perpendicular	1	kg m/s^2	1	reaction	1
persuade	1	measurement	1	static friction	1
pull	1	natural force	1	strong	1
star wars	1	Newton's first law	1	system	1
torque	1	Newton's third law	1	unit	1
work	1	push	1	work	1
xy coordinate plane	1	spring	1		
TOTAL:	136	TOTAL:	135	TOTAL:	142

Table 3.10: Frequencies of occurrence of standardized response terms to the prompt term "force" (all subjects) for the three sessions in which it was presented in the p151s99 study.

Session/Task	Subjects entering "acceleration"	Total
B1_TPTE	2, 4, 6, 7, 8, 11, 12, 13, 151, 16	10
C2_TPTE	4, 5, 6, 7, 8, 9, 11, 12, 13, 14, 16	11
J2_TPTE	1, 3, 4, 5, 6, 8, 9, 12, 13, 14, 15	11

Table 3.11: Identification of subjects entering "acceleration" as a response to the prompt term "force" in the TPTE tasks of the three sessions in which it was presented.

Session/Task	Subjects entering "trigonometry"	Total
C2_TPTE	1, 7, 10, 13, 15	5
D2_TPTE	1, 2, 10, 13, 15	5
J2_TPTE	1, 3, 6, 10, 15	5

Table 3.12: Identification of subjects entering "trigonometry" as a response to the prompt term "inclined plane" in the TPTE tasks of the three relevant sessions.

Table 3.11 indicates which subjects entered "acceleration" as a response to "force" for the three sessions. Comparing rows, one finds that five subjects entered "acceleration" in all three sessions, seven entered it in two of the sessions, three entered it in one of the sessions, and only one did not enter it in any. Table 3.12 shows the same information for the response "trigonometry" to the prompt "inclined plane", as an example of a response with consistent frequency values lower than those of the previous example. The table indicates that three subjects entered the response in all three sessions, one entered it in two of the sessions, four entered it in one of the sessions, and eight did not enter it in any.

Other cases analyzed (not presented here) follow the same general pattern, revealing that the occurrences of response terms in frequency tables is largely but not completely due to the same subjects in each session.

3.3.3.3. Suggestions for Further Study

For further study, a larger sample population is clearly crucial to provide better statistics and to better separate noise from significant changes in the frequencies of specific responses. A more sophisticated term-mapping scheme is also desirable to reduce the impact of subjective term standardization decisions, and to allow frequency counting at various levels of specificity (e.g. all energy terms, all terms describing forms of energy, any mention of potential energy). Different criteria for terminating term-entry for a prompt might reduce the level of noise in the data; if the first few terms carry the most information, for example, eliminating the later ones could help reveal important patterns. Making the termination criteria more stringent could be simulated in after-the-fact analysis; relaxing or qualitatively changing them could not.

The information provided by this kind of response-frequency analysis seems likely to be of more use to curriculum diagnosis, in which the efficacy of a course's treatment of domain material is studied, than to student diagnosis, in which probing individual subjects' comprehension is the goal. If changing patterns of response choice can be linked to subjects' changing understanding of physics, it might also be useful in determining how TPTE data can be used to reveal subject-specific information of cognitive and pedagogic value.

3.3.4. TPTE Similarity Values

The previous subsection considered all of the standardized response terms to a specific TPTE session and prompt, and examined how many subjects entered each. It was found that for all prompt terms and sessions considered, a few terms were entered by many or most of the subjects, while the majority of terms were idiosyncratic to one or two subjects. This raises the question of whether some subjects are generally more "normal" than others; and if so, whether one might be able to define a measure to represent how normal a subject is, in the sense of how similar the contents of his/her response list is to that of the other subjects in the study. To that end a quantity called *similarity* was defined which measured the overlap of a subject's response list with the lists of other subjects in the study. The degree to which such similarity values were characteristic of individual subjects or prompts, or varied systematically session to session throughout the study, was then investigated.

3.3.4.1. Definition of "Similarity"

Given each subject's standardized (cf. Section 3.3.1) response list for one prompt of one task, define a *raw similarity score* for each list relative to other subjects' lists as follows: each term in the list contributes a number of points equal to the number of other subjects who also included that term. Thus, if a subject included two terms that were also included by five other subjects, one term included by two other subjects, and three terms included by one other subject, that subject's response list would have a raw similarity score of $(2\cdot5 + 1\cdot2 + 3\cdot1) = 15$. This score has a theoretical maximum value equal to $M \cdot (N - 1)$, where M is the maximum possible number of responses (ten in this study) and N is the number of subjects.

The raw similarity score just defined suffers from two obvious defects. For one thing, it depends on the number of subjects in the sample, making cross-study comparisons difficult. For another, all subjects could supply exactly the same response list and still get a less-than-maximum similarity score if the subjects entered fewer than the maximum allowed number of response terms, which violates the desired notion of “similarity”.

The first defect is easily remedied by dividing each subject’s calculated score by (N – 1). The second can be addressed by dividing each calculated score by the average number of response terms provided by all subjects on that task and prompt. We term the resulting measure the *normalized similarity* (just *similarity* for short) for the subject’s response list. The normalized similarity can be considered to range between zero and one, although it is theoretically possible that the value could be as large as $1 \ / \ (1 - o(1/N))$ in unrealistic circumstances. If all subjects entered exactly the same list, all would have a normalized similarity score of one.

Normalized similarity values have been calculated for all 16 subjects on 48 of the 52 session/prompt combinations. The four omitted cases were non-physics prompt terms presented during Session A to familiarize subjects with the task. The mean of all 768 values is 0.25, and the standard deviation is 0.16. The maximum, median, and minimum are 0.71, 0.22, and 0 respectively.

Note that calculated values of similarity are strongly dependent upon the term-standardization mapping scheme used. If similarity values are calculated without any term-correction or mapping, then even trivial differences and misspellings will cause two response terms to be interpreted as “different”, and they won’t contribute towards similarity. At the other extreme, a strong mapping scheme that maps large numbers of similar terms to one equivalent will result in much higher similarity values. The mapping scheme used for the analysis in this section was weak, implementing only spelling correction, adjustment of tense and phrase ordering, and other “trivial” difference corrections, as described in Subsection 3.3.1.

3.3.4.2. Are Similarities Characteristic of Individual Subjects?

The set of normalized similarities calculated for each subject for each of the session/prompt combinations was plotted by subject in Figure 3.42. It appears that some subjects tend to have consistently higher similarity values than others, as measured by mean and standard error, but the range of similarities obtained for any one subject is broader than the variation of the subject means, and outliers are common.

To see whether mean similarity values correlate with subject exam performance, Figure 3.43 plots each subject’s mean similarity across the 48 task/prompt combinations against the sum of their in-course exam scores. Error bars indicate the standard error of the mean.

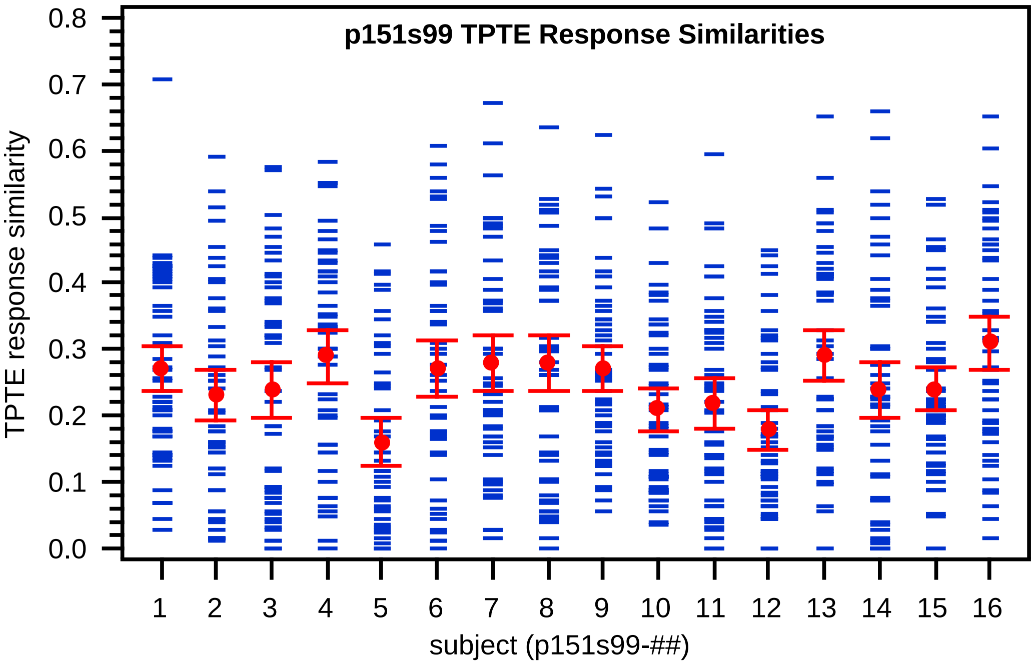


Figure 3.42: TPTE similarities by subject. Dashes represent similarity values for one response list; dots represent subjects' mean similarity values, with error bars indicating the associated standard error.

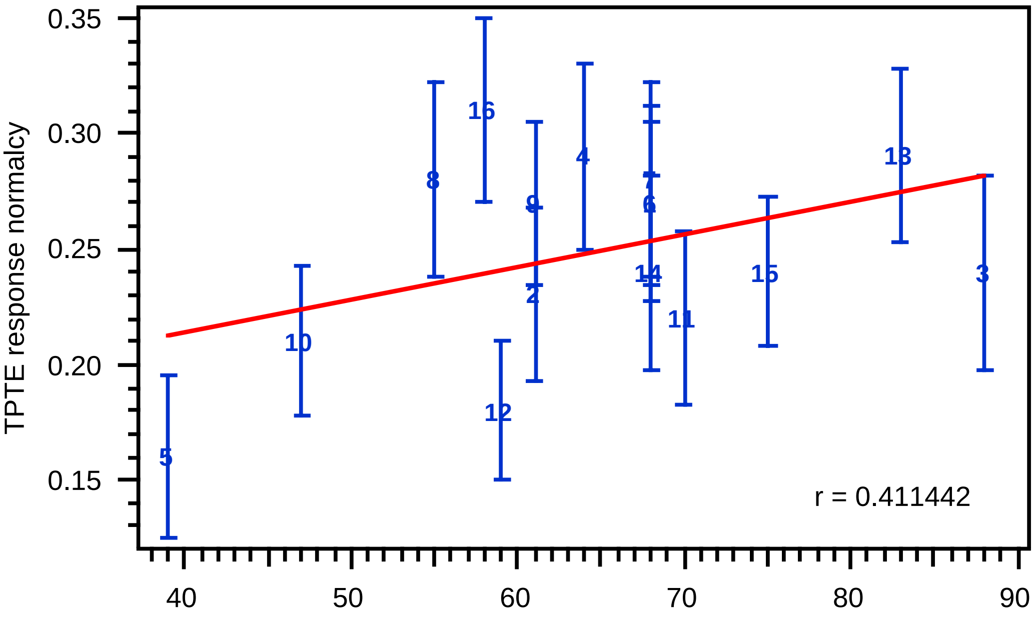


Figure 3.43: Subjects' mean TPTE response similarities vs. overall exam performance. The data markers are numbers indicating which subject is represented. Error bars indicate the associated standard errors. The best-fit line from which the coefficient of correlation (Pearson's r-value) was calculated is indicated.

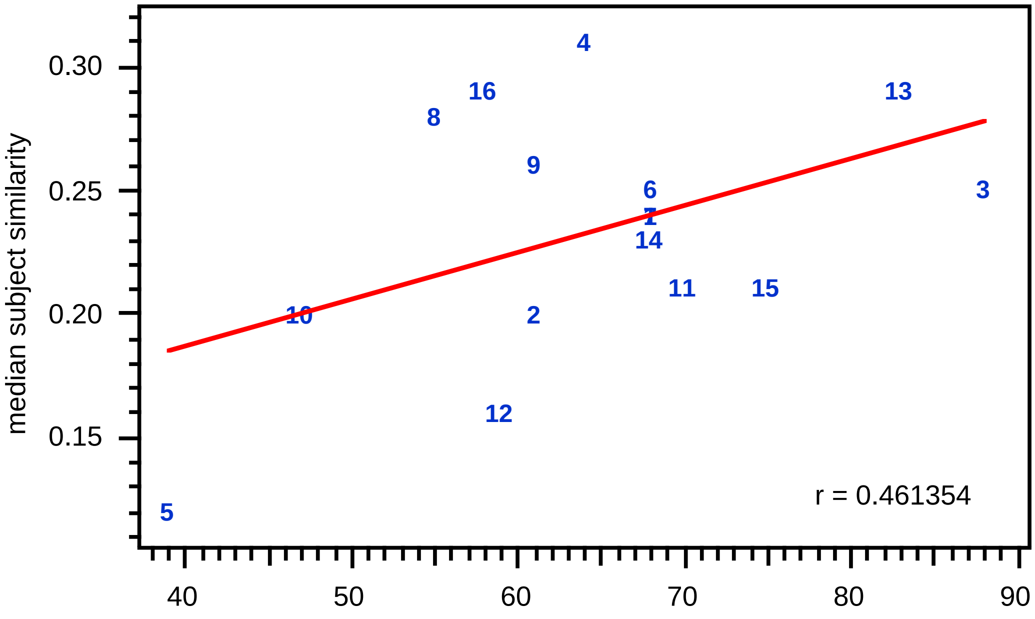


Figure 3.44: Subjects' median TPTE response similarities vs. overall exam performance. The data markers are numbers indicating which subject is represented. The best-fit line used for calculating the coefficient of correlation (Pearson's *r*-value) is indicated.

The plot suggests a weak and noisy but statistically significant positive correlation between mean similarity values and exam performance. A linear fit to the data produces a Pearson's *r*-value of 0.411, where 0.13 is the statistical significance threshold for 16 data points. This suggests that normalized similarities are to some extent characteristic of individual subjects, although to be meaningful they must be averaged over many TPTE prompts.

Because outliers exist and the mean is sensitive to them, subjects' median TPTE response similarities were also plotted against overall exam performance, as displayed in Figure 3.44. The plot shows that using the median rather than the mean does not significantly change the distribution of points, and the *r*-value rises only slightly to 0.461.

For these investigations of correlation, the noise level of the data might be reduced by omitting similarity values for certain "outlier" prompts with atypical response patterns. For example, the first time "impulse" was presented as a prompt term, the subjects had not yet encountered the topic in the associated course. Looking at the response terms subjects entered, it is clear few had any idea of the physics meaning of "impulse"; many entered non-physics associations such as "drive", "fast", "natural", and "urge", and most entered many fewer terms than was typical of other prompts. Although this behavior is informative about subjects' understanding of impulse, it obscures a study of the more typical response patterns for other prompts.

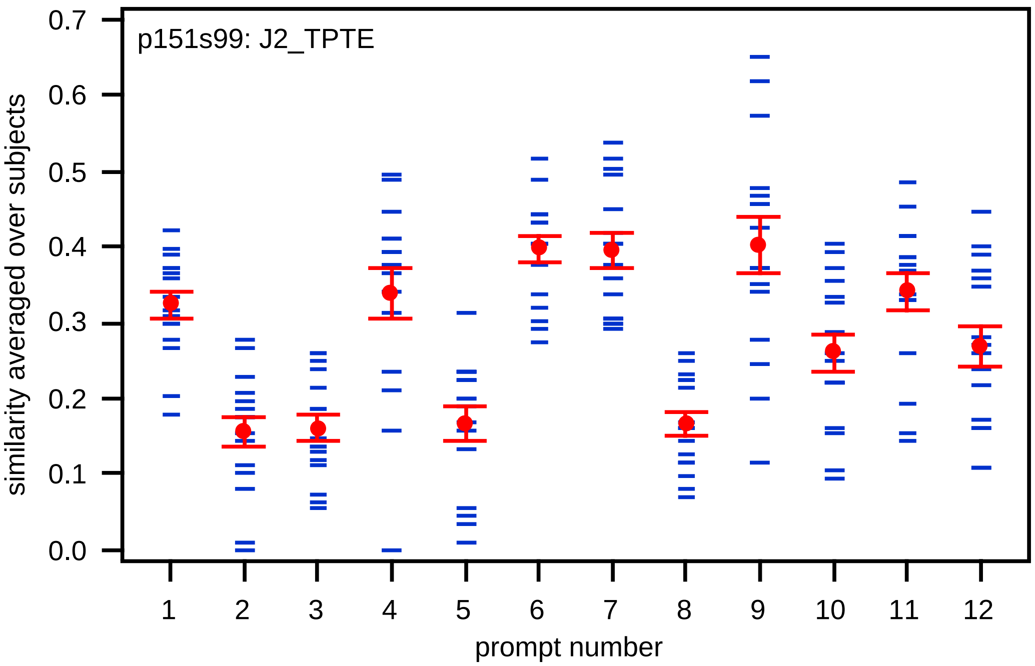


Figure 3.45: Similarity value vs. prompt, for task J2_TPTE of study p151s99. Blue markers indicate values for individual subjects; red points indicate mean across subjects, with error bars indicating standard error of the mean. Table 3.13 gives the mapping between prompt number and prompt term.

3.3.4.3. Are Similarities Characteristic of Individual Prompts?

The previous subsubsection demonstrates that similarity values are somewhat characteristic of individual subjects, in that some subjects tend to manifest consistently lower similarity values than others. An analogous question can be asked about prompt terms: do some prompt terms tend to evoke systematically lower or higher similarity values from subjects than other prompt terms? To investigate this, mean similarity values were calculated for each prompt term of task J2_TPTE, averaged across subjects. To minimize the number of variable factors in the comparison, only similarity values for prompts used in Session J were compared. Figure 3.45 shows the similarity values plotted by prompt, with mean and standard error indicated. Prompts are indicated by number rather than by the prompt term itself, where prompt 1 was the first presented in the task, prompt 2 was the second, etc. Table 3.13 indicates what the prompt term for each prompt number was.

prompt #	prompt term	prompt #	prompt term
1	inclined plane	7	force
2	conservative	8	graph
3	rotation	9	spring
4	vector	10	free-fall
5	displacement	11	friction
6	energy	12	velocity

Table 3.13: Mapping between prompt number and prompt term for Figure 3.45.

It is clear from the plot that despite variations among subjects, some prompt terms elicit consistently higher similarity values than others. The smallest similarity values of the sixth and seventh prompts (“energy” and “force”), for example, were larger than the largest ratings for the third and eighth prompts (“rotation” and “graph”). Other prompts, notably the fourth and ninth (“vector” and “spring”), elicited a particularly wide range of similarities from the subjects. The TPTE task clearly probes something which is sensitive to different prompts or kinds of prompts, and further study is warranted.

Since the prompts are numbered in the order they were presented to subjects during the task, the plot shows no correlation between similarity values and term order, suggesting that variations are due to the meanings the prompt terms hold for the subjects.

3.3.4.4. Do similarities increase as the semester progresses?

Figure 3.46 shows the same set of similarity values as Figure 3.42, but plotted vs. session rather than subject. A weak trend towards increasing similarity values over the course of the study is apparent. Fitting a line to the plot produced a correlation coefficient of $r = 0.758$, where 0.214 is the threshold for statistical significance. The best-fit line had a slope of 0.0101 ± 0.0015, indicating a statistically significant positive slope.

It is possible that similarity values increased because of the different sets of prompt terms given in different sessions. To check this, mean similarity values were compared between sessions for occasions when the same prompt term was given in two different sessions. The TPTE tasks in sessions E and H used exactly the same list of prompt terms, in the same order. Many of the terms were related to momentum ideas, which were covered in class between the two sessions. Therefore, if class coverage causes similarity values to increase, session H similarity values should be generally larger than those for session E. Table 3.14 displays the difference between session H and E similarity values, by subject and prompt term. Averages and standard errors by subject, by prompt term, and overall are indicated.

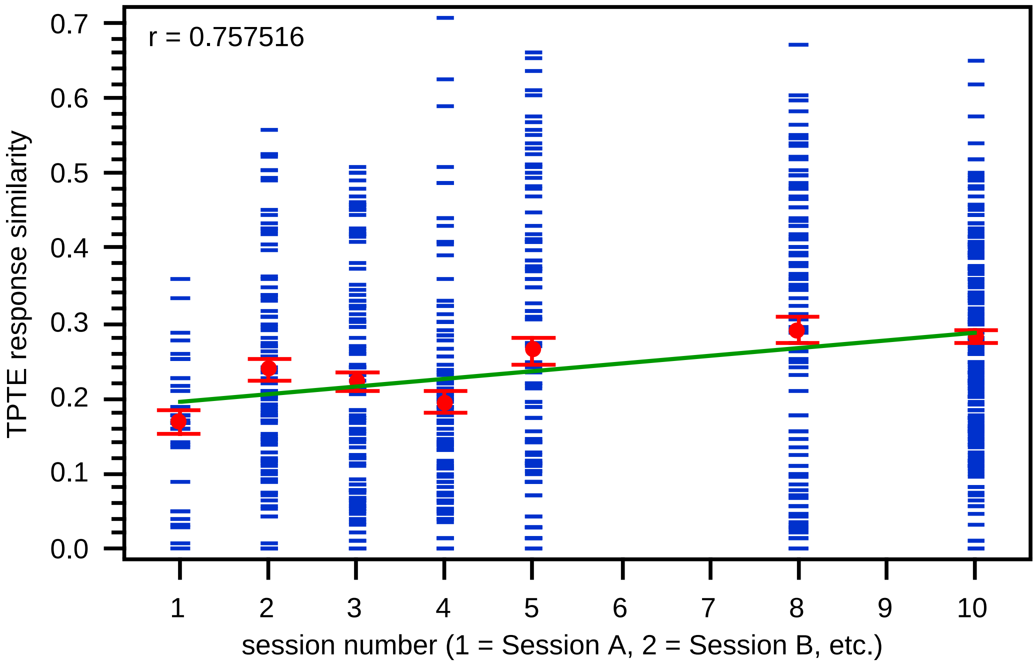


Figure 3.46: TPTE response similarities vs. session number. Dashes represent similarities for individual response lists. Dots with error bars represent the mean of all similarities for the session, with standard error. The best-fit line is indicated.

In general, similarity values increased by an average of 0.027, 1.7 standard errors above zero. This indicates a general trend towards larger similarity values, with moderately strong statistical significance. Although significant, the change is small: 0.027 is 10% the overall two-session mean similarity value of 0.275.

For the mean differences by subject (averaged across prompts), five are greater than one standard error from zero, and all of those are positive. Only one of those was greater than two standard errors from zero. Of the remaining eleven subjects, four had negative means, indicating overall decreasing similarity values. Overall, this suggests the possibility that some subjects are more inclined than others to demonstrate increased similarity values, but the evidence is weak. The largest changes are of order 0.1, which is nontrivial compared to the mean similarity value of 0.275.

For the mean differences by prompt term (averaged across subjects), the statistics are clearer: similarity differences were more than seven standard errors from zero for one prompt (“impulse”), and approximately three to four standard errors from zero for three more prompts (“collision”, “power”, and “range”). The magnitudes of those changes are 0.244, 0.100, –0.074, and –0.054, nontrivial compared to the mean similarity value. Two of the apparently significant changes are negative, indicating that subjects were more likely to have lower similarity values for those prompt terms on the later session.

Δsim	Prompt Term									
subj.	center of mass	collision	impulse	momentum	potential	power	range	mean	stdErr	ratio
1	0.038	0.308	0.315	0.030	-0.183	0.006	-0.101	0.059	0.072	0.82
2	0.027	0.218	0.388	0.426	-0.231	-0.059	0.029	0.114	0.091	1.25
3	-0.020	0.152	0.168	-0.248	0.102	0.083	0.014	0.036	0.054	0.67
4	-0.038	0.117	0.265	0.173	0.001	-0.050	-0.101	0.053	0.051	1.03
5	0.016	0.197	0.009	-0.311	-0.054	-0.087	-0.014	-0.035	0.057	-0.61
6	0.022	-0.065	0.257	-0.200	0.074	-0.099	-0.115	-0.018	0.057	-0.32
7	0.013	0.130	0.234	0.081	0.061	-0.013	0.000	0.072	0.033	2.19
8	0.057	0.171	0.150	-0.204	-0.131	-0.095	-0.058	-0.016	0.054	-0.29
9	-0.096	0.113	0.351	-0.051	0.009	-0.154	-0.072	0.014	0.065	0.22
10	0.025	0.135	0.022	-0.006	0.042	-0.135	-0.043	0.006	0.031	0.18
11	-0.010	0.024	0.314	0.075	0.357	-0.148	0.014	0.090	0.069	1.30
12	-0.130	0.038	0.051	-0.046	0.199	0.051	-0.115	0.007	0.043	0.16
13	0.157	0.074	0.313	0.319	-0.094	0.094	-0.058	0.115	0.061	1.88
14	-0.084	0.197	0.469	-0.167	-0.221	-0.288	-0.173	-0.038	0.103	-0.37
15	-0.005	-0.192	0.332	0.044	0.086	-0.162	-0.101	0.000	0.068	0.00
16	0.092	-0.014	0.263	0.051	-0.442	-0.137	0.029	-0.023	0.083	-0.27
mean	0.004	0.100	0.244	-0.002	-0.027	-0.074	-0.054	0.027		
stErr	0.018	0.030	0.033	0.050	0.048	0.025	0.015		0.016	
ratio	0.22	3.29	7.37	-0.04	-0.56	-2.95	-3.53			1.73

Table 3.14: Change in TPTE response similarities from Session E to Session H. "ratio" indicates the ratio of a mean to its corresponding standard error.

The case for the prompt "impulse" is particularly clear, and is easy to interpret. The mean similarity value across subjects for that term was 0.089 for session E and 0.333 for Session H. The concept of "impulse" was introduced to students in the associated course during the time between those two sessions. Looking at the specific response terms subjects entered, it is obvious that most subjects had no understanding of "impulse" in the physics sense during Session E, and resorted to random and non-physics responses like "urge" and "sudden"; by Session H, however, they could provide a reasonable set of responses. At least in this extreme case, similarity values do seem to be sensitive to subjects' general familiarity with a prompt term.

In summary, there is no evidence for a significant overall trend to increasing similarity values, but there is evidence that similarity values for some specific terms either increase or decrease significantly, and there is weak support for the hypothesis that some subjects show a tendency towards increased similarity values.

3.3.5. TPTE Response List Scoring

The previous section developed and investigated a measure of how similar a subject's TPTE response list for a prompt was to the lists of other subjects in the study. This section develops a procedure for scoring a subject's response list based on the quality of response terms entered, as judged by domain experts. The measure developed is thus much more like a conventional "grade" for an assignment.

A response term is deemed to be "high quality" relative to the prompt term if it indicates a concept related in a meaningful way to the concept described by the prompt term, and "low quality" if it is irrelevant, trivial, or otherwise not something an expert would consider significantly related to the prompt.

Scoring of lists was carried out for only one prompt term, "force", which was used as a TPTE prompt during p151s99 study sessions B, C, and J.

3.3.5.1. Expert Ratings and Student Scores

A panel of five physics experts — four physics professors and one advanced graduate student — was formed. Four of the five had detailed knowledge of the ongoing ConMap research project, so the panel cannot be considered representative of any general population of physics experts. To familiarize the expert panelists with the TPTE task and to acquire some data for later comparisons, the experts were all assigned a 16-prompt TPTE session which included the prompt "force".

A master list was constructed which consisted of every response term given by every subject in the p151s99 study to the TPTE prompt "force" in each of the three sessions in which it was presented, and also every response term given by each of the five expert panelists to the prompt "force". Basic term mapping (cf. Subsection 3.3.1) was carried out on this list, resulting in a set of 80 standardized terms. This set was alphabetized and presented to each of the expert panelists. The experts were instructed to rate the quality of each term as a TPTE response to the prompt "force", and assign to it a "2", "1", or "0", according to the following scale:

2: Good/valuable/important. "This student knows his/her stuff."

1: Has some merit. "Not an unreasonable response."

0: Irrelevant, worthless. "Reveals no knowledge of a nontrivial relationship."

The five experts' ratings were averaged for each response term, resulting in a *quality* value between 0.0 and 2.0 for that term relative to the prompt term.

The *score* for a response list was defined to be the sum of the quality values for each term in that list, which can range from zero to twenty. Such a score was calculated for each of the p151s99 subjects' response lists to the prompt "force" in each of the three sessions. The resulting set of 48 scores (three each for sixteen subjects) ranged from 4.2 to 15.2, with a mean of 11.8 and a standard deviation of 2.9.

For each subject, a mean score for the "force" prompt was calculated by averaging their scores for each of the three sessions' response lists. The resulting set of 16 mean scores ranged from 8.0 to 14.5, with a mean of 11.8 and a standard deviation of 2.0.

3.3.5.2. Score Correlation with Exam Performance

To investigate whether such TPTE response list scores correlate with subject exam scores (and thus, presumably, with "expertise"), scatterplots of "force" response list score vs. exam performance were constructed for each of the three sessions and for the three-session mean scores. These are displayed in Figure 3.47, Figure 3.48, Figure 3.49, and Figure 3.50. Pearson's "*r*-value" coefficient of correlation is indicated in each plot; the threshold for statistical significance for 16 data points is 0.13.

Although all four plots have a "statistically significant" *r*-value, only one of the three single-session plots, Figure 3.48 shows a convincingly strong correlation. The plot of the three-session subject averages displays a noticeably stronger correlation. Overall, there is some evidence for the correlation of list score with exam performance, at least for the prompt term "force", although the correlation appears quite noisy unless scores are averaged across multiple presentations (sessions) of the prompt.

3.3.5.3. Comparison with Expert Scores

Using the same scoring mechanism and term quality ratings, the five expert panelists' response lists received scores of 14.2, 13.4, 11.6, 9.4, and 8.4, for a mean of 11.4 and a standard deviation of 2.5. It is interesting to note that the mean of the experts' scores was slightly lower than that of the students in the p151s99 study, in apparent contradiction to the hypothesis that higher scores indicate greater domain expertise.

This result does not necessarily rule out the hypothesis, however, as other factors may be responsible. The conditions under which the experts performed the TPTE task were quite different from conditions for the p151s99 study. Perhaps most significantly, the experts' data came from their first encounter with

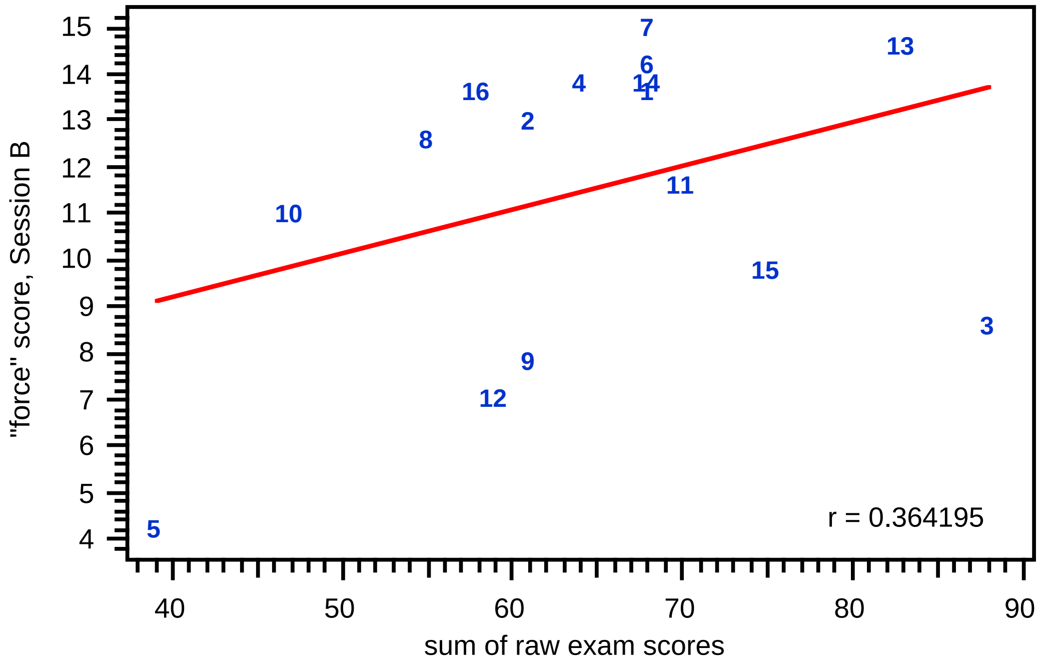


Figure 3.47: P151s99 B1_TPTE scores for "force" vs. course exam performance. The data point markers are numbers indicating the subject represented by the point. The best-fit line is shown.

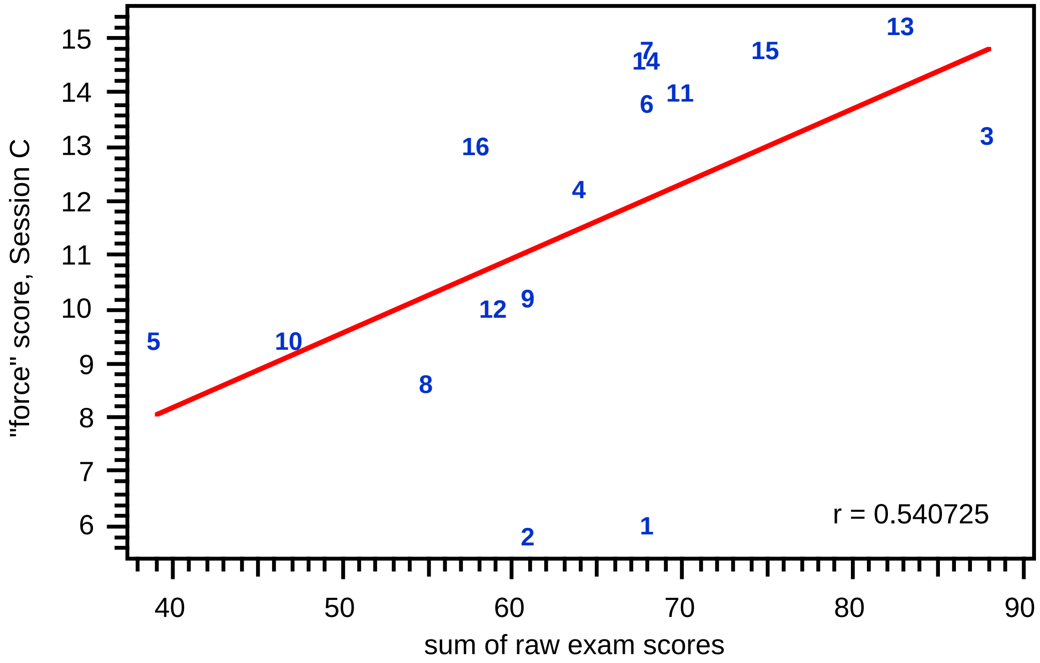


Figure 3.48: Same as Figure 3.47, but for task C2_TPTE.

the TPTE task, and for some it was their first experience with any ConMap task; whereas for the p151s99 subjects, the sessions relevant to this analysis were the

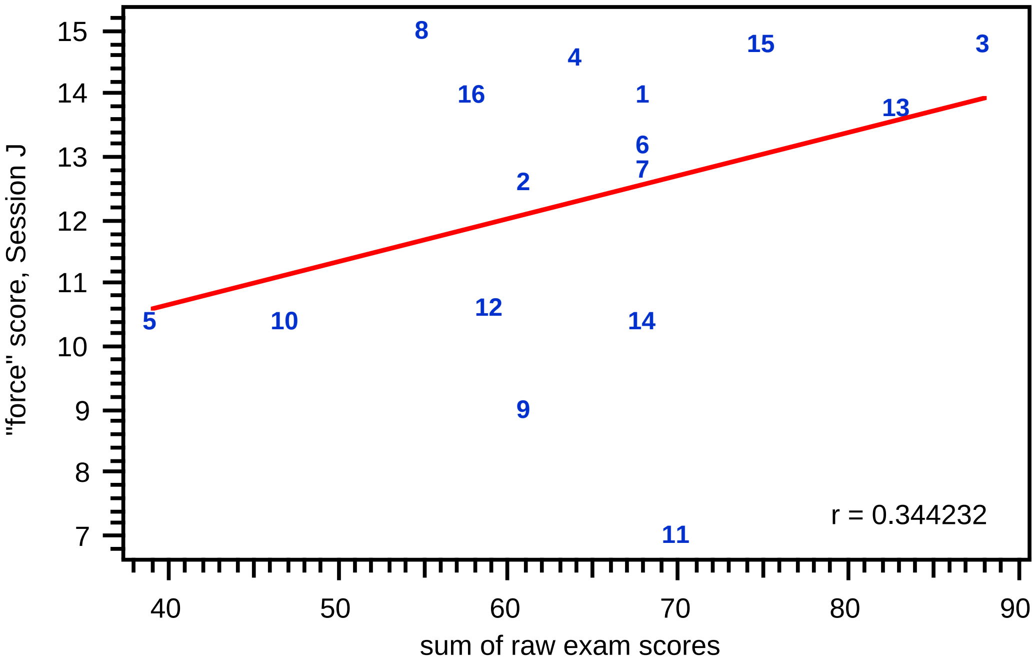


Figure 3.49: Same as Figure 3.47, but for task J2_TPTE.

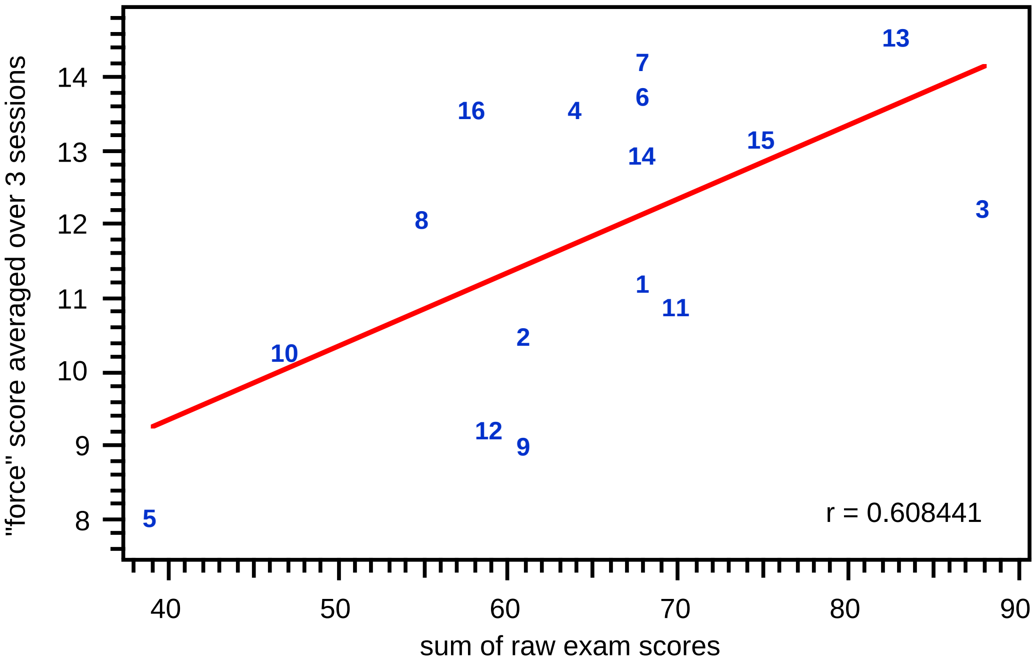


Figure 3.50: Same as Figure 3.47, but for the average of the three session scores for each subject.

second, third, and tenth of the study, so that all had prior TPTE experience. In addition, most of the experts had participated in extensive discussions about the

ConMap study, and likely brought to the task a different set of preconceptions than did the student subjects.

3.3.5.4. Future Research

A repeat of the above analysis for prompt terms other than "force" is warranted, to see whether the patterns observed are unique to that prompt or common to all.

It is the intention of the TPTE task design that a subject's responses to a prompt term should reveal concepts in the subject's conceptual knowledge store that are linked to the concept described by the prompt term. Therefore, response list scores might correlate more strongly with exam performance if only questions depending on an understanding of that concept are considered. Questions from the course exam which depend on a conceptual understanding of "force" could be identified, and a partial exam score for those questions could be calculated and compared to TPTE response scores for the "force" prompt term. Unfortunately, the course exams used during the p151s99 study do not lend themselves well to this kind of analysis, as few target such a conceptual understanding. Future studies should include a "problem-solving" task in which carefully selected problems are presented to subjects to analyze and solve, and a set of problems targeting understanding and conceptual relationships involving "force" and other significant prompt terms should be included.

An improved set of term "quality" ratings should be generated. The current set was produced by a small set of experts, and subsequent discussions revealed that the experts had different strategies and criteria for assigning values to terms, resulting in significant disagreement in ratings. A better-defined procedure for determining term values, applied to a larger sample of domain experts, should reduce the noise level in term scoring. When such experts disagree in their ratings, discussions between them might help to refine and make explicit the criteria to use for rating terms.

Further research comparing experts' and students' performance on TPTE tasks is warranted, to determine what signatures of domain expertise the task is sensitive to. Results of such research might suggest improvements in the scoring criteria to make TPTE score more sensitive to domain knowledge, or perhaps even improvements in the TPTE task design.

3.3.6. Comparison of TPTE With HDCM Data

The use of hand-drawn concept maps (HDCM) for teaching and assessment has received much attention in the educational research literature, as discussed in Subsubsection 1.2.3. Consequently, it would be valuable to know the extent to which the information elicited by ConMap tasks is similar to that elicited by a

HDCM. The TPTE task seems particularly likely to produce information similar to a HDCM, given that it, like the HDCM, presents subjects with a specific term or idea and asks them to specify related terms/ideas. (Of the vast range of variants of the HDCM discussed in the literature, the TPTE appears more like those in which subjects must generate their own terms to add to the map, and less like those in which the terms are provided and only the relationships, or only relationship descriptors, need be added.)

Therefore, one intent of the ConMap study was to compare the terms subjects entered as TPTE responses to the terms they chose as elements in a HDCM of the same prompt (starting) term. Subjects were given a HDCM task during four different sessions of the p151s99 study. The sessions and task numbers and the prompt terms for each HDCM task are listed in Table 3.15.

Session/Task	Prompt Term
B2_HDCM	Force
G2_HDCM	Energy
H2_HDCM	Momentum
J4_HDCM	force

Table 3.15: Prompt terms used for HDCM tasks during p151s99 study.

For B2_HDCM, H2_HDCM, and J4_HDCM, a TPTE task had been given earlier during the same session, and one of the prompt terms in the TPTE was the same as the HDCM prompt. No TPTE was given during Session G, so subjects' maps for G2_HDCM were compared to their response lists for the same prompt term from J2_TPTE.

3.3.6.1. Quantification of data

For each subject's map on each HDCM task, every node was assigned a "level" depending on how removed it was from the prompt term's node. Nodes directly linked to the prompt term node were defined to be level one; nodes directly linked to level one nodes but not to the prompt term node were defined to be level two; and so on.

Some subjects included one or more terms more than once on a map, resulting in duplicate nodes with different sets of links. Such duplicate nodes were counted as one node only, with level equal to the lowest of the level numbers that would be given to any of the instances. That is, if a term occurred in a node at level one and elsewhere on the map in another node at level three, it was counted as one level one node only; and all nodes linked to *either* instance were assigned level two (unless the were also linked directly to the prompt term node, in which case they would also be level one).

In cases where subjects drew invalid map structures, such as Y-shaped links that connected three nodes together, a valid structure was substituted which reflected what the subject seemed to be attempting to represent, and analysis proceeded with that substitution.

As indicated above, each map was associated with a TPTE response list. Each term in the TPTE response list was matched with an HDCM node containing an equivalent term, if one existed, and the level of that node was noted. Duplicate terms in the TPTE response list were ignored.

Because subjects were free to choose their own phrasing and spelling, inexact matches were common. We considered a TPTE response term to match a map node term if their meanings were equivalent, whether or not the terms were identical. For example, "gravity" and "gravitation" were considered matches, as were "F_N" and "normal". Contextual clues from adjacent nodes were sometimes used to aid in identifying the intended meaning of map terms.

On occasion, a TPTE response term did not appear by itself as a map node, but did appear as part of a compound term in a map node: for example, "acceleration" might appear in the TPTE response list and not on the HDCM map, but "mass × acceleration" might appear on the map. In such cases, the term was counted as appearing on the map, with the level of the compound term containing it.

comparison	mean	standard deviation
B2_HDCM vs. B1_TPTE	0.54	0.22
G2_HDCM vs. J2_TPTE	0.61	0.19
H2_HDCM vs. H1_TPTE	0.64	0.18
J4_HDCM vs. J2_TPTE	0.47	0.18
All 4 combined	0.57	0.20

Table 3.16: Mean and standard deviation across study subjects of the fraction of level 1 HDCM terms appearing in the corresponding TPTE response list, for each of the four HDCM/TPTE sets.

3.3.6.2. First-Level Map Terms in TPTE

For each subject's map on each HDCM task, the fraction of level 1 map terms that appeared in the corresponding TPTE response list was calculated. Results are displayed in Table 3.16.

On average, slightly more than half of the terms from each subject's first-level map nodes also appear in the subject's corresponding TPTE response list. A few list-to-map pairings showed an atypically low fraction of common terms; when such pairings were inspected in detail, it was often found that the subject had "categorized" several of the terms from the TPTE response list and used that category as a first-level node on the HDCM, causing the terms themselves to appear at the second level. For example, a subject might have listed several kinds of forces as TPTE responses to the prompt "force", but might have categorized kinds of forces into "contact" and "at a distance" on the HDCM, with the two category names directly linked to the central "force" node and the specific forces connected at level two.

comparison	mean	standard deviation
B2_HDCM vs. B1_TPTE	0.19	0.20
G2_HDCM vs. J2_TPTE	0.13	0.13
H2_HDCM vs. H1_TPTE	0.13	0.14
J4_HDCM vs. J2_TPTE	0.28	0.12
All 4 combined	0.18	0.16

Table 3.17: Mean and standard deviation across study subjects of the fraction TPTE response terms not appearing on the corresponding HDCM, for each of the four HDCM/TPTE sets.

3.3.6.3. Fraction of TPTE Terms Not on Map

For each subject map and associated TPTE response list, the fraction of TPTE response terms not appearing anywhere on the map was calculated. Table 3.17 displays the results. For all the subjects and tasks, between 65% and 100% of a subject's TPTE response terms are likely to appear on his/her corresponding HDCM.

3.3.6.4. Discussion

Despite the fact that the HDCM is a considered, reflective task and the TPTE is a spontaneous, impulsive one, TPTE data sets seem to provide a subset of the information provided by a HDCM. Specifically, a subject's TPTE response list typically contains slightly more than half of the first-level terms appearing in the corresponding HDCM, and few of the TPTE responses are entirely absent from the HDCM. The TPTE thus seems useful for probing the "core structure" of the subject's CKS, while the HDCM allows gathering of more widespread structural information.

Speculating, one might consider whether the fact that some of a subject's level one HDCM terms appear in the TPTE response list and some do not indicates anything fundamental about the subject's knowledge, rather than indicating that the TPTE task is noisy. Perhaps the HDCM level one terms which also appear in the TPTE are those to which the subject has automated, instant access, while those which don't appear are only accessible to the subject upon conscious reflection.

In further study, one might compare branchings from subsidiary nodes of a HDCM to response lists when the subsidiary node term is used as a TPTE prompt. It might be possible to predict significant portions of a subject's HDCM from TPTE response data for a set of prompt terms; if so, the TPTE task might provide a basis for an easier-to-administer, easier-to-evaluate equivalent to the much-studied HDCM assessment.

3.3.7. Summary of TPTE Findings

Section 3.1 began by introducing a mechanism for standardizing terms to deal with the idiosyncratic variation in subjects' spelling and word choice. It then presented a statistical description of TPTE response data for the p151s99 study. It was found that some subjects tend to enter more response terms ("response count") than others, on average, but that the variations in each subject's response counts from prompt to prompt is larger than the variation between subjects' mean response counts. Subjects' mean response counts were found to correlate only very crudely with their overall course exam performance. There does appear to be a systematic variation in response count vs. session during the study, when averaged over subjects.

For each prompt term of each session, each response term's "frequency" — the number of subjects entering it or a term that mapped to it — was calculated. It was found that for each prompt, approximately one-quarter of the response terms were common to several subjects, and the majority were unique to one or perhaps two. Typically, the most popular four responses accounted for approximately one third of all responses from all subjects. Some prompt terms were found to elicit more agreement among subjects (higher frequencies) than others. For prompt terms that were given in multiple sessions of the study, it was found that many of the more popular response terms had close to the same frequency in all sessions, although there were a few notable exceptions. For such terms with approximately constant frequencies in all sessions, the subjects responsible were identified, and it was found that the set of subjects was largely but not completely the same for the different sessions.

In order to compare a subject to the norm defined by other subjects in the study, a quantitative measure called "similarity" was defined that quantifies a response list’s overlap with a pool of other response lists. It was found that over all prompts and sessions, some subjects tend to have higher mean similarity values than others, but that the range of similarities exhibited by an individual subject is larger than the range of differences between subjects’ mean similarity values. Subjects’ mean similarity values, over all sessions and prompts, have a statistically significant but weak correlation with course exam performance. Similarity values seem to be more strongly characteristic of the prompting term, such that the lowest of all subjects’ similarity values for some prompts were higher than the largest of all for some other prompts. Average similarity values for prompts did not appear to depend on the order of the prompts during a session. Over the course of the study, however, there does appear to be some increase in average similarity values by session. Detailed analysis shows this to be attributable to a few specific prompt terms and perhaps to a subset of the subjects.

As a measure of the overall “quality” of a subject’s response term list, a procedure was developed for assigning to each list a score based on a panel of experts’ judgments of how relevant and insightful various response terms were to the given prompt. Only responses for the prompt term “force” were analyzed, for the three sessions in which it was presented. It was found that for any one session, subjects’ list scores correlate weakly with course exam performance; when averaged over all three sessions, the correlation is significantly stronger. There is thus reason to believe that TPTE measures could be developed which provide some of the same information that traditional exams do.

TPTE response lists were compared to subjects’ hand-drawn concept map (HDCM) structures for the same prompt term, and it was found that a subject’s TPTE responses typically had significant overlap with node terms in his/her map, especially the node terms directly connected to the starting node. Few of subjects’ TPTE responses didn’t appear in their maps at all. Therefore, the TPTE task would seem to probe similar aspects of subjects’ knowledge structure as the HDCM task.

Overall, the TPTE task shows promise as a probe of subjects’ conceptual knowledge structure. Information obtained appears similar to that gained through HDCM methods, and also has some correlation with exam scores. Further, more intensive study is recommended.

Two shortcomings of the present analysis should be rectified. One is the poor statistics available due to the small size of the sample; further studies should aspire to a significantly larger population of subjects. The other is the arbitrariness, subjectivity, and probable inconsistency introduced when attempting to deal with the vagaries of subjects’ term choices. Spelling and typographical mistakes and arbitrary choices of word ordering and pluralization are simple to correct; deciding which of an almost continuous range of possible terms should or should not be considered “the same” is not. Although all ConMap tasks encounter this difficulty, it is more significant for the kinds of analysis done so far with TPTE data. It is not at present clear how the difficulty should best be approached, but careful thought is needed prior to the design of any follow-up study.

Also, some possible confounding problems need to be addressed. If a subject is presented with a prompt term that has already been presented in a previous session, his/her responses might be impacted (inter-session memory); this impact should be assessed and a methodology for its minimization developed. Similarly, for prompt terms other than the first presented during one TPTE task, a subject’s memory of the immediately previous prompt and response terms might impact their responses (intra-session memory); this impact should also be considered.

3.4. Problem-Prompted Term Entry (PPTE) Data Analysis

The Problem-Prompted Term Entry (PPTE) task is identical to the Term-Prompted Term Entry (TPTE) task, except for the fact that subjects are presented with prompt problems (or occasionally questionless problem situations or isolated diagrams) instead of prompt terms. Subjects were instructed to read a prompt problem on paper, and then begin the term-entry portion of the task. The response term and timing information collected was identical to TPTE data; consequently, much of the analysis of PPTE data follows that for the TPTE.

The mechanism of mapping subjects’ idiosyncratic terms to standardized forms that was developed for TPTE analysis was applied here as well. For each prompt problem whose response lists were analyzed, a mapping table was constructed by the author. As with the TPTE, if a prompt problem was presented during multiple sessions, all subjects’ response terms for that prompt from all sessions were gathered into one list, for which standardized forms and mappings were determined.

Section 3.3 presents the results of analysis on PPTE data gathered during the p151s99 study, the only study employing the PPTE task and including a large enough subject population for meaningful discussion. Subsection 3.4.1 analyzes the statistics of response counts (the number of term responses entered) by subjects as a function of subject and session. Subsection 3.4.2 examines term frequencies (the number of subjects entering a term for one prompt of one session’s PPTE task). Subsection 3.4.3 departs from pure phenomenology and investigates whether the evolution of subjects’ physics knowledge, presumed to occur as a result of participation in a physics course, is manifest in response lists as a higher occurrence of response terms related to the “right” or “expert-like” solution to the prompting problems. Subsection 3.4.4 seeks correlations between subjects’ choice of response terms and their success or failure on the prompting problem when encountering it on a course exam. Finally, Subsection 3.4.5 summarizes the section’s findings.

3.4.1. Response Counts

In the p151s99 study, 16 subjects were presented with 40 PPTE prompts distributed among seven of the ten study sessions. For the 40 prompts, 30 unique prompt problems were employed: 23 appeared once, four appeared twice, and three appeared three times. Of the 23 that appeared once, five were diagrams with no problem description or question associated; these constituted the only five prompts for the Session I task, I1_PPTE. Of the ones that were repeated twice or three times, given numerical values were sometimes altered between presentations. For two of the problems that appeared twice, the question was removed from the problem for one of the appearances, resulting in a “problem situation”. Seven of the problems appearing once were taken verbatim from course exams, and given shortly after subjects took the exam containing them.

With 16 subjects and 40 response lists per subject, the total number of TPTE response lists collected throughout the study was 639. (Due to a procedural error during the study, one subject failed to respond to one of the prompts.) The maximum number of responses allowed per prompt term was ten. The minimum was generally three; for the three prompt problems of Session A, however, the minimum was one. For all subjects, PPTE tasks, and prompts in the study, the mean number of responses per list was 6.71, with a standard deviation of 2.39; the median was six. Figure 3.51 shows a normalized histogram of the response counts.

Comparing Figure 3.51 to the equivalent histogram for the TPTE task (Figure 3.38 on page 85), it is evident that the shape of the distribution is similar for the two types of task, but that a slightly larger fraction of response lists have a full ten terms for the TPTE task than for the PPTE task (27% vs. 23%). The shape of the remainder of the distribution is also somewhat different for the two: the PPTE distribution is peaked near the center (5 responses), while the TPTE distribution is skewed and peaks to the left (4 responses).

Figure 3.53 displays each subject’s mean number of responses per prompt, averaged over all prompts in all PPTE tasks in the study. Error bars indicate the associated standard deviation (not standard error). The plot looks very similar to the equivalent plot for the TPTE task (Figure 3.39 on page 85), even down to the approximate location of each data point. Figure 3.55 shows a plot of each subject’s mean PPTE response count against his/her mean TPTE response count;

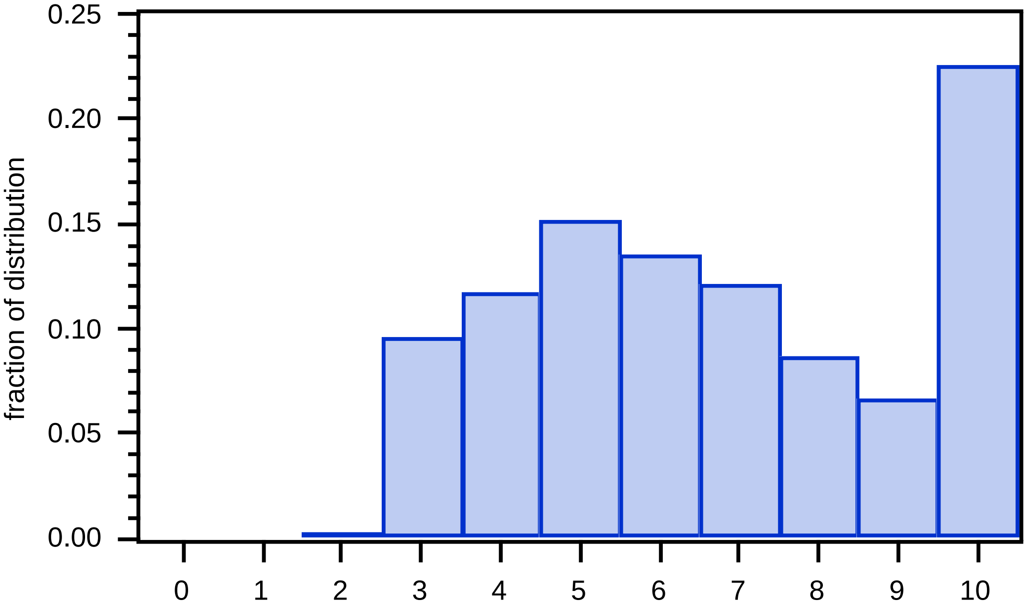


Figure 3.51: Histogram of response counts for all PPTE tasks and prompts in the p151s99 study. Total number of counts was 639.

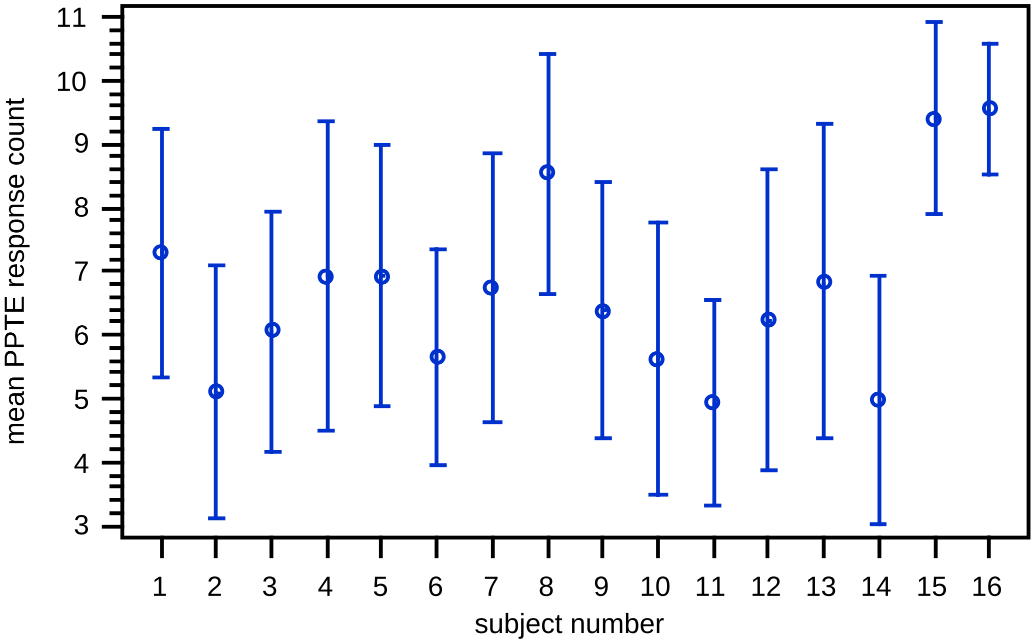


Figure 3.53: Mean PPTE response counts by subject (p151s99 study). Error bars indicate the standard deviation of the set of term counts.

the coefficient of correlation is $r = 0.859$, where 0.13 is the statistical significance threshold. It is clear that subjects’ response counts on PPTE and TPTE tasks are strongly correlated. The correlation of subjects' PPTE response counts with exam

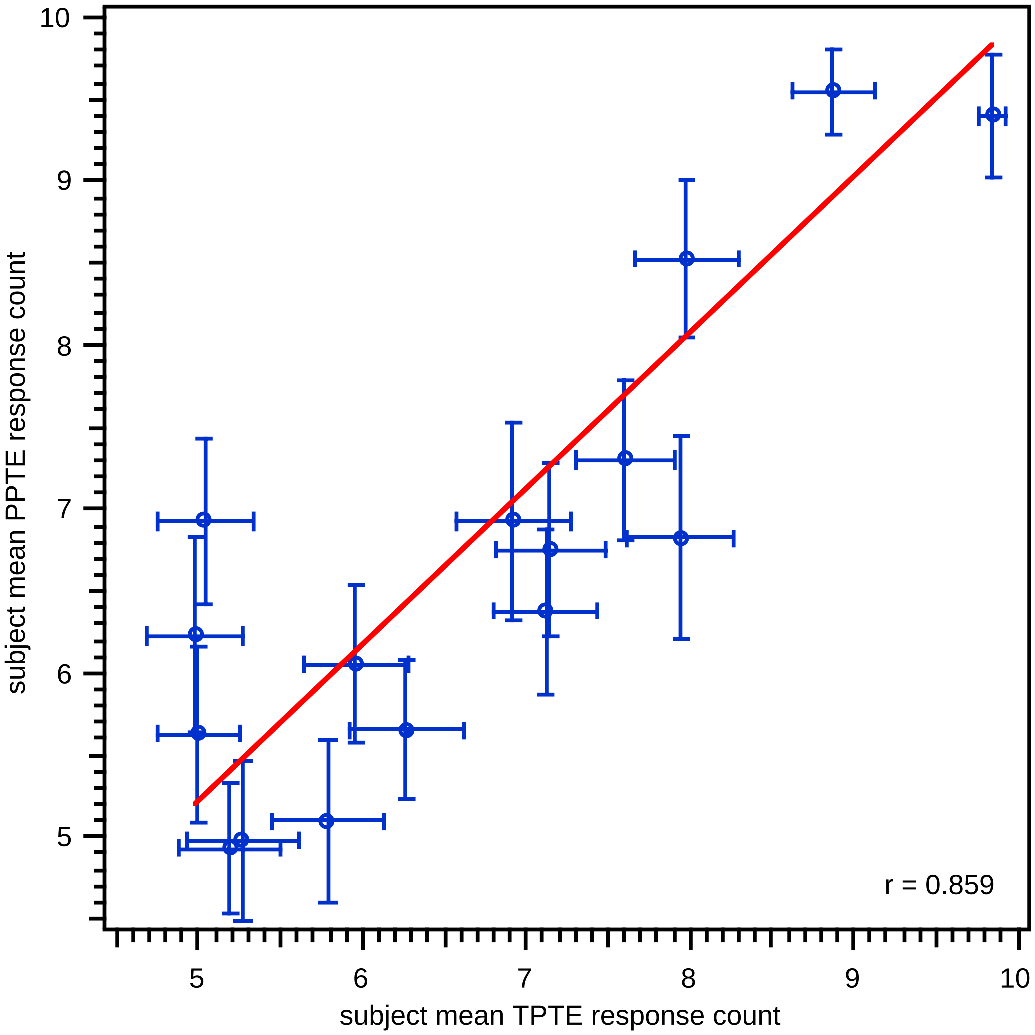


Figure 3.55: Subjects' mean PPTE response count against mean TPTE response count, with standard errors in the means (*not* standard deviations) indicated, for all prompts and sessions of the p151s99 study.

scores should therefore be approximately the same as was found for the TPTE task, and will not be presented here.

Figure 3.56 shows PPTE response term counts by session of the study, averaged over prompts and subjects. The plot is essentially level, although the Session A mean is slightly larger than the others, and the Session J mean is slightly smaller. Because both of these sessions were atypical in the study — the first was largely a "warm-up" session to get subjects familiar with the tasks, and the last was unusually long, contained many tasks, and included domain topics from much of the course — the plot suggests that there was no overall trend towards decreasing response count over study. This differs from the corresponding plot for the TPTE task (Figure 3.41 on page 86), which showed a smoothly decreasing mean response count over the first five sessions, and then an increase for the eighth and tenth sessions. (No TPTE was given during the sixth, seventh, and ninth sessions.)

Overall, conclusions from the analysis of PPTE response counts are similar to those from analysis of TPTE response counts: evidence suggests that response counts might be characteristic of individual subjects, and therefore might serve as a useful probe of subject knowledge. Further study is warranted, with attention paid to the effect of different task termination criteria.

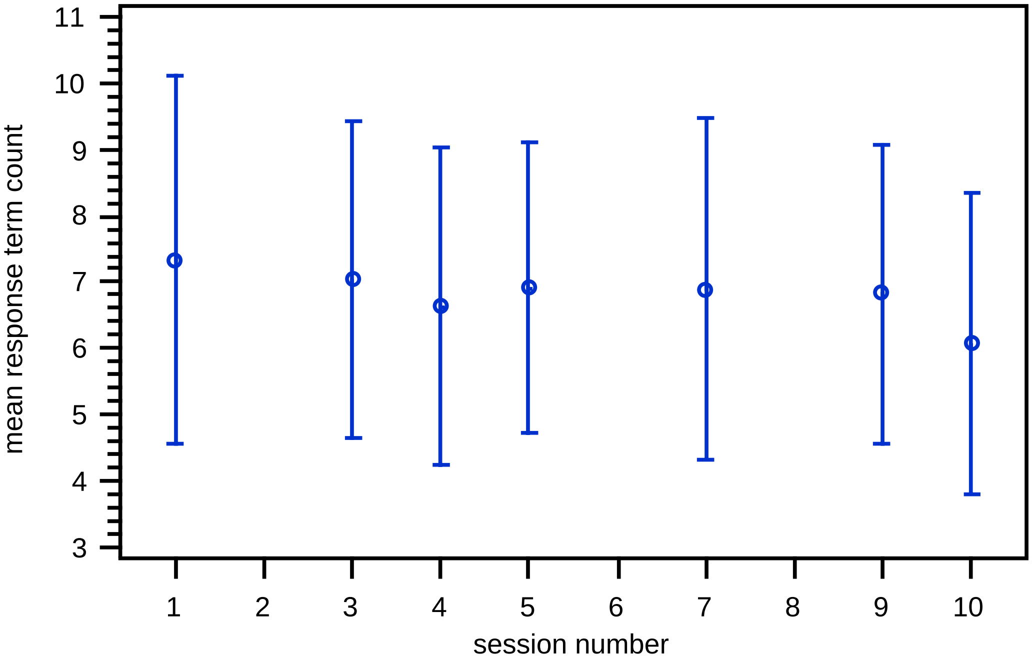


Figure 3.56: Mean number of response terms by session, averaged across subjects and prompts. Error bars indicate standard deviations. (Session 1 = "Session A", etc.)

3.4.2. Response Term Frequencies

For PPTE data, as for TPTE data, response term frequencies — the number of subjects entering each standardized response term — can be calculated. Due to the great variation among the prompt problems, which included isolated diagrams and questionless problem situations as well as problems from the course exams and problems designed specifically for the study, it is not expected that patterns of frequencies will be as uniform across prompts for the PPTE as they were for the TPTE.

Because of the time necessary to create term-standardization mapping tables for the analysis, response term frequencies were only calculated for two prompt problems. These were relatively standard problems without associated pictures, not drawn from the course exams. Both of the problems were presented as prompts during three different sessions' PPTE tasks (C1_PPTE, D1_PPTE, and J3_PPTE), in identical form each time except for variations in some of the given numerical values.

The overall pattern of the response frequencies resembled the pattern for TPTE frequencies: a small number of responses were entered by half or more of the subjects, and roughly half of the responses were idiosyncratic to one subject. The three or four most popular responses accounted for approximately one-quarter of all response counts each time a prompt problem was presented. The two problems analyzed showed a different signature in that for one (problem "C1", shown in Section 3.4.3.2), the most popular terms had frequencies of 14, 11, and 10 for Sessions C, D, and J respectively, while for the other (problem "C4", also shown in Section 3.4.3.2), the most popular terms had frequencies of 8, 7, and 8. By this measure, problem C1 produced more uniformity among subjects than problem C4.

General comments made in Subsection 3.3.3 about the utility of TPTE response frequencies and suggestions for future study also apply here.

3.4.3. PPTE Response Evolution

One of the primary aims of the ConMap project is to develop tasks capable of detecting the evolution of a subject's conceptual knowledge store as he or she learns physics. In an attempt to determine whether the PPTE task is sensitive to such evolution, certain prompt problems were repeated during multiple sessions, and subjects' response terms to those prompts examined for indications of increased domain expertise. Evidence was found that after covering relevant material in class, subjects are more likely to include among their responses terms corresponding to the key principle required for optimal solution of the prompt problem.

A comment on notation: "problem D6" means "the sixth prompt problem given during the Session D PPTE task". A full notation would be "D1_PPTE, Problem 6", where "D" signifies the session, "1" signifies the task during the session, and "PPTE" signifies the type of task. However, since no session had more than one PPTE task, "problem D6" is sufficiently specific within the context of this section. All subjects were given prompt problems in the same order, so problem D6 was identical for all subjects.

Two PPTE prompt problems were given during three different sessions of the p151s99 study, and four others were given during two different sessions. The two problems given three times each (C1=D6=J3 and C4=D2=J5) were given during Sessions C, D, and J. Domain material relevant to the problems was covered in the concurrent physics course between Sessions C and D, and an exam on the material was given during the same week as Session D, so it is reasonable to assume that subjects spent time studying the material during the week between Sessions C and D. The data from these two prompt problems were therefore examined as a test of the hypothesis that the PPTE task can detect conceptual change resulting from lecture coverage and exam studying.

Two of the problems given two times each (I2=J4 and I5=J1) were separated by only one week, and both sessions were significantly later in the semester than the relevant material was covered. These two cases served as a control test by providing a measure of how consistent subjects' PPTE responses were for two consecutive sessions, in the absence of directly relevant course coverage.

The other two problems which were used during two sessions (C2=J2 and D1=J7) did not lend themselves well to this analysis. The crux of each problem lay not in choosing the right conceptual approach, but rather in applying it correctly. Term responses therefore seemed inadequate for judging whether subjects' expertise relevant to the problems had increased, and these problems were omitted from the analysis.

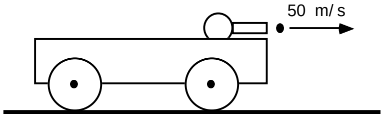
3.4.3.1. Control: Consecutive Weeks, No Relevant Coverage

PPTE prompts I2 and J4 used the same prompt problem (taken from the third course exam, given during the week before Session H). The same is true of prompts I5 and J4. For each occurrence of each prompt, subjects who responded with terms indicating the key concept(s) needed to solve the problem were identified, and a comparison was done to see how consistent subjects were in this regard across the two sessions.

<u>I2 and J4</u>

Problems I2 and J4 both read:

> A cannon mounted on top of a wagon fires a cannonball horizontally at a muzzle speed of 50 m/s, as shown. The mass of the wagon and cannon is 100 kg, and the mass of the cannonball is 5 kg. The system is initially at rest prior to the cannonball being fired. What is the final speed of the wagon and cannon immediately after the cannonball is fired?



For I2 and again for J4, each subject who included "momentum" or "conservation of momentum" among his or her responses was binned as "positive". Table 3.18 shows the resulting counts as a two-way table (Moore and McCabe 1989).

	J4: neg	J4: pos	J4: any
I2: neg	2	1	3
I2: pos	0	13	13
I2: any	2	14	6

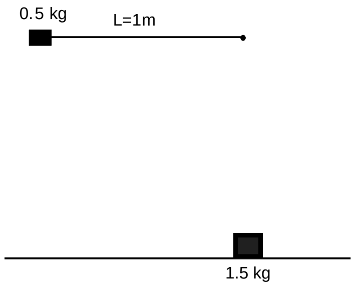
Table 3.18: Two-way table of subjects who included “momentum” or “conservation of momentum” among PPTE responses to I2 and J4 prompts.

It is immediately clear that there was great consistency between the two sessions: only one subject in sixteen changed categories. A chi-squared test against the null hypothesis that subjects’ binnings for the two sessions are uncorrelated (preserving the same ratio of positive to negative as the data yields for each session) confirms this: the *P*-value for the data against such a hypothesis is 0.0016, which means the null hypothesis is extremely unlikely.

I5 and J1

Problems I5 and J1 both read:

> A pendulum is made by attaching a mass of 0.5 kg to a string 1 m long. The pendulum is released from rest with the string horizontal as shown. When the pendulum mass gets to the bottom of the swing, it collides, and sticks to, another mass of 1.5 kg. How high above the ground do the two masses rise after the collision?



For I5 and again for J1, a subject was binned as positive if he or she included “momentum”, and also included “conservation of energy”, “conservation of mechanical energy”, or both “kinetic” and “potential” energy. Table 3.19 shows the resulting counts as a two-way table.

As with the comparison of I2 and J4, there is great consistency between the two sessions; once again only one subject in sixteen changed categories. (That subject was a marginal “negative” for I5, and an argument could be made for placing him or her in the positive bin, which would mean no subjects changed bins at all.) A chi-squared test against the null hypothesis that subjects’ binnings for the two sessions are uncorrelated (preserving the same ratio of positive to negative as the data yields for each session) confirms this: the *P*-value for the data against such a hypothesis is 0.00087.

	J1: neg	J1: pos	J1: any
I5: neg	3	1	4
I5: pos	0	12	12
I5: any	3	13	16

Table 3.19: Two-way table of subjects who included "momentum", and also included "conservation of energy", "conservation of mechanical energy", or both "kinetic" and "potential" among PPTE responses to I5 and J1 prompts.

3.4.3.2. Lecture Coverage and Exam Studying

The two problems given during Sessions C, D, and J are optimally solved with the work-energy theorem, but can be solved (with much more labor) through other paths: kinematics for one problem; kinematics, vector resolution, and the definition of work for the other. When subjects were presented with the problems during Session C, they had been introduced to work and energy concepts, but the lecture instructor had not completed his treatment of the work-energy theorem. During Session D, one week later, coverage of energy topics was essentially complete, and subjects were taking an exam on the material. Most of the subjects were taking the exam later on the same day as their ConMap session; a few took the session two days after the exam. The Session J presentation occurred significantly later, at the end of the semester, near the last day of classes.

It was hypothesized that additional lecture and homework coverage focused on the material and preparation for the exam would impact the way subjects responded to the prompt problems. Specifically, it was anticipated that more students would respond with terms indicating an inclination to consider the work-energy theorem for solving the problems during Session D than during Session C. For the Session J responses, two outcomes seemed plausible, assuming that the hypothesis about Sessions C and D turned out to be correct: if the increase from C to D was due to short-term immersion in work and energy course material (i.e. subjects had those terms on their minds), then the fraction of positively-binned subjects should decrease from D to J; or, if the increase was due to a real change in subject's conceptual reaction to the problems, then the rate for J should be comparable to the rate for D and significantly higher than the rate for C.

C1, D6, and J3

Problem C1 read:

> An object is launched directly upward with an initial speed of 18 m/s. What is the object's speed after rising 8 meters?

Problem D6 was identical to C1 except that "18 m/s" was changed to "12 m/s" and "8 meters" was changed to "5 meters". Problem J3 was identical to problem D6.

For each subject's Session C list of response terms to the prompt problem, the subject was binned as "positive" if he or she included "work" or "energy" as a response term or part of a response term. Subjects who mentioned neither were binned as "negative". Each subject was binned according to the same criterion for Sessions D and J, resulting in three binnings per subject. For each of the three sessions, the 16 subjects' binnings were assembled, and the fraction of subjects binned as positive for that session and prompt was calculated. The results are displayed in Table 3.20.

Session & Problem:	C1	D6	J3
Fraction "positive":	1/16	7/16	6/16

Table 3.20: Fraction of subjects including "work" or "energy" among their PPTE responses for prompts C1, D6, and J3 (same problem).

To compare C1 to D6, the null hypothesis tested was that subjects have the same probability of responding with "work" or "energy" on D6 as on C1; the one-sided alternative hypothesis was that subjects are more likely to so respond on D6 than on C1. Using a *z*-test for before/after task comparisons (Moore and McCabe 1989), it was found that the data had a *P*-value of 0.00715 with respect to the stated null and alternative hypotheses. The 90% confidence interval on the difference of the two probabilities was 0.375±0.252. This indicates that the observed data is quite unlikely to result if the null hypothesis is true. In other words, the data suggests that the increase from 1/16 to 7/16 is statistically significant. Note that the statistical test employed is not considered trustworthy for such a small sample.

To compare D6 to J3, the null hypothesis was that subject probabilities for being binned positive were the same for D6 and J3, against the two-sided alternative hypothesis that the two probabilities were non-equal (in either direction). Using the same standard analysis, a P-value of 0.719 was calculated. The 90% confidence interval on the difference of probabilities was 0.063±0.286. Again assuming that small-sample effects did not intrude, the data are consistent with the null hypothesis of no change.

To investigate whether the subjects binned as positive on D6 were the same as those binned positive on J3, Table 3.21 was tested against the null hypothesis that there was no correlation between subject binning on the two sessions, using a chi-squared test. The resulting *P*-value was 0.15, indicating that it was only somewhat unlikely that the data would result if subject binning was uncorrelated. The table shows that of the six positive subjects in J3, only four were positive in D6. While the overall probability of subjects responding positively to the task remained roughly constant from Sessions D to J, approximately one in three subjects changed bins.

	J3: neg	J3: pos	J3: any
D6: neg	7	2	9
D6: pos	3	4	7
D6: any	10	6	16

Table 3.21: Two-way table for subjects including “work” or “energy” among their PPTE responses for prompts D6 and J3 (same problem).

As a sensitivity check, the null hypothesis that the probabilities for subjects binning as positive were the same for C1 and J3 was tested, against the one-sided alternative hypothesis that the probability was higher for J3 than C1. The resulting *P*-value was calculated to be 0.016, contradicting the null hypothesis relatively strongly. The 90% confidence interval on the difference of probabilities was 0.31±0.24.

For the comparison of C1, D6, and J3, the data support the hypothesis that for the prompt problem used, an ensemble of subjects equivalent to our study population would have been more likely to include “work” or “energy” in their list of responses during Sessions D and J than during Session C, and that they would have been equally likely for Sessions D and J.

C4, D2, andJ5

Problem C4 read:

> A 30 kg box starts from rest on a frictionless horizontal floor. A force of 200 N is applied to the box, pushing down at an angle of 45°. How much work must the applied force do to get the box moving at 1 m/s?

Problem D2 was identical except that “30 kg” was changed to “25 kg”, “200 N” was changed to “320 N”, and “1 m/s” was changed to “1.5 m/s”. Problem J5 was identical to problem D2.

For each subject’s list Session C of response terms to the prompt problem, the subject was binned as “positive” if he or she included “energy” or “work-energy theorem” as a response term or part of a response term. Subjects who mentioned neither term were binned as “negative”. Subjects who merely entered “work” were binned as negative, since the problem itself explicitly asks for the work to be determined. The same calculation of fractions as for the set C1, D6, and J3 was carried out.

One of the subjects binned as negative for D2 did in fact produce the response term “conservation” for D2, but also entered “Newton’s 3rd” and “f = ma”. This was judged to be an insufficiently clear reference to the work-energy theorem or energy conservation to be counted. Were this reference considered sufficient, the fraction positive for D2 would be 7/16 rather than 6/16.

Session & Problem:	C4	D2	J5
Fraction “positive":	3/16	6/16	6/16

Table 3.22: Fraction of subjects including “energy” or “work-energy theorem” among their PPTE responses for prompts C4, D2, and J5 (same problem).

To compare C4 to D2, the null hypothesis used was that subjects have the same probability of meriting the positive bin for D2 as for C4; the one-sided alternative hypothesis was that subjects are more likely to be binned as positive for D2 than for C4. Using the same analysis as above, and testing against the data of Table 3.22, a *P*-value of 0.119 with respect to the stated null and alternative hypotheses was calculated, an inconclusive result. The 90% confidence interval on the difference of the two probabilities was 0.188±0.261.

To compare D2 to J5, the null hypothesis used was that subject probabilities of being binned positive were the same for D2 and J5, against the two-sided alternative hypothesis that the two probabilities were non-equal (in either direction). Using the same standard analysis, the calculated *P*-value was 1.0. The 90% confidence interval on the difference of probabilities was 0.00±0.28. The data are completely and obviously consistent with the “no change” hypothesis.

To investigate whether the subjects binned as positive on D6 were the same as those binned positive on J3, Table 3.23 was tested against the null hypothesis that there was no correlation between subject binning on the two sessions, using a chi-squared test. The resulting *P*-value was 0.062. As with the comparison of D6 and J3, four of the six positive subjects for J3 were also positive for D6. While the overall probability of subjects responding positively to the prompt remained constant from Sessions D to J, approximately one in three subjects changed bins for this prompt also.

As a sensitivity check, the null hypothesis that the probabilities of subjects binning as positive were the same for C4 and J5 was tested, against the one-sided alternative hypothesis that the probability was higher for J5 than C4. Since the fractions for D2 and J5 are the same, the resulting *P*-value and confidence interval are identical to the C4 to D2 test.

For the comparisons on the prompt problems C4, D2, and J5, the data are inconclusive: although the pattern is similar to that seen for prompts C1, D6, and J3, the results are not statistically significant. In light of the results for the C1/D6/J3 tests, the results for C4/D2/J5 should be taken as weak corroboration.

	J5: neg	J5: pos	J5: any
D2: neg	8	2	10
D2: pos	2	4	6
D2: any	10	6	16

Table 3.23: Two-way table for subjects including "energy" or "work-energy theorem" among their PPTE responses for prompts D2 and J5 (same problem).

3.4.3.3. Conclusions and Discussion

The "control" comparison data support the hypothesis that a subject's proclivity to respond to a PPTE prompt with terms indicating the "correct" conceptual approach to the prompt problem remains the same for two consecutive sessions if the relevant physics is not addressed by an ongoing course during the interim.

The "lecture coverage and exam studying" comparison data support the hypothesis that subjects were more inclined to respond with work/energy terms during Session D than Session C, and about equally likely to respond with work and energy terms during Sessions D and J. In light of the control comparison, this suggests that a PPTE task can be sensitive to the learning that occurs in students on the time scale of a week.

While statistically significant, the support is weak, and the analysis of statistical significance is itself questionable because of the small sample size. Further studies with larger samples and a larger, more varied set of prompt problems is warranted.

It is open to question whether subjects would have been more inclined to respond with work and energy terms during Session D than during session C for prompt problems for which work and energy ideas were *not* appropriate to the solution. Since subjects had recently been heavily exposed to work and energy ideas during the course, they might have been more inclined to respond with such terms for any prompt problem. The similarity of "positive" rates between Sessions D and J suggest that the increase from C to D was not such a temporary aberration, but it is a weakness of the study design that no prompt problems were given in Sessions C and D for which work-energy ideas were not the best solution, to be used as another control set.

3.4.4. PPTE Responses vs. Exam Problems

The previous subsection examined whether PPTE responses can reveal the presumed evolution of subjects' conceptual knowledge store due to experiences associated with their physics course. This section investigates the PPTE task as a probe of subject expertise by inquiring whether subjects' PPTE responses to a problem correlate with their performance on the problem as part of a course exam. The results are generally inconclusive, due largely to an inadequate study design.

Two of the problems from Exam 1 of the course were presented as PPTE prompts during Session C. Five of the problems from Exam 2 were given as PPTE prompts during Session E. Five of the problems from Exam 3 were given as PPTE prompts during Session I. Subjects' PPTE term response lists were compared to their exam results to determine whether, for each problem, the presence or absence of significant terms in the PPTE response list correlated with success or failure of the corresponding exam problem. Unfortunately, many of the exam problems were too easy, in the sense that only a small fraction of the study subjects did not get them correct. This prevented the extraction of statistically meaningful information from the analysis.

For a given problem, terms related to the "right" approach to the problem were identified in subjects' response lists, and also some terms which related to a likely incorrect approach. For each subject, an identification was made of which of these terms appeared in his or her PPTE response list for that problem prompt. For each problem, a table of subjects vs. possible terms was then constructed, indicating which subjects included which terms in their response list, and also which subjects correctly answered the problem on the exam.

Table 3.24 displays an example of such a table. Each "x" in the "Exam problem" column indicates a subject who chose the correct answer when the problem appeared on a course exam. Each "x" in the "Terms in response…" columns indicates a subject who entered that term, or a term with equivalent meaning, in his or her list of PPTE responses for the problem. For subject p151s99-15, the two "~" characters indicate that the subject entered a response term ("centripetal") which approximated both of the column-heading terms equally; the presence of this response term was judged to be significant, and so a "partial" match was indicated on the table.

subject	Exam problem E2A.17	Terms in response to PPTE E4								
		"centripetal acceleration", "v^2/ r"	"centripetal force", "mv^2/ r"	"work", "energy"	"free-body diagram"	"tension"	"weight", "gravity"	"net force"	"Newton's 2nd law", "F=ma"	"acceleration"
p151s 99-01					x	x	x	x	x	x
p151s 99-02				x		x	x			
p151s 99-03	x	x	x				x	x	x	
p151s 99-04	x			x		x	x			x
p151s 99-05					x	x	x			
p151s 99-06	x		x	x			x			
p151s 99-07				x	x	x	x			
p151s 99-08		x			x	x	x			
p151s 99-09	x					x	x			x
p151s 99-10				x		x	x			
p151s 99-11		x			x	x	x			
p151s 99-12	x			x		x	x			
p151s 99-13	x	x					x		x	
p151s 99-14				x			x			
p151s 99-15	x	~	~	x	x	x		x	x	x
p151s 99-16	x			x		x	x			
Count:	8	4	2	9	6	12	15	3	4	4

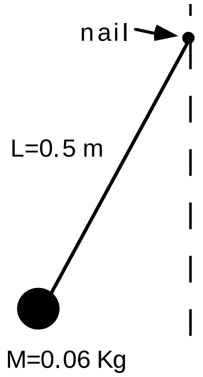
Table 3.24: Exam success vs. presence of various terms in corresponding PPTE response list, by subject.

No significant pattern is evident in this or any of the other problems: for no problem does there appear to be a term or combination of terms whose presence or absence in a PPTE response correlates significantly with exam correctness. The statistical significance of a few example cases is analyzed below with a chi-squared test.

3.4.4.1. E4: Exam Problem 2A.17

PPTE prompt problem E4 and course exam 2 problem 17 read:

> A pendulum is formed by attaching a mass, M=0.06 Kg, to the end of a string 0.5 m long. The other end of the string is attached to a nail so that the pendulum can swing freely in a complete circle in the vertical plane. Assume no friction between the nail and the string. Use g=10 m/s^2. If the pendulum is set in motion so that its speed at the bottom of the swing is 2 m/s, what is the tension in the string at this point?



Solving the problem requires subjects to use the idea of centripetal force (mv^2/r) or centripetal acceleration (v^2/r). Table 3.25 compares subjects who correctly answered the exam problem to subjects who entered "centripetal", "v^2/r", or "mv^2/r" as all or part of a PPTE response term.

	exam correct	exam wrong	total
term present	4	2	6
term absent	4	6	10
total	8	8	16

Table 3.25: Comparison of presence of "centripetal", "v^2/r", or "mv^2/r" among PPTE response terms to exam correctness for problem E4 (Exam 2A, problem 17).

For the null hypothesis that the presence or absence of the terms is uncorrelated with the correctness or incorrectness of the exam problem, a chi-squared test yields a *P*-value of 0.30, meaning the data are not inconsistent with the null hypothesis. The data do not provide significant evidence that the presence of one of the terms "centripetal", "v^2/r", or "mv^2/r" correlates with success on the exam problem.

The strongest correlation found for this problem was a negative one between the presence of the term "free-body diagram" in a subject's PPTE responses and the subject's correctness on the exam problem. The data is shown in Table 3.26. For the "uncorrelated null hypothesis", a chi-squared test yields $P = 0.039$, which indicates that the null hypothesis is most likely false. However, the chi-squared test makes assumptions that are not safe for a sample this small (Moore and McCabe 1989). To illustrate the sensitivity of the *P*-value calculation to fluctuations in a sample of this size: if the *P*-value is recalculated for the hypothetical case that one of the subjects in the "term present, problem incorrect" bin had instead answered the problem correctly on the exam, then $P = 0.15$ results, which indicates a very weak statistical significance. The results here should be interpreted as only suggesting a possible correlation.

Such a correlation does not seem likely from a theoretical perspective, either. Drawing a free-body diagram should be a useful tool for solving the given problem, since improper identification of the forces acting on the pendulum is a likely mistake. It seems unlikely that the observed correlation does in fact indicate a pattern that would be observed in a large population, as opposed to a fluke of small-sample statistics. Correlations between exam correctness and the presence or absence of the terms "work" or "energy", "tension", "weight" or "gravity", "net force", "Newton's 2nd law" or "F = ma", and "acceleration" were all checked for; considering the number of possible terms considered, it is not surprising that a "possibly significant" correlation for one of them was found.

	exam correct	exam wrong	total
term present	1	5	6
term absent	7	3	10
total	8	8	16

Table 3.26: Comparison of presence of "free-body diagram" among PPTE response terms to exam correctness for problem E4 (Exam 2A, problem 17).

3.4.4.2. Discussion

Subject's rate of correctness on the majority of the problems with exam counterparts that were not discussed above was uniform enough — almost all correct, or almost all incorrect — to make any possible correlation undetectable. For the cases where there were enough correct and incorrect answers to make analysis worthwhile, all problems proved to be as unenlightening as the case discussed above, and none showed a correlation as strong as that found for "free-body diagram".

One possible interpretation of this is that the PPTE task does not reveal information about a subject's knowledge structure relevant to solving problems of the sort found on the exams from the p151s99 course. Another is that the task is sensitive to relevant knowledge structure information, but that the particular comparison done in this study is too noisy to reveal it.

One likely source of noise is that subjects took the study session with the PPTE prompts approximately a week after taking the corresponding exam. Much could have happened in that intervening week to influence their PPTE responses. A probable influence is from the exam itself. The exam consisted of two phases: an "individual phase" wherein students completed the exam on their own and submitted their answer sheet (and from which results for this study were drawn), and a subsequent "group phase" where students discussed the same exam questions in groups of three and then submitted a group answer sheet. The discussion that took place during the group phase could very easily influence study subject's thinking about the problem and thus their subsequent PPTE responses.

Another likely source of noise comes from the fact that the course exams were multiple-choice, which means subjects who did not know how to do the problem could still get marked as correct by guessing the proper answer from the list. If a subject could eliminate some possibilities via reasoning, he or she could even guess with a nontrivial probability of success.

A better study to investigate the relationship of students' PPTE responses to problem-solving facility is recommended. Such a study should have subjects respond to a PPTE prompt problem and then immediately thereafter solve the problem in an open-ended, non-multiple choice format. Detailed comparison of subject's written solutions with their PPTE responses could shed light on the significance, if any, of PPTE responses. For statistical purposes, the problems chosen should span a range of difficulty, with many at a level such that subject success rate is approximately 50%. And, obviously, a larger student sample is necessary so that meaningful statistical analysis can be performed.

3.4.5. Summary of PPTE Findings

Section 3.3 began with a phenomenological study of the PPTE data. A statistical description of response counts indicated that when each subject's overall mean PPTE response count is calculated, the variation within the resulting set of subject means varies only slightly compared to the variation within each subject's set of response counts. Nevertheless, it was found that the subject means correlate very strongly with the corresponding means for the TPTE task. Averaged over prompts and subjects, mean response counts by session do not show a noticeable trend over the study, unlike their TPTE counterparts.

When response term frequencies were calculated, the resulting set of frequencies showed a similar pattern to that of TPTE response term frequencies. Specifically, only a few of the response terms were entered by half or more of the subjects, and the most popular three or four responses were responsible for approximately one quarter of all response counts for all subjects.

By comparing subjects’ response sets when the same prompt problem was presented during multiple sessions, and looking for response terms corresponding to the key concepts necessary for the “right” solution to the prompt problem, evidence was found that the PPTE task can be sensitive to learning that occurs as a result of subjects’ physics course. Although the result is well supported by the existence of appropriate control cases, its statistical significance is open to challenge due to the fact that the study population was too small to meet the requirements of the statistical tests employed.

When subjects’ response terms were compared to their performance on the prompt problem when given on a course exam, no statistically meaningful results were found due to small sample size and inadequate study design.

Because of the great variety of problem types employed as prompts in the study, many more questions can be asked of the data than have been discussed here. For example, response patterns to questionless problem situations can be compared to those for standard problems and to those for isolated diagrams. The analysis done so far, however, suggests that the small size of the study population will prevent any findings of statistical significance. Such questions should be investigated in the existing data, but only as a guide to suggest design improvements for any follow-up studies.

Overall, analysis of PPTE data has revealed some tantalizing results, the most significant being evidence that the task is sensitive to changes occurring in subjects’ understanding due to course coverage. Since the overarching goal of the ConMap project is to develop assessment methods that probe students’ conceptual knowledge structure for changes which occur during learning, this result suggests that the goal might be achievable. Further study is definitely indicated, with a larger study population to enable meaningful statistical analysis. In addition, the addition of a problem-solving task designed to allow comparisons of problem-solving performance on conceptually focused problems with PPTE responses would be helpful.

3.5. Term-Prompted Statement Entry (TPSE) Data Analysis

For the Term-Prompted Statement Entry (TPSE) task, subjects were presented with a prompt term and asked to respond with statements about that prompt term which said something important about the term. A statement was defined to be a single sentence expressing a single idea or relationship. Subjects were allowed to enter as many statements as they could in four minutes, up to a maximum of nine.

	F1_TPSE			I2_TPSE		
subject	accel.	force	energy	friction	momentum	mean
p151s99-01	9	7	5.5	5.5	6.5	6.70
-02	6	6	4	6.5	5.5	5.60
-03	6	7	7	5	7	6.40
-04	6	8	6	5	6	6.20
-05	4	5.5	4	4.5	4	4.40
-06	5.5	6.5	6	5	4.5	5.50
-07	8	8	7	5	6	6.80
-08	8	9	5.5	9	8	7.90
-09	7.5	6.5	8	7	6	7.00
-10	8	8	4.5	6	7	6.70
-11	7	7.5	6	6	5	6.30
-12	7	6.5	5.5	6	4	5.80
-13	9	9	8	9	7	8.40
-14	6	5	5	4	6	5.20
-15	9	9	9	9	9	9.00
-16	7.5	7	4.5	7	6.5	6.50
mean	7.09	7.22	5.97	6.22	6.13	

Table 3.27: Number of response statements entered by each term for each TPSE prompt of p151s99 study. If a subject's final statement was cut off before it was complete enough to be clearly identifiable, it was counted as 0.5 of a response.

The p151s99 study was the only ConMap study to include the TPSE task. Table 3.27 displays the number of statements entered by each subject for each TPSE prompt in that study. Some of the entries indicate the entry of a non-integer number of statements (e.g. 5.5 or 7.5). When the allotted time had passed, the task administrator sometimes cut subjects off while they were in the middle of writing a statement. If the resulting partial statement was complete enough to allow unambiguous determination of what the complete statement was intended to be, it was counted as a full response. If it was not complete enough, it was counted as one-half statement, even if only one or two words were written.

Despite receiving explicit instructions to include only one idea (proposition) per statement, subjects displayed a distressing inclination to lump more than one idea together in a statement. Sometimes subjects used a semicolon to fit two distinct statements into one sentence and call it one statement. In future studies with the TPSE task, subjects should be given a training period during which such tendencies can be noticed and chastised by an administrator. In addition, subjects who forget the prohibition against compound statements should be corrected after each prompt/response set. In a computer-administered version of the task, some level of human vigilance must be maintained.

It is regrettable that subjects were occasionally cut off in the middle of writing a statement; it seems a waste of valuable data to throw away partially complete statements that the subject had presumably fully thought out. Future studies should allow subjects to complete any statement they are in the process of writing when time expires.

One line of analysis of TPSE data that has not yet been pursued is a comparison of the physics terms that appear within subjects' response statements to the terms that appear in their TPTE response term list for the same prompt term, and whether the relationships between terms described by TPSE statements coincide with relationships indicated by HDCM links. The TPTE to TPSE comparison could shed light on the kinds of relationships that occur between prompt term and response terms (and possibly between response terms) in the TPTE. The HDCM to TPSE comparison could help illuminate subjects' interpretation of the HDCM task and what conceptual connections they consider to be a "link".

3.6. Summary of Results from ConMap Analysis

FTE analysis characterized the distributions of the data's thinking and typing times and the evolution of term entry rate during task completion. A distinct spike-plus-peak distribution of thinking time logarithms was discovered. It was found that shorter thinking times tend to occur between pairs of terms relatively unconnected in meaning, as judged by a domain expert. Subjects' jump rates showed some correlation with their course exam performance.

Analysis of HDCM data provided weak evidence that the number of nodes, the number of links, and the ratio of links to nodes in a subject's map might correlate with exam performance.

TPTE mean response counts were found to be weakly characteristic of individual subjects, when averaged over prompts and sessions. When averaged over prompts and subjects, they were found to vary slightly but systematically by session, decreasing monotonically for the first five sessions of the study and then increasing for the eighth and tenth. For any given prompt term, a small subset of the set of all responses provided by all subjects was common to many or most subjects, and was responsible for most of the population's response counts. When the same prompt was presented in different sessions, most response terms' frequencies were relatively unchanged, but a few changed dramatically, suggesting the impact of course-related learning.

Similarity, a measure of the overlap between a subject's TPTE response list and the rest of the study population's lists, was found to be weakly characteristic of individual subjects, and to have a weak but statistically significant correlation with exam scores. Similarity was found to be more strongly characteristic of individual prompt terms. There was some evidence that similarity values tended to increase for a subset of subjects and prompt terms as the semester progressed.

Subjects' TPTE response lists were given a *score* by domain experts according to the merit of the response terms contained, and for the one prompt term analyzed the resulting scores were found to correlate significantly with the subjects' exam scores when averaged over three sessions' presentation of the same prompt. Subjects' TPTE response lists were found to overlap significantly with their HDCM maps for the same prompt term, especially with the subset of map nodes closest to the prompt node.

The phenomenology of PPTE response data was found to be similar in many ways to that of TPTE response data: response counts were also weakly characteristic of individual subjects, although no trend by session was found. The pattern of response term frequencies was similar but slightly more varied. Evidence was found that subjects' response lists "improved" on the time scale of a week when relevant topic material was covered in class during that week, where "improvement" means the lists were more likely to include terms indicating the key concepts in the prompt problem's solution. Without contemporaneous course coverage, lists did not show such a change. This suggests the PPTE task can be sensitive to learning that occurs in domain material. A comparison of PPTE response lists to subjects' success rates for the same prompt problem on an exam revealed no statistically meaningful results.

Overall, results from the analysis of ConMap data suggest that the various ConMap-style probes are sensitive to at least some aspects of students' expertise and the changes that occur in it as learning occurs. Because of the preliminary nature of the study design and the small sample populations employed, few of the results should be taken as more than suggestive. Further studies, however, are clearly indicated. In addition, phenomenological descriptions of patterns found in the data can serve as a first target during the construction of theoretical models of subjects' underlying cognitive processes.

In most of the analyses of task data, weak evidence was found for the correlation of various measures of subject productivity — response counts of various sorts — with exam performance. This may indicate nothing more than a hypothetical "conscientious subject effect": more conscientious students tend to perform better in courses, and also tend to apply themselves more earnestly to study tasks, resulting in longer response lists, more elaborate concept maps, and the like. It does not necessarily follow that these various measures of ConMap task productivity are sensitive to subject’s domain expertise. Any further studies must confront this possibility directly, and design in methods to separate such an effect from more significant causal connections.

4. ConMap Modeling

In physics and many others scientific disciplines, the first step in investigating a system is to devise experimental probes, and the second step is to gather a set of measurement data. The third is to look for patterns which characterize the data (phenomenology). The fourth is to construct theoretical models of the system which "explain" or "predict" the observed patterns. When a relatively comprehensive and successful model exists, the system is often said to be "understood".

This prescription is of course oversimplified; real research involves frequent iteration through the various steps, and it is often difficult to separate them quite so cleanly. Furthermore, the patterns one observes and attempts to model often depend on the preconceived model one brings to the research. Nevertheless, the paradigm has demonstrated its merits. With the results of Chapter 3's analysis of ConMap data, the time seems propitious for initial forays into the construction of models of knowing, learning, and thinking which can "explain" the findings.

This chapter introduces two approaches to modeling the temporal aspects of Free Term Entry (FTE) data. Section 4.1 presents the *random distribution model*, which is essentially a unified phenomenological description with no insight into cognition or knowledge. Section 4.2 presents the *matrix walk model*, which does begin with a quantitative representation of cognitive structure. Both are as successful as could be expected at simulating FTE temporal data, although room for improvement is found and indicated. The parameter choices necessary to make the matrix walk model work, however, are intuitively unsatisfying and suggest that a different approach to constructing a knowledge-representing "link matrix" is necessary.

For these initial attempts, only the temporal aspects of FTE data was considered. Formalizing and quantifying the meanings of entered terms presents formidable problems, and other tasks do not present the large data sets desirable for confident statistical analysis. Extending the two models to TPTE data should not be overly difficult.

4.1. Random Distribution Model for FTE Times

Subsection 3.1.1 presented a phenomenological description of the FTE data gathered during ConMap studies. It was argued that thinking time sets obtained from subjects' FTE task data typically show a logarithmic "spike plus peak" distribution: when a histogram of the logarithms of a subject's thinking times is constructed, it looks like a narrow, tall spike superimposed on the leading edge of a broad, approximately Gaussian peak. When data sets from all subjects of one study were rescaled to a common mean and width and then aggregated into one large data set, the histogram of that set's logarithms could be well fit by a superposition of two Gaussian peaks.

Successor correlation plots for thinking times in FTE data did not show any correlation between successive thinking times. The only correlation found among thinking times was a general tendency for longer thinking times to appear nearer to the end of the set, causing subjects' average term entry rate to decrease as the task progressed.

Similarly, typing times were found to be approximately describable by a log-normal distribution (i.e. histograms of the typing time values' logarithms resembled the probability density function of a Gaussian distribution). Typing times also showed no correlation between successive values, and, unlike thinking times, did not show a noticeable trend towards larger values as the task progressed.

These observations suggest that temporal aspects of FTE task data might be modelable by suitable random number distributions. The trend to decreasing term entry rate (increasing thinking times) would require a random number distribution whose parameters vary with elapsed task time, a complication which will be ignored for the moment. This section presents such a model of FTE temporal data. It is a purely phenomenological model, in that it incorporates no assumptions about the cognitive processes or knowledge structures underlying FTE task performance and does not attempt an explanation at that level; it merely attempts to predict as many of the statistical features of FTE data sets' time values as possible. This model will be referred to as the *random distribution model*.

4.1.1. The Model Defined

Thinking time sets from individual subjects' data proved too noisy to allow a satisfactory fit of a function with five parameters, required for a double-Gaussian or other similarly general two-peaked function. As a result, no functional form was found to describe the spike-plus-peak shape characteristic of FTE thinking time histograms. Instead, the random distribution used for the model was the logarithmic double-Gaussian, whose probability density function fit the data set produced by standardizing and aggregating all subjects' data from a study.

According to the model, each successive thinking time in a FTE data set is drawn independently from a *logarithmic double-normal* random distribution, whose probability density function (PDF) is described by

$$p_{\mathrm{LDN}}(x) \equiv \frac{\alpha}{\sigma_1\sqrt{2\pi}}\exp\left(\frac{-[\ln(x)-\mu_1]^2}{2\sigma_1^2}\right)+\frac{1-\alpha}{\sigma_2\sqrt{2\pi}}\exp\left(\frac{-[\ln(x)-\mu 2]^2}{2\sigma_2^2}\right) \quad \text{(Eq. 4.1),}$$

where μ_1, σ_1, μ_2, σ_2, and α are parameters of the model. This is equivalent to taking the exponential of numbers drawn independently from a (non-logarithmic) *double-normal* random distribution, whose PDF is

$$p_{\mathrm{DN}}(x) \equiv \frac{\alpha}{\sigma_1\sqrt{2\pi}}\exp\left(\frac{-[x-\mu_1]^2}{2\sigma_1^2}\right)+\frac{1-\alpha}{\sigma_2\sqrt{2\pi}}\exp\left(\frac{-[x-\mu 2]^2}{2\sigma_2^2}\right) \quad \text{(Eq. 4.2).}$$

A histogram of the logarithms of thinking times generated by such a model should therefore have the desired dual-Gaussian shape.

Similarly, each successive typing time in a FTE data set is drawn independently from a *log-normal* random distribution, whose PDF is described by

$$p_{\mathrm{LN}}(x) \equiv \frac{1}{x\sigma_3\sqrt{2\pi}}\exp\left(\frac{-(\ln(x)-\mu_3)^2}{2\sigma_3^2}\right) \quad \text{(Eq. 4.3),}$$

where μ_3 and σ_3 are model parameters. This is equivalent to taking the exponential of numbers drawn independently from a (non-logarithmic) *normal* (*Gaussian*) random distribution, whose PDF is

$$p_{\mathrm{N}}(x) \equiv \frac{1}{\sigma_3\sqrt{2\pi}}\exp\left(\frac{-(x-\mu_3)^2}{2\sigma_3^2}\right) \quad \text{(Eq. 4.4).}$$

A histogram of the logarithms of typing times so generated should therefore have the desired Gaussian shape.

A *synthetic data set* is generated from the model by producing a set of *start times* and *enter times* according to the following prescription: the first event's start time is defined to be zero, and its enter time (equal to typing time, for the first term) is drawn from the log-normal distribution of Equation 4.3. Each successive event's start time is equal to the previous event's enter time plus a thinking time drawn from the logarithmic double-normal distribution of Equation 4.1, and its enter time is equal to its start time plus a typing type drawn from the log-normal distribution of Equation 4.3. The prescription is repeated until data for the desired number of events have been generated.

4.1.2. Comparing Model Output to Subject Data

By construction, the model will produce data whose thinking times and typing times obey the distributions found in Subsection 3.1.1, so obedience these distributions cannot be used as a test of the model. It is not expected to show the tendency to decreasing term entry rate found in real data. In the absence of other statistical tests with which to compare synthetic model-generated and real study-gathered data, timeline plots were generated for several sets of synthetic data and compared by eye to timelines for real subjects' data.

Specific values of model parameters were required. For comparison to subjects' data from the p151s99 study, values for μ_1, σ_1, μ_2, σ_2, and α were taken from the fit to aggregated data done in Subsection 3.1.1 (cf. Figure 3.11), but these values were for data that had been *standardized* to a mean of zero and a width of one. The model values were "un-standardized" by the inverse of the transformation used to standardize subject p151s99-01's data. Values for μ_3 and σ_3 were taken directly from a normal-distribution fit to subject p151s99-01's set of typing times (listed in Table 3.2). A number of terms was generated equal to the number of terms in the subject's data set. Subject p151s99-01 was chosen because his or her FTE data was typical in most respects, but contained a relatively large number of terms, resulting in less noise and better fits.

For the timeline comparisons done, synthetic data sets did not show the increasing sparseness evident in real subjects' data, as expected. Apart from this, the synthetic data sets were not visibly different from real data sets except for the

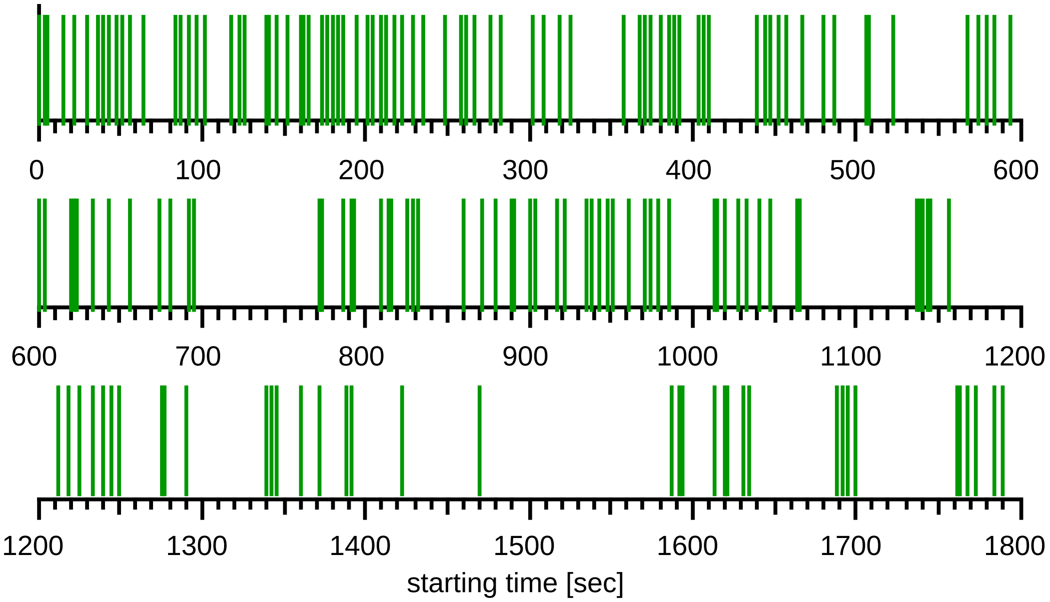


Figure 4.1: Timeline of FTE entries for study p151s99, subject 01 on task J1_FTE.

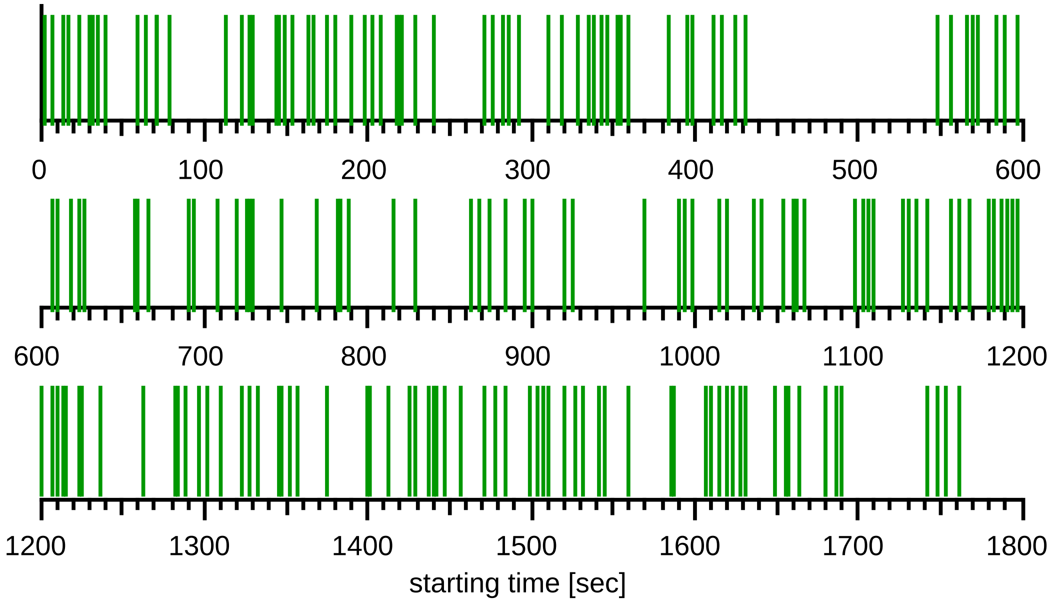


Figure 4.2: Timeline for synthetic data set generated from random distribution model, with model parameters taken to match J1_FTE data of subject p151s99-01.

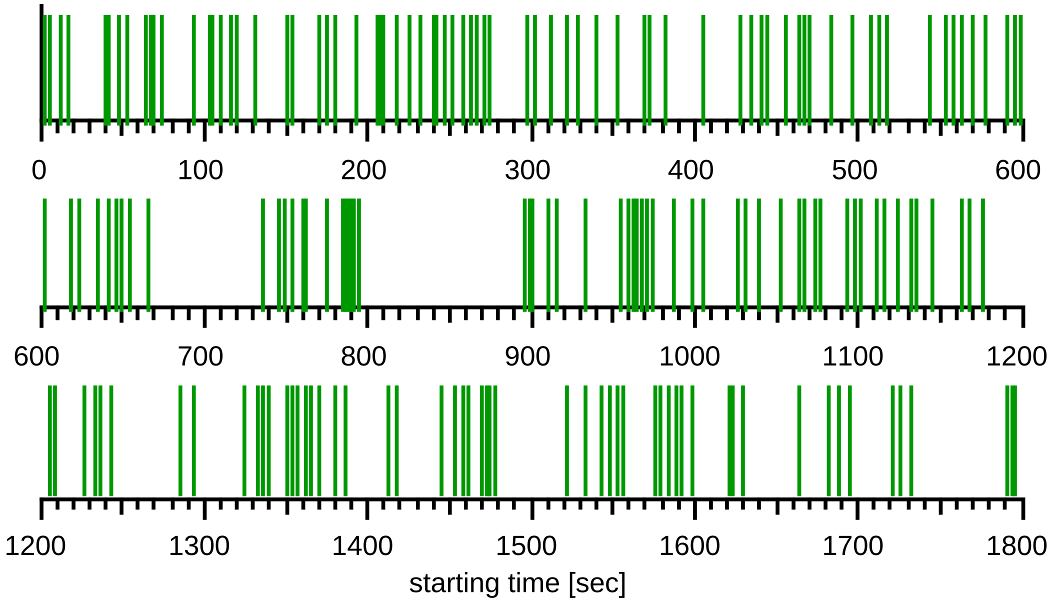


Figure 4.3: Same as Figure 4.2, with a different seed to the random number generator.

expected and nonsystematic effects of randomness. As an example, Figure 4.1 shows a timeline for subject p151s99-01 on task J1_FTE; Figure 4.2 and Figure 4.3 show timelines for two synthetic data sets generated as described above.

4.1.3. Decreasing Term Entry Rate

The model can be elaborated so that it exhibits a decreasing term entry rate during simulated FTE tasks. For data generated from the model, the rate of term entry events has a statistical average (for an appropriately sized moving window) equal to the inverse of the mean time between events, which is the expectation value of the thinking time plus the expectation value of the typing time for their respective distributions:

$$r = \langle \Delta \mathrm{t} \rangle^{-1} = \left(\langle t_{\mathrm{think}} \rangle + \langle t_{\mathrm{type}} \rangle \right)^{-1} \qquad \text{(Eq. 4.5)}$$

The expectation value for typing times drawn from the log-normal distribution of Equation 4.3 is

$$\langle t_{\mathrm{type}} \rangle \equiv \int_0^\infty \mathrm{d}x \; x \; p_{\mathrm{LN}}(x) = \exp\left(\mu_3 + \frac{\sigma_3^2}{2} \right) \qquad \text{(Eq. 4.6)}$$

and the expectation value for thinking times drawn from the logarithmic double-normal distribution of Equation 4.1 is

$$\langle t_{\mathrm{think}} \rangle \equiv \int_0^\infty \mathrm{d}x \; x \; p_{\mathrm{LDN}}(x) = \alpha \exp\left(\mu_1 + \frac{\sigma_1^2}{2} \right) + (1-\alpha) \exp\left(\mu_2 + \frac{\sigma_2^2}{2} \right) \qquad \text{(Eq. 4.7)}$$

The term entry rate *r* can therefore be expressed as a function of the model parameters:

$$r = \left[\alpha \exp\left(\mu_1 + \frac{\sigma_1^2}{2} \right) + (1-\alpha) \exp\left(\mu_2 + \frac{\sigma_2^2}{2} \right) + \exp\left(\mu_3 + \frac{\sigma_3^2}{2} \right) \right]^{-1} \qquad \text{(Eq. 4.8)}$$

If *r* is to vary with elapsed task time, then one or more of the model parameters α, μ_1, σ_1, μ_2, σ_2, μ_3, and σ_3 must depend on time. As discussed in Subsection 3.1.1, subjects' thinking time distributions evolved noticeably over the course of a FTE task, with longer times being significantly more likely to occur later in the task, but typing time distributions appeared to remain relatively unchanged. It therefore seems reasonable to keep μ_3 and σ_3 constant and allow the other parameters to vary.

Which of those parameters should vary and what the functional form of their time dependence should be are questions that have not yet been investigated. The primary objective of the modified model should be to demonstrate term entry rate vs. elapsed task time behavior statistically equivalent to real subjects' data. Given that the distribution for thinking times has five parameters which can be assigned a completely arbitrary time dependence, and given that calculations of term entry rate for subject data are extremely noisy even with large moving-window averages, parsimony considerations alone are unlikely to suggest how the time dependence should be incorporated. Further statistical analysis of thinking time distributions could help, perhaps by constructing and fitting histograms for subsets of the data drawn from early, middle, and late periods of the task. Additional studies designed to gather data specifically for this purpose might be required.

4.1.4. Summary and Discussion

It has been shown that a relatively simple statistical model of FTE data, in which thinking times and typing times are independently drawn from appropriately parameterized random distributions, can produce synthetic data which resembles real subjects' data in most respects. The primary failing of the model is that it does not predict that the rate of term entry during a task decreases with elapsed task time. A strategy for extending the model to address this failing was presented.

Because the model is purely phenomenological, it is not particularly illuminating; it is more a summary of the statistical behavior observed in the data than an explanation for that behavior. It does not address the mechanisms of knowledge, knowledge access, learning, cognition, and task performance required to "understand" FTE data, but buries all such detail within the randomness of the distribution.

4.2. Matrix Model for FTE Times

In an initial and exploratory attempt to construct a model more meaningful than the random distribution model of the previous section, the *matrix walk model* was developed. The model is intended to be a simple yet reasonable description of some of the cognitive processes underlying FTE task performance. It can produce a distribution of thinking times that is approximately log-normal. The log-double-normal or "spike plus log-normal" distribution found in the real FTE data have not been attempted yet, but are in principle possible. The model naturally demonstrates decreasing term entry rate throughout a simulated FTE task.

4.2.1. Description of the Model

A subject's conceptual knowledge store for a particular domain is represented as an *N* by *N link matrix* **L** of real numbers, where *N* is the number of *knowledge elements* in the structure. These knowledge elements are assumed to represent concepts and their associated terms. Each matrix element $\mathrm{L}_{\mathrm{i,j}}$ ($i \in \{1, 2, \ldots, N\}$, $j \in \{1, 2, \ldots, N\}$) represents the strength of the "link" from element *j* to element *i*: the degree to which element *j* "triggers" or "brings to mind" element *i*. **L** is not necessarily symmetric, and the diagonal elements $L_{i,i}$ are irrelevant. The model does not ascribe a specific meaning to any of the knowledge elements.

The "dynamics" of the model is an algorithm called "the walk" which represents performance of an FTE task, and which produces simulated FTE data analogous to the data collected from subjects. The algorithm is a set of rules for generating a list of the knowledge elements $\{a_k\}$ representing a term list, given a specific matrix and a choice for the initial "prompt" term. A thinking time τ_k is determined for each element. The model does not address typing times; on the assumption that typing times depend primarily on the letters appearing in terms and are relatively independent of terms' cognitive associations, a thinking time for each term can be drawn from a random distribution. The distribution used for this was the same as the typing time distribution for the random distribution model of Section 4.1. Once lists of elements, thinking times, and typing times have been generated, they can be used to construct corresponding lists of start times and term entry times, completing the construction of a simulated data set.

The walk is defined by the following steps:

1. The first "active" element is arbitrarily chosen as element one: $a_0 = 1$. There is no associated thinking time. This element is ineligible for future selection.

2. Given a currently active element ak, the next active element is chosen to be the one for which the link strength L_{a_{k+1},a_k} is maximal, excluding previously-active elements a0, a1, …, ak. In other words, the next element chosen is the one linked to most strongly from the set of all not yet chosen.

3. The thinking time for the recall process of step 2 is defined to be $\tau_{i+1} = \tau\left(L_{a_{k+1},a_k}\right)$, where $\tau(s)$ is a recall function which will be discussed below.

4. Update the "counter" variable k → k + 1.

5. Go back to step 2 and repeat, unless a criterion for task termination (e.g. total number of terms or elapsed task time) has been met, in which case the task is finished.

To complete the model, values for the matrix elements of **L** must be specified, and the recall function $\tau(s)$ must be specified. (In the notation used here, $\tau(s)$ represents a mathematical function, while τ_k represents one particular thinking time value.)

In order to keep initial investigations of the model as simple as possible and facilitate the statistical analysis that follows, the link matrix was populated with link strengths randomly drawn from a distribution uniform between zero and one, except for diagonal elements which were set to zero. The merits of this choice will be discussed below.

4.2.2. Choice of Recall Function

For the model as described above, the thinking time associated with a particular FTE response element (term) is uniquely determined by the link strength to that element from the previously entered element. The recall function $\tau(x)$ determines the mapping between link strength value and thinking time. The following criteria were set for the recall function:

- It should tend to small values (zero or some specified minimum possible thinking time) for link strengths approaching their maximum possible value;

- It should tend to infinity for link strengths approaching their minimum possible value;

- It should be a relatively simple, well-behaved mathematical function amenable to analysis.

The first two criteria reflect the intuitive notion that strong conceptual connections should produce quick responses, while weak connections should produce long responses. Since FTE thinking times in the study data extended over two orders of magnitude, with vanishingly small frequencies of the largest times, having the recall function tend asymptotically to infinity seemed appropriate. The third criterion enforces parsimony and analytical convenience.

For link strengths between zero and one, two candidate recall functions are the *logarithmic recall function*

$$\tau(x) = -\alpha \ln(x) \qquad \text{(Eq. 4.9)}$$

and the *power-law recall function*

$$\tau(x) = \alpha \left(x^{-\gamma} - \beta\right) \qquad \text{(Eq. 4.10),}$$

where α, β, and γ are arbitrary parameters within the ranges $\alpha > 0$, $0 < \beta \leq 1$, and $\gamma \geq 1$. The logarithmic recall function is simpler (has fewer parameters); the power-law allows more tuning via parameters. For both functions, the parameter α sets the time scale. For the power law, β determines the minimum possible thinking time, and γ controls the relative abundance of long vs. short thinking times.

It is not necessary to generate many data sets with the model, trying various wild guesses for the parameter choices for both recall functions, in order to get a sense of which recall function and what approximate parameter values might best fit the observed FTE data. Instead, it is possible to work backwards from the desired distribution of thinking times to find an appropriate recall function. For initial investigations, a log-normal distribution of thinking times was chosen as the goal. Since the FTE data was crudely log-normal, producing such a distribution with the model seemed a valuable demonstration. Predicting the "spike plus peak" appearance of subjects' FTE thinking time sets with a more complicated recall function has not yet been attempted.

Finding a recall function which provides the appropriate distribution of thinking times is done by deriving an expression for the probability density function (PDF) of the thinking times produced by the model, for a given distribution of link strengths and choice of recall function. A simplification is made to remove the complication imposed by the fact that the distribution changes as elements are chosen and the pool of "eligible" elements dwindles: it is assumed that whenever an element is chosen and made ineligible, a new element with new randomly-chosen link matrix elements replaces it. This is equivalent to assuming that the number of elements chosen during a complete FTE task is much smaller than the total number N of available elements. Another way to look at it is that the distribution derived only describes the first thinking time of a task for each of an ensemble of model subjects. Since the calculation is only intended to aid in the selection of a recall function and reasonable first approximations for the recall function's parameters, this simplification is not a major compromise.

The model prescribes that an element is chosen for "term entry" by finding the j for which $L_{i,j}$ has the largest value, given the previously chosen (or initially seeded) element i. This is equivalent to choosing the largest of a set of $N - 1$ numbers drawn randomly and independently from a distribution uniform between zero and one. The probability distribution for the result is therefore

$$\mathrm{p}(s) = (N-1)\, s^{N-2} \qquad \text{(Eq. 4.11),}$$

where *s* is the value of the selected link strength (matrix element). This distribution has the distinctive feature that for large *N* it very strongly weights strengths close to one and assigns almost zero probability to other values. The implications of this feature are important and will be discussed below.

Given a recall function τ(*s*), the distribution of thinking times q(τ) corresponding to the distribution of selected link strengths p(*s*) is described by the relation

$$\mathrm{q}(\tau)\,d\tau = \left|\mathrm{p}(s)\,ds\right| \quad \Rightarrow \quad \mathrm{q}(\tau) = \mathrm{p}\big(\mathrm{s}(\tau)\big)\left|\frac{d\,\mathrm{s}(\tau)}{d\tau}\right| \qquad \text{(Eq. 4.12),}$$

where s(τ) is the recall function τ(*s*) inverted. If the desired distribution q(τ) is known, the recall function can be solved for by integrating Equation 12 to get

$$\int_0^\tau d\tau'\,\mathrm{q}(\tau') = \int_s^1 ds'\,\mathrm{p}(s') \qquad \text{(Eq. 4.13),}$$

performing the integrals, and solving for τ in terms of *s*. Note that the limits of integration have been chosen so that the low end of the τ range corresponds to the high end of the *s* range, since τ is intended to be a decreasing function of *s*.

To produce thinking times with a log-normal distribution,

$$\mathrm{q}(\tau) = \frac{1}{\sqrt{2\pi}\,\sigma\,\tau}\exp\left(\frac{-\left(\ln(\tau)-\mu\right)^2}{2\sigma^2}\right) \qquad \text{(Eq. 4.14)}$$

where μ and σ describe the mean and width of the peak, respectively, on a logarithmic plot. Inserting this and Equation 4.11 into Equation 4.13, integrating, and solving for τ results in

$$\tau(s) = \exp\left(\mu - \sigma\sqrt{2}\ \mathrm{erf}^{-1}\left(2s^{N-1}-1\right)\right) \qquad \text{(Eq. 4.15),}$$

where erf^{-1}() is the inverse of the error function defined by

$$\mathrm{erf}(x) = \frac{2}{\sqrt{\pi}}\int_0^x dt\,\exp\left(-t^2\right) \qquad \text{(Eq. 4.16).}$$

The inverse error function can be evaluated numerically but is analytically problematic, suggesting that one of the candidate recall functions presented in Equations 4.9 and 4.10 would be a better choice for the model if it could reasonably approximate Equation 4.15 with the right parameter choices.

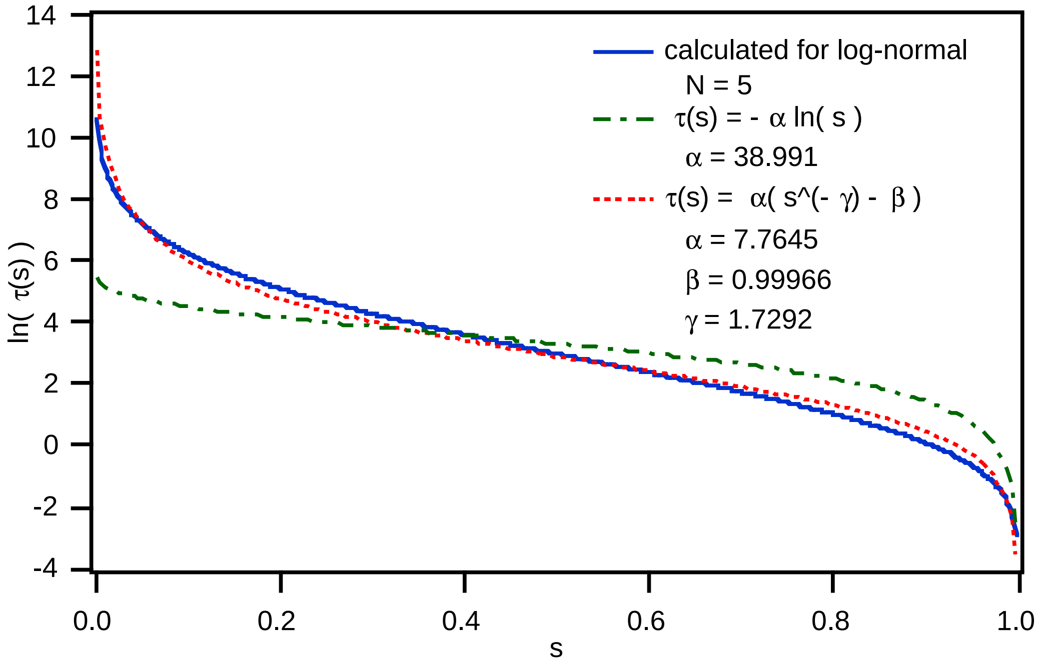


Figure 4.4: Comparison of two candidate recall functions with the recall function derived for a log-normal distribution of thinking times. Curves for the candidate functions are the result of a chi-squared fit, with best-fit parameters shown.

Figure 4.4 compares the recall function in Equation 4.15, derived from a log-normal distribution, with the logarithmic and power-law recall functions of Equations 4.9 and 4.10. For the derived function, $N = 5$ has been used for illustrative purposes. $N \geq 200$ would be more realistic, but the range of values produced by the function for large N is so extreme that numerical overflow problems prevent the construction of an accurate plot. The parameter values for the logarithmic and power-law functions have been chosen for maximum agreement with the derived function, according to a chi-squared curve fit.

It is clear from the plot that at least for small N, the power-law recall function can better approximate the log-normal-derived recall function than can the logarithmic recall function. Furthermore, the best-fit value of β is effectively unity. These two results hold for other relatively small values of N (not shown), suggesting that the best choice for a model recall function would be the power-law, simplified by fixing $\beta = 1$.

Using Equation 12, the approximate distribution of early thinking times generated by the model (before element ineligibility becomes a significant effect) can be determined:

$$\mathrm{q}(\tau) = \frac{N-1}{\alpha\,\gamma}\left(\frac{\tau}{\alpha} + \beta\right)^{-\left(\frac{N-1}{\gamma}+1\right)} \qquad \text{(Eq. 4.17).}$$

As discussed during analysis of the study data, it is more convenient to work with the logarithms of the thinking times. The distribution of thinking time logarithms for the model can be derived from Equation 17 with a simple change of variables, yielding

$$\mathrm{r}(z) = \eta \, e^{z-z_0}\left(e^{z-z_0} + \beta\right)^{-(\eta+1)} \qquad \text{(Eq. 4.18)},$$

where $z \equiv \ln(\tau)$, $z_0 \equiv \ln(\alpha)$, and $\eta \equiv (N-1)/\gamma$. This distribution ought to resemble a Gaussian curve, and it does, as seen in Figure 4.5. With β fixed at unity, the model only provides two parameters to control the shape of the distribution, since N and γ always appear in the same combination (labeled η). α determines z_0, a horizontal axis offset (equivalent to setting the time scale for thinking times). η controls the peak width and also impacts the location of the peak maximum, as demonstrated by Figure 4.5. With parameters to control both peak width and location, the distribution ought to be capable of modeling real FTE data.

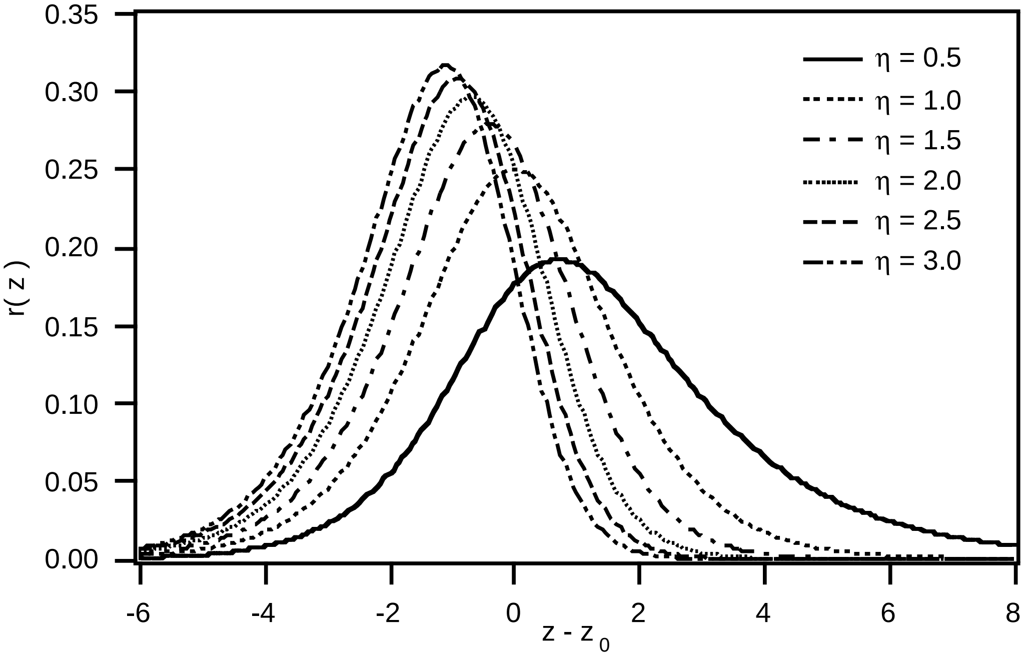


Figure 4.5: Probability distribution function r(z) for logarithms of thinking times generated by power-law recall function, for a range of values of the parameter η and for $\beta = 1$.

4.2.3. Comparison of Model Results with Study Data

The model, with the power-law recall function, was used to generate synthetic FTE data. The distribution function of Equation 4.17 was used to estimate parameter values which would result in a distribution comparable to subject p151s99-01's distribution of thinking time logarithms for task J1_FTE. This subject was chosen because he or she demonstrated an unusually large number of response terms but appeared otherwise typical, resulting in less noisy data than that of most other subjects. Because the goal was to demonstrate that the model is capable of matching the general characteristics of FTE data, not to model the details of individual subjects, choosing one subject as an archetype introduces no compromises.

This provided initial guesses for η and α. Additional information is required to determine N and γ from η. The choice of N affects how term entry rate changes as the task progresses: if N is much larger than the total number of terms C entered during the task, then the term entry rate does not change significantly; but if N is only slightly larger than C, term entry rate drops drastically, since only a few elements remain near the end. One can think of the parameter η changing throughout the task, with

$$\eta_k \equiv \frac{N-1-k}{\gamma} = \eta_0 - \frac{k}{\gamma} \qquad \text{(Eq. 4.19)}$$

being the value after k term entry events. As η_k drops throughout the task, the distribution of resulting thinking times spreads and moves to higher values (see Figure 4.5), causing the rate of term entry to drop. Using this behavior to determine a reasonable estimate for N would require more precise data on FTE term entry rates than is currently available. Therefore, values of N in the range of 200-500 were explored numerically.

With initial parameter guesses in hand, the model was implemented on a computer and run to generate a synthetic data set. The simulation was terminated after 175 events had been generated, since the subject data being compared to consisted of 174 events. This was repeated many times for varying parameter values until a parameter set was found which produced data reasonably similar to the target subject's FTE data. More specifically, a quantile plot was constructed for each model data set, and fit with a log-normal cumulative distribution function; the parameter set chosen produced the same best-fit parameters as the same fit to the target subject's data. Since the intent was to demonstrate the model's general capability to produce reasonably realistic data, this was considered sufficient tuning of the parameters.

The final set of parameter values chosen was $N = 300$, $\gamma = 75$, and $\alpha = 15$, with β still fixed at 1. The resulting quantile plot and histogram for one instantiation (i.e. one randomly generated link matrix) are shown in Figure 4.6 and Figure 4.8, along with a best-fit log-normal curve (actually a best-fit normal curve to the logarithms of the thinking times). For comparison, the equivalent plots for subject p151s99-01's FTE data are shown in Figure 4.7 and Figure 4.9.

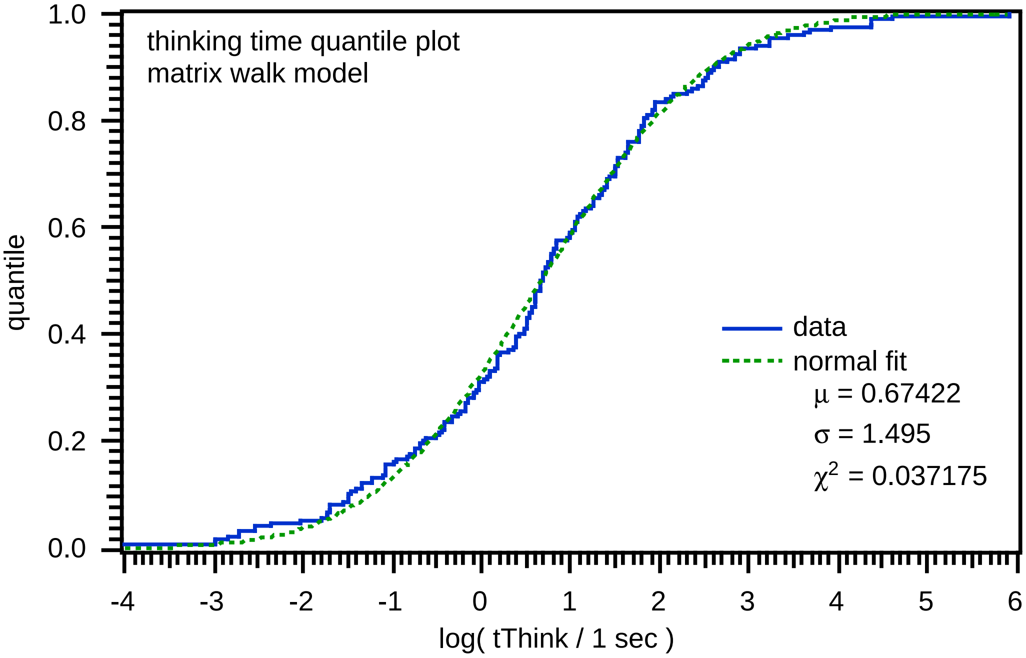


Figure 4.6: Thinking time quantile plot for model-generated data, using $N = 300$, $\gamma = 75$, $\alpha = 15$, and $\beta = 1$. The best-fit normal (Gaussian) CDF to the thinking time logarithms is shown.

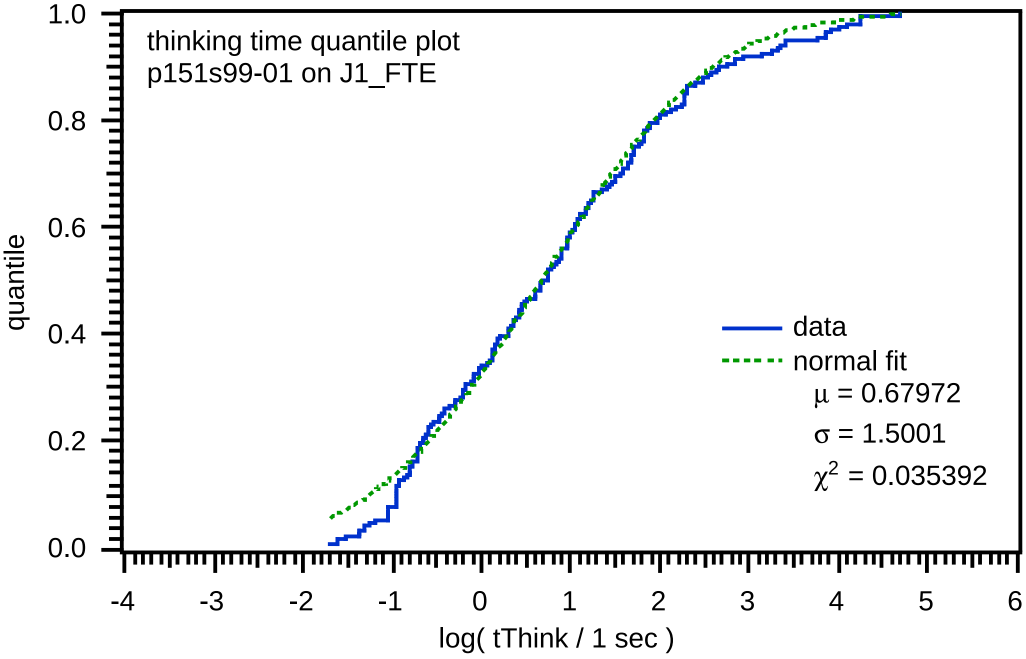


Figure 4.7: Thinking time quantile plot for subject p151s99-01 on task J1_FTE, with best-fit normal (Gaussian) CDF.

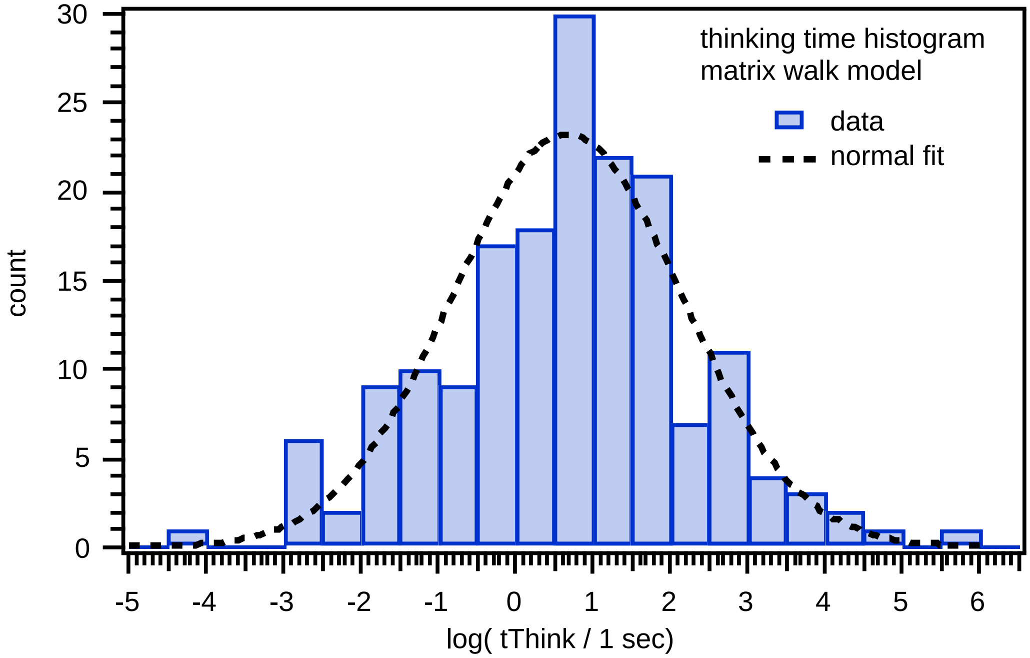


Figure 4.8: Histogram of logarithms of thinking times for the model data displayed in Figure 4.6, with PDF curve for the best-fit normal distribution.

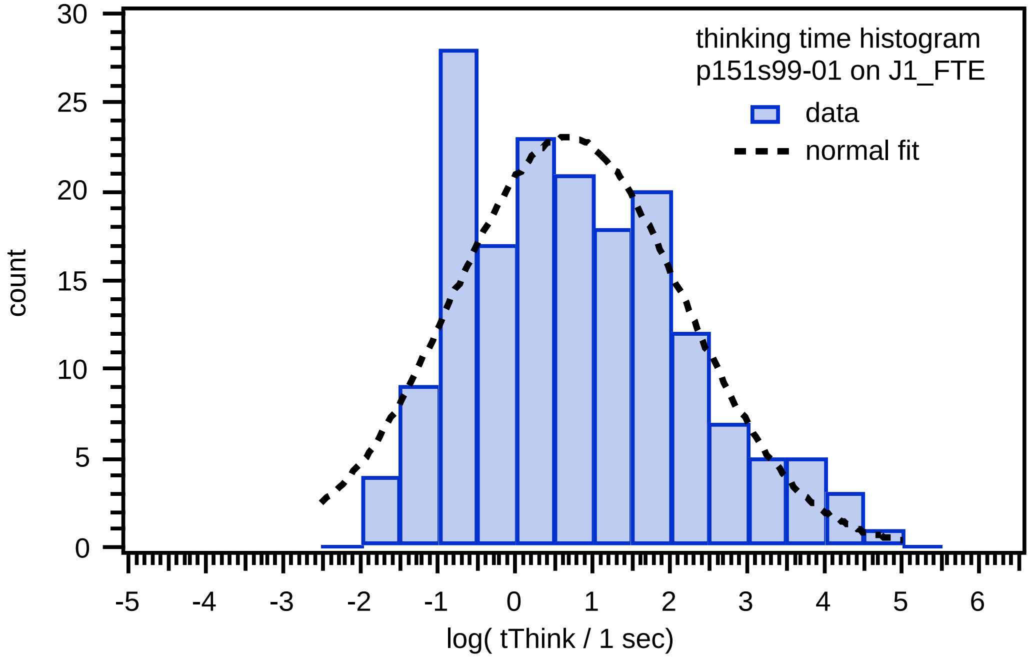


Figure 4.9: Histogram of logarithms of thinking times for the subject data displayed in Figure 4.7, with PDF curve for the best-fit normal distribution.

It is evident from these plots and from similar plots for other model runs that the model is capable of producing thinking time distributions that resemble real FTE thinking time distributions, but without the leading spike seen in most subject's data. The model generally produces more outliers on both sides of the distribution than is seen in subject data. Low-end outliers could perhaps be eliminated by choosing a value of β lower than 1.

Figure 4.10 shows a timeline for the same model-generated data (including randomly-generated typing times); for comparison, Figure 4.11 shows a timeline for subject p151s99-01's data. The two timelines are not unreasonably different, aside from one long gap in the model timeline due to one of the aforementioned long-time outliers. The model data does show a tendency towards decreasing term entry rate as the task progresses, and longer thinking times are significantly more likely to be found during the later part of the task.

Figure 4.12 and Figure 4.13 compare plots of term entry rate vs. number of terms entered for the model and subject data. The plots are not qualitatively dissimilar, aside from the dip to zero that the model-generated data takes, corresponding to the long thinking time outlier. This confirms that the model demonstrates decreasing term entry rates throughout an FTE, as seen in the study data.

4.2.4. Discussion

The matrix walk model presented here has proven capable of modeling most of the general features of FTE thinking time data. It can reproduce the general log-normal statistical distribution of thinking times, although with the

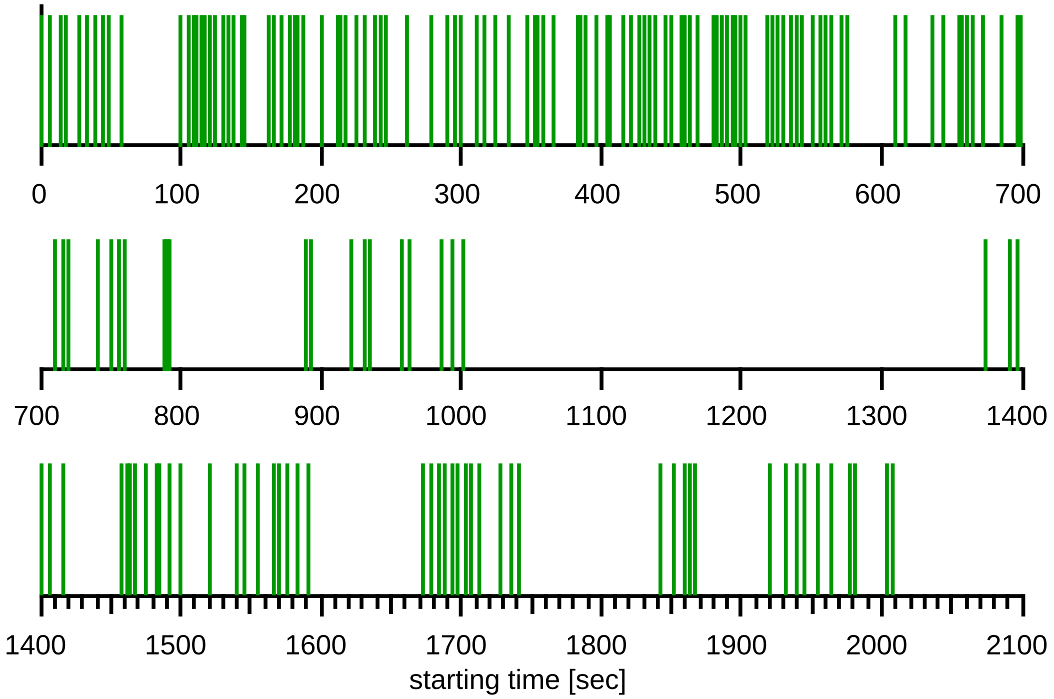


Figure 4.10: Timeline for model-generated data of Figure 4.6.

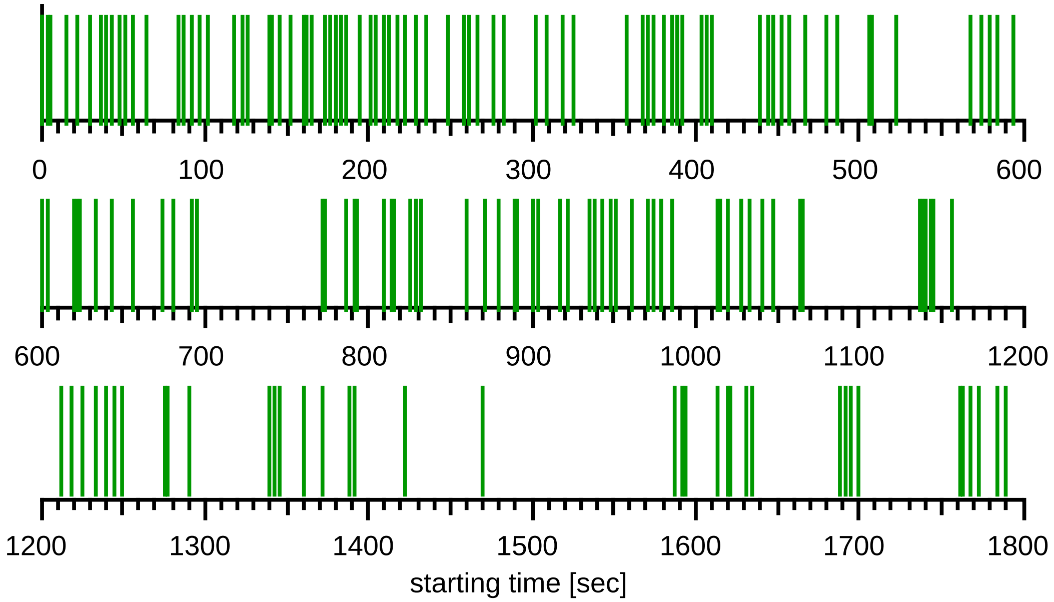


Figure 4.11: Timeline for subject p151s99-01 data on task J1_FTE.

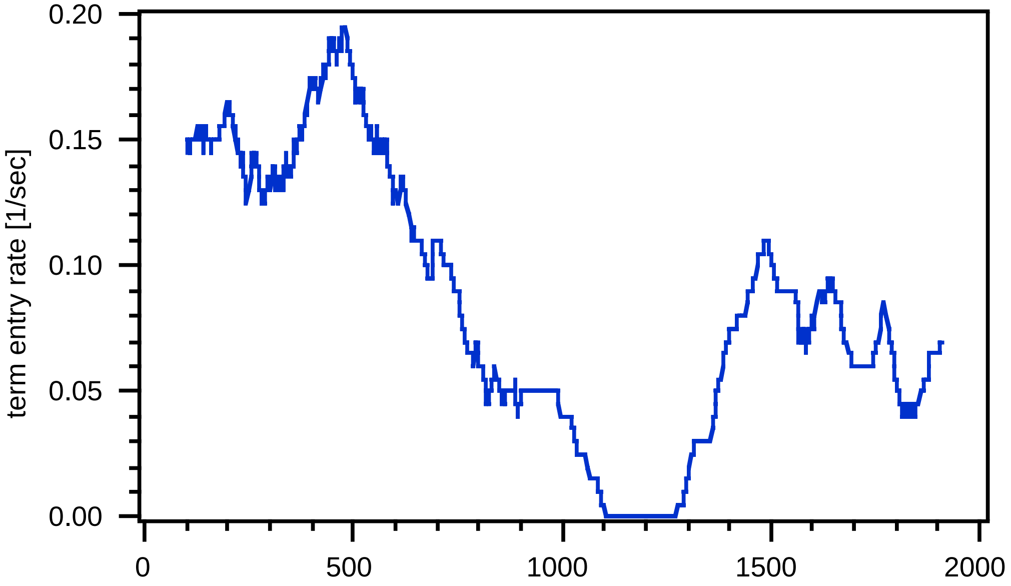


Figure 4.12: Term entry rate vs. number of terms entered for model-generated data of Figure 4.6.

simple recall functions considered so far it cannot reproduce the log-spike-plus-normal or log-double-normal distributions which more accurately describe thinking times. It naturally produces larger thinking times later in a simulated task, causing term entry rate to generally decrease, in accord with the behavior of

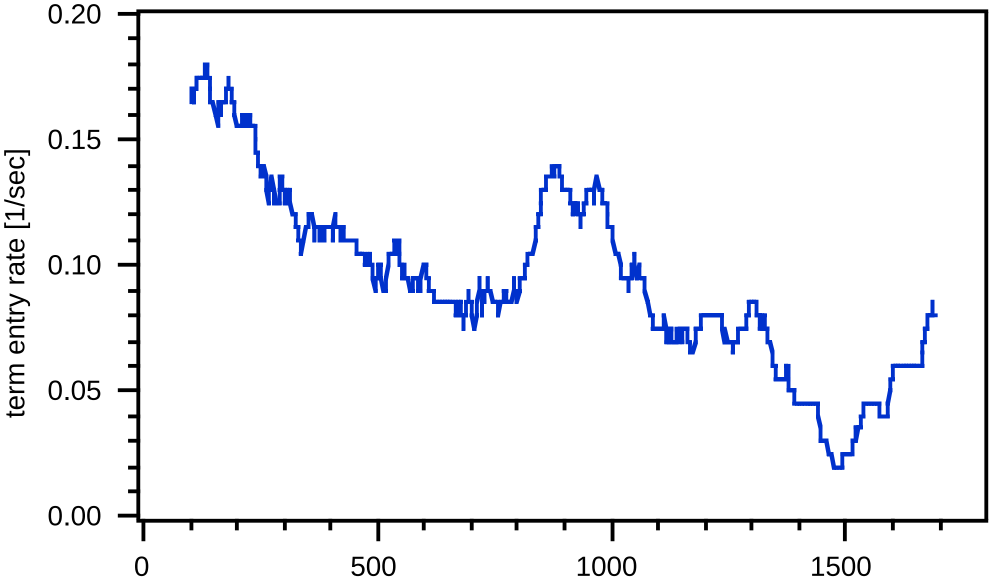


Figure 4.13: Term entry rate vs. number of terms entered for subject p151s99-01 on task J1_FTE.

subject data. The model seems to produce more distant outliers than are found in subject data.

Thus, the model appears relatively successful, and might even be modifiable to produce a double-normal or spike-plus-normal distribution of thinking time logarithms. However, the model is unsatisfying in one significant aspect: the size of the parameter γ needed to produce a reasonable thinking time distribution. For plausible model output, it was found that $\eta = (N - 1)/\gamma$ must be close to 4.0, and N must be approximately 300, which requires γ to be approximately 75. This can be interpreted in terms of distributions. The distribution of link strengths chosen according to the model's next-element selection algorithm is given by Equation 11. For large N, this distribution is extremely heavily weighted in favor of values very close to one: the mean value of the distribution is $(N - 1)/N$, which is 0.9967 for $N = 300$. This makes sense: for each element selection, the strongest of a set of $N - 1$ strength values is being selected.

Because all strengths responsible for thinking times are therefore extremely close to one, with very little variation, a hypersensitive recall function is required to produce an acceptable range of thinking times. Thus γ, the exponent in the power law recall function of Equation 10, must be huge. This is unacceptable for a model that aspires to be cognitively interpretable.

Inspection of FTE response terms entered during the studies shows that many terms entered by subjects are connected only tenuously if at all to the previous term, and analysis showed that these connections tend to correlate with longer thinking times. It therefore seems reasonable to insist on a model in which longer thinking times result from weak links, not from links almost equal to the strongest links.

This suggests that the link matrix representing connections between knowledge elements should not be filled with numbers randomly drawn between zero and one, but according to a different scheme. One possibility is to fill a randomly-chosen subset of the elements with numbers between zero and one, and set the rest to zero: a “sparse matrix”. If the number of nonzero elements in each column is the same, defined to be *M*, then all of this section’s calculations for the uniformly-filled matrix walk model hold except for Equation 4.19, with *M* replacing *N*. Effectively, this modification allows one to reduce the parameter *N* to a less problematic value without “running out of elements” too soon in the task.

If the statistics of the resulting thinking times are still not appropriate, one might consider filling the matrix with strength values in a correlated way, resulting in “clusters” of strongly interconnected elements, with sparser, weaker connections between elements of different clusters. The resulting FTE “walk” through the matrix should encounter all of the elements in such a cluster, and then follow a weaker link to a new cluster. Such an approach would prevent the random-distribution-based analysis performed above, but also seems to capture current qualitative understanding of how physics knowledge is structured. FTE subjects certainly perceive themselves to enter a set of closely-related terms (e.g. on “circular motion” or “types of forces”), and then follow a weaker connection to another tightly-associated set. Fractal algorithms for generating such a link matrix have been considered but not yet implemented.

In fact, the “large γ problem” found for the uniformly-filled matrix may be interpreted as evidence that a uniform matrix of randomly chosen numbers is not a good model of a subject’s conceptual knowledge store, and that other, sparse or correlated-element models should be investigated. This is good: as experimental evidence constrains our choice of model, it shapes our understanding of the system.

5. Conclusion

The purpose of the ConMap research project has been to investigate the utility of a particular set of proposed assessment tools — brief, computer-administered tasks for eliciting conceptual associations — for probing the quality and extent of a physics student's conceptual knowledge structure in an introductory physics domain. Three component goals were identified and pursued:

1. Devise and test possible strategies for probing physics students' knowledge structures (probe design);

2. Search for potentially meaningful patterns and correlations in data provided by the probes (data analysis); and

3. Develop quantitative models of knowledge structure and access consistent with the gathered data (modeling).

A number of interesting and promising results are summarized below.

5.1. Significant Findings

5.1.1. Free Term Entry (FTE) Task

Because of the size of FTE data sets, statistical patterns the set of subjects' response times could be well identified. It was found that subjects' thinking time sets and typing time sets both obey a generally log-normal distribution, and that thinking time distributions have a sharp leading spike. Knowledge of these distributions provides a target for cognitive modeling.

Two models for FTE timing data were constructed. The *random distribution model* demonstrated that most of the statistical behavior of FTE times could be reproduced by randomly chosen values drawn from appropriate distributions, without specifying any detailed mechanism. The model did not reproduce the trend towards decreasing term entry rate with elapsed task time, but additional model parameterizations were suggested which should allow the model to fit this behavior as well. The *matrix walk model* was based on a deterministic algorithm acting on a link matrix representation of knowledge structure, and was also able to match general statistical aspects of FTE temporal data; this model naturally produced a decreasing term entry rate. The method chosen to fill the link matrix — uncorrelated random values all drawn from the same uniform distribution — required an intuitively unsatisfactory choice of value for one of the model parameters. This fact was interpreted as a positive result, because it indicates that modeling can provide information about the structure a link matrix must have to provide a realistic representation of knowledge structure. Suggestions for more sophisticated link matrix structures were made.

It was found that FTE response terms which are “jumps” — relatively unrelated to immediately prior response terms — are associated with significantly longer thinking times than are non-jumps. This indicates that thinking time can be used to reveal subjects’ conceptual associations, and suggests that thinking time measurements can be used to construct efficient, automated ConMap-style tools for measuring students’ knowledge structures. General measures of subjects’ overall “jump rate” were found to correlate significantly, if noisily, with course exam performance.

5.1.2. Hand Drawn Concept Map (HDCM) Task

Suggestive evidence was found that various quantitative summary measures of students’ hand-drawn concept maps — number of nodes, number of links, and ratio of links to nodes — might correlate with the students’ course exam performance.

5.1.3. Term-Prompted Term Entry (TPTE) Task

The impact of course-related learning was detected in TPTE responses by looking at the set of all response terms entered by the study population to a particular prompt term, and the number of subjects entering each term: it was found that when the same term was presented during different sessions, most of the response terms appeared with approximately the same frequency, but a few drastically changed frequencies. Subjects’ response lists also became slightly more similar to one another as the course progressed, and dramatically so for a few prompts.

When subjects’ response lists were given a score based on the response terms present, according to domain experts’ judgments of the relevance and importance of each response, it was found that the resulting scores correlate significantly with course exam performance. This indicate that a TPTE-style task has the potential to serve as a useful assessment of domain expertise.

The terms in subjects’ TPTE response lists were also found to overlap strongly with their HDCM terms for the same prompt, especially with terms in nodes directly connected to the prompt term’s node, indicating that a TPTE-based equivalent of the time-consuming concept map assessment could be developed.

5.1.4. Problem-Prompted Term Entry (PPTE) Task

The fraction of subjects who entered a PPTE response term which was descriptive of the fundamental principle required to solve the prompt problem was found to increase after a week of course lecturing, homework, and studying on directly related topics, but to remain essentially the same without that coverage. This shows that the PPTE task is sensitive to changes in student knowledge resulting from course coverage.

5.2. Areas of Concern

Experience gained throughout the study has revealed some potential pitfalls in the use of the ConMap tasks and analysis of resulting data, and has led to some suggestions for more effective use of the tasks.

Much of the difficulty arises from the fact that the ConMap tasks allow subjects to choose their own responses. Subjects must therefore be cautioned and regularly reminded to keep their responses within the spirit of the task: avoiding, for example, terms that wander too far from the designated domain area, long terms that describe a relationship rather than a concept or "thing", and TPSE statements that contain two distinct propositions. It has become clear with experience what specific kinds of errors subjects tend to make, so these can be explicitly addressed in future studies.

Subjects' choices of term spelling, tense, and phrasing tend to be idiosyncratic and inconsistent, so any analysis methods that focus on the meaning of entered responses rather than their timings alone must either involve careful human judgments or sophisticated computer prescriptions for term comparison. This is not a weakness of the task designs, but rather a necessary consequence of the fundamental ConMap philosophy that subjects' spontaneous responses, not reactions to a set of comparisons, should be elicited.

5.3. Directions

Throughout this dissertation, several suggestions have been made about specific ways in which follow-up studies could further various lines of analysis and modeling. This section will propose a few major directions for follow-up research.

The studies conducted were designed to gather a wide variety of data with many different factors varied. More information on the consistency and reproducibility of the data would be valuable. For example, if the same FTE task is repeated to the same subject at different times, how similar are the timing patterns? How similar is the list of response terms, in terms of both global position of the terms (which are near the beginning, which near the end, etc.) and of local structure (do the same mini-sets show up each time)? How reproducible are a subjects' TPTE responses to a particular prompt, and how does this depend on recent experience such as previous prompts? For these kinds of questions, the thorny issue of subject memory must be confronted, since a subject's memory of tasks done will likely influence his or her behavior on subsequent tasks.

It is an open question how significantly the physical aspects of task performance — specifically, term typing on a keyboard — affects the gathered data. It seems likely that subjects with different typing abilities might be influenced qualitatively differently: subjects' self-observations expressed during informal interviews suggest that some subjects can think ahead to their next term while typing a term in, but others find typing to be so cognitively demanding that after they complete entering a term they must think back to recall their train of thought, or look at the screen to remind themselves of the prompt term. Alternative implementations of the term-entry tasks, perhaps involving the audio recording of spoken responses, should be illuminating. This could involve significant technical challenges; for example, extracting sufficiently accurate and precise time data on term starting times and "entering" times might require computerized auditory analysis of the recorded signal.

The TPTE task has been shown to provide information similar to that of a concept map, especially the portion of the map closest to the prompt or "starting" term. This suggests that a computer program could use repeated TPTE tasks to build up a concept map representation of a subject's knowledge structure, by using his or her responses to initial prompts as subsequent prompts and keeping track of all the interconnections that are indicated. If the problem of idiosyncratic spelling and phrasing choices can be solved, a useful "automatic concept mapping tool" might be possible. As a further elaboration, TPTE results might be used to guide the selection of term pairs for a TPJ task, reducing the number of pairings presented and helping to solve the "N^2 problem".

Attempts to validate ConMap measures by comparison with exam scores are of questionable value since in-course exam scores are a poor indicator of the kinds of expert-like knowledge structuring being probed for: if exam scores were reliable and properly sensitive, new assessment tools would be unnecessary. Comparison of ConMap task performance with performance on carefully designed, conceptually-focused problems should be much more effective at revealing the sensitivity of ConMap measures. Studies specifically designed to validate proposed ConMap measures of domain expertise are in order.

Studies which can provide data for further modeling are recommended. Specifically, the variation of FTE term entry rate with elapsed task time has shown itself to be important for the construction and testing of models. Having subjects repeat an FTE task multiple times, as suggested above, could be helpful in this regard. Modeling of TPTE time data was not attempted yet because TPTE data sets are short, and aggregating many together to get sufficient data points for meaningful analysis is dangerous given the number of factors varying. TPTE data with more repetition would be of benefit.

Finally, it would be interesting and valuable to investigate why subjects associate pairs of terms: in other words, to investigate what happens, cognitively, when one term brings to mind another. Comparing TPSE response statements with TPTE response terms might be helpful, and so might other avenues of information like focused interviews and perhaps some variant of talk-aloud protocol.

Bibliography

Anderson, J. R. (1993). Rules of the Mind. Hillsdale, New Jersey, Lawrence Erlbaum Associates.

Bara, B. G. (1995). Cognitive Science: A Developmental Approach to the Simulation of the Mind. Hillsdale, New Jersey, Lawrence Erlbaum Associates.

Bialek, W. and A. Zee (1988). "Understanding the Efficiency of Human Perception." Physical Review Letters **61**(13): 1512-1515.

Britton, B. K. and P. Tidwell (1995). Cognitive structure testing: A computer system for diagnosis of expert-novice differences. Cognitively Diagnostic Assessment. P. D. Nichols, S. F. Chipman and R. L. Brennan. Hillsdale, New Jersey, Lawrence Erlbaum Associates: 251-278.

Chi, M. T. H., P. J. Feltovich, et al. (1981). "Categorization and representation of physics problems by experts and novices." Cognitive Science **5**: 121-152.

Chipman, S. F., P. D. Nichols, et al. (1995). Introduction. Cognitively Diagnostic Assessment. P. D. Nichols, S. F. Chipman and R. L. Brennan. Hillsdale, New Jersey, Lawrence Erlbaum Associates: 1-18.

Cliburn, J. W., Jr. (1990). "Concept maps to promote meaningful learning." Journal of College Science Teaching(February): 212-217.

Cooke, N. M., F. T. Durso, et al. (1986). "Recall and measures of memory organization." Journal of Experimental Psychology: Learning, Memory, and Cognition **12**(4): 538-549.

Corbett, A. T., J. R. Anderson, et al. (1995). Student modeling in the ACT programming tutor. Cognitively Diagnostic Assessment. P. D. Nichols, S. F. Chipman and R. L. Brennan. Hillsdale, New Jersey, Lawrence Erlbaum Associates: 19-41.

Corter, J. E. (1995). Using clustering methods to explore the structure of diagnostic tests. Cognitively Diagnostic Assessment. P. D. Nichols, S. F. Chipman and R. L. Brennan. Hillsdale, New Jersey, Lawrence Erlbaum Associates: 305-326.

Donahoe, J. W., J. E. Burgos, et al. (1993). "A selectionist approach to reinforcement." Journal of the Experimental Analysis of Behavior **60**(1): 17-40.

Gerace, W. J. (1992). Contributions from cognitive research to mathematics and science education. Workshop on Research in Science and Mathematics Education, Cathedral Peak, South Africa, University of Natal, Pietermaritzburg, South Africa.

Gerace, W. J., W. J. Leonard, et al. (1997). Concept-based problem solving: Combining educational research results and practical experience to create a framework for learning physics and to derive effective classroom practices. Amherst, University of Massachusetts Physics Education Research Group**:** 26.

Goldsmith, T. E., P. J. Johnson, et al. (1991). "Assessing structural knowledge." Journal of Educational Psychology **83**(1): 88-96.

Gonzalvo, P., J. J. Cañas, et al. (1994). "Structural representations in knowledge acquisition." Journal of Educational Psychology **86**(4): 601-616.

Hardiman, P. H., R. J. Dufresne, et al. (1989). "The relation between problem categorizatin and problem solving among experts and novices." Memory and Cognition **17**(5): 627-638.

Hertz, J., A. Krogh, et al. (1991). Introduction to the Theory of Neural Computation. Redwood City, California, Addison-Wesley.

Holland, J. H. (1990). "Concerning the emergence of tag-mediated lookahead in classifier systems." Physica D **42**: 188-201.

Holland, J. H., K. J. Holyoak, et al. (1986). Induction: Processes of inference, learning, and discovery. Cambridge, Massachusetts, The MIT Press.

Hopfield, J. H. (1982). "Neural networks and physical systems with emergent collective computational abilities." Proceedings of the National Academy of Sciences USA **79**(April): 2554-2558.

Hopfield, J. J. (1984). "Neurons with graded response have collective computational properties like those of two-state neurons." Proceedings of the National Academy of Sciences USA **81**(May): 3088-3092.

Johnson, P. E. (1964). "Associative meaning of concepts in physics." Journal of Educational Psychology **55**: 84-88.

Johnson, P. J., T. E. Goldsmith, et al. (1995). Similarity, structure, and knowledge: A representational approach to assessment. Cognitively Diagnostic Assessment. P. D. Nichols, S. F. Chipman and R. L. Brennan. Hillsdale, New Jersey, Lawrence Erlbaum Associates**:** 221-250.

Ju, T.-P. (1989). The development of a microcomputer-assisted measurement tool to display a person's knowledge structure. Pittsburgh, University of Pittsburgh**:** 219.

Kohonen, T. (1993). “Physiological interpretation of the self-organizing map algorithm.” Neural Networks **6**: 895-905.

Konold, C. E. and J. A. Bates (1982). “The episodic/semantic memory distinction as an heuristic in the study of instructional effects on cognitive structure.” Contemporary Educational Psychology **7**: 124-138.

Larkin, J. H. (1979). Information processing models and science instruction. Cognitive Process Instruction. J. Lochhead and J. Clement. Philadelphia, Pennsylvania, The Franklin Institute Press**:** 109-118.

Lisman, J. E. and M. A. P. Idiart (1995). “Storage of 7±2 short-term memories in oscillatory subcycles.” Science **267**(10 March): 1512-1515.

Martin, J. and K. VanLehn (1995). A Bayesian approach to cognitive analysis. Cognitively Diagnostic Assessment. P. D. Nichols, S. F. Chipman and R. L. Brennan. Hillsdale, New Jersey, Lawrence Erlbaum Associates**:** 141-165.

Mashhadi, A. and B. Woolnough (1996). Cognitive mapping of advanced level physics students' conceptions of quantum mechanics. Conference on Educational Research, Singapore.

McClelland, J. L. and D. E. Rumelhart (1986). Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Cambridge, MA, MIT Press.

McDermott, L. and E. F. Redish (1999). “RL-PER1: Resource letter on physics education research.” American Journal of Physics **in press**.

Mestre, J. and J. Touger (1989). “Cognitive research — what's in it for physics teachers?” The Physics Teacher **27**(6): 447-456.

Mestre, J. P. (1991). “Learning and instruction in pre-college physical science.” Physics Today **44**(9): 56-62.

Mestre, J. P., R. J. Dufresne, et al. (1993). “Promoting skilled problem-solving behavior among beginning physics students.” Journal of Research in Science Teaching **30**(3): 303-317.

Mislevy, R. J. (1993). Foundations of a new test theory. Test Theory for a New Generation of Tests. N. Frederiksen, R. J. Mislevy and I. I. Bejar. Hillsdale, New Jersy, Lawrence Erlbaum Associates**:** 19-39.

Mislevy, R. J. (1995). Probability-based inference in cognitive diagnosis. Cognitively Diagnostic Assessment. P. D. Nichols, S. F. Chipman and R. L. Brennan. Hillsdale, New Jersey, Lawrence Erlbaum Associates**:** 43-71.

Miyamoto, S., S. Suga, et al. (1990). “Methods of digraph representation and cluster analysis for analyzing free association.” IEEE Transactions on Systems,Man, and Cybernetics **20**(3): 695-701.

Moore, D. S. and G. P. McCabe (1989). Introduction to the Practice of Statistics. New York, W. H. Freeman and Co.

Nakhleh, M. B. (1994). “Chemical education research in the laboratory environment: How can research uncover what students are learning?” Journal of Chemical Education **71**(3): 201-205.

Novak, J. D. and D. B. Gowin (1984). Learning How to Learn. Englewood Cliffs, New Jersey, Prentice-Hall.

Redish, E. F. (1994). “Implications of cognitive studies for teaching physics.” American Journal of Physics **62**(9): 796-803.

Regis, A., P. G. Albertazzi, et al. (1996). “Concept maps in chemistry education.” Journal of Chemical Education **73**(11): 1084-1088.

Rice, D. C., J. M. Ryan, et al. (1998). “Using concept maps to assess student learning in the science classroom: Must different methods compete?” Journal of Research in Science Teaching **35**(10): 1103-1127.

Ruiz-Primo, M. A. and R. J. Shavelson (1996). “Problems and issues in the use of concept maps in science assessment.” Journal of Research in Science Teaching **33**(6): 569-600.

Tatsuoka, K. K. (1995). Architecture of knowledge structures and cognitive diagnosis: A statistical pattern recognition and classification approach. Cognitively Diagnostic Assessment. P. D. Nichols, S. F. Chipman and R. L. Brennan. Hillsdale, New Jersey, Lawrence Erlbaum Associates**:** 327-359.

ter Braak, C. J. F. (1995). Ordination. Data Analysis in Community and Landscape Ecology. R. H. G. Jongman, C. J. F. Ter Braak and O. F. R. Van Tongeren. Cambridge, Cambridge University Press.

Trehub, A. (1991). The Cognitive Brain. Cambridge, MA, MIT Press.

Trowbridge, J. E. and J. H. Wandersee (1996). “How do graphics presented during college biology lessons affect students' learning?” Journal of College Science Teaching **1996**(September/October): 54-57.

Watkin, T. L. H. and A. Rau (1993). "The statistical mechanics of learning a rule." Reviews of Modern Physics **65**(2 (April)): 499-556.

Young, M. J. (1993). Quantitative measures for the assessment of declarative knowledge structure characteristics. Pittsburgh, University of Pittsburgh**:** 294.

Zajchowski, R. and J. Martin (1993). "Differences in the problem solving of stronger and weaker novices in physics: Knowledge, strategies, or knowledge structure?" Journal of Research in Science Teaching **30**(5): 459-470.