Technical Report 472

The Structure of Mathematical Knowledge

Edwina Rissland Michener

MIT Artificial Intelligence Laboratory

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY

A.I. TECHNICAL REPORT NO. 472

August 1978

THE STRUCTURE OF MATHEMATICAL KNOWLEDGE

by

EDWINA RISSLAND MICHENER

This report describes research done at the Artificial Intelligence Laboratory of the Massachusetts Institute of Technology. Support for the laboratory's artificial intelligence research is provided in part by National Science Foundation Grant 77-19083 SED.

THE STRUCTURE OF MATHEMATICAL KNOWLEDGE

by

EDWINA RISSLAND MICHENER

ABSTRACT

This report develops a conceptual framework in which to talk about mathematical knowledge. There are several broad categories of mathematical knowledge: *results* which contain the traditional logical aspects of mathematics; *examples* which contain illustrative material; and *concepts* which include formal and informal ideas, that is, definitions and heuristics.

Just as results can be structured by the relation of *logical support* in which A ---> B means that result A is used to prove result B, examples and concepts can also be organized by relations. Examples can be ordered by the relation of *constructional derivation* in which A ---> B means that example A is used in the construction of example B. Concepts can be structured by the judgement that concept A should be introduced or examined before concept B, which defines a *pedagogical ordering.*

The three item/relation pairs -- *results/logical support, examples/constructional derivation,* and *concepts/pedagogical ordering* -- establish three representation spaces for a mathematical theory: *Results-space, Examples-space,* and *Concepts-space.* They are best shown as directed graphs, *representation graphs,* where the direction matches the predecessor-successor ordering inherent in the relations.

When we consider a theory item, we first decide whether we want to classify it as a result, example or concept and then we fit it into its representation space by determining its predecessors and successors. In addition we can also consider items outside of its representation space to which it is related. *Dual relations* concern these inter-space associations. The *epistemological dual* of a result consists of examples motivating it, concepts needed to state and prove it, and concepts and results derived from it. The dual items of an example are ingredient concepts and results needed to discuss or construct it, and concepts and results motivated by it. The dual items for a concept are examples motivating it and results laying the groundwork for it, and examples illustrating it and results proving things about it.

While the placement of an item within its own representation graph determines one definition of closeness, consideration of its dual items leads to additional definitions. For instance, two results are related or close in the *example dual sense,* if they share common examples. The power of the dual idea is that it provides a way to describe the intuitive notion of what it means for two items to be related or close in one's understanding of a theory.

Not all examples, results and concepts serve the same function in one's understanding of a theory. We single out those that play special roles and group them into *epistemological classes*.

In the class of examples for instance, there are perspicuous *start-up* examples which we can grasp immediately; *reference examples* which we use repeatedly; *model examples* which are paradigm situations that suggest to us the essence of a result or concept; *counter-examples* which limit the use and validity of other items.

In addition to definitions, *Concepts-space* contains the heuristic advice that we give to ourselves (and to others) while working in a theory. *Mega-principles* are postive imperatives and interpretations. *Counter-principles* offer warnings.

Results-space also has many subclasses of items: *basic results* establish initial basic facts in a theory; *key* results are frequently used results; *culminating* results are goal results towards which the theory drives; *technical* results establish technical details; *transitional* results provide logical stpping-stones.

The three main types of items - examples, results, and concepts -- have enough in common so that they can be represented by essentially the same framework which can then be fine-tuned to reflect their special features. The resulting representation and its interconnections provide a rich representation for mathematical theories which allows us to build data bases of mathematical knowledge and to discuss many of the ingredients and processes involved in understanding mathematics.

We illustrate these ideas by mapping out some of the knowledge in three important domains of mathematics from the undergraduate curriculum: calculus (specifically, the Mean Value Theorem), linear algebra (matrices and eigenvalues), and real analysis (convergence, compactness and open sets).

In the last chapter we analyze the understanding of mathematics in terms of our conceptual framework. We present some questions that probe understanding and which can be used as a heuristic for how to understand a theory or item. We report on applications of these ideas to teaching.

Acknowledgements

I thank my thesis committee of Seymour Papert, Marvin Minsky and Ted Martin for their continued support and steadfast interest.

I happily acknowledge Seymour Papert and Oliver Selfridge who have contributed substantially to this revision and extension of my doctoral thesis. Their conscientious readings of drafts of this document helped clarify my thinking and writing.

I thank Ken Hoffman for sharing his wisdom of real analysis and how it is taught, and also, for discussions of his textbook and how it relates to Chapter 6 of this document.

Of course, without the generous support of the M.I.T. Mathematics Department and the Division for Study and Research in Education, this report would never have been written.

Again to all of my friends and fellow mathematicians acknowledged in my dissertation, I once again offer my sincere appreciation.

August 1978 *Edwina Rissland Michener*

Foreword to the Reader

This document is written so that the mathematics is as modularized and distinct from the rest of the text as possible. Many of the examples are taken from elementrary number theory, calculus, linear algebra and real analysis; most of these are taken from several widely used undergraduate textbooks: [Halmos; Hoffman; Ireland and Rosen; Rudin; Strang; Thomas]. Whenever possible extended examples from mathematics are set off from the main body of the text; thus if an example is not understandable or appealing, it can be skipped without too much effect on the presentation. However, this is a monograph on the structure of mathematics, and one cannot talk about mathematics without considering some examples from mathematics.

Chapter 1. INTRODUCTION

When a mathematician says he understands a mathematical theory, he really knows much more than the deductive details of definitions, axioms, and theorems and their proofs. This paper is concerned with the other, often extra-logical, knowledge that is critical to his understanding. The goal is to understand the understanding of mathematics, in order to improve how one learns, teaches, and does mathematics.

One fundamental aim is to develop a conceptual framework in which to talk precisely about the knowledge actually involved in the understanding of mathematics. This problem is largely epistemological, but it is a prerequisite to trying to mechanize or support that understanding.

The research presented here was first reported in Part I of Michener's doctoral dissertation [1977]. Parts II and III of that work described a computer based interactive environment to aid expert mathematicians, and an auxiliary one to aid neophytes. Although we shall not describe those systems in detail here, we shall take advantage of some of its terms and its metaphors, which are helpful in making certain vague notions more precise.

Since we are concerned with the understanding of mathematics, a natural question is "What is a mathematical theory?" In the narrowest sense it is just a collection of definitions, axioms, theorems and proofs. But those are merely its deductive aspects. A mathematician uses other resources: the stock of examples he finds useful, and their organization; certain rules of thumb or heuristics, telling which are good ideas to try and warning him of trouble; his rules of inference. He also has images of how all his knowledge hangs together. In short, he knows and uses a great deal more than purely logical deductive knowledge and this is the sense in which we think of a mathematical theory.

To understand a body of mathematics, one must be able to travel freely through it, experiment with its elements, examine its connections, survey its mathematical topography, and follow threads of associations. One must deal with examples, theorems and heuristics; perturb contexts and hypotheses; and shift the levels of concern from detail to overview, and vice versa, with facility and spontaneity.

Thus understanding is a very active process. It is as if what is to be understood is a multi-faceted prism that must be held in the hand, rotated, viewed from many perspectives, and sliced through along many different planes. Polya and Szego [1972] describe it:

> One should try to understand everything: isolated facts by collating them with related facts, the newly discovered through its connection with the already assimilated, the unfamiliar by analogy with the accustomed, special results through generalization, general results by means of suitable specialization, complex situations by dissecting them into their constituent parts, and details

by comprehending them within a total picture.

> There is a similarity between knowing one's way about a town and mastering a field of knowledge; from any given point one should be able to reach any other point. One is even better informed if one can immediately take the most convenient and quickest path from the one point to the other. If one is very well informed indeed, one can even execute special feats, for example, to carry out a journey by systematically avoiding certain paths which are customary...

> There is an anology between the task of constructing a well-integrated body of knowledge from acquaintance with isolated truths and the building of a wall out of unhewn stones. One must turn each new insight and each new stone over and over, view it from all sides, attempt to join it on to the edifice at all possible points, until the new finds its suitable place in the already established, in such a way that the areas of contact will be as large as possible and the gaps as small as possible, until the whole forms one firm structure.

Understanding has several aspects. One is the ability to solve problems; this has been investigated extensively by Polya and others and will not be discussed in this monograph. Another is the process of building up and enriching a knowledge base with all of its elements and associations; that is the aspect which concerns us here.

In Chapter 7, we present several questions. Being able to answer them should be regarded as symptoms of understanding. The computer systems mentioned before were designed to support dealing with them, and to help establish the modes of thought that might contribute to the users' needs in understanding. Indeed, the problems of such support are themselves illuminating to the general conceptual questions.

We have several motivations for this work: first, there is an intellectual curiosity that tries to understand better what we know and do; second, it seems obvious that a successful attack here can be useful in education. Students, teachers, and mathematicians generally, ought to be aware of the ingredients of their understanding. In particular, computer assisted instruction (CAI) is not likely to have a broad impact on our educational system unless we understand better what we know and how we know it. Finally, we have a long range aim of the mechanization of mathematics itself.

There are several steps. One is epistemological: to identify the key elements of mathematical knowledge and examine their interrelationships. Another is to represent these elements in a coherent way, which captures the major features of their content and function. Only after those steps can we then start to mechanize and experiment.

As mathematicians, we analyze the structure and epistemology of mathematics, and use our analysis to help us know how we undertake and understand mathematics. The insights we may gain can provide new approaches which may be especially useful in further progress. Understanding our own understanding processes can enable us to design a computer system custom-tailored to our mathematical research efforts.

From a student's point of view, knowing how to understand mathematics helps him to understand and assimilate mathematics, and to do it spontaneously. It also develops that much talked about but hard to define quality of mathematical maturity. It helps him to become a good question asker, to see the forest for the trees, and to learn how to organize mathematical knowledge in a coherent way.

From the computer scientist's point of view, the domain of mathematics is a key area to explore and leads to questions of representation and understanding. We need to capture the knowledge of expert understanders (i.e., mathematicians), so that it can be used by other researchers in fields like automatic theorem proving and analogy programs.

We are going to talk about many things that will be familiar to mathematicians, but which they rarely discuss. We shall try to make explicit some of the many intuitive notions that every good mathematician has. Such knowledge is usually unconsciously natural, and mathematicians, like everyone else, tend to under-estimate the amount they have.

Once such knowledge is remarked upon, it may seem completely obvious. The attitude that "anything I know must be trivial" is not only silly but also detrimental. There is a great need to disambiguate and clarify this knowledge for as Hadamard says [1954], "how else can we then see what the consequences of our knowledge are". Explication is also a prerequisite for mechanization, and it is critical to the improvement of teaching, learning, and doing of mathematics.

In doing this work, Hadamard, Lakatos, and Poincare were valuable references; but the most so was Polya. While much of Polya's work is concerned with problem solving and the teaching of problem solving -- like, for instance, *How to Solve It* and *Induction and Analogy in Mathematics* -- rather than understanding understanding and teaching skills of understanding, it is complementary to what we are trying to do here. Doing and understanding are the two sides of the coin. Thus we gratefully acknowledge our debt to George Polya for the spirit and content of his work.

may gain can provide new approaches which may be especially useful in further progress. Understanding our own understanding processes can enable us to design a computer system custom-tailored to our mathematical research efforts.

From a student's point of view, knowing how to understand mathematics helps him to understand and assimilate mathematics, and to do it spontaneously. It also develops that much talked about but hard to define quality of mathematical maturity. It helps him to become a good question asker, to see the forest for the trees, and to learn how to organize mathematical knowledge in a coherent way.

From the computer scientist's point of view, the domain of mathematics is a key area to explore and leads to questions of representation and understanding. We need to capture the knowledge of expert understanders (i.e., mathematicians), so that it can be used by other researchers in fields like automatic theorem proving and analogy programs.

We are going to talk about many things that will be familiar to mathematicians, but which they rarely discuss. We shall try to make explicit some of the many intuitive notions that every good mathematician has. Such knowledge is usually unconsciously natural, and mathematicians, like everyone else, tend to under-estimate the amount they have.

Once such knowledge is remarked upon, it may seem completely obvious. The attitude that "anything I know must be trivial" is not only silly but also detrimental. There is a great need to disambiguate and clarify this knowledge for as Hadamard says [1954], "how else can we then see what the consequences of our knowledge are". Explication is also a prerequisite for mechanization, and it is critical to the improvement of teaching, learning, and doing of mathematics.

In doing this work, Hadamard, Lakatos, and Poincare were valuable references; but the most so was Polya. While much of Polya's work is concerned with problem solving and the teaching of problem solving -- like, for instance, *How to Solve It* and *Induction and Analogy in Mathematics* -- rather than understanding and teaching skills of understanding, it is complementary to what we are trying to do here. Doing and understanding are the two sides of the coin. Thus we gratefully acknowledge our debt to George Polya for the spirit and content of his work.

Chapter 2. OVERVIEW OF THE EPISTEMOLOGY

2.1 Three Representation Spaces

2.1.1 Examples, Results, Concepts

When one analyzes the mathematical knowledge used by students and professionals to do and explain mathematics, it becomes clear that there are several kinds of mathematical knowledge: (1) clusters of strongly bound pieces of information, such as the statement of a theorem, its name, its proof[s], a diagram, an evaluation of its importance, which can be taken together to comprise a single *item*, such as a theorem; and (2) relations between the items, such as the logical connections between theorems. One can distinguish at least three major categories of items: *results*, which contain the traditional logical-deductive elements of mathematics, i.e., theorems; *examples*, which contain illustrative material; and *concepts*, which contain mathematical definitions and pieces of heuristic advice.

Results can be naturally organized according to their logical connections. For results, the relation of *deduction* or *logical support* written as A--> B means that result A is needed or used to prove result B.[1] Since we are as interested in the relation as the results themselves, results together with the relation of logical support are said to compose *Results-space*. For instance in the theory of unique factorization, in order to prove that every integer has a unique factorization, one must first prove supporting results on the Euclidean algorithm and the greatest common divisior [Ireland and Rosen, Chapter 1]. (See Figure 2.)

Examples and concepts can each also be organized by relations. Examples can be ordered by the relation of *constructional derivation* in which A --> B means that example A is needed in a construction of example B.

For instance, the unit interval is used in the construction of the Cantor set, which in turn is used in the construction of the Cantor function [Hoffman] (see Figure 1). The relation of constructional derivation often reflects the development of increasing complication between an example and its derivates.

---

[1] A distinction can be made between needed to prove and used to prove: the first represents some sort of logical necessity, whereas the second, just says that the proof of B can be done this way. We shall allow the latter interpretation. Thus, needed here means result A enters into in the proof of B (as found in the presentation or knowledge being mapped out).

Figure 1

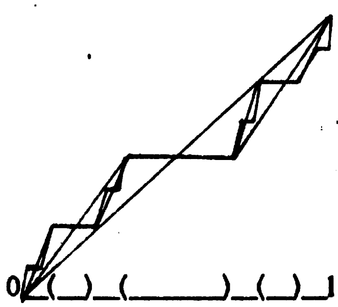
0______________________1 *the unit interval*

define sequence of sets by deleting middle thirds

0_______( )_______1

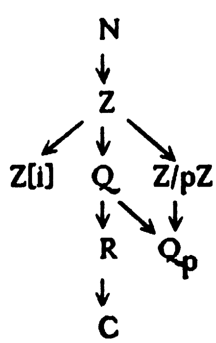
0_( )_( )_( )_1 *limiting set is the Cantor set, P*

define sequence of piecewise linear functions, flat on P^c



limit is the Cantor function

Another familiar set of examples which can be structured according to their constructional derivation starts with the natural numbers N. These beget the integers Z (by closure with respect to subtraction), which beget the rationals Q (by forming quotients), which beget the real numbers R (by completion of Cauchy sequences), which beget the complex numbers C (by algebraic closure). Many more examples, such as the Gaussian integers Z[i], the field of integers modulo a prime Z/pZ, and the p-adic numbers Q_p, can also be organized according to their constructional relations:



The p-adic numbers have arrows coming from both Q and Z/pZ since either can be used to construct Q_p [2] The above diagram could show the examples constructed as intermediate steps between Z/pZ and Q_p (e.g., Z/p^2Z, Z/p^3Z); however, the point is that there are two constructional routes leading to Q_p (and thus two representations available: Q_p from lim Z/pZ and as completion of $(Q, \|.\|_p)$. Thus a directed graph, and not merely a tree, is needed to show the relations.

In this way, the relation of constructional derivation has allowed the collection of examples to be coherently organized. Examples together with the relation of constructional derivation make up *Examples-space*, which can be pictured as a directed graph, the ***examples-graph***.

Concepts include formal and informal ideas, that is, definitions and heuristics. Concepts can be structured by the pedagogical judgement that concept A should be introduced before concept B; this relation is called *pedagogical ordering*. Sometimes it simply reflects the fact that concept A enters into the definition of concept B and at other times, expository tastes.

In this way, the three item/relation pairs -- examples/constructional derivation, results/logical support, and concepts/pedagogoical ordering -- establish three representation spaces: *Examples-space*, *Results-space* and *Concepts-space*. They are best shown as directed graphs where the nodes represent the items and the arrows, the relations between them with the direction matching the predecessor-successor ordering inherent in the relations [3].

Directed graphs are used not only because they can show multiple routes, but also because they give equal emphasis to nodes and arrows, much as category theory treats objects and morphisms and in this epistemology relations are as important as items. The relations represent much of the evolutionary aspects of the knowledge; to de-emphasize them would be to forget some of the most important aspects of "descriptive" and "genetic" epistemology [Piaget], [van Steenberghen].

---

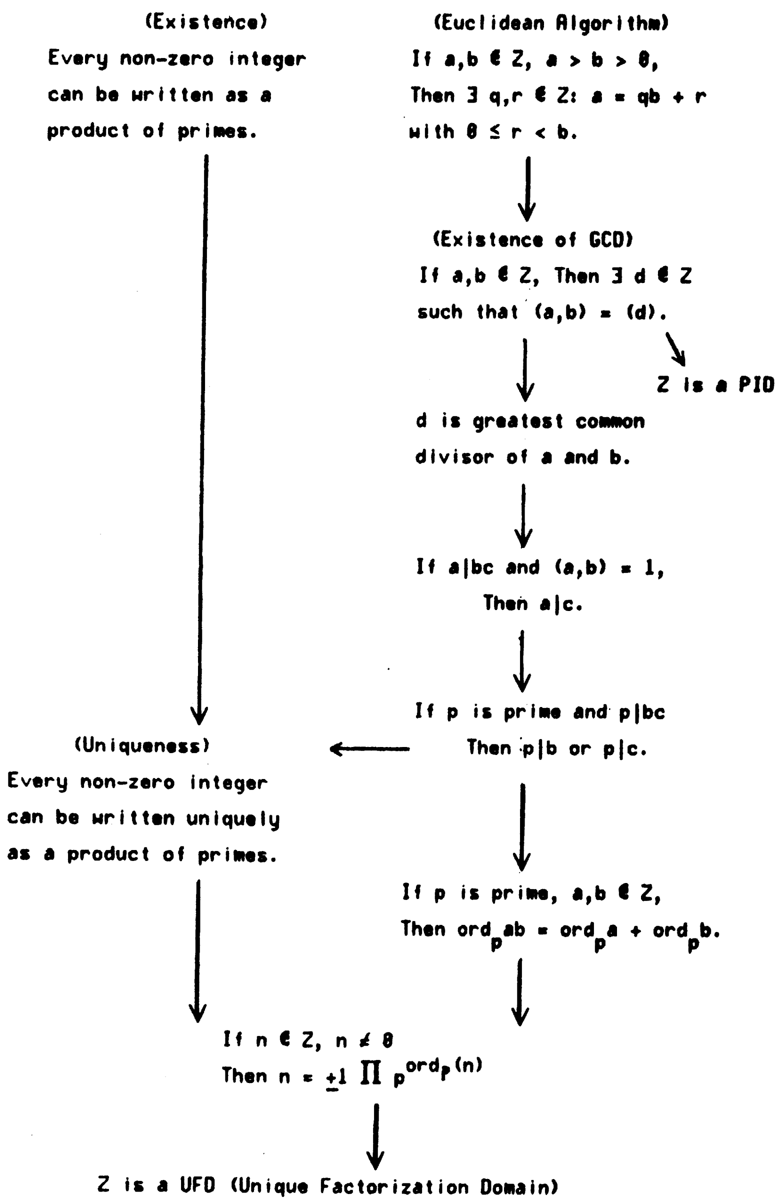
[2] Q_p can be constructed from Q by "completing" Q with respect to the p-adic metric in a fashion completely analogous to the construction of the reals R from Q with respect to the absolute value [Boravitch and Shaferavitch, Chapter I, Section 3]. Q_p can also be constructed from Z/pZ by taking the direct limit of the rings $Z/pZ \subset Z/p^2Z \subset \ldots \subset Z/p^kZ \subset \ldots$ and then forming the field of fractions [Eilenberg and Steenrod].

[3] Graph nodes can be further described by their positions in the graph: "starting nodes" are at the top with no predecessor nodes, and "end nodes" at the bottom with no successor nodes. These two types of nodes are worth noting because starting nodes are places where one can usually start reading into or building a theory, and end nodes represent culminating items or the "state-of-the-art".

2.1.2 Examples of Representation Graphs

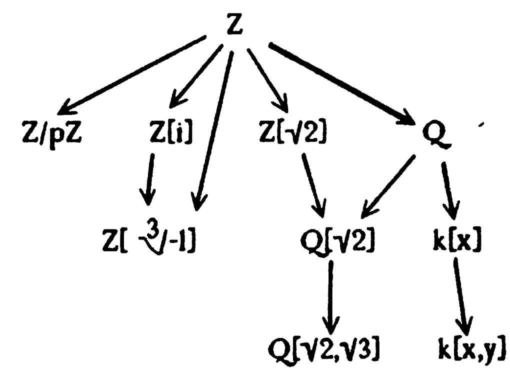
To illustrate this representation scheme, the following are the three representation graphs for the elementary theory of unique factorization as presented by [Ireland and Rosen].

Figure 2



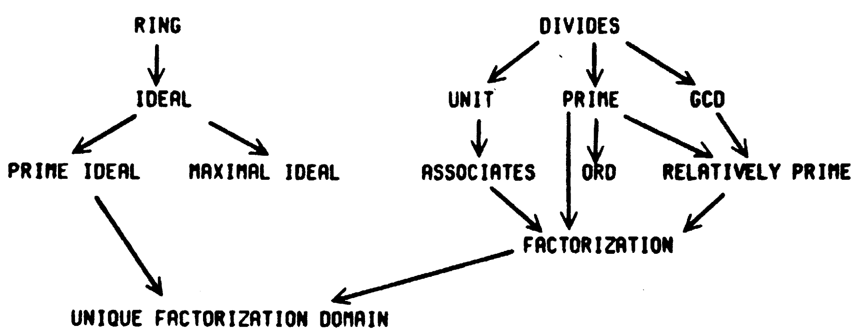
Results-space represents the logical aspects of the theory: A -> B means A is used to prove B.

Figure 3



This examples-graph contains examples for the theory of unique factorization and organizes them according to the relation: A -> B if A is used to construct B.

Figure 4



In concepts-space, A -> B indicates that A is used in the formulation of B or that one should know about A before learning about B.

2.1.3 Relation to Other Representations

Although mathematicians have not traditionally been concerned with representing their knowledge, particularly that outside of the formal logic, some books do provide a skeletal classification of knowledge. In mathematical texts, some authors consciously attempt to classify the things about which they write. Some restrict their categories to definitions and theorems; others include examples.

For instance, in his classic book on real analysis, Rudin [Rudin 1964] uses four subject headings -- definitions, theorems, examples and discussions. He thus uses the three categories of this epistemology, but he makes no further distinctions between the subcategories. More importantly, while he does structure his mathematics, he does not exploit this structure to teach. He certainly does not attempt to make the *student aware* of how structured knowledge can help him learn mathematics or even of how mathematics can be systematically structured. Nevertheless, his text shows that mathematicians, consciously or unconsciously, use some of the epistemological ideas discussed here.

Other authors organize definitions and topics by pedagogical ordering to indicate the conceptual dependencies in their presentations. Dunford and Schwartz display their concepts-graph in the first volume of their encyclopedic work on functional analysis [Dunford and Schwartz]. The conceptual organization of the British Open University series is used as a cover illustration for some of their instructional modules.

Directed graphs have often been used effectively to represent different aspects of mathematics. The graph representation has been used to represent the interdependence of ideas and sections of their books [Royden; Dunford and Schwartz]. Some have used these representations of Concepts-space to show the main routes, *pedagogical trails*, through their texts [Hartley and Hawkes]. Other researchers have used directed graphs to represent the formal logic of proofs [Chester 1976]. However, no one -- to this author's knowledge -- has used more than one network simultaneously to structure and represent a mathematical theory as a coherent whole. The tripartite representation scheme of this report allows different cognitive strands to be isolated, examined and played off against one another. By analysis and synthesis of mathematical knowledge in this way, we are able to explore the multi-layered fabric of mathematical understanding.

Only one recent A.I. program addresses the representation problem for mathematics [Lenat 1976]. However, it essentially puts everything into one large semantic network and does not represent many of the other relations, like the constructional connections between the examples themselves. While it does use a relation for is-a and is-an-example of, it does not consider the relations between examples; it simply hangs an example off of a concept, in what would be called a post-concepts-dual relation in this paper. Also, Lenat uses examples in the narrow sense of instantiation whereas examples are used in this work in the broad sense of any item that illustrates or motivates another; we also include in the category of examples counter-examples and limiting examples that show what another item isn't. (See Chapter 3.)

Although more needs to be learned about representation schemes from the psychological point of view, it does seem that a tripartite classification of knowledge is supported by some researchers in cognitive psychology [Bruner 1971]:

> "Human beings have three different systems, partially translatable one into another, for representing reality. One is through action....A second way of knowing is through imagery and those products of mind that, in effect, stop the action and summarize it in a representing ikon.... Finally, there is the representation by symbol."

It seems reasonable to match the symbolic elements of this description with result items and the ikonic with examples. Action elements correspond to the heuristic imperatives of Concepts-space and with the procedural formulation that is attached to each result, example and concept.

2.2 The Dual Idea

In this section, we shall try to capture the intuitive notion of what it means for two items to be related or close in one's understanding of them, to propose relations and structures to model one's ability to freely associate one mathematical item with another that is not immediately connected to it through the intra-space relations of the representation spaces, and to explain how items distant in the sense of intra-space space relations can be so easily chained together.

2.2.1 Dual Items

A theory item is related to other items in its representation space through the space's predecessor-successor relation. In addition, an item is related to items outside of its representation space. The *dual idea* concerns these *inter*-space relations.

Specifically, *dual items* are defined as follows:

> *The dual items of an example are: (a) the ingredient concepts and results needed to discuss or construct it; and (b) the concepts and results motivated by it.*

> *The dual items of a result consist of: (a) the examples motivating it and the concepts needed to state and prove it; and (b) the concepts and examples that are derived from it.*

> *The dual items of a concept are: (a) the examples motivating it and the results laying the groundwork for it; and (b) the examples illustrating it and the results proving things about it.*

Thus each item has two subsets of dual items:

dual (an example) = *{results}*, *{concepts}*
dual (a result) = *{examples}*, *{concepts}*
dual (a concept) = *{examples}*, *{results}*

The subset of examples in the dual set of an item is called the *examples-dual*, the subset of results, the *results-dual*, and the subset of concepts, the *concepts-dual*.

The item and the two associated subsets of its dual make up a structural *triad* of items. Such a triad represents a closely interconnected cluster of items in our representation scheme and our understanding.

Each of the associated subsets of the item's dual can be further sub-categorized into those items that precede the item in the understanding or development of a theory, which are called the *pre-dual* items, those that come after the item, the *post-dual* items, and those that have neither a strong "pre" nor "post" flavor, the *neutral-dual*. The pre-dual items motivate the item, or in Polya's words, are "suggestive", and the post-dual items bear witness for the item or are "supportive" [Polya, I and A, p.7]. In the above definition of dual items, the items listed under (a) are usually in the pre-dual, and those of (b), in the post-dual.

2.2.2 Relation via the Dual Idea

Two items are said to be *related via the dual idea* if they share common dual items. For instance, the examples of the real and p-adic numbers are related via the shared concept of completion.

The mathematical world is full of relations via the dual idea. The following are examples of this sort of relation:

(1) examples that share a dual concept:

R (real numbers) and Q_p (p-adic numbers) via completion;
Q (rationals) and P (Cantor set) via measure zero;
the matrix $\begin{pmatrix} 1 & a \\ 0 & 1 \end{pmatrix}$ and exp(a) via group characters;
Tschebyscheff, trigonometirc, Hermite polynomials
and almost periodic functions via complete orthonormal sets;

Fibonacci numbers and perfect squares via finite differences;

(2) results that share a dual concept:

spectrum of idempotent, nilpotent and permutation matrices via the power idea;

(3) concepts that share a dual example:

stability, roots of unity, power idea via the equation $x^n = 0$
measure and length via the ordinary generic open interval (a,b);
countability and measure zero via the Cantor set;
fixed point and power idea via $\cos^n x$;
continuity and differentiability via absolute value function;
bounded variation and absolute continuity via Cantor function;

(4) results that share a dual example:

Parseval's Identity and Pythagorean Theorem via R^2;
Jordan Normal Form and Cayley-Hamilton Theorems via diagonal matrices;

(5) concepts that share a dual result:

symmetric and diagonalizable via "Symmetric matrices are diagonalizable.";
equicontinuity and compactness via Ascoli's theorem;
ascending chain condition, existence of maximal element, and finitely generated via Noetherian characterization result;
quadratic extensions and straight-edge and compass constructions via Galois theorem;

(6) examples that share a dual result:

90^o rotation and "counter identity" matrices via the result "$A^n = I \Rightarrow \lambda(A)^n = 1$.";
$L_2(R)$ and $l_2(R)$ via the Riesz-Fischer Theorem;

Some of these dual relations contain relations by analogy, such as the analogous construction of Q_p and R by completion of Q. This could be written as:

$$Q_p{:}Q \quad as_{<completion>} \quad R{:}Q.$$

However, the dual relation is broader, because two items can be related whether or not they share the same sense of analogy. For instance, the concepts of bounded variation and absolute continuity are related via the example of the Cantor function; the first because the Cantor function is an example of a function of bounded variation; the second, because the Cantor function is an example of a function not absolutely continuous. While these two concepts are related via the dual idea, it would be difficult to describe this particular relation as an analogy. One doesn't think of analogy as existing between two items because one has a quality and the other doesn't:

$$BV : \text{Cantor fcn} \quad as_{<instance>} \quad \text{not } AC : \text{Cantor fcn.}$$

Thus, while one does not usually speak of these two concepts as being analogous, one can easily speak of them as being related via the dual idea.

2.2.3 Notation

To facilitate discussion of these ideas, some notation is in order. We introduce this notation not because we are going to prove theorems using these formalisms, but rather because they help to abstract some of the mathematical and information processing ideas that underlie the dual idea. They help us to think about these ideas by pointing out the functions and objects involved in our analysis.

The dual of an item I is denoted by $D(\mathrm{I})$. The examples-dual of an item I (when I is a result or concept) is denoted by $E(\mathrm{I})$; the concepts-dual of I (a result or example), by $C(\mathrm{I})$; and the results-dual of I (a concept or example), by $R(\mathrm{I})$. Thus for instance, the dual of an example E is $D(\mathrm{E}) = R(\mathrm{E}) \cup C(\mathrm{E})$. The capital letters I, C, E, and R will always be used generically for the words *item, concept, example,* and *result.* The italized letters *D, C, E* and *R* will always be used for the *function of taking* the dual, examples-dual, concepts-dual or results-dual of an item.

Also, to distinguish between the pre-, post- and neutral-duals, the symbols $<$, $>$, $=$ are used respectively. Thus,

$$R(\mathrm{I}) = (<R)(\mathrm{I}) \cup (>R)(\mathrm{I}) \cup (=R)(\mathrm{I})$$
$$C(\mathrm{I}) = (<C)(\mathrm{I}) \cup (>C)(\mathrm{I}) \cup (=C)(\mathrm{I})$$
$$E(\mathrm{I}) = (<E)(\mathrm{I}) \cup (>E)(\mathrm{I}) \cup (=E)(\mathrm{I})$$

and by extending the definitions so that $<$, $>$, or $=$ of a union is the union of the $<$, $>$, $=$, we also have:

$$D(\mathrm{I}) = (<D)(\mathrm{I}) \cup (>D)(\mathrm{I}) \cup (=D)(\mathrm{I})$$

A dual item is classified into the pre- or post-dual only if it clearly belongs before or after the item in the development or understanding of the theory. Dual items which could either come before or after the item because they are so intimately bound to it that their order is somewhat arbitrary, or because a decision about membership in the pre- or post-dual has not been made, are placed in the "=" dual.

Note that the following is a good rule of thumb:

$$R \in (>R)(C) \text{ whenever } C \in (<C)(R)$$
$$E \in (>E)(C) \text{ whenever } C \in (<C)(E)$$

and so on. Thus, in general:

$$I_1 \in (>D)(I_2) \text{ whenever } I_2 \in (<D)(I_1)$$

and

$$I_1 \in (=D)(I_2) \text{ whenever } I_2 \in (=D)(I_1)$$

It is not a proper theorem because of the fuzziness of the "=" dual.

Notice for instance, that:

$$C \in C(R(C))$$

or in general that:

$$I \in D(D(I))$$

Thus a theory item is contained in its double dual [4]. Going a step further, it is clear that a theory is contained in its double dual:

$$T \subset D(D(T))$$

where T, the theory, is the union of its sets of example, concepts and result items, and where D of a union is the union of the D's. Hence, in the case where there is equality -- i.e., $T=D(D(T))$ -- the items don't point outside of the theory and thus the theory is in some sense self-contained.

With this notation, *relation via the dual idea* can now be written as:

$$A \sim B \text{ iff } D(A) \cap D(B) \neq \phi$$

If more precision of description is needed, one can say that items A and B are:

conceptually dual if they share common concepts: $C(A) \cap C(B) \neq \phi$
illustratively dual if they share common examples: $E(A) \cap E(B) \neq \phi$
deductively dual if they share common results: $R(A) \cap R(B) \neq \phi$

[4] This is reminiscent -- at least symbolically -- of the natural embedding of a Banach space in its second dual.

Often it is useful to indicate which dual items are shared and are being used as the basis of the dual relation. Items A and B are said to be *related modulo an item or items* <I>, if <I> is contained in D(A) ∩ D(B). This is written as:

$$A \sim_{<I>} B$$

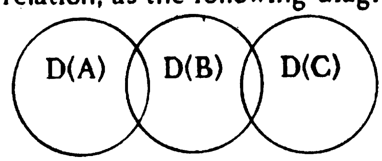
(This is read "A is related to B mod item[s] I".)

In the preceeding list of items related through the dual idea, the pairs of items are related modulo the "via" item. Thus, for the examples of the real and p-adic numbers,

$$R \sim_{<\text{completion}>} Q_p.$$

This degree of precision is useful for pinpointing threads of associations.

While relation via the dual idea (~) is trivially reflexive and symmetric, it is not transitive and is thus not an equivalence relation, as the following diagram indicates:



However, there is a stronger notion, the *equivalence of items*. Two items in different spaces are *dual equivalent* if their duals are equal [5]:

$$A \text{ "=" } B \text{ iff } D(A) = D(B).$$

This is an equivalence relation.

Two items which are dual equivalent are very similar since they completely share their duals. For instance, two dual equivalent results will have identical sets of dual concepts and examples; they will be motivated and supported by the same collection of illustrations and ideas. Two such results are in some sense are the *same* and should be "identified" [6].

Slight modification of dual equivalence can be used when there is only a partial equivalence, e.g., when only the example-duals are equal. Items A and B are:

> *conceptually equivalent* if their concept-duals are the same: C(A) = C(B)
> *illustratively equivalent* if their example-duals are the same: E(A) = E(B)
> *deductively equivalent* if their results-duals are the same: R(A) = R(B)

---

[5] This does not mean that there is equality between the pre-, post-, and neutral duals.

[6] "Identified" in the mathematical sense of belonging to the same equivalence class under the relation of dual equivalence.

2.2.4 Knitting the Larger Fabric

In reality, all three graphs belong to a larger graph encompassing the whole theory. However, to emphasize the use of different relations within a theory, the representation spaces are pictured and treated separately. The dual idea describes one way in which all three spaces are connected together. In one's memory or in a computational representation, all three spaces are linked together. They are obviously all part of one (huge!) semantic network which has not just one type of connective relation, but many. By separating them, this analysis hopes to untangle their relationships and make their interconnections clear.

The dual space idea can be used to topologize the representation spaces. One says that two items are *close in the dual sense* if their dual items are close in their representation spaces (using the standard graph metric, for instance) or if their dual sets share a significant overlap (using the symmetric difference). There are several alternative definitions for such a dual metric and its topology. In general, the graph and dual metrics generate very different topologies or senses of closeness. For instance, within Results-space, the Jordan Normal Form and Cayley-Hamilton Theorems are not deductively near each other, but because of many shared examples, they are close in the examples-dual sense.

The power of the dual idea and the sense of closeness it induces is that it provides a good approximation of the human notion of what it means for two items to be related or close in one's understanding. Items distant within their representation may be quite closely related in the dual sense. The dual ideas defines new neighborhoods in which to look for illustrative examples, elucidating results, and relevant concepts, and so provides a new source of hints in problem solving and new regions of exploration. It emphasizes associative referencing in retrieving information.

To organize mathematical knowledge by means of these categories and relations, several judgements must be made. First, the representation space for an item must be chosen (e.g., Q, the rational numbers, could alternatively be presented as a definition or an example), and secondly, the item must be tied into its chosen space by naming its predecessors and successors (e.g., Q points back to Z, and ahead to R and Q_p). Thirdly, an item must also be linked to its dual items (e.g., Q can be linked to the concepts of division, completeness, density, field, fractions, and to the results on Archimedian principle, cardinality, and the irrationality of $2^{1/2}$, etc.). Fourthly, the dual items can also be ordered.

A particular representation can clearly reflect certain mathematical, pedagogical, esthetic, and historical biases. However, the representation scheme is perfectly general. It helps organize mathematical knowledge by providing a framework of structures and relations, serves as a basis against which to compare the knowledge of several mathematicians, and gets one started on knowing what one knows and what others know.

2.3 Epistemological Classes of Items

Items can be classified in several ways, some of which are by their: (1) role in the logic, illustration or pedagogy of the theory; (2) mathematical content; (3) importance in the theory; (4) role in learning and understanding of the theory. Each of these criteria represents a different "cut" through mathematical knowledge.

The *epistemological classes* summarized here and discussed in subsequent chapters address the role of items in one's learning and understanding of a theory. Other classifications are probably needed to capture other aspects, such as the importance of items which are addressed by systems of worth ratings, such as the *Michelin* scheme [7].

(1) Examples

(E1) *Start-up* examples are perspicuous, easily understood illustrations.
(E2) *Reference* examples are used throughout the theory.
(E3) *Counter-examples* exhibit the falseness and limits of an item.
(E4) *Model* examples are paradigms or generic illustrations.
(E5) *Anomolous* examples are cases that don't fit in with expectations.

(2) Concepts

(C1) *Definitions* are formal mathematical definitions and procedures.
(C2) *Mega-principles* are heuristic kernels of wisdom.
(C3) *Counter-principles* are heuristic words of warning.

(3) Results

(R1) *Basic* results establish first basic facts of a theory.
(R2) *Key* results establish frequently used results.
(R3) *Culminating* results are goal results of the theory.
(R4) *Technical* results establish technical details.
(R5) *Transitional* results provide logical stepping-stones in the theory.

These classes are not necessarily disjoint because an item can serve more than one function in one's understanding. For instance, an example may serve as both a point of reference (E2) and as a counter-example (E3), e.g., the Cantor set. Also, in this classification, one must realize that the boundaries are not defined absolutely: one man's example can be another man's theorem. For instance, real finite dimensional vector spaces -- an example of Banach spaces to a functional analyst -- is a vast subject in itself for a beginning student of linear algebra.

---

[7] The Michelin rating assigns from no to four *'s to an item as follows: * for an interesting item, wortth noticing; ** for an important item, worth a "stop"; *** for a very important item, worth a "detour"; **** for an extremely important item, worth a "journey" in itself.

Also, this classification is not exhaustive since there are no doubt other types of items.[8] It is not static either, since items tend to migrate between classes as one's understanding of a theory deepens.

2.4 The Setting of Items

2.4.1 The Setting

The *setting* of an item is the domain in which the item is demonstrated, that is, stated, defined, constructed or proved. It is the context in which the item is known and discussed.

Typical settings in analysis are:

R	the real numbers
C	the complex numbers
R^n	real Euclidean n-space
$l_2(R)$	the Hilbert space of square summable sequences
C([0,1])	the continuous functions on the unit interval
$M_n(F)$	nxn matrices with entries from F

A setting does not necessarily have to be a particular space, such as R; it can be a general type of space, such as:

nls	normed linear space
ms	metric space
vs	vector space
fdvs	finite dimensional vector space
top sp	topological space
H	Hilbert space
B	Banach space
L_p	L_p-space
C(S)	Continuous functions on a set S
$C_o(S)$	Continuous functions on S with compact support
B(S)	Bounded functions on S
BL(X,Y)	Bounded linear operators from X to Y

An extensive list of the special settings used in analysis appears in sections 2 and 16 of Chapter IV of Dunford and Schwartz.

[8] For instance, this taxonomy of results does not include classes for doubted, suspected or conjectured results.

The setting of an item arises very naturally when stating it. For instance, the statement of the Bolzano-Weierstrass Theorem is:

In R^n, every bounded sequence has a convergent subsequence.

The setting of this theorem is R^n. Formalistically, the setting can be absorbed into the hypothesis (in this theorem as a condition on the sequence), but in fact, one doesn't think of settings this way. For one thing, they are often set off differently than the rest of the hypotheses by use of certain key words such as *In* and *For*. For another reason, the hypotheses and conclusion tend to be grouped together and treated as a unit; for example, there is a table in Dunford and Schwartz [p.372] that shows the validity of eight basic analytic if-then statements in a list of settings used in analysis; One moves an if-then statement around like a domino which can be placed in different settings. Lastly, declarations of settings are often conspicuous in their absence; one is rarely so careless as to omit a hypothesis or conclusion when stating a theorem, but one often neglects to mention the setting. The setting comes as a default assumption. Thus the setting-hypothesis distinction is embedded in the structure of mathematical knowledge. Also, notice that the setting is really a "common factor" to both sides of an implication. Hence, for these various reasons, settings are treated separately.

Pinpointing the setting has several benefits. It makes explicit the domain of discussion and clarifies a whole host of implicit assumptions and defaults thus eliminates one potential source of ambiguity. It facilitates variation of contexts, especially with regards to lifting a statement to a more general one.

2.4.2 Settings and Disambiguation

Omitting the setting allows for great variation in the interpretation of an item and potentially "fatal" ambiguity. For instance, consider a result statement such as:

On the unit sphere, a continuous real-valued function assumes its extreme values.

This result depends implicitly on the compactness of the unit sphere, a condition which holds only in finite dimensional vector spaces [Hoffman]. Thus, this statement is false in any infinite dimensional setting. A student might wonder, "Am I supposed to think of this result in the plane or three-space, where everything pretty much agrees with my geometric intuitions, or should I consider the statement in a space such as $l_2(R)$ where some funny things happen?" This result can flip-flop from true to false depending on the choice of setting.

Lack of declared setting has lead to an *instability* not only in this result's logical validity, but also in the understanding of it. In this paper, an item is *well-stated* only if its statement explicitly mentions its setting.

Declaration of setting is important not just for results, but for concepts and examples as well. For instance, the statement that "2 is a prime" can be true or false. Most students' reaction is that it is true, because they normally think of the number 2 as residing in the integers, Z. However, if the domain of discussion is the Gaussian integers Z[i], as is often the case for examples in some topics of number theory such as quadratic reciprocity, the statement is false, since

$$2 = (1+i)(1-i).$$

Omitting the setting forces the reader to guess or infer it. The first alternative is logically unsound, since guesswork is totally inappropriate. The second is pedagogically bad since it breeds frustration for the reader by forcing him to search for clues to the intended setting. Resolution of ambiguous settings can lead to prolonged backchaining, such as occurs when one is reading the proof of a result whose logical predecessor is stated without setting; one goes back to the predecessor to verify applicability and finds that this can not be decided without more backchecking. Explicit declaration of setting also keeps the statement of items in sharp focus by constraining the setting to the least structured setting -- i.e., most general -- that supports the item. Omission on the other hand, leads to overly specified settings, since one would rather choose a powerful setting and be safe, than risk a deficient one and be sorry. Thus declaration of setting is essential, wise and efficient.

2.4.3 Settings and Generality

It is often possible to nest settings according to their increasing generality. For example:

$$C_0(S) \subset C(S) \subset B(S)$$

$$l_2(X) \subset H \subset \text{inner product space}$$

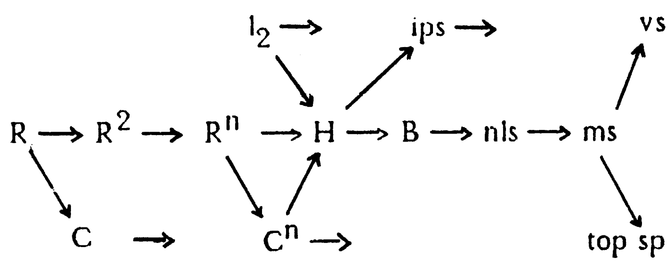
(These chains may be read as: "continuous functions with compact support are (a subclass of) continuous functions are bounded functions" and " an l_2 space is a Hilbert space is an inner product space".) Such *chains* exhibit an ordering of settings from less to more general by *is-a* relations: *is-an-instance, is-a-subset, is-a-subclass, is-a-type*, etc. Settings are thus intimately related to the generality of items and can be organized in a traditional is-a hierarchy.

A particular setting may occur in several chains, emphasizing different directions of generalization. For instance the chain with *H* (Hilbert space) shown above stresses the inner product idea; each of the settings is an inner product space. *H* also occurs in the chain which emphasizes the metric concept:

$$R^n \subset H \subset B \subset \text{nls} \subset \text{ms}$$

("R^n is a Hilbert space is a Banach space is a normed linear space is a metric space.") This second chain tells you that you can think of items residing in R^n as also residing in Banach spaces, for instance. Such a re-thinking sometimes leads to simplifications as in the case of thinking of linear transformations on R^n not as nXn matrices but as linear operators on a Banach space. This second chain also reminds one that any normed linear space can be turned into a metric space (by a standard trick of measuring distance between elements as the length of the vector joining them). The chain also serves to note that one cannot usually make assumptions in both directions of the chain; for instance, not all normed linear spaces are Banach spaces.

Each chain is usually a subset of some complex graph. For example, the two chains containing *H* can be embedded in the following graph:



Most chains stress a distinct flavor. For example, in topology there is a ranking of spaces according to the separability of points: T_0, T_1, T_2, T_3, T_4, where for instance, T_2-spaces are Hausdorff spaces and T_4-spaces are normal Hausdorff spaces. The chain is [Dugundji or Kelley]:

$$T_4 \subset T_3 \subset T_2 \subset T_1 \subset T_0$$

Chains of settings most frequently used in the EIGEN data base [Michener 1977] are:

$$R \rightarrow R^n \rightarrow C^n \rightarrow \text{fdvs} \rightarrow \text{vs}$$

and

$$M_2\{(0,1)\} \rightarrow M_2(Z) \rightarrow M_2(R) \rightarrow M_3(R) \rightarrow M_n(R)$$
$$\rightarrow M_n(C) \rightarrow M_n(F) \rightarrow BL(X,Y)$$

The first chain stresses the vector space idea and the latter, the linear operator idea.

Thus in addition to the three representation spaces already discussed, there is really a space of settings. In *Settings-space* the relation is *is-a*. Since this report focuses on the understanding of the mathematics within particular theories, there are usually not large numbers of settings involved, and so it is not really necessary to get involved in a discussion of Settings-space; most of the settings will be related through simple chains. However, if the discussion is broadened to address how knowledge in various theories is related then inclusion of a Settings-space is in order.

2.5 Representation of Individual Items

2.5.1 Well-Stated Items

There are three ingredients of information in the declarative statement of a result: the *setting*, the *hypotheses*, and the *conclusions*. [9] The statement of a example contains two ingredients: the setting and a *caption* which states what the example exemplifies. The statement of a concept contains its setting and the *declarative* statement of its definition or heuristic. For an item to be *well-stated* it must contain these ingredients.

The requirements for well-stated examples, results and concepts can be summarized as:
for an example: the setting and the caption;
for a result: the setting, hypothese and conclusions;
for a concept: the setting and the declarative statement;

The following are examples of well-stated items:

In Z[i], 2 is an example of a rational prime that ramifies (i.e., is no longer prime).

In R^n, if a sequence is bounded, then it contains a convergent subsequence.

In finite dimensional vector space, a set of vectors is called a basis if it spans the space and is linearly independent.

2.5.2 Procedural and Declarative Aspects of Items

An item can be presented in more than one way. For instance, it can be expressed by a declarative statement or by a procedure or the result of a procedure; declarative statement and procedure are different aspects of an item. Concepts are often presented as declarative definitions or in terms of procedures. An example can be represented by its caption and picture or as the result of its construction. A result can be presented as a statement or the result of a chain of syllogisms. In this way, all three types of items have declarative statement/ procedural representations:

[9] In the case of an equivalency result, the convention in this discussion is that the phrase that precedes the "iff" or equivalency arrow (<=>) is called the hypotheses, and that following, the conclusions, although logically there is no distinction between the two. The hypothesis can be null, as in the identity $\sin^2(x) + \cos^2(x) = 1$.

result: statement/ proof;
example: caption and picture/ construction;
concept: declarative statement/ procedural formulation.

For instance, the Cantor set can be presented as the outcome of applying the procedure of "deleting middle thirds" or or with the caption "the Cantor set is an example of a uncountable set which has measure zero" or s the summarizing ikon of Figure 1. The concept of eigenvalue in finite dimensional vector spaces may be ***defined*** as:

> λ is an *eigenvalue* and a non-zero vector x is an ***eigenvector*** of a linear transformation A if $Ax = \lambda x$.

It may also be expressed as the result of the *procedure:*

> SOLVE the characteristic equation: $\det(A - \lambda I) = 0$; the roots are eigenvalues.

Besides declarative and procedural aspects, items can have other aspects such as: static or kinetic diagrams and sketches; discourses in natural language; symbolic notation such as formulas and equations.

2.5.3 The Item Frame

In deciding what knowledge about examples, results and concepts to represent, (in particular, in the design of the Grokker System/Grokker Learning Advisor (GS/GLA) [Michener 1977]) it became evident that all three types shared a great many similarities. They are all represented fundamentally by the same frame structure, modified to reflect their special needs. As an overview, the representation for each theory item contains:

(1) *HEADER* information, such as *ID, NAME*, epistemological *CLASS*, the *Michelin RATING* (from one to four ☆'s) which describes the importance of the item relative to the theory as a whole, and other high-level descriptors;

(2) *STATEMENT* information which explicity declares the mathematical *SETTING* and includes the declarative formulation of the item: in the case of a result, its if-then *STATEMENT*; in the case of a concept, its mathematical *DEFINITION*; and for an example, a *CAPTION* stating what the example illustrates;

(3) *DEMONSTRATION* information which includes the procedural aspect of an item: a *PROOF* in the case of a result; *CONSTRUCTION* for an example; and a *PROCEDURAL* formulation for a concept;

(4) a *PICTURE* which is a static or dynamic (i.e., sequence of) pictures;

(5) *IN-SPACE POINTERS* to the item's predecessors and successors;

(6) *DUAL-SPACE POINTERS* to its two associated subsets of dual items;

(7) *PEDAGOGICAL* data indicating which teachers use this item and when;

(8) *REMARKS* on the item, such as when and how to use it;

(9) *EXTRAS* which fine-tune the representation for an example, concept, or result;

(10) Additional data such as bibliographic references and useful applications.

The complete specification for the item frame, including the data types used in each slot, can be found in [Michener 1977]; included there is also the EIGEN data base for the study of eigenanalysis, an important topic in linear algebra. This data base is also discussed in Chapter 6 of this report.

Figure 5 shows part of the representation for the "Basic 16" example, an important reference example consisting of the sixteen 2x2 matrices whose entries are 0's and 1's. Entries in the pointer fields are ID's of other items from the EIGEN data base. MP(0/1) is the mega-principle that suggests substituting 0's and 1's; MP(n=2) suggests examining the 2x2 case.

Figure 5

ID E100	CLASS Reference RATING *** NAME **Basic 16**
STMNT	SETTING $M_2(R)$ CAPTION The BASIC 16 2X2 matrices illustrate the properties and problems of eigenanalysis for general matrices.
DEMON-STRA-TION	AUTHOR ERM MAIN-IDEA Apply MP(0/1) and MP(n=2) to general NXN matrices CONSTRUCTION: Examine each of the 16 matrices by calculating their spectrum and eigenvalues
PICTURE	For each matrix, a picture in R^2 of its spectrum, eigenvectors, and image of the unit circle.
REMARKS	Caution: These matrices have (over-) simplified spectra since most are unitary or singular. This example is a good source of counter-examples.
EXTRAS	LIFTINGS: Let n>2. Let entries be {0,1,-1}, Z, Q, R, C.
PEDAGOGUES (ERM,9)	
IN-SPACE POINTERS	BACK E90 FORWARD E101, E102, E103, ..., E116
DUAL-SPACE POINTERS	DSP1 C10(Def. eigenvalue/vector), C1(MP(0/1)), C2(MP(n=2)) DSP2 R10, R20, R40, R50, R60, R120

Chapter 3. EXAMPLES-SPACE

3.1 Classification of Example Items

In the collection of all the examples used to illustrate a theory, there are special groups whose elements share a salient and similar functional role in one's understanding.

3.1.1 Start-up Examples

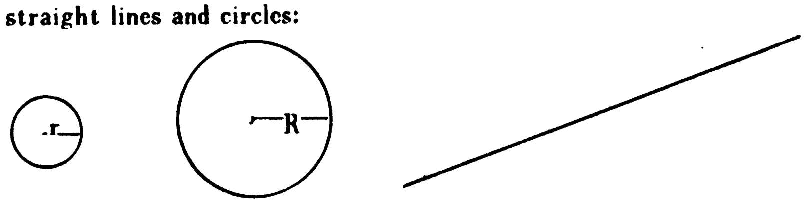
Start-up examples are perspicuous examples that can be grasped immediately when one studies a theory for the first time. They can be understood on a *stand-alone* basis: that is, in order to understand them, one does not need to understand many predecessor examples or pre-dual items. They strongly motivate their post-dual concepts and results and help one get started by setting up the right kinds of intuitions. They are constructionally uncomplicated, and in fact, the less complicated, the better. Thus, they often occupy starting nodes in Examples-space and are often dual to starting nodes of the other representation spaces.

Since start-up examples are highly suggestive of the central ideas and questions to be studied and motivate the basic definitions and results of the theory, they are often *projective* objects, in the sense that their relation to the studied phenomena can be lifted from their particular situation to a more general case. When the culminating items in a theory have been reached, a projective start-up example serves as a concrete special case that at best captures the essence and that at least is easily understood.

From a data structures point-of-view, start-up examples are rich in forward pointers, and lean in backpointers. By comparison, culminating theorems (Chapter 5) are heavy with backpointers. Both are mostly one-sided in their pointer flow, but their emphasis is antithetical.

An excellent, stand-alone, projective start-up example in the study of the curvature of plane curves is the example of "circles and lines". It can be easily lifted to a general definition [Spivak, Chapter 1]:

Consider straight lines and circles:



One can agree that: (1) a straight line does not curve at all; and (2) a circle of radius $R > r$ curves less than a circle of radius r. One can then make a trial definition of curvature for these two special cases:

(1) $K = 0$ for straight line

(2) K = 1/r for circle of radius r

This definition can be lifted to the general case by relating circles to plane curves. The method is to locally define curvature at a point by fitting a circle (the "osculating circle") to the curve in a neighborhood of that point [1]:

Then one defines:

K(t) = 1/(radius of the osculating circle)

By methods of differential calculus and inner product geometry one can then develop this definition and its consequences. In the process one derives the other, perhaps less intuitive, definition of curvature as the length of the differentiated tangent vector. The circle-line example has the advantage of always providing a concrete example (K(circle) = 1/r) and a lifting method (osculating circle) for getting a "handle" on curvature for plane curves.

Another good example of a start-up example is Strang's use of the simple ordinary differential equation

$$x'(t) = ax(t)$$

to introduce the concept of eigenvalue [Strang, pp.172-174]:

> "The first step is to understand what eigenvalues are and how they can be useful. One of their applications, the one by which we want to introduce them, is to the solution of a system of ordinary differential equations. We shall not assume that the reader is an expert on differential exquations; if he can differentiate the usual functions like x^n, sin x, and e^x, he knows more than enough. As a specific example, consider the coupled pair of equations..."
>
> The substitution of $u = e^{\lambda t}$ into $du/dt = Au$ gave $\lambda e^{\lambda t}x = Ae^{\lambda t}x$, and the

[1] More precisely, one considers curves, parametrized by t, c(t) continuously differentiable, and three points $c(t_1)$, $c(t_2)$ and $c(t_3)$, and sees if the circles defined by them approach a limiting circle as $t_1, t_2, t_3 \longrightarrow t$; if they do, it is known as the osculating circle.

cancellation produced

$$Ax = \lambda x$$

This is the fundamental equation for the eigenvalue λ and the eigenvector x."

Some other well-known start-up examples and the items they motivate are:

Z, the integers (for the concepts of group and ring and the theory of factorization);
P[x], the polynomials of one variable (for concepts of ring and algebras);
B(X;r), the open ball of radius r about X (for concept of open set);
R^2, Euclidean space (for concepts of vector and Hilbert spaces);
C([0,1]), continuous functions on the unit interval (for concept of function algebras);
$M_2(R)$, the real 2x2 matrices (for concepts of linear operators, non-commutativity);
D, the unit disc (for concepts of open and closed sets);
A(D), the analytic functions on a disc (for concept of a sheaf);

These start-up examples are all instances of the concepts they motivate. This need not be so: a start-up example can motivate a property by failing to have it. For instance, the Cantor function is a start-up example for the study of absolutely continuous ("AC") [2] functions, because it, itself, fails to be AC in a way that pinpoints what an AC function should do.

Each field of mathematics has its own special start-up examples. For instance, the scaling operator (cI) and rotations on R^2, and the differential operator (D) are start-up examples from the theory of eigenvalues in finite dimensions [Strang]. The differential operator is also a good start-up example for general spectral analysis. In the study of analytic functions, z and e^z are start-up examples [Knopp].

3.1.2 Reference Examples

Reference examples are examples which are useful throughout the theory. One refers to them repeatedly while wending one's way though a theory as test situations against which to guage new concepts and results. They tie together various items of a theory by emphasizing common illustrative situations. Thus, they simplify one's understanding of a theory by providing a common dual node through which many results and concepts can be linked together. In fact, linking is one of the primary functions of reference examples in understanding. Another is to provide a touchstone to which one can always go back.

---

[2] An AC function has a very nice relationship with its (Lebesgue) integral:

$$f(b) - f(b) = \int_a^b f'(x)dx$$

The Cantor function on the unit interval fails to satisfy this equation:

$$1 - 0 = f(1) - f(0) \neq \int_0^1 0dx = 0.$$

For instance, no matter how knowledgeable one is in algebraic number theory, one invariablly looks at Z, the integers, to test things out. In his books *Induction and Analogy* and *How To Solve It*, Polya frequently references the following standard triangles:

3-4-5
30-60-90
45-45-right
isosceles
equilateral

In his exposition of Euler's formula in *Proofs and Refutations*, Lakatos often refers to cubes and tetrahedrons, reference examples in the realm of polyhedra.

The example $l_2(R)$, the Hilbert space of square-summable sequences of real numbers [3] is a very important reference example in analysis. It is important not only because it is a specific example of an l_p space and a model for all Hilbert spaces, but also because it is any easily constructed example of an infinite dimensional normed linear space. It provides an easily studied situation in which to study properties and conjectures regarding infinite dimensional spaces .

For instance it is often used to expose a statement's dependence on finite dimensionality; this makes it especially useful since so much intuition is thoroughly rooted in finite dimensions. Consider the famous theorem:

> Bolzano-Weierstrass: In R^n, a bounded sequence has a convergent subsequence.

One can ask if this result is true in any normed linear space; the answer is *no*. The reference example l_2 provides the needed test situation [Hoffman]:

> Consider the following sequence of sequences in $l_2(R)$:

$$p_1 = (1, 0, 0, ...)$$
$$p_2 = (0, 1, 0, ...)$$
$$p_3 = (0, 0, 1, ...)$$
$$\vdots$$
$$p_n = (0, 0, 0, ... , 1, 0, ...)$$
$$\vdots$$

[3] A sequence $X = \{x_n\}$ of real numbers is "square-summable" if $\|X\| = (\Sigma x_n^2)^{1/2}$ is finite.

> One sees that $||p_i|| = 1$ (and thus each sequence p_i is in the unit ball), but $||p_i - p_j|| = 2^{1/2}$ for $i \neq j$ and thus the sequence has no limit points, hence no convergent subsequence. The problem is that there are infinitely many independent directions in which to move and that the sequence $\{p_i\}$ can wander with no two elements coming near each other. Therefore the unit ball in l_2 is not sequentially compact.

Other familiar reference examples are:

> Z[i], the Gaussian integers;
> Z/pZ, the integers modulo a prime number p;
> R^2, real Euclidean 2-space;
> C([0,1]), the continuous functions on the unit interval with the sup-norm;
> BL(X,Y), the bounded linear operators from space X to space Y;
> [0,1], the unit interval;
> P, the Cantor set;

In linear algebra, the 2x2 matrices whose entries are 0's and 1's which is called here the ***Basic 16*** -- denoted as $M_2(\{0,1\})$ -- is an important reference example. It is an example that shows many of the "good" properties of matrices as well as many of the things that can go wrong with them. For instance, the Basic 16 exemplifies the following kinds of matrices:

> singular
> repeated eigenvalues and diagonalizable
> repeated eigenvalues and not diagonalizable
> symmetric
> non-symmetric
> non-symmetric and diagonalizable
> non-symmetric and not diagonalizable
> unitary
> orthogonal
> circulant
> permutation
> projection

(All of these concepts would be included in the concepts-dual of the Basic 16, and each of the concepts would contain the Basic 16 in its examples-dual.)

Notice that the examples Z and R^2, besides functioning as start-up examples, also serve as reference examples; they are referenced throughout algebra and analysis. Thus, in one's understanding, an example may at first be thought of as a start-up example, then as one acquires more knowledge, one sees the example as a reference example. This conversion from start-up to reference might occur either because one sees it used so frequently or because one recognizes its fundamental relavance to the whole theory since so much knowledge is linked throught it.

The conversion process of start-up to reference might proceed faster if these classes were recognized in teaching.

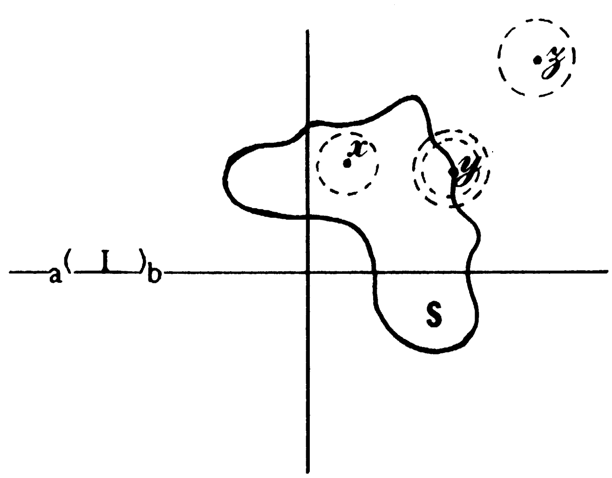
3.1.3 Model Examples

Model examples are general, or as mathematicians say "canonical", illustrations. A model example is often referred to as *generic* since "it represents to you the general case" [Polya, I & A, p.23, exercise 10]. Model examples contain the *essence* of a situation in the sense that they very strongly emphasize its outstanding feature usually by means of a simple picture or schematic diagram. Model examples contain some of the most important illustrative material of a theory and as such, can be considered "theorems" of Examples-space.

Model examples are flexible and adaptable. They are often used as first approximations to a situation, which are then fine-tuned to meet the specifics. They provide a canonical rack on which to "hang one's hat". Whereas reference examples are used as-is as test situations, model examples are custom tailored by embellishment or adjustment, often of their pictorial or schematic elements. Pictorial and schematic elements, and in general the representation, of a model example are extremely important.

For instance in his analysis book, Hoffman explicitly presents the model that mathematicians have for sets, especially with regards to the concepts of closed, open, boundary, etc.:

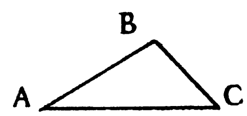
> "EXAMPLE 18. Let's look at some subsets of R^2. In order to understand closure, interior, and boundary, one usually begins by drawing a set S as below. Indicated are a point X in the interior of S, a point Y on the boundary of S, and a point Z which is *not* in the closure of S. The boundary of (this particular) S is a curve, and it does not matter whether S is the open region bounded by the curve or the closed region bounded by the curve...."



Note the strong use of R^2 as a reference situation for this example.

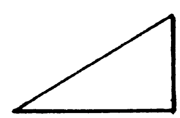
Knowing what representation to use and when is important knowledge. Acquiring models and recognizing them is an important step in gaining understanding. To make this step, one must understand the regularities, assumptions, and expectations that the model example expresses. Annotating the model as to its appropriateness, good points, and limitations is another important step.

A familiar model from plane geometry is that whereby one draws a triangle as:



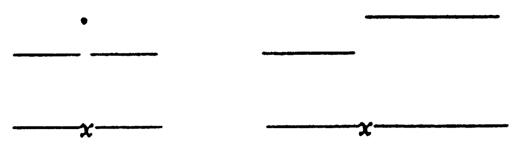
when setting up the "givens" in a plane geometry problem. Such a diagram was used by Gelernter [Gelernter 1963] with great success as a "diagram filter" in his geometry proving program.

It is a reasonable initialization step, but it is by no means a universally valid representation for all triangles since it does not depict obtuse angles, for instance. However, if the triangle model were a right triangle:



used in discussions only about right triangles, the model would in some sense be universal in this context. So the context or setting of a model is extremely important. Notice in the picture, the right angle is all that matters, not the other angles; that is, the values for the lengths of the sides may be filled in later to represent a specific situation, such as a (3, 4, 5)-triangle. Thus, it is important that the model not be suggestive of things it shouldn't: e.g., the right triangle model should not be isosceles, and the model for isosceles should not be equilateral.

Model examples often provide prototypical models for situation. For instance, in the study of real-valued functions, the following diagrams indicate the kind of behavior a function has at a point where it has a simple discontinuity. The diagram on the left represents a function with an "aberration" discontinuity at *x*, i.e., the right and left hand limits exist and are the same, but the function has the "wrong" value at *x*, and that on the right, a "jump" discontinuity, i.e., right and left limits exist but are not the same:



jump aberration

Observe that the specific measurements in these pictures are unimportant; what counts is that they capture the essence of the situation. magnitude of the jump or aberration is almost of no interest; it's the topology of the diagram that counts. The same is true in the other models; specific measurements of the pictures don't matter as much as the general shape. Thus, as can be seen from the above examples, model examples have a strong pictorial representation which can be adjusted to meet specific situations and the graphical elements of the picture are similar to slots or place-holders for information. In this way, model examples are a frame [Minsky 1975] of what to expect or consider standard or reasonable in a theory.

In linear algebra, diagonal and upper triagular matrices are important models:

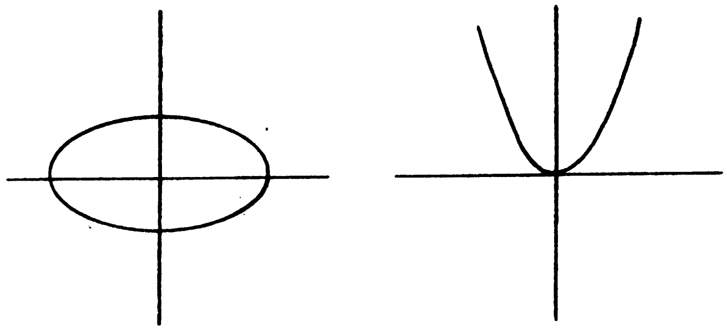
$$\begin{pmatrix} * & & & & & \\ & * & & & & \\ & & * & & & \\ & & & * & & \\ & & & & * & \\ & & & & & * \end{pmatrix} \qquad \begin{pmatrix} * & * & * & * & * & * \\ & * & * & * & * & * \\ & & * & * & * & * \\ & & & * & * & * \\ & & & & * & * \\ & & & & & * \end{pmatrix}$$

Other models are tri-diagonal and block diagonal matrices, both of which are derivates of the diagonal model.

The model status of the upper triangular example is established by the Jordan Normal Form Theorem. It retains its validity throughout finite-dimensional eigenanalysis and is thus universally indicative of the general case. It can be said to be a "global" model in that domain. It is fine-tuned by plugging in values for the elements.

In the more restricted setting of real symmetric matrices, diagonal matrices are the only story and are a global model, whereas in the context of general matrices, the diagonal model is not the universal canonical form. However in the context of general matrices, the Gerschgorin Circle Theorem shows that the diagonal model for eigenvalues is not so far from the truth when the off-diagonal elements are small, for in this case, the eigenvalues are near the diagonal elements [Strang]. Thus the diagonal model is a projective item with the Gerschgorin or diagonalizability results providing the lifting, or generalizing, map.

Very familiar models for the conic sections are those figures centered at the origin and whose principal axes are aligned with the x- and y-axes:



These figures can easily be adjusted to reflect off-center or rotated sections.

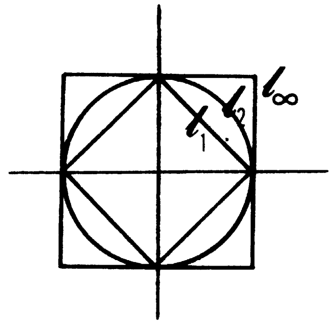
The globalness of a model example is often demonstrated by WLOG (without loss of generality) reduction arguments and classification theorems. In WLOG arguments one shows that the model is good enough for the task at hand. For instance, in calculus one learns how to reduce conic sections to canonical forms with the x- and y-axes as principal axes.

Examples are sometimes shown to be models by classification theorems, which provide pigeon holes or equivalence classes for the space of possible situations. For instance, the Jordan Normal Form Theorem, stating that all matrices are similar to upper triangular matrices of a special form, and the result that all symmetric matrices are unitarily equivalent to diagonal matrices show the universality and sufficiency of the upper triangular and diagonal models.

These types of results justify restriction to model situations. There is no need for constructing a lifting each time the model is used. When the models are projective, such as these last two matrix models, one implicitly applies the lifting procedure when using the example in a global sense.

Somewhat orthogonal to this approach is that whereby one investigates why the model is not good enough, or in other words, under what assumptions does the model fail. For instance, the model of the unit ball as a circle is not valid in L_p-spaces wth $p \neq 2$ where it is not circular (see below). Knowing the limitation of the circle model would prevent one from making the incorrect assumption that all unit balls have no flat spots, i.e., are strictly convex.

However, local models can often be pasted together to represent a universal situation (see Patch Proofs, in Section 5.3). In calculus and differential geometry one uses the model of locally-linear (the examples of lines and planes) to provide a global description. This patching of local models to obtain a global description is a fundamental technique throughout mathematics and particularly in calculus, differential geometry and analytic function theory.



The unit ball in l_1, l_2, and l_∞.

3.1.4 Counter-examples

Counter-examples are examples used to show a statement is not true or to sharpen a statement. Any item used in a *counter* or *delimiting* manner is a counter-example.

One can limit the truth of a statement by investigating the effect of the setting on its validity. Providing a counter-example within whose setting the statement is false can place an "upper bound" on its *generality*. For instance, the generality of the result that "2 is a prime number" is bounded by the counter-example of "2 = (1-i)(1+i)" set in the Gaussian integers [4]:

$$\begin{array}{ccc} R(Z,2,prime) & \not\to & R(Z[i],2,prime) \\ Z & \subset & Z[i] \end{array}$$

A counter-example that restricts the generality of a statement is called a counter-example *of setting*.

On the other hand, some counter-examples work within a given setting to sharpen a statement, negate a conjecture, disprove the converse, show that a hypothesis is necessary, or show that a conclusion cannot be expanded. Such counter-examples are found throughout mathematics. They help one to differentiate between concepts and are all uses of examples to *sharpen* one's understanding.

For instance in the setting of finite dimensional vector spaces, the relation between the concepts of symmetric and diagonalizable can be sharpened with the following 2X2 matrix

[4] Recall that in the R(S,H,C) notation for a result, S = the setting, H = the hypotheses, C = the conclusions.

(from the Basic 16). It is a matrix which is non-symmetric, but which is diagonalizable, thus showing that symmetric is not a necessary condition for diagonalizability although it is sufficent: [5]

$$\begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}$$

This matrix has distinct eigenvalues -- 0 and 1 -- and thus is diagonalizable.

This type of counter-example restricts the truth set of a statement within the universe of the given setting; this is related to the strength of results, discussed in Chapter 5.

A great many counter-examples are *pathological* constructions. Many of these *technical* constructions are, as pointed out by Freudenthal [1973], *hapax legomena* -- one shot atoms of information used once to establish a point, perhaps like the "clamshell" example mentioned below. While this example could be someone's favorite example and thus not so "one-shot" in his understanding, there are many people for whom it is.

Certainly, there are certain examples that come up very seldom and thus are isolated -- not richly linked -- in one's knowledge. Such counter-examples are close to Examples-space analogues of *technical results* of Results-space; their function in our understanding is very limited.

On the other hand, some counter-examples recur repeatedly, as reference examples, and become part of a mathematician's standard bag of tricks, for instance, the Cantor set and the Cantor function. The Cantor set itself has a rich genealogy of descendent examples [Gelbaum and Olmstead]. Another standard function to any student of analysis is the following:

$$f(x) = \begin{cases} 1 \text{ for } x \text{ in } Q \\ 0 \text{ for } x \text{ not in } Q \end{cases}$$

This is very far from a hapax legomenon.

---

[5] Two of the most basic results in the study of eigenvalues [Strang], state that "symmetric ==> diagonalizable" and that "distinct roots ==> diagonalizable". In fact one can set up a matrix in which the i,j-th entry has property i and property j:

	diagonalizable	not diagonalizable
symmetric	by result above	impossible by result
non-symmetric	(0 1)	(1 1)
	(0 1)	(0 1)

Counter-examples are one epistemological class of example that has long been recognized by mathematicians. Several books have been published that are compilations of counter-examples: *Counterexamples in Analysis* [Gelbaum and Olmstead] and *Counterexamples in Topology* [Steen and Seebach].

3.1.5 Anomolies and Pathologies

Some examples are anomolous. They are in some sense "strange" or surprising by going against one's expectations or intuitions. They don't fit in with one's understanding. This in itself might make them noteworthy even if they ar not well linked with the rest of one's knowledge. However, some are simply anomolous and remain largely unconnected. Anomolies are the anti-thesis of model examples.

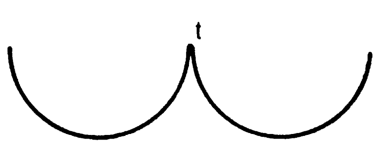
3.2 Constructional Derivation

Examples exhibit a constructional derivation: one builds a new example from old ones. Construction, which is the Examples-space analogue of formal deduction of Results-space, has a very strong procedural nature; it often heavily uses pictorial elements.

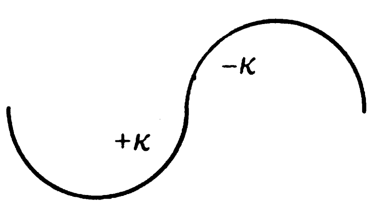
As mentioned in the last chapter, a good example of constructional derivation is the Cantor function. It is fabricated pictorially and procedurally from the Cantor set which is in turn constructed from the unit interval. The Cantor star [Hocking and Young, p.157] is another fabricational descendent of the Cantor set.

3.2.1 Some Simple Constructions

Another instance of constructional derivation is illustrated by the following pictures, built up from the start-up example for curvature. To investigate troublesome situations for curvature that arise due to lack of smoothness in a curve, one can "paste" circles together:



A curve with a discontinuous derivative at t but clearly constant curvature everywhere it is defined; there is an aberration discontinuity.



A curve with a discontinuous curvature (it jumps from $+K$ to $-K$); there is a jump discontinuity.

(These pictures may be compared with those for aberration and jump discontinuity discussed in Model Examples.) These situations of pointiness and inflection are paradigms for the difficulties curves may exhibit regarding their curvature. In the above illustration, the start-up example for curvature (a cicle) has become not only a reference but a model example.

Another instance of deriving a new example from an old one can be found in the continuation of Hoffman's EXAMPLE 18 from Section 3.1.3 above [Hoffman, p.64]:

> "In the same figure, let T be the set obtained by deleting from S that part of the real axis which lies in S. The points on the deleted line segment are in the boundary of T. Yet, somehow they seem "interior" to T in a weak sense.. They are not in the interior of T, because they are not even in T. But those points are in the interior of T closure."

Again the derived example is offering limits on the use of its predecessor as a model. In the trouble-with-curvature example and this last example, the models have been further manipulated to create a troublesome situation -- for the mathematical definition or the reader's intuitions. This last example can also be used to sharpen the strong reliance of the open-ness of a set on its setting. (What is open in R^1 may not even have an interior when considered as "living" in R^2.)

Thus it can be seen that there is a strong relationship between start-up and model examples and between model examples and counter-examples. A start-up example can be generalized to a model, and then the model can be used to create a counter-example to sharpen the limits of the concepts for which it is a model.

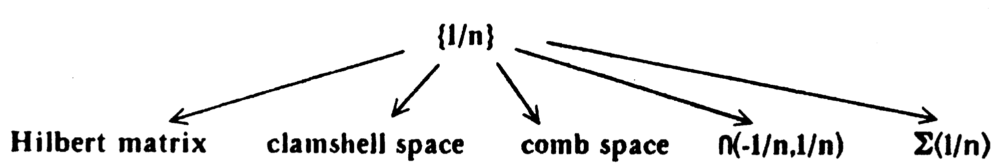
The start-up example of the harmonic sequence {1/n} -- from the theory of sequences and limits -- together with the standard construction technique of taking intersections gives rise to the counter-example to the statement "the countable intersection of open sets is open":

$$\cap\,(-1/n,\ 1/n) = \{0\}$$

To show the non-sufficiency of the "nth term approaching 0" for the summability of a series, one uses the example of {1/n} with nothing other than a counting argument:

$$1 + (1/2 + 1/3) + (1/4 + 1/5 + 1/6 + 1/7) + \ldots\ < 1 + (1/2 + 1/2) + (1/4 + 1/4 + 1/4 + 1/4) + \ldots$$

The harmonic sequence also spawns the examples of the Hilbert matrix [Ortega, p.32], and the comb and clamshell spaces [Dugundji, p.325, Example 8].



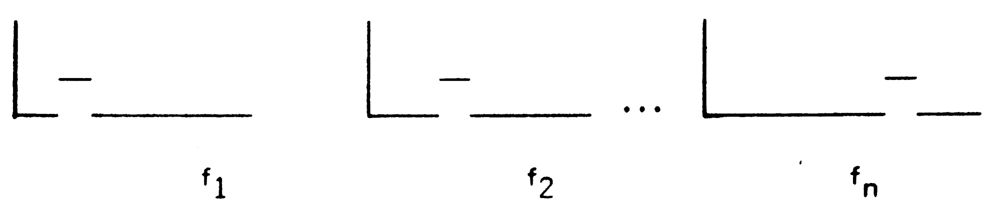
3.2.2 A Plethora of Constructions

The combination of the harmonic sequence with step and hat, or tent, functions gives rise to a plethora of function sequences that are the stockpot of counter-examples for a wealth of topics in analysis, e.g., convergence, interchange of limits, smoothness [Rudin; Gelbaum and Olmstead]. Such sequences are often built by simply *sliding* or moving a function along (to the right, for example) by defining each f_n in terms of an n or $1/n$ entering into its domain of definition, such as in $[n, n+1]$ or $[n, \infty)$. Obviously, such a sequence of functions is a constructional derived from harmonic series and sequences, and general models of step and hat functions.

The following are all examples of such sequential construction techniques. In the following examples of function sequences, each of the functions in the sequence is constructed by sliding a constant function along the x-axis. Often the function used is the simplest of all functions, the constant function 1, or the next simplest, the characteristic function of a set S, χ_S, which by definition is 1 on the set S and 0 elsewhere. Characteristic functions are themselves derived from the exceedingly general idea of "1 for *on, in* or *yes*" and "0 for *not on, not in* or *no*".

SEQUENCE 1. For instance, the following sequence of functions sharpens the need for care in interchanging limit and integration operations (i.e., the necessity of uniform convergence):

$$f_n(x) = \chi_{[n,n+1]}(x)$$



The function sequence converges pointwise everywhere to the function f = 0, but the integral of each function is 1. Thus:

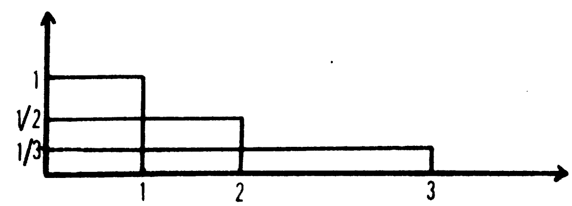
$$\lim_n \int_0^\infty f_n(x)dx = 1 \text{ which is not } \int_0^\infty \{\lim f_n(x)\}dx = 0.$$

This example also arises in connection with certain convergence theorems (e.g., Riesz's Theorem) in measure theory. This sequence converges everywhere, and thus almost everywhere to the function f=0, but not in measure or uniformly or almost uniformly.

SEQUENCE 2. By varying both the support and the function value to keep the area under each function equal to 1 by using the inverse relationship for function value and function support derived from:

$$n \times 1/n$$

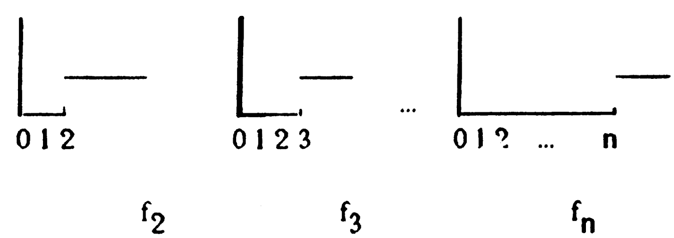
one gets the following sequence which is another counter-example highlighting the danger of interchanging limits and integration. Again, the limit function is the constant function 0 whose integral is 0 and the integral of each of the f_n is 1:



Sliding the support of the characteristic function leads to the following sequence of functions, another frequently used example in measure theory:

SEQUENCE 3. On the set $(0,\infty) \subset R$, define the sequence:

$$f_n(x) = \begin{cases} 0 & \text{if } x \in (0,n) \\ 1 & \text{if } x \geq n \end{cases}$$



This sequence converges pointwise everywhere, therefore almost everywhere, to $f(x) = 0$, but this sequence does not converge almost uniformly to 0 or anything else since one can exhibit a set $[k, k+1]$ outside of which the convergence is not uniform (the rate of convergence -- the n -- depends on where x is).

"Hat" or "tent" functions are the next most complicated sort of function after the characteristic and step functions. Such a function is simply a piecewise-linear function that rises linearly from 0 to an apex point and then falls linearly back to 0, and is 0 everywhere else.

SEQUENCE 4. A sequence which arises in the study of uniform convergence [Hoffman, p. 172] is the following:

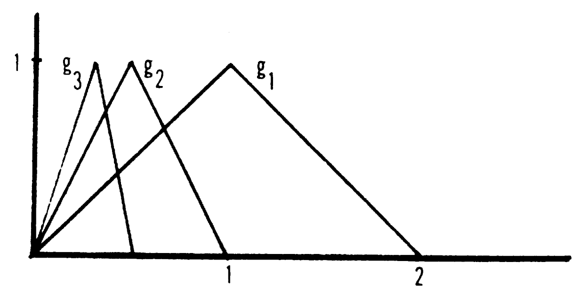
$$f_n(x) = \begin{cases} n^2 x, & 0 \leq x \leq 1/n. \\ n - n^2(x - 1/n), & 1/n \leq x \leq 2/n. \\ 0 & \text{otherwise.} \end{cases}$$

"In other words f_n is a tent function which rises linearly from 0 to n on the interval [0, 1/n], falls linearly from n to 0 on $[1/n, 2/n]$ and is 0 elsewhere. The sequence $\{f_n\}$ converges pointwise to 0; however, the convergence is not uniform..."

SEQUENCE 5. The trouble with SEQUENCE 4 is boundedness, which can be fixed by introducing a $1/n$ as an ameliorating fudge factor to generate a new sequence:

$$g_n = (1/n)f_n.$$

The g_n look like:



There are lots of variations on the tent function sequence

AND SO ON. An obvious way to generate additional sequences of functions is to make each of the functions smoother. The smoothest such functions would be infinitely differentiable functions, such as:

(i) $n \sin nx$.
(ii) $1/n \sin nx$
(iii) $n^{1/2} \sin nx$.

By chopping off part of these functions -- i.e., using them *joined* with the 0 function outside of certain intervals, like (0, $1/n$) or (0, n) -- one generates functions looking very similar to the hat functions but which are smoother. (See [Hoffman, p.260] for a picture showing this with (iii).)

Note how one can specify a hat function in terms of its apex. Also notice that one can construct a hat function by integrating a step function, and thus integration is used as a fabricational technique.

3.2.3 General Construction Techniques

Some examples of other general techniques for the construction of examples are:

integration
differentiation
summation
passing to the limit
union
intersection
cone constructions
one-point compactification
orthogonal projection
Cartesian products
identification topologies

One obviously can use any concept or result of the domain -- specifically the procedural aspect -- in a construction. A construction technique can be applied to any of the known examples, especially model or reference examples. One also has ways of combining and cleaving apart examples.

Let us remark that most examples are constructed for a purpose: to instantiate, reinforce, refute. These goals often impose certain constraints on the examples being sought (e.g., smoothness, integral = 1). The constraints can be used as conditions not only to generate examples (e.g., by tuning model examples) but also to restrict the search either directly through Examples-space or indirectly through the examples-duals of the concepts involved in the constraints. Work on such construction tasks, which we call *Constrained Example Generation* or *CEG*, is currently being carried out by the author [Michener 1978].

There has not been much previous attention paid to the generation of examples. However, some powerful generators have been built by Bledsoe and by Ballantyne. A technique for constructing counter-examples in topology by building up a space and its open sets by starting with one point and then adding others has been successfully programmed by Ballantyne [1975]. A technique for finding the largest set satisfying prescribed conditions by paring down the setting space, so as to instantiate existential quantifiers, has recently been automated by Bledsoe's theorem proving program [1977].

3.2.3 General Construction Techniques

Some examples of other general techniques for the construction of examples are:

- integration
- differentiation
- summation
- passing to the limit
- union
- intersection
- cone constructions
- one-point compactification
- orthogonal projection
- Cartesian products
- identification topologies

One obviously can use any concept or result of the domain -- specifically the procedural aspect -- in a construction. A construction technique can be applied to any of the known examples, especially model or reference examples. One also has ways of combining and cleaving apart examples.

Let us remark that most examples are constructed for a purpose: to instantiate, reinforce, refute. These goals often impose certain constraints on the examples being sought (e.g., smoothness, integral = 1). The constraints can be used as conditions not only to generate examples (e.g., by tuning model examples) but also to restrict the search either directly through Examples-space or indirectly through the examples-duals of the concepts involved in the constraints. Work on such construction tasks, which we call Constrained Example Generation or CEG, is currently being carried out by the author [Michener 1978].

There has not been much previous attention paid to the generation of examples. However, some powerful generators have been built by Bledsoe and by Ballantyne. A technique for constructing counter-examples in topology by building up a space and its open sets by starting with one point and then adding others has been successfully programmed by Ballantyne [1975]. A technique for finding the largest set satisfying prescribed conditions by paring down the setting space, so as to instantiate existential quantifiers, has recently been automated by Bledsoe's theorem proving program [1977].

Chapter 4. CONCEPTS-SPACE

4.1 Classification of Concept Items

4.1.1 Definitions

Definitions present mathematical ideas in a formal manner. Included as definitions in this epistemology are ideas specified by procedures, processes or algorithms, such as:

Cantor Diagonalization process
Horner's scheme
Sieve of Erastosthenes
the square root algorithm
Newton's method

The distinguishing feature of this epistemological class is definitive precision. Heuristics, on the other hand, often have some leeway. Definitions can be expressed in English, symbolic notation, or a mixture of the two.

Definitions can have either or both of declarative and procedural presentations. For instance, the concept of eigenvalue is a concept that has both declarative and procedural formulations; the Gram-Schmidt process is an example of a concept that is most naturally defined as a procedure.

Declarative definitions, the usual type of definitions in mathematics, are dense in any mathematics text and procedural definitions, while somewhat less frequent, are also abundant. Procedural and declarative formulations will be discussed in more detail in Section 4.2.

4.1.2 Mega-Principles

Mega-principles are big ideas expressed informally as kernels of wisdom. They are positively-oriented heuristics that provide powerful directives or suggestions. In their universality and importance, they stand head-and-shoulders above the bulk of ideas discussed in a theory. They are like "proverbs" [Polya 1973]. Mega-principles are frequently dual to important results and examples. They are the ideas or "flavors" of a theory, remembered long after the details have been forgotten.

Some well-known examples of MP's are:

"Try the 2X2 case." (in matrix theory)
"Always write down a basis first." (in linear algebra)
"Write the number in terms of its prime factors." (in number theory)

"Symmetric matrices are nice." (in numerical analysis)
"Polynomial time means reasonable time." (in complexity)
"Continuity means you can draw it without lifting your pencil." (in analysis)
"Do things component-wise." (in linear algebra)
"Examine extreme points." (in analysis)

Mega-principles are generalities which are useful in much the same way that model examples are, that is, as broad, suggestive, initial descriptions or expectations.

Mega-principles often express the contents of theorems and definitions as heuristic advice. For instance, the above MP on symmetric matrices is really a synopsis of several perturbation theorems for matrices [Ortega]. Another example is "to work with integrals, consider differential boxes of width δx". The MP "2 is almost always an interesting prime" is a condensation of many results in elementary number theory.

There are two types of mega-principles: (1) *interpretive* and (2) *imperative*. An interpretive MP offers a way of thinking about an concept, such as the MP's on symmetric matrices, polynomial time and continuity. It is a "folksy" paraphrase of other more formal statements. Imperative MP's suggest approaches or procedures to try, such as the MP's on 2X2 cases, basis, and prime factors.

Mega-principles, like model examples, have their limitations. For example, the MP suggesting trying the 2X2 case must be tempered with the remark that one should not be too hasty in jumping to conclusions for the general n-dimensional case and at the very least, one should check out the 3X3 case.

The MP that suggests doing things component-wise has its limitations in infinite dimensional settings. It even has limitations in finite dimensional ones (see Section 4.1.3). For instance, when examining the convergence properties of sequences in finite dimensional spaces, it is sufficient to do the analysis componentwise, i.e., a necessary and sufficient condition that a sequence of vectors converge is that the sequences of the components converge [Hoffman]. This is not the case in infinite dimensional spaces. One has only to look in $l_2(R)$, the infinite dimensional space of square summable sequences of real numbers, and actually only in the set of sequences consisting of a *1* as the *nth* term, and *0*'s elsewhere to find a counter example (See Section 3.1.2).

Thus, the requirement of explicit setting or domain of applicability is as necessary for MP's as for other items, and perhaps even more so since the informality of heuristics often leaves more ambiguity as to how and where to apply them. Also, MP's should be annotated with regards to the appropriateness of their application. As one learns more about the MP, one acquires more knowledge about what it is good for and when to use it: "the circles and arrows and a paragraph on the back of each one" [Guthrie 1970].

However, many mega-principles do wield far-reaching influence; MP's are often found to be much more widely applicable than one thought when one first learnt them. For instance, "Try the 2x2 (2-dimensional) case" and "Examine extreme points" are valid and useful throughout algebra, analysis, and topology. Such "trans-theoretic" mega-principles are very general principles of mathematics. The MP that "Symmetric matrices are nice." is associated with functional analysis of operators, eigenvalue analysis of matrices and stability of numerical process; its presence in these areas reflects the fact that in each symmetric matrices are an easy case. Such an MP allows linkages to be made between areas of mathematics which on the surface can seem to be quite distant, but on closer inspection can be seen to be addressing some of the "same" questions.

4.1.3 Counter-principles

Counter principles are *cautions* that alert one to possible sources of blunders and confusion, such as the statements:

> "Double roots are troublesome."
> "Watch out for division by 0."
> "Be careful with limit interchanges." (in analysis)
> "Funny things happen in infinite dimensions." (in analysis)
> "Watch out for over-specified problems." (in boundary value problems)
> "Be careful about 'post and rail' bugs." (in arithmetic)
> "Don't forget to set up the correct limits." (in calculus)
> "Watch out for this:
> det(A + B) ≠ det(A) + det (B) (where det represents the determinant)"
> "When changing the variable of integration, don't forget to
> recalculate the *dx*." (in calculus)
> "Nth term going to zero ==/=> convergence." (in analysis of series)

As can be seen, counter principles (CP's), like MP's, come in interpretive and imperative varieties.

CP's are warnings to the reader. One often adds a CP to one's knowledge base to try to prevent the occurance of known, perhaps personal, bugs and to temper the application of some MP's. For instance, in relation to the MP of the last section that suggests doing things componentwise, one could add a CP reminding one to be careful about this heuristic in infiinite dimensional settings. "Dangerous curves" in Bourbaki's ***Elements de Mathemetique*** serve similar functions.

Counter-principles are closely linked to certain counter-examples. The counter-example provides the *raison d'etre* for the CP. The following is taken from Hoffman [p.52] [1]:

---

[1] Note this CP also places further restrictions on the "component-wise" MP within that MP's intended setting.

"Beware of working with coordinates when discussing accumulation points."

"Consider in R^2

$$X_n = (0, 1), \quad \text{n odd}$$
$$X_n = (1, 0), \quad \text{n even.}$$

The sequence of the first coordinates is 0,1,0,1,...which has two accumulation points in R, 0 and 1. The sequence of second coordinates is 1,0,1,0,... and it has the same accumulation points. In particular, 0 is an accumulation point for the first and for the second coordinates. We cannot conclude that (0,0) is a point of accumulation of the sequence in R^2.

In some sense, CP's are like negatively biased results -- results stating that something is not true or not to be expected. Such wisdom is rarely given result status, since the "proofs" are often mere counter-examples.

Like other items, CP's need declaration of setting, but perhaps fewer annotations than MP's since they are warnings that don't usually send the reader off to try certain procedures and approaches. They prune rather than add to one's agenda of things to try.

As with mega-principles, some counter principles are trans-theoretic in their domain of applicability. For instance, the CP "Double roots are troublesome" is relevant in finite dimensional eigenanalysis as well as in the theory of numerical root finding. Another very general CP is the warning to "Be careful with interchanges of operations"; this caution applies to limit processes, functions, mappings, and operations like multiplication (e.g., in the domain of matrices), and just about any situation in which there is a composition of operators.

4.2 Procedural and Declarative Aspects of Concepts

As indicated previously (in Sections 2.5.2 and 4.1), a concept can be stated in more than one way: it can be expressed by its declarative statement or as a procedure or as the result of a procedure. Declarative and procedural formulation are different aspects of a concept.

For instance, the concept of eigenvalue in finite dimensional vector spaces may be defined as:

λ is an *eigenvalue* and a non-zero vector x is an *eigenvector* of a linear transformation A if

$$Ax = \lambda x.$$

An eigenvalue may also be expressed as the outcome of the procedure:

> SOLVE the characteristic equation: $\det(A - \lambda I) = 0$; the roots are eigenvalues.

The first presentation is the declarative definition of the eigenvalue concept, denoted as *DEC(eigenvalue)*, while the second is the procedure for it, *PROC(eigenvalue)* [2].

Even declarative definitions that don't seem intrinsically procedural can often be put in a procedural form. Familiar to all students of calculus are the definition and restatement of continuity:

> Def. A function f is *continuous at x* if for any $\epsilon > 0$, there exists a δ, such that
>
> $$|f(x) - f(y)| < \epsilon \text{ for all } |x - y| < \delta.$$
>
> Proc. A function f is *continuous at x* if whenever I give you an ϵ, you can find a δ (in closed form, if you are sufficiently clever), such that whenever $|x - y| < \delta$, we have $|f(x) - f(y)| < \epsilon$.

Other concepts such as the Gram-Schmidt idea are most naturally expressed as procedures. For these concepts, expresssion as a formal declarative definition obfuscates the idea. For instance, the following is the procedure known as the *Gram-Schmidt process*:

> Let $\{a_n\} \subset$ be a countable, linearly independent set in a Hilbert space.
>
> Let $b_1 = a_1 / \|a_1\|$. (get started)
>
> Let $t_2 = a_2 - \langle a_2, b_1\rangle\, b_1$ (orthogonalize).
>
> Let $b_2 = t_2 / \|t_2\|$. (normalize)
>
> Let $t_3 = a_3 - \langle a_3, b_2\rangle\, b_2 - \langle a_3, b_1\rangle\, b_1$.
>
> Let $b_3 = t_3 / \|t_3\| \ldots$
>
> Let $t_n = a_n - \Sigma_{i=1,\ldots,n-1} \langle a_n, b_i\rangle\, b_i$.
>
> Let $b_n = t_n / \|t_n\|$.
>
> Then $\{b_n\}$ is an orthonormal set and $\mathrm{span}\{b_n\}_{n=1,\ldots,K} = \mathrm{span}\{a_n\}_{n=1,\ldots,K}\ \forall K$.

[2] The identification of the DEC and PROC, which is established very early on in any study of eigenvalues, allows a geometric problem with no obvious "handles" to be turned into a routine algebraic, i.e., procedural, task [Halmos].

The following is its reformulation in delcarative form:

> If $\{a_k$ for $k = 1 \ldots n\}$ is a basis in R^n, then $\{b_k$ for $k = 1 \ldots n\}$ will be the "Gram-Schmidt" of this basis if the b_k have norm = 1, are mutually orthogonal, and the subspace spanned by the first K b_k is equal to the subspace spanned by the first K a_k (for $K = 1 \ldots n$) and also that the inner product of a_k and b_k is positive.

Given this definition, it is not even clear that such b_k exist.

It often seems that it is easier to move from a declarative to a procedural formulation, and that the procedural formulation is easier to understand. This is especially true for concepts that have a computational aspect such as eigenvalue or maxima-and-minima. For example, the restatement of conditions for critical points in terms of calculating the zeros of first and second derivatives is a reformulation that most calculus students take for granted as soon as they learn about it. It makes the location and determination of maxima, minima and inflection points an almost mechanical process (of course, limited to the case where these derivatives exist).

However, concepts that just have a procedural formulation are often slighted as concepts since they don't fit into the usual mathematical pattern of declarative definitions. Recognition of concepts that are mostly procedural is important so that they can be retrieved as individual entities and not just in conjunctuon with other items. Establishing their modularity makes it easier to reference and invoke them.

While the restatement of a concept in procedural form is often a trivial transformation for an experienced mathematician, it is often difficult for a neophyte. Notice that the procedures can be as abstract or symbolic as the definitions themselves; procedures do not make the ϵ's and δ's disappear. However, they use them like variables in a computer program. The point is that procedures present the abstraction as something to do, not just something to contemplate, and they impart to the ingredient concepts and conditions an order of execution and verification. Thus, a procedural presentation of a concept can be useful in tutoring situations since it provides a way of working with an idea. It can force a student to consider his problem one step at a time. Such approaches help him build confidence, and confidence seems to be a necessary (but not sufficient) condition for expertise.

For these reasons, the representation of a concept item contains both formal definition and explicit procedural information. This seems to be more important for concepts than for examples or results. To use an example or a result one doesn't need to know or remember how to construct or prove it, whereas to use a concept, one needs to do more than just cite the definition.

4.3 The Arrows of Concepts-space

As mentioned earlier one of the reasons for grouping definitions and heuristic principles together in Concepts-space was the desire to keep track of how they evolve from one another. This is a basic concern of what Piaget calls "genetic" aspects of epistemology. It is also stressed by Polya [MD].

Pedagogical ordering is but one level of describing the relations between concepts. One can describe more fully the relation that is summarized by the arrow. In particular, one can describe the process by which a concept is derived from another and thus give a deeper reason for the ordering.

Sometimes there is nothing very deep about the relation A --> B other than A is used in the formulation of B. Other times, the relation is more intrinsic in that B is evolved from A by a "genetic" process such as specialization, generalization, induction or analogy, to name a few of the famous ones described by Polya.

In his book *Proofs and Refutations*, Lakatos discusses concept formation. He shows how a concept can be generated from analysis of (failed) proofs and conjectures. In particular, he describes "monster barring", a process in which one refines a concept by defining certain troublesome (pathological) cases out of the definition, and "exception barring" in which one narrows the class (or setting) of objects for which the concept is to be applied.

One of the most common ways to form a principle is to *paraphrase* a definition (into a principle) or a principle (into a definition). Other ways are to abstract the heuristic content of results and to summarize experience with examples. Paraphrase is a transformation that occurs within Concepts-space; the others involve dual relations to the other spaces.

Many MP's are *folksy* restatements of formal definitions. For instance, saying that "continuity means you don't have to lift your pencil" is a very informal recasting of the formal definition of continuity [3]. In his chapter on eigenvalues, Strang presents many paraphrases of the eigenvalue definition (See Chapter 6, Section 2). Some concepts are even derived by formalizing certain heuristics.

[3] Historically the genesis was the other way. But today the concept is defined first and then explained.

The genetic information carried by the arrows -- generalization, specialization, analogy -- are very important in mathematics. They are also important in other domains such as the rule-based domains considered by Goldstein [Goldstein 1978]. He singles out certain relations between rules -- generalization, specialization, refinement, debugging -- and uses them to create his "genetic graph".

4.4 Very General Concepts

Many of the concepts -- ideas and principles -- that belong to one theory do in fact belong to many theories. Such concepts are very general and could be said to be *trans-theoretic*.

4.4.1 Ubiquitous Themes in Mathematics

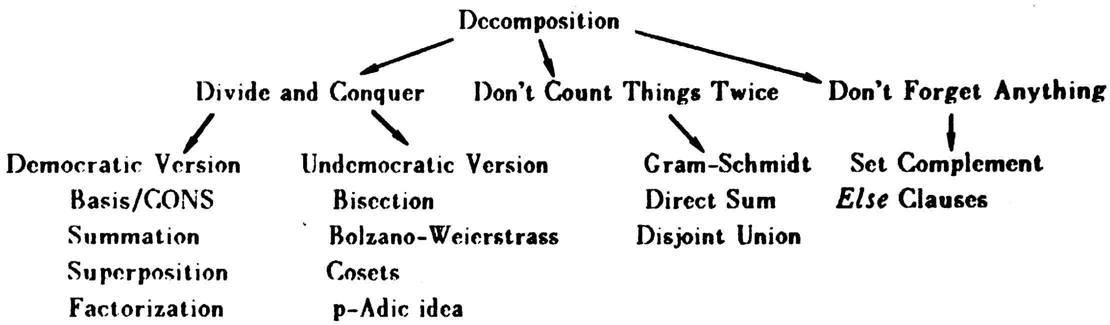
Some concepts while made originally in a particular theory, are later found in other theories that are quite different in focus and flavor. Bourbaki discusses certain very general ideas, "mother structures", which pervade all of mathematics [Bourbaki1950]. The "group" concept is one of these. Originally an idea of algebra, it has found its way into analysis and topology. According to Bourbaki it is one of the three most fundamental ideas of mathematics; the others are "order relation" and "topology". The latter includes the basic concepts of "closeness" and "continuity". Piaget has found these concepts to be ubiquitious in cognitive development as well [Piaget 1968]. Such general concepts are truly *meta-level* ideas and "lie above" many theories, or said differently, they are at the very foundations of mathematics.

In addition to the three general concepts cited by Bourbaki, there are many others, such as:

Decomposition
Do it again
Beg the question
Closeness
Perturb/Change
IVP vs. BVP (Initial Value vs. Boundary Value Problem)

Such general ideas spawn many concepts in mathematics. For instance, *Do It Again* leads to the concepts of Iterate, Recurse, Induct, and Pass-to-the-Limit. Perturb/Change has Continuity/Jump, Stability, and Homotopy as descendent ideas. Closeness leads to the ideas of Distance/Metric, Length/Norm, and Area/Measure.

The Decomposition idea has the following rich genealogy:



4.4.2 Meta-level Principles

Some very general principles are what are often called "control" or "strategic knowledge" [Brown 1977]. An example of such a general principle is the heuristic "Check things out on a reference example", "Try applying the MP's you know, if you are stuck", "Pay more attention to the culminating theorems than the technical ones". Less control oriented, but yet of a very general nature, is "Extreme points are almost always interesting". This heuristic was one of the prime pieces of knowledge in Lenat's recent program [Lenat 1976]. These very general imperatives and interpretations are part of the knowledge that one has and employs in order to gain understanding. We will come back to them when we discuss a model for understanding in Chapter 6.

One particularly powerful piece of strategic advice in any theory is the:

> Restriction Principle - *Refine or limit the current domain of discussion to a more restricted setting or a specific item.*

Many important examples, principles and test situations are the results of applying this idea.

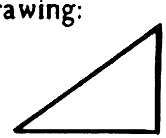
For instance, one can produce the following nested sequence of settings and specializations by successively applying the restriction idea:

$$BL(X,Y) \text{ -> } BL(X,X) \text{ -> } BL(R^n,R^n) \text{ -> } M_n(R) \text{ -> } M_2(R) \text{ ->} M_2(\{0,1\})$$

The restriction onto $M_2(R)$ (from any bigger context) is in actuality the mega-principle "Try the 2x2 case," and thus, it -- MP(2X2) -- may be considered generated by the restriction principle. When the restriction has been carried further, the outcome is the "Basic 16" reference example of 2x2 matrices whose entries are 0's and 1's.

Thus the Restriction Principle maps general settings and items onto ones of increased restriction. This principle is often used in tandem with a nested chain of settings (see Chapter 2). However, the emphasis of restriction is downward to more specific settings whereas that of nesting is upward to more general settings.

Instantiating a model example can be considered an application of the Restriction Principle. For example, specifying the Pythagorean triple (3,4,5) for the right triangle model leads to the specific drawing:



In number theory, the restriction idea leads to the advice "Check out a conjecture for prime numbers" and more specifically to "Check the cases of p = 2, 3, 5 and 7."

In some cases the "image" of the restriction mapping does in fact logically span the general situation. The sufficiency is often stated by major theorems of Results-space. For instance, the Chinese Remainder Theorem guarentees that the prime cases (Z/pZ) do represent the general case (Z/mZ). The equivalency theorem stating that all linear spaces of dimension n are isomorphic indicates the generality of restriction to the standard basis. These instances are consequences of the *projective* nature of certain restricted cases (they are generalizable and one knows the generalizing map). In mathematics, the key to many problems is to apply the:

> Projection Principle - *Restrict onto projective situations and then lift back up.*

4.4.3 Analogy and Identification Between Theories

Some of the most striking realizations in acquiring mathematical understanding occur when one makes a link between items in different theories. Often this association is established by a shared concept item, or in other words, by a dual relation that operates between theories.

Often the very general meta-level ideas of the last section are at the root of such associations. These identifications often take the form of statements such as "this is a stability problem," or "the basic plan is decomposition".

If enough ideas are shared between diverse theories, the theories could be said to be related through the dual idea operating on the meta-level. Such shared concepts are often the basis of striking anaolgies.

There are several types of correspondences and analogies that operate between items in diverse theories. One common form is that the two items in the different theories are both instantions of some more general concept. For instance, the CONS (Complete Orthonormal Set) idea allows identification of the method of Gauss sums in number theory, Fourier series in Hilbert spaces, and ordinary bases in finite dimensional vector spaces.

Analogy is often obtained through the procedural formulations of the concepts whose symbols, steps and plans can be correlated through a correspondence map. For instance, the Gram-Schmidt process allows a correspondence between a basic result in Hilbert space theory with a result on subadditivity from measure theory:

> Result 1 - In a Hilbert space, any countable, linearly independent set can be orthonormalized.

This next result is strikingly similar to Result 1. In fact, its proof is nothing more than the Gram-Schmidt process in disguise; here "Gram-Schmidt" (see Section 4.2) is done on a collection of sets rather than vectors..

> Result 2 - In a measure space, any countably additive measure is subadditive.

Proof:
Let m be a countably additive measure.
Let $\{A_n\} \subset X$ be a countable family of sets.
Let $B_1 = A_1$.
$B_2 = A_2 - (A_2 \cap B_1)$.
$B_3 = A_3 - (A_3 \cap B_1) - (A_3 \cap B_2)$.
...
$B_n = A_n - \cup_{i=1,\ldots,n-1} (A_n \cap B_i)$
Then $\{B_n\}$ is a disjoint family of sets and $\cup_{n=1,\ldots,K} B_n = \cup_{n=1,\ldots,K} A_n \ \forall K$.
Thus $m(\cup A_n) = m(\cup B_n) = \Sigma m(B_n) \leq \Sigma m(A_n)$.

When one recognizes such striking similarites between entire theories, one often says, "Aha, these theories are really *the same*." Such recognition of commonality could be said to be identification via the dual idea, but on a higher level: i.e., not just between two items in the same theory, but between two theories. For example, the theories of plane geometry and analytic geometry are strongly related; they share R^2 as their setting and a large stock of examples of triangles, parallelograms, and conic sections.

The theory of quadratic reciprocity in number theory [Ireland and Rosen, Chapters

6 and 8] and the theory of Hilbert spaces [Rudin 1966, Chapter 4] share a strong same-ness relation. One of the connecting links is through the Complete Orthonormal Set idea; another is through concepts dealing with measure and integration. In view of these ideas their central techniques of Gauss sums and Fourier series, respectively, are really the same:

For f a real-valued, 2π-periodic function, in $L_1(-\pi,\pi)$,

the Fourier coefficients of f are:

$$\hat{f}(n) = (1/2\pi) \int_{-\pi}^{\pi} f(x)\, e^{-inx} dx \qquad (n \in Z)$$

and the Fourier series of f is: $f(x) = \Sigma_n \hat{f}(n)\, e^{inx}$

A Gauss sum on the finite field Z/pZ belonging to the character χ is:

$$G_a(\chi) = \Sigma_t\, \chi(t)\, \zeta^{at} \qquad (a \neq 0)$$

where the sum is over all t in Z/pZ, and $\zeta = e^{2\pi i/p}$.

For f:Z->C a p-periodic function (i.e., f(n+p)=f(n)), such as the character χ,

the Fourier coefficients of f are: $\hat{f}(a) = (1/p)\Sigma_t\, f(t)\, \zeta^{-at}$

and the "finite" Fourier series of f is: $f(t) = \Sigma_a \hat{f}(a)\, \zeta^{at}$

It can be seen that $\hat{\chi}(a) = (1/p)G_{-a}(\chi)$.

Thus, the Gauss sum "is" the finite Fourier transform of the character χ.

Identifying these objects from number theory with their counter-parts from analysis, and thinking in terms of inner product spaces gives them motivation and structure and suggests a whole cluster of ideas to investigate.

Chapter 5. RESULTS-SPACE

5.1 Classification of Result Items

This section addresses the classification of results based on their role in understanding. The logical dependency of results, i.e., which results are needed to prove other results, is represented by the deductive relation of the results graph. Other logic-related ideas such as *generality* and *strength* are treated separately (Sections 5.3 and 5.4).

5.1.1 Basic Results

Basic results establish the fundamental properties of concepts and examples. They are frequently starting nodes in Results-space and thus are analogous to start-up examples. In a presentation of a theory, basic results closely follow the introduction of new concepts; they elaborate on definitions and often provide procedural formulations. Thus basic results flesh out one's knowledge of a concept and are preparation for further work. Basic results also build dual connections between the three representation spaces. Building of links is one of the primary functions of basic results. For instance, they often show that certain concepts are related via the dual idea. In other words, they knit new items into one's established knowledge and thus, are often the first steps taken to meld new with old knowledge.

Many basic results define procedural aspects of concepts. For instance, in the theory of finite dimensional eigenanalysis, the following very important result establishes the procedural formulation of eigenvalue:

$$\text{For } A \in M_n(C),\ \lambda \in \text{Spec}(A) \text{ iff } \det(A-\lambda I)=0.$$

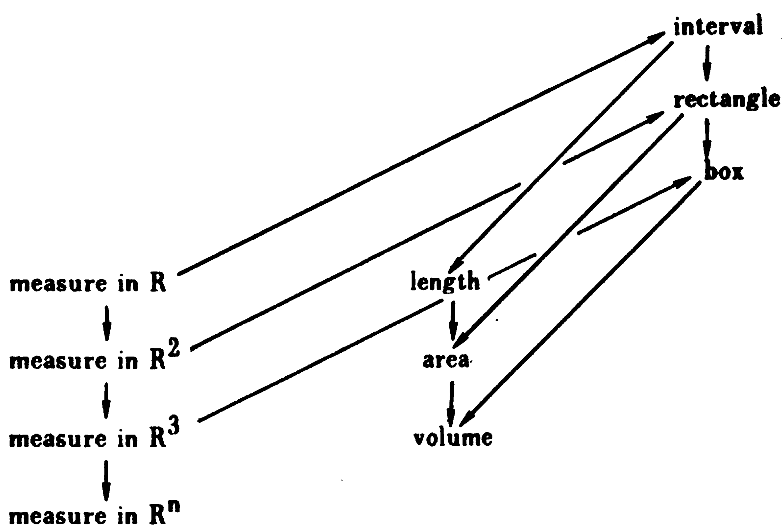
This result transforms a declarative definition to a well-defined computation, in this case, by reducing a geometric criterion to a purely algebraic manipulation of determinants. Basic results which are procedural in character are important since they establish well-defined methods and procedures with which to work with a concept.

Linking is a very important aspect of basic results; it is the way in which they obtain their basic or even start-up quality. The links can be between different aspects of the same item, as in the last result, or between two items, as in the next result.

A basic result linking the concept of outer measure with the fundamental reference example of intervals on the real line is [Royden, p.54]:

Proposition - The outer measure of an interval is its length.

This result knits the concept of outer measure into the network of knowledge about intervals and length. This connection is important for understanding the concept of measure since the concept of outer measure is usually formulated in terms of open sets and this result establishes the fact that length, one of the most basic of all mathematical ideas, is at the root of the idea. It is an excellent example of the dual idea at work: it identifies two branches of Concepts-space via a reference example. Since the concept of length leads directly to the concepts of area and volume, one is lead, through the interval example, to a analogous path that ties measure in higher dimensions to areas and volumes. It also suggests tying in the constructional chain of intervals-rectangles-boxes as examples of these ideas:



5.1.2 Key Results

Key results establish the fundamental, underlying facts of a theory. They are used repeatedly once they have been proved. Thus they are analogous to reference examples. Key results provide intermediate goals for one to reach in one's understanding of a subject.

Like principal cadence points in music, they are points at which the work can be tied together and summarized before it is carried further. Key results can link together different entry points into a theory. (An entry point is any item at which the study of a theory can begin.) Thus many key results are equivalency results that show two concepts are mathematically equivalent.

Key results provide items at which to pause and recast one's knowledge. This quality of results was remarked upon by Hadamard [1954]. Thus they are excellent results with which to conclude a lecture. They also motivate reviews of one's knowledge. They provide a temporary cap to a deductive sequence. In a sense, then, they serve to release some of the tension that is built-up when one is driving to complete a chain of reasoning. Key results are the kind of result one goes back to in order to regenerate a deductive sequence of results.

The result given before as an example of a basic result, is also a very important key result in eigenanalysis. It ties together the geometric ($Ax=\lambda x$) and the algebraic ($\det(A-\lambda I)=0$) definitions (entry points). It is an example of a basic result that also functions as a key result in much the same way as a start-up example can serve as a reference example.

Other examples of key results are the Riesz-Fischer Theorem which states that L_p-spaces are complete and caps an introductory sequence of results on L_p-spaces. The Bolzano-Weierstrass Theorem is a key result from real analysis, which already has been used several times in this report as an example (e.g., in Section 3.1.2)

A key equivalency result tying together many different entry points into the study of projective modules is the Wedderburn Structure Theorem [Jans, p.12]: The exact meaning of the terms in this result are unimportant for this discussion; what is worth noticing is the variety and multiplicity of ways in which the concept of "projective module" can be handled:

> Theorem (Wedderburn) - For R, a ring with identity, the following statements are equivalent:
> (1) Every R-module is projective.
> (2) Every short exact sequence of R-modules splits.
> (3) Every R-module is injective.
> (4) Every non-zero R-module is a direct sum of simple R-modules.
> (5) R is a direct sum of a finite number of left ideals generated by a set of orthogonal idempotents...
> (6) R is a direct sum of two-sided ideals, each of which is isomorphic to a matrix algebra over a division ring.

5.1.3 Culminating Results

Culminating results are the results to which the theory and its presentation drive. They are goals of both the deductive reasoning, pedagogical exposition and one's understanding. This coalescing of logical and pedagogical purposes is one of the reasons that culminating results are so outstanding.

To test if a result is a culminating result, one asks, "If this result is omitted, has the main point of the theory been missed?" If the answer is "yes," the result is a culminating theorem. If a theory is extensive it may have more than one culminating result.

Granted this question is hard to ask while one is in the midst of learning a theory, but it is a question that can, and probably should, be asked by teachers and by students who are reviewing and trying to understand on a deeper level.

Culminating results are special in one's understanding: without them one's understanding is incomplete. Culminating results are the punch lines of a theory. Like the final cadence a piece of music, they tie together what has been stated and developed before.

Examples of some well-known culminating results from various fields of mathematics are:

The Fundamental Theorem of Calculus
The Riesz Representation Theorem (analysis)
Law of Quadratic Reciprocity (number theory)
Cauchy Integral Formula (complex analysis)
Jordan Normal Form Theorem (linear algebra)
The Spectral Theorem (functional analysis)

Because they tie one's knowledge together, it is natural that many culminating results are equivalency or classification results. *Classification* results provide pigeon holes into which objects may be sorted. An example of such a result is the following theorem which provides three exhaustive classes for all "real division rings" [Herstein]:

Theorem (Frobenius) - D a division ring
If D is algebraic over R, the real numbers,
Then D is isomorphic to one of the following:
R, the field of real numbers
C, the field of complex numbers, or
the division ring of real quarternions.

5.1.4 Technical Results

Technical results treat technical points in a theory; they work out nitty-gritty details. When they are used in a way preliminary for another result, i.e., as lemmas, they provide technical "scaffolding" [Davis 1972, p.259] from which to prove succeeding results. Technical results usually do not have the potential for adding very much to one's understanding since their focus is so narrow. Like some counter-examples, their limited use makes them *hapax legomena* [Freudenthal 1973]. They are some of the first results to fade from memory or to be dropped from discussion when an overview is taken.

For instance, at the beginning of the study of L_p-spaces, one defines conjugate p's and q's [1] and then proves the following technicality [Royden, p.112]:

Lemma - Let A, B, s be real numbers, such that $A,B \geq 0$ and $0 < s < 1$.
Then $A^sB^{(1-s)} \leq sA + (1-s)B$ with equality only if $A = B$.

[1] Real numbers p and q are conjugate if $(1/p + (1/q) = 1$ where $1 < p,q < \infty$.

This is a result whose statement and proof are technical. Neither in itself adds very much to one's understanding of L_p-spaces (or real numbers). Its main function is to lay the technical groundwork for proving Hölder's Inequality ($\|fg\|_1 \leq \|f\|_p \|g\|_q$).

Technical results are often subsumed in the proof of the results of which they are immediate logical predecessors. For instance, the above lemma cited from Royden is the first paragraph of the proof of Hölder's Inequality in Dunford and Schwartz [p. 119]. However, isolation of technical results is useful because it facilitates their omission which enables one to separate the main steps of the proof from the low-level details and the technical scaffolding. It also makes it easy to re-arrange the order of proof, presentation or perusal.

Since technicalities serve a narrow function in the deduction and understanding of a theory, they may not be worth much effort to understand them thoroughly (see Section 11.5). Whereas culminating results were at the focus of logical and pedagogical considerations, technical results are antithetically removed from them. This is why so many technical results in themselves are unimportant in the theory as a whole.

5.1.5 Transitional Results

Transitional results lay the logical groundwork for future results; they are the deductive *stepping-stones* of a theory. They point forward to key results further down the results graph and derive their importance by helping to establish these target results.

If one were to prune a key result (and its successors) from the Results-graph, the transitional results leading deductively to it would be left hanging because their deductive sequence could not attain its logical or pedagogical goal (and thus in some sense one may as well also prune the transitional result). Key results are often reached by splicing together (e.g., with *modus ponens*) a series of transitional results.

There exists an analogy between the Basic-Key-Culminating spectrum of results and the Sart-up-Reference-Model spectrum of examples.

5.2 Vertical and Horizontal Generality

It is often useful to distinguish two kinds of comparisons between results that often both fall under the name of *generality*: *vertical* and *horizontal* generality. (The spatial metaphors come from picturing settings arranged in a manner in which the higher placed setting is more general.) The term *horizontal* is used when two results stated *within the same setting* are compared; the terms *stronger* and *weaker* also apply is this case. *Vertical* refers to comparison of two results *in two different settings, one of which is more general than the other.* (See Section 2.4.3 on Settings.)

Thus, if we represent one result as $R_1(S_1,H_1,C_1)$ and another as $R_2(S_2,H_2,C_2)$, there is a possibility for a:

(1) horizontal generality when $S_1 = S_2$;
(2) vertical generality when $S_1 \subset S_2$.

5.2.1 Vertical Generality

If two results have essentially the same hypotheses and conclusions but the setting of Result 2 is more general than the setting of Result 1, then Result 2 can be said to be ***more general*** than Result 1, which we indicate as:

$$\text{Result}_1(S_1,H_1,C_1) < \text{Result}_2(S_2,H_2,C_2) \qquad \text{(where } S_1 \subset S_2\text{)}$$

or simply as $\text{Result}_1 < \text{Result}_2$

We leave the relation between the hypotheses and conclusions of the two results loosely specified by saying that roughly speaking they should deal with the same ideas. The simplest case of course is when they are exactly the same: $H_1 = H_2$ and $C_1 = C_2$. Another nice case is where H_1 implies H_2 (H_1 ==> H_2) and C_2 implies C_1 (C_1 <== C_2):

$$\begin{array}{lccc} \text{Result}_2\text{:} & S_2, & H_2 & \Longrightarrow C_2 \\ & \cup\!\!\uparrow & \Uparrow & \Downarrow \\ \text{Result}_1\text{:} & S_1, & H_1 & \Longrightarrow C_1 \end{array}$$

The next examples involve generalization with respect to the following chain of settings:

fdvs $\subset$ Hilbert space $\subset$ Banach space $\subset$ normed linear space $\subset$ vector space

Consider the two following classification results on vector and Hilbert spaces [Halmos 1942, p.15], [Dunford and Schwartz, p.254]:

> **Result 1** - All finite dimensional vector spaces over a field F of the same dimension n, are isomorphic to F^n. Hence when F = R or F = C, to R^n or C^n.

> **Result 2** - All Hilbert spaces over a field F of the same dimension are isometrically isomorphic and hence equivalent to an l_2-space of the suitable dimension.

The main part of these two results can be written as:

Result 1 - In fdvs, $\dim_F X = \dim_F Y$ ===> $X \simeq Y$ (as fdvs).

Result 2 - In H-sp, $\dim_F X = \dim_F Y$ ===> $X \simeq Y$ (as H-sp).

In this case, the two hypotheses and the two conclusions are really the same, and we have the simplest type of comparison. [2]

As another example of vertical generalization, consider the two following results which characterize compactness [Dunford and Schwartz]:

> Result 1 (The Heine-Borel Theorem) - In a finite dimensional vector space, a set is compact iff it is closed and bounded.
>
> Result 2 - In a Banach space, a set is compact iff closed and totally-bounded.

The setting of Result 2 is more general than that of Result 1 since every finite dimensional vector space is a Hilbert space; their hypotheses are the same; and the conclusions of Result 2 are stronger than those of Result 1 since totally bounded implies bounded.

5.2.2 Horizontal Generality

Instead of examining statements in different settings, one can hold the setting fixed and compare related results. The *strongest* result concludes the most, but requires the fewest hypotheses.

To strengthen a given result, one deletes hypotheses and/or adds conclusions and then proves the new statement. In a given setting, to achieve the strongest possible result one deletes as many hypotheses as possible--i.e., minimizes the hypotheses--while adding as many conclusions as possible--i.e., maximizes the conclusions. Thus, achieving the strongest result is a kind of *mini-max* problem.

At the opposite end of the spectrum are the *weakest* results which conclude the least but require the most. Mathematicians do not tend to be concerned with weakening results, except perhaps in the context of generating homework problems.

---

[2] Note how close this discussion is getting to some ideas from category theory, such as that of an isomorphism within the category of objects being discussed.

Attempts at strengthening are often done in an incremental fashion--one hypothesis or conclusion altered at a time--and very often involve only the hypotheses. Sometimes however, breathtaking breakthroughs can be achieved by bold vertical generalization followed by restriction; the generalization may allow one to invoke very powerful general tools. Thus, while strength is an attribute of results within a a given setting, strengthening is a process which does not necessarily take place within one setting.

5.3 The Architecture of Proof

In classifying results, one cannot avoid noticing that there are also broad categories of proof techniques. This section discusses some of these techniques.

There are several aspects of a proof. There is the external perspective describing how this proof fits in deductively with its predecessors and there is the internal description of how the proof, itself, is structured. The internal structure can be described on a surface level by its logical plan of attack and the main idea of the technique executing this plan. Its fine-structure can be described by the proof's individual steps and the specific reasoning used to establish these steps.

5.3.1 Some High Level Descriptors

The overall internal structure of a proof can be described by its *logical attack*, for example, as a direct proof, indirect proof or proof by contradiction, contrapositive proof, proof by induction, proof by cases, proof by exhaustion.

Among the adjectives providing the highest level description of the proof and its external relation to other theory items, are *stand-alone*, *splicing*, and *corollary*.

A *stand-alone* proof establishes its result with little or no reference to the result's logical predecessors. A stand-alone proof is often accomplished by direct calculation (as in the binomial theorem) or by matching both sides of an equality (as in trigonometric identities).

Splicing proofs [Davis 1972, p. 259] build deductively on predecessor results by splicing two results together usually with *modus ponens*: i.e., the conclusions of the first result as hypotheses of the next: $H_1 \Longrightarrow C_1$ and then $H_2 \Longrightarrow C_2$ where $C_1 = H_2$; many results from plane geometry are of the splicing variety (e.g., see [Jacobs].)

In some sense stand-alone and splicing describe opposite ends of a spectrum; most proofs combine aspects of both. Also, different levels of detail convey different senses of the degree of splicing or isolation between the internal steps of a proof.

Corollary results fall out of their predecessors with slight modification of the statement or proof of the predecessor. There are several types of corollary results. A corollary may:

(1) isolate and extract a result or procedure developed in the proof of its predecessor, such as the Gram-Schimdit porcess.

(2) restate its predecessor in the specific instance of an example. An example of such an *instantiation* is the following [Rudin 1964, Chapter 2]:

> Theorem - The Cartesian product of countable sets is countable.
> Corollary - The rational numbers $Q \subset Z \times Z$ are countable.

(3) restate its predecessor in a more restricted setting. The following is an example of such a *restriction* [Rudin, p.77, theorems 4.14 and 4.15]:

> Theorem - In topological spaces, the continuous image of a compact set is compact.
> Corollary - An R^n-valued continuous function on a compact set is bounded.

The last result also involved some restatement and weakening of the concept of compactness: first, compactness is reformulated as "closed and bounded" (which is an equivalency valid in the setting R^n) and secondly, only the "bounded" part of compactness retained. Rephrasing and weakening of the predecessor is very common in the statement of corollaries.

5.3.2 Vertical Techniques

Vertical techniques occur when the proof is established by working in two or more different settings that belong to the same chain of generalization.

Bootstrapping and *wlog reduction* proofs both establish a result by a nesting of sub-results according to the generality of their setting. This can be pictured as the ladder:

R(S,H,C)
|
R'(S',H,C)
|
R"(S",H,C)

where upward direction indicates increasing generality of the setting ($S'' \subset S' \subset S$). Note that the hypothesis and conclusions, H and C, are the same. This is a special case of the situation discussed previously for vertically generalizing results. The particular *ladder* pictured here has three stages.

A *bootstrapping* result works up the ladder and proves increasingly more general statements. For instance, it would use R" plus some lifting technique to establish R', and then R' to prove R.[3] A *WLOG reduction* (without loss of generality) proof works in the opposite downward direction by showing that it is sufficient to prove a more restricted result. This is often established in effect by exhibiting the lifting technique of the bootstrapping process at each reduction step.

These two antithetical techniques may be pictured as:

$$\text{bootstrapping}\ \Big\uparrow \quad \begin{array}{c} R \\ | \\ R' \\ | \\ R'' \end{array} \quad \Big\downarrow\ \text{wlog reduction}$$

The Riesz Representation Theorem is often established by these vertical techniques:

> <u>Riesz Representation Theorem</u> - In $L_p^*(X,M,\mu)$ where μ is a σ-finite measure and $1 \leq p < \infty$, if F is a bounded linear functional on $L_p(X,M,\mu)$, then there exists a unique element $g \in L_q(X,M,\mu)$, (where $(1/p)+(1/q)=1$) such that
> $$F(f) = \int fg\, d\mu, \text{ for all } f \in L_p(X,M,\mu).$$
> Also $\|F\| = \|g\|_q$.

For instance, Royden's proof [p.246] uses a bootstrapping approach in which he first proves the result for μ a finite measure and then lifts this result via the technique of a sequence and a convergence theorem to establish the case of μ a σ-finite measure:

$$\begin{array}{c} L_p^*(X,M,\mu\ \sigma\text{-finite}) \\ \uparrow \\ \text{via } X_n \to X \text{ and Monotone Convergence Theorem} \\ | \\ L_p^*(X,M,\mu \text{ finite}) \end{array}$$

The same proof also works in the particularized setting of Lebesgue measure on the real line (R,M,m), where some mathematicians do it top-down to produce a wlog reduction proof [Banks, p. 134].

[3] In some sense, an induction proof can be considered bootstrapping from the n^{th} to the $(n+1)^{st}$ case, especially when there is a dimension argument involved, as is the case in many proofs of results in linear algebra.

Several other theorems from measure theory can also be proved via vertical techniques. For instance, the Radon-Nikodym Theorem can be established by bootstrapping through the following four-runged ladder [Banks, p. 162; Rudin 1966, p. 124]:

λ,μ both σ-finite signed measures
λ,μ both σ-finite measures; λ signed measure, μ positive measure
λ,μ both finite; λ signed measure, μ positive measure
λ,μ both finite, positive measures

In summary, vertical proof techniques can be described as an ordered pair:

(*technique, lifting*)

such as (bootstrapping, convergence argument), (wlog, dimension argument).

It is important that one know these techniques, not only so that one can understand the proof of an established result more readily, but also so that one can use them to prove new results.

5.3.3 Horizontal Techniques

Among the techniques most often used within a fixed setting are *divide and conquer* and *patch* proofs. These techniques are *horizontal* in nature since they are set within one level of setting.

5.3.3.1 Divide and Conquer Proofs

A *divide and conquer* proof splits the problem into clearcut independent subproblems whose "union" is the original problem, proves each piece separately or in parallel, and then recombines the component results to generate a proof of the original result. A *clearcut* subdivision is defined as a partition whose pieces are disjoint or intersect only at their boundaries, as for example in the decomposition of the real numbers into positive (> 0) and non-positive (≤ 0) subsets or subsets of positive, negative and zero [4].

The requirement of independence means that in the case of a non-disjoint decomposition, boundary or edge effects cancel. The recombination is a simple "anding" or concatenation of the individual sub-proofs. The number of subpieces is almost always small, finite or countable.

[4] Clearcut could be said to be 'almost-disjoint', that is, disjoint except on a set of measure zero (for some suitable measure), as is the set $\{0\}$ in the second dividing up of the real numbers.

The componentwise treatment given vector space results is a model of the divide and conquer approach. For example, to prove two linear transformations are equal on a space, it suffices to show they are the same on a basis (i.e., in each component).

Another good example is the proof of the following result which describes the units (invertible elements) of Z/mZ in terms of those of the sub-pieces Z/p^aZ [Ireland and Rosen, p.45]:

> Theorem - Let $m = 2^a p_1^{a_1} p_2^{a_2} \dots p_k^{a_k}$ be the prime decomposition of m.
>
> Then $U(Z/mZ) \sim U(Z/2^aZ) \times U(Z/p_1^{a_1} Z) \times \dots \times U(Z/p_k^{a_k} Z)$

The sub-pieces of the proof are characterizations of $U(Z/2^aZ)$ and $U(Z/p^aZ)$, which are recombined via the Chinese Remainder Theorem. All the $U(Z/p^aZ)$ are done generically in parallel.

A divide and conquer approach which treats each subpiece with equal emphasis is further described as a *democratic* divide and conquer approach. Direct sum and product decompositions, such as in the last result, typify the democratic treatment.

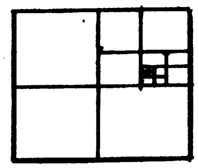
A divide and conquer proof which places unequal emphasis on the subpieces is called an *undemocratic* divide and conquer proof. Typically, one piece bears the onus of the entire proof.

An excellent example of an undemocratic divide and conquer strategy is demonstrated by the Bolzano-Weierstrass Theorem [Rudin 1964, p. 35]:

> Bolzano-Weierstrass Theorem - In R^k, if E is a bounded set with an infinite number of points, then E has a limit point in R^k.

The proof in the two dimensional case of R^2 begins with a WLOG argument that reduces the proof to consideration of a rectangle or unit box (a "2-cell"); it then proceeds as follows:

> Divide the box in four smaller boxes by halving each side. In one box there must be an infinite number of points, else the total number of points would be finite.
>
> Divide that box into four; one of these four must have an infinite number of points.
>
> Continue this process until an infinite number of points is trapped within a very small box. That is, the points are all within some "epsilon" of one another and therefore there must be a limit point.

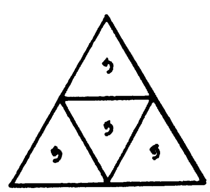


The very same undemocratic process underlies the Goursat proof of Cauchy's Theorem on a triangle of which the first few lines follow [Nevanlinna and Paatero, p. 118]:

Cauchy's Theorem - In C the complex plane, if the function f(z) is analytic in a triangle T, then

$$\int_T f(z)\,dz = 0.$$

Proof (due to Goursat):
Let $I = \int_T f(z)dz$ and assume $I \neq 0$.
Decompose T into four congruent triangles (by joining the midpoints of the sides of T).
Then $\int_T f(z)dz = \Sigma_{i=1,2,3,4} \int_{Ti} f(z)dz$ where the integrals are taken in the positive sense with respect to the enclosed areas of the triangle and sub-triangles:



(Notice boundary effects cancel.)

At least one of the four triangles in the subdivision - call it T_1 - must be such that the integral over it is non-zero and in particular, such that: $|I_1| = | \int_{T1} f(z)dz| \geq |I|/4$, i.e., $|I| \leq 4\,|I_1|$

By induction, we obtain a sequence of nested similar triangles:

$$T \supset T_1 \supset \ldots \supset T_n \supset \ldots$$

with boundaries decreasing by a factor of two ($\partial T_i = \partial T\, 2^{-i}$), areas decreasing by a factor of 4 ($A_i = A\, 4^{-n}$),

By making estimations and using compactness, a contradiction is pushed and thus the origianl integral, I must indeed be zero.

The general Cauchy-Goursat Theorem is then produced by arguing (a horizontal generalization) from the case of a triangle, to that of a disk, to that of a simply-connected region. Some authors prefer to start the process from the case of a rectang: [Ahlfors, p.109 and Nehari, p.82] in which case the picture is exactly the one already shown for the Bolzano-Weierstrass Theorem.

This undemocratic process of the Cauchy-Goursat and Bolzano-Weierstrass Theorems (which we shall refer to as the B-W trapping process) is used in R^1 (together with the Mean Value Theorem) to produce the numerical root finding method of the binary chop or bisection method [Acton, p. 179]. The B-W trapping process is an example of a specific technique used so frequently that it deserves to be called *standard* [5].

All of the above proofs using the B-W process put off to the (n+1)st stage what could be attempted at the nth: that is, they all *beg the question*. They all work because the sequences they generate run into the wall of finiteness of some kind (i.e., compactness) and are trapped.

Involved in many of these undemocratic divide and conquer situations is the Dirichlet *pigeon hole principle* [Herstein, p.90] which says essentially, "If there are K pigeon holes containing (K+1) pigeons, there must be at least one pigeon hole with more than one pigeon". In the Goursat proof, the conclusion that there is at least one triangle T_1 such that $|I_1| \geq |I|/4$ is a consequence of such reasoning. The selection of the next box in the Bolzano-Weierstrass theorem is made by the DPHP variant: if $E = \cup E_n$ is infinite, at least one E_n is also infinite.

Thus far, all the divide and conquer situations presented have been *explicit*, that is, all the subpieces are labeled or referenced at least once, even though some are then immediately forgotten as in the B-W trapping technique. An *implicit* divide and conquer approach divides the universe at hand into complimentary sets E and E^C, the haves and have-nots of whatever is being considered, and mentions only one of them. When all the emphasis is placed on one of these, one has an undemocratic strategy, as in the following result:

> Proposition - In a measure space, (X,M,μ)
> If f is a nonnegative, measurable function such that $\int_X f\, d\mu = 0$,
> Then $f = 0$ a.e.
>
> Proof (by contradiction):
> Else there exists a measureable subset of X, call it E, of finite measure, such that $\mu(E) \neq 0$ and $f \neq 0$ on E, and ($f=0$ on E^C).

[5] Other such "standard tricks" are: integration by parts (IBP), geometric series, and Taylor series expansion. These three techniques are so powerful that it sometimes seems that one could do "all" of mathematics with them alone.

Notice that the undemocratic version of divide and conquer is particularly suited to proof by contradiction, whereas the democratic version is usually employed in direct proofs.

5.3.3.2 Patch Proofs

A *patch* proof splits the problem into subpieces which are neither disjoint nor independent, works each piece separately, and then patches the subresults together so that they agree on their overlaps or match-up at their joins.

The hallmark of a patch proof is the necessity of matching-up subpieces, i.e., the mutual dependence of subproblems, whereas that of a divide and conquer proof is that the matching-up is not necessary (because of disjointness or mutual cancellation), i.e., the independence of subproblems.

Many patch proofs are found in complex analysis. In particular, the Principle of Analytic Continuation is a patch process [Nevanlinna and Paatero, Chapter 12, p. 213.]:

> Principle of Analytic Continuation -
> Let a regular function $w_1(z)$ be defined in a region G_1.
> Let a regular function $w_2(z)$ be defined in a region G_2.
> Let $G = G_1 \cap G_2$ be non-empty and connected.
> If $w_1(z) = w_2(z)$ on an infinite number of points of G,
> Then $w_1(z) = w_2(z)$ in the whole domain G,
> and they are partial representatives of the same function w(z):
>
> $$w(z) = \begin{cases} w_1(z) & z \in G_1 \\ w_2(z) & z \in G_2 \end{cases}$$
>
> where w(z) is regular in $G_1 \cup G_2$.

This principle implies that analytic contintuation can be carried out in a succession of patches: subdivide the domain into overlapping pieces (which are typically open discs) and show that the individual solutions agree on the overlaps. Such a proof will be called an *overlap* patch.

The definition of a "sheaf" embodies the idea of the overlap patch process [Godement, p. 109]:

> (F1) Uniqueness
> Let $\{U_i\}_{i \in I}$ be a family of open sets of X whose union is U
> Let s', s" be two elements of F(U)
> If the restrictions of s' and s" to each U_i are equal,
> Then s' = s".

(F2) Existence
Let $\{U_i\}_{i \in I}$ be a family of open sets of X whose union is U
Given $s_i \in F(U_i)$, such that for all i,j
the restrictions of s_i and s_j to $U_i \cap U_j$ are equal,
Then there exists an $s \in F(U)$ whose restriction to each U_i is s_i.

Property (F2) states that if the elements agree locally *on the overlaps*, there exists a global solution.

A subdivision of a domain into clearcut pieces (typically closed), in which boundary conditions do not cancel but rather must be matched is a particular kind of patch proof deserving of a name of its own: *boundary* patch proof.

Examples of this method can be found in numerical analysis and boundary value problems of differential equations. For instance, the technique of finite elements uses what is known as a "patch basis" [Strang and Fix]. Other examples can be found in the method of matched asymptotic expansions used in boundary layer theory. Using boundary layer theory for instance, the hydrodynamic flow around an island in a channel is described by one very involved six-fold patch proof [Bender and Orszag].

In most patch proofs, the individual patches are not identical in size nor of the same generic type. However, in cases in which they are, the patches are called *uniform*. For instance, proofs using the definition of totally bounded usually need a covering by discs all of the same radius. Patches in complex function theory are often uniform since all the patch elements are discs, although not necessarily of the same radii.

In differential calculus and differential geometry, many proofs are uniform patches of the *global-local* variety.: An example of this (mentioned in Chapter 2) is the osculating circle definition of curvature which is applied locally at each point of a plane curve to obtain a global definition of curvature. The interpretation of the derivative in terms of the tangent vector in a differential box of size δx by δy is another local characterization. Clearly sheaves are examined locally in a generic way in order to create a global picture. Regarding local-global patches, it is a super-principle that

local knowledge + something else => global knowledge

and that the something else is typically patching information.

As in vertical proof techniques, one does not get something for nothing and thus horizontal techniques are also specified by a two-tuple:

(*horizontal technique, joining information*) such as (divide and conquer, sorting) or (boundary point, match boundary conditions).

such as (divide and conquer, ending) or (boundary patch, match boundary conditions).

Chapter 6. EXAMPLES

In this chapter we discuss topics from three bodies of mathematical knowledge that are standard in the undergraduate curriculum: calculus, linear algebra and real analysis. We analyze topics within these theories using the framework we have developed in the previous chapters.

The first section (on real analysis) will serve to illustrate how we distil ingredients of the epistemology from a standard mathematics text. The second section (on linear algebra) presents a knowledge base that was compiled by examination of a half-dozen or so texts and later refined through discussions with students learning the material. The third section (on calculus) presents another example of how we build a knowledge base in our representation. The last section will touch briefly on another domain (from plane geometry) and discuss some of the problems encountered in it. The sections are basically independent and so if a topic is not familiar to the reader, it can be skipped with little effect on the others. However, the discussion, especially the section on calculus, should be accessible to anyone with an undergraduate background in mathematics.

In order to avoid reproducing large sections of standard presentations, we shall constantly refer the reader to three widely used textbooks by Thomas, Strang and Hoffman, and in fact, shall assume that the reader has these texts in front of him while he reads this report. This report can be read as a guide to the relevant sections of texts.

6.1 A Paradigm Example from Real Analysis

A textbook which provides an illustration *par excellence* of the epistemology we have been developing is Hoffman's *Analysis in Euclidean Space*. Almost any section of this book illustrates our points.

Because of its accessibility (i.e., it doesn't depend on much previous work in analysis) and its importance (e.g., it covers the Bolzano-Weierstrass Theorem), we shall examine Section 2.4, entitled "Sequential Compactness", and Section 2.5., "Closed and Open Sets" [pp.51-61]. It should be kept in mind that it takes time for the richly interconnected fabric of mathematics to be woven and that in the space of the very few pages which we shall examine, not much development -- especially as shown by the representation graphs -- will be apparent. What will be striking is the richness of dual relations and certain epistemological classes.

6.1.1 Sequential Compactness

Hoffman starts Section 2.4 by rephrasing the concept of convergence as:

The sequence $\{X_n\}$ converges to the point X if X_n is near X for all sufficiently large n.

He then points out that in many circumstances one doesn't need to know that a sequence converges, but only that it accumulates:

"i.e., X_n is near X for infinitely many values of n."

These two informal statements are the predecessors for the formal definition of accumulation point which is given first in terms of neighborhoods and then in terms of epsilons.

"**Definition.** *The point* X *is a* **point of accumulation (accumulation point)** *of the sequence* $\{X_n\}$ *if every neighborhood of* X *contains* X_n *for infinitely many values of* n.

"We can say it another way: X is an accumulation point of $\{X_n\}$ if, for each $\epsilon > 0$ and each positive integer n, there exists $k > n$ such that

$$|X - X_k| < \epsilon.$$

The next remark is really an easy basic result relating this concept to the predecessor concept of convergence:

"If $\{X_n\}$ converges to X, then clearly X is the unique point of accumulation of the sequence.

Thus if we were to start filling in the slots of the framework (see Figure 1) for the accumulation point concept, it would contain in-space backpointers to the concepts of convergence, neighborhoods, limits, etc., and the informal (MP-like) statements of convergence and accumulation point. Its results-dual would contain a (post-) dual pointer to the last remark on uniqueness. Of course, either or both of the two definitions would be included (in the DEClarative statement). Also, the setting is R^m.

The next order of business is to make some connections to the examples-dual of this new concept. The first example is the standard reference example of the positive rational numbers enumerated in a special way:

"EXAMPLE 10. Let $r_1, r_2, r_3, \ldots$ be the sequence which consists of the positive rational numbers, enumerated according to the scheme:

1/1	1/2	1/3
2/1	2/2	2/3
3/1	3/2	3/3

Then, every non-negative real number is an accumulation point of the sequence $\{r_n\}$."

Figure 1. A partially filled out framework for the accumulation point concept.

ID CLASS Definition RATING ** NAME Accumulation Point

STMNT SETTING R^m
DEF'N The point X is a point of accumulation of the sequence $\{X_n\}$ if every neighborhood of X contains X_n for infinitely many values of n.

DEMON-STRA-TION AUTHOR
MAIN-IDEA
PROC

PICTURE

REMARKS Caution: Nothing is assumed about the non-repetitiveness of the X_n.

EXTRAS

PEDAGOGUES HOFFMAN

IN-SPACE POINTERS BACK neighborhood, limit, convergence, MP(convergence), MP(accumulates)
FORWARD

DUAL-SPACE POINTERS RESULTS uniqueness, Bolzano-Weierstrass,
EXAMPLES Example 10, Example 11,

Even though it is labeled an example, the next item is really a counter-principle coupled with a counter-example illustrating the kind of problems the CP warns against:

> "EXAMPLE 11. Beware of working with coordinates when discussing accumulation points."
>
> "Consider in R^2
>
> $$X_n = (0, 1), \quad \text{n odd}$$
> $$X_n = (1, 0), \quad \text{n even.}$$
>
> The sequence of the first coordinates is 0,1,0,1,...which has two accumulation points in R, 0 and 1. The sequence of second coordinates is 1,0,1,0,... and it has the same accumulation points. In particular, 0 is an accumulation point for the first and for the second coordinates. We cannot conclude that (0,0) is a point of accumulation of the sequence in R^2.

This CP would be entered in the FORWARD pointers slot of the item frame for accumulation point. Note how this counter-example is done by considering a two dimensional case involving only 0's and 1's.

Next comes the definition of subsequence and then a lemma, which is a basic type of result, showing that accumulation points correspond to limits of subsequences:

> **Lemma.** *The point* X *is a point of accumulation of the sequence* $\{X_n\}$ *if and only if some subsequence of* $\{X_n\}$ *converges to* X.

The first major result item is the Bolzano-Weierstrass Theorem which is introduced by the comments that summarize the essence of what is to follow:

> "The completeness of the real number system guarantees that bounded sequences in R^m have accumulation points. A sequence can wander aimlessly; however if it stays in a bounded part of R^m, it must accumulate somewhere. This property is usually called the "sequential compactness" of bounded parts of R^m."
>
> **Theorem 5 (Bolzano-Weierstrass).** *Every bounded sequence in* R^m *has a point of accumulation. Equivalently, every bounded sequence in* R^m *has a convergent subsequence."*

The image of aimless wandering is one that Hoffman uses again [p.274] (also see Chapter 3) in conjunction with another sequence which has problems with convergence.

Hoffman offers two proofs of the Bolzano-Weierstrass Theorem. The first is a bootstrap/wlog argument working from R^1:

> "We shall work with coordinates, and, as we noted in Example 11, we must exercise some care. ...Suppose we have proved the theorem for bounded sequences in R^1. The proof for R^m could then be given this way..."

The second proof (given after the corollary) is the undemocratic divide and conquer proof we mentioned in Chapter 5, but given in terms of specific boxes. To be precise this argument contains a wlog argument which allows the proof to be performed in a unit box in the first quadrant (by scaling and translation). The rectangular schematic diagram (see Chapter 5) is also included. Hoffman also remarks that there are two variants on the undemocratic divide and conquer proof: the proof can be nailed home either using the nested interval property or a Cauchy criterion. Thus if we were to enter the Bolzano-Weierstrass theorem in our data base we would catalogue two principal proofs with a remark that one of these has two slight variations.

The corollary further relates convergence and points of accumulation.

> **Corollary.** *A bounded sequence in* R^m *converges if and only if it has precisely one point of accumulation.*

Thus at the conclusion of this section of Hoffman, the three representation spaces are:

Concepts-space

MP(convergence) - The sequence $\{X_n\}$ converges to the point X if X_n is near X for all sufficently large n.

↓

MP(accumulates) - The sequence $\{X_n\}$ accumulates at X if X_n is near X for infinitely many values of n.

↓

DEF(Accumulation Point)

↓

CP(Working coordinatewise with accumulation points)

↓

DEF(Subsequence)

Results-space

lemma (accumulation points)

↓

Theorem 5 (Bolzano-Weierstrass)

↓

Cor (B-W)

Examples-space

E10 (Q^+) **E11: {(0,1), (1,0)}**

6.1.2 Open and Closed Sets

The next section of Hoffman, Section 2.5, discusses "two very special classes of sets": open and closed sets. Hoffman begins with the usual definition of open sets in terms of neighborhoods [p. 55]:

> **Definition.** *The set* U *is* **open** *(in* R^m*) if it is a neighborhood of each of its points.*

(Neighborhoods were defined two sections earlier in terms of open balls $B(X;r)$ where $B(X;r) = \{X;\ |X-X_0| < r\}$.)

Next is the key result:

> **Theorem 6.** *The union of any collection of open sets is open. The intersection of any finite collection of open sets is open.*

Even though it is "virtually trivial", it is considered *key* since:

> "it states properties of open sets which are used so often."

Interestingly, Hoffman does not give the usual counter-example (cf. [Rudin]) here (or in the exercises) of the intersection of intervals (-1/n,1/n) to show the necessity for specifying "finite" and not arbitrary intersections.

Next comes a sequence of three examples. The first is a model example for an open set that harks back to the foundation concept of this chapter, namely open balls. The second gives supporting examples and counter-examples and the last is a reference example which he uses throughout the book to link analysis with linear algebra.

> EXAMPLE 12. Every open ball $B(X,r)$ is an open set.
>
> ...
>
> Thus, the union of any collection of open balls is open... Futhermore, every open set is of the last type.
>
> EXAMPLE 13. Let us look at open sets in R^1. Each open interval (a,b) is an open set in R^1. On the other hand, an interval $(a,b]$ is not open in R^1, because $b \in (a,b]$ but no open interval about b is contained in $(a,b]$. The unbounded interval $(a\ \text{ffl})$ is open in R^1.

In the text, Example 11 continues with what is actually a useful technical result:

> Every open set in R^1 is a union of open intervals (a,b). In this 1-dimensional case, the open set U can be expressed as a union of intervals in a very special way.....Thus every open set in R^1 is uniquely expressible as the union of countable collection of open intervals which are pairwise disjoint.

This result shows the sufficiency in many cases (in a WLOG sense) of working with only intervals. Although he doesn't show the last result this way, it can be demonstrated with a Gram-Schmidt type argument (See Chapter 5).

> EXAMPLE 14. Let's look at the space of kxk matrices (real or complex entries). Let U be the set of invertible matrices... To summarize, the set U of invertible matrices is open because, if $A \in U$, then U contains the open ball of radius $|A^{-1}|^{-1}$ about the point A.

Embedded in Example 14 is a result on the invertibility of matrices; it takes its logical support from another predecessor item that showed every matrix near the identity matrix is invertible, i.e., the validity of geometric series expansions for $(I - T)^{-1}$ when $|T| < 1$. Also, note that while there is a choice of setting ($M_k(R)$ or $M_k(C)$), the setting is made explicit.

Hoffman then goes on to define a cluster point, prove a basic equivalency result about this new concept and discuss its relation to the previous concept of accumulation point:

> **Definition.** *The point* X *is a* cluster point *of the set* S *if every*

neighborhood of X *contains a point of* S *which is different from* X.

Lemma. *Let* S *be a subset of* R^m *and let* X $\in$ R^m. *The following are equivalent (all true or all false).*

(i) X *is a cluster point of the set* S.

(ii) *Every neighborhood of* X *contains infinitely many points of* S.

(iii) *There exists a sequence* $\{X_m\}$ *in* S *such that* $X_m \neq X$ *and* $X = \lim_n X_n$.

At this point, Hoffman tries to help the reader keep straight the two concepts of cluster and accumulation point. He points out that accumulation points are for sequences and cluster points are for sets. Consideration of a sequence of 0's and 1's -- a standard reference to check out ideas -- is offered to sharpen the difference between the two concepts [p.58]:

> "The reader may have noticed the similarity of the concepts of "cluster point of a set" and "accumulation point of a sequence". It is important to be clear about the relationship between the two ideas.... A simple example should make this clear. The sequence of real numbers
>
> 0, 1, 0, 1, 0, 1, ...
>
> has two points of accumulation, 0 and 1. The image of the sequence is $S = \{0, 1\}$, and it has no cluster points at all."

The preceding discussion actually includes the standard "hack" to convert discussions of accumulation points of sequences to that of cluster points by looking at the associated image set of the sequence.

Next is a definition of closed sets and then the theorem (Theorem 7) that open sets are complements of closed sets and vice versa. The corollary to Theorem 7 is the analgous theorem for the combination of closed sets through set intersection and finite unions:

Definition. *The set* K *is* closed *if every cluster point of* K *is in* K.

Theorem 7. *A set* S *is open if and only if its complement (complementary set) is closed.*

Corollary. *The intersection of any collection of closed sets is closed. The union of any finite collection of closed sets is closed.*

He then remarks that one inherits a whole host of examples of closed sets from those for open sets:

> "In view of Theorem 7, there is no need for a separate list of examples of closed sets...But, the human mind being what it is, doesn't follow that just because we know about open sets we'll recognize a closed set when we bump into it.
>
> EXAMPLE 15. Let's look at a famous closed set -- the **Cantor set**. We shall refer to it often.

Example 15 presents the usual construction of the Cantor set on the unit interval by deleting middle thirds (see Chapter 1). The alternative characterization of the Cantor set in terms of ternary expansions is also given.

We now come (once again) to the culminating theorem of this section, the Bolzano-Weierstrass Theorem, which for emphasis is reformulated in two more different ways:

> **Theorem 8** (Bolzano-Weierstrass). *Every bounded and infinite set of* R^m *has a cluster point.*
>
> **Theorem 9.** *let*
>
> $$K_1 \subset K_2 \subset K_3 \subset \ldots$$
>
> *be a nested sequence of bounded closed sets in* R^m. *If each* K_m *in non-empy, then the intersection*
>
> $$\cap K_m$$
>
> *is non-empty.*

Hoffman remarks [p. 61] that:

> "one might think of Theorem 9 as a slighly more geometrical way of stating the Bolzano-Weierstrass theorem. If (in Theorem 9) one knows that diam(K_m) converges to 0, then the intersection of all the K_m's will consist of precisely one point. That result is weaker than Theorem 9. It is (essentially) a reformulation of the fact that each Cauchy seuqence in R^m converges."

By these restatements and perturbations of the Bolzano-Weierstrass Theorem, we are given a feeling for its strength.

The last illustration, Example 16, is an "amusing" application of this result to the intersection of the medians of a triangle.

Figure 2. A partially filled out item frame for the Bolzano-Weierstrass Theorem.

ID CLASS Key, Culminating RATING **** NAME Bolzano-Weierstrass Theorem

---

STMNT SETTING R^m

SENT1 Every bounded sequence has a point of accumulation

SENT2 Every bounded sequence has a convergent subsequence.

DEMON-STRA-TION AUTHOR Hoffman

MAIN-IDEA WLOG to R^1; diagonalization process

PROOF1

Let the sequence be $X_n = (x_{m1},...,x_{nm})$.

.... (from p.53)

AUTHOR Hoffman

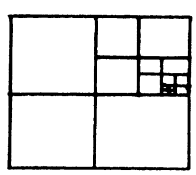
MAIN-IDEA Undemocratic Divide and Conquer

PROOF2

WLOG assume the set is a box.

Divide the box into four boxes, ...

PICTURE



REMARKS Proof2 is a paradigm undemocratic divide-and-conquer proof.

EXTRAS

PEDAGOGUES Hoffman

IN-SPACE POINTERS BACK Lemma on accumulation points

FORWARD Corollary

DUAL-SPACE POINTERS CONCEPTS Accumulation point, bounded, subsequence

EXAMPLES Unit box, unit ball in $l_2(R)$

6.1.3 Recapitulation of this example

This example from real analysis might have seemed long, and perhaps tedious, but it is the kind of exercise one must go through to build up a knowledge base. Even within the space of a mere ten pages, there is a fantastic amount of material. If it seems a large task to comment on it (and supposedly, the author and hopefully, the reader have seen it before and know it to some extent), consider the amount of work a neophyte student must do to learn it for the first time. Yet it does happen.

To give a further idea of what the knowledge base would look like, Figure 2 contains a partially filled out framework for the culminating result of these sections, the Bolzano-Weierstrass Theorem.

The fact that the Bolzano-Weierstrass Theorem is visited several times would not be reflected in the representation graphs. Rather, it would be part of the pedagogical knowledge associated with this mini-domain. For instance, the PEDAGOGUE's field of the B-W item would have several entries for the pedagogue HOFFMAN and Hoffman's pedagogical trail (or "PTRAIL" [Michener 1977] would have the B-W theorem occuring several times in the list.)

This extended example has also shown that the representation scheme of this report is fairly adequate for representing textbook knowledge of mathematics. Its main deficiency in that regard is the awkwardness of encoding extended comments in natural language, such as introductory and summary remarks; shorter ones can sometimes be viewed as MP's and CP's. Thus, a more faithful representation of textbooks would need a better representation for such text, perhaps as a "hypertext" [van Dam].

6.2 The EIGEN Domain

The theory of eigenvalues for matrices, which can be considered operators on finite dimensional vector spaces, is an important and interesting area. It has applications in many fields, not only in mathematics (e.g., differential equations, numerical analysis) but also in other disciplines as well (e.g., in quantum mechanics, electrical circuits). It also generalizes to more abstract settings such as Hilbert and Banach spaces, where it is then usually called "spectral theory".

The knowledge base presented here was built by first examining a dozen or so undergraduate level linear algebra texts; then choosing three or four of these [Strang], [Halmos], [Shilov], and [Ortega] and encoding the knowledge presented within their chapters on eigenvalues as the EIGEN data base and revising the representation after experience teaching this area to undergraduates ([Michener August 1978], See Chapter 7).

6.2.1 Concepts-space

Instead of going back to the texts used to build up this knowledge base, we shall simply "walk" the reader through the representation graphs so that he can get an idea of the kind of knowledge bases we are interested in building. As a bonus, this perusal should help him acquire an overall feeling for how this mini-theory hangs together. Specific details of this knowledge base can be found in [Michener 1977, Appendix A] (which contains a large number of the item frameworks) as well as in the textbooks themselves.

First of all, let us examine the concepts-graph. The primary parent item for the EIGEN knowledge base in C10 which is the definition of "eigenvalue" (and "eigenvector", the two going hand-in-hand); it is formulated both declaratively as $Av=\lambda v$ as well as procedurally as $\det(A-\lambda I)=0$. Included as predecessor nodes are the two very general mega-principles C1, *Try 0's and 1's*, and C2, *Try the 2X2 case*. These are entered as predecessors not just because one ought to know about them for C10, but also because one should know about them for the whole theory and by pointing back to them from C10, they are inherited by C10's successor items, which in this case is the entire theory. Note how the concepts-graph is not only a connected graph, but also that it develops from one starting node, C10.

Having defined "eigenvalue/vector", one can then give names to a few very closely related ideas such as the "characteristic equation", C20, and the "spectrum" of a matrix, C30. More importantly one can then paraphrase the eigenvalue idea in two useful ways: *On its eigenvectors, the matrix acts like scalar multiplication by the eigenvalue*, which is C40, the MP labeled as "like multiplication"; and C50, *Eigenvalues represent singular shifts of the matrix*, another MP. (Both of these MP's come from Strang's text, which offers about five paraphrases of the eigenvalue concept in the form of MP's.) One can also now define what an "eigenspace" is. One is also in a position to consider "upper triangular form" (in which the eigenvalues are displayed along the main diagonal).

From the concepts-graph, one can see that it is now possible to go on to define many other concepts:

C25 - Definition of the algebraic multiplicity

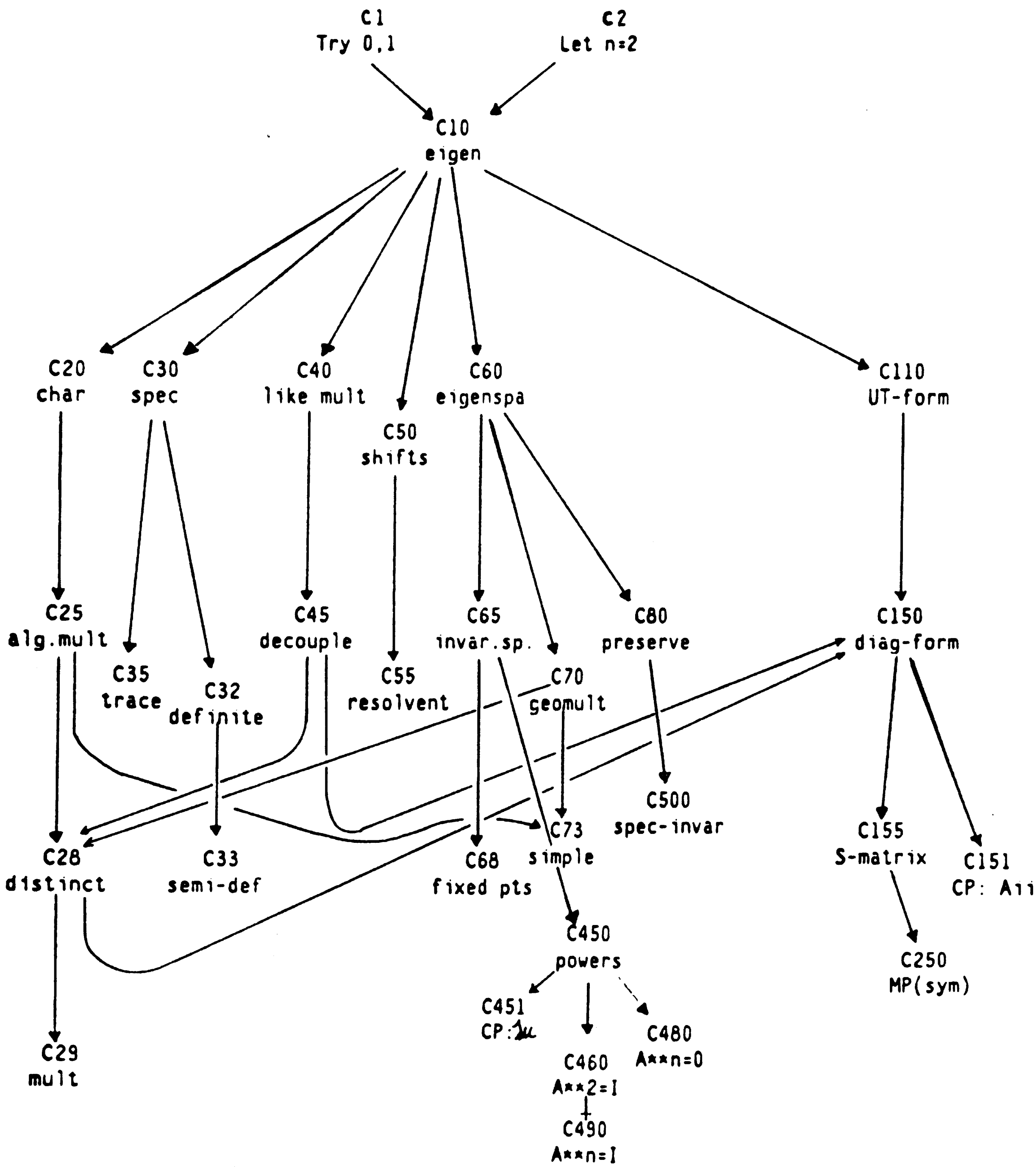
and its sucessor concepts of:

C28 - MP: *Distinct eigenvalues are good.*
C29 - CP: *Multiple eigenvalues are troublesome.*

Succeeding the node C30, one has:

C35 - Definition of the trace
C32 - Definition of positive definite

Figure 3. The Concepts-graph for the EIGEN domain.



and then

C33 - Definition of semi-definite

The important MP C40 has another MP as successor:

C45 - MP: *De-couple the system through its eigenvalues.*

The idea of singular shifts, C50, leads to:

C55 - Definition of the resolvent

Once one knows about eigenspaces, one can then consider many more ideas, such as the "power idea" MP::

C450 - MP: *To learn about a matrix A, look at its powers,* A^n.

After defining diagonal form and the "S" matrix (i.e., the S of $S^{-1}AS = \Lambda$), one has the very important MP on symmetric matrices. (The actual definition of a symmetric matrix is assumed to be inherited from knowledge of matrices.)

C250 - MP: *Symmetric matrices are nice.*

To see the exact statement of concept items, please refer to Appendix A.

The concepts-graph for EIGEN is given in Figure 3. A summary of the concept items is listed below, where for each item we list its ID, "Michelin rating", epistemological CLASS, and NAME. The complete item frames can be found in Appendix A of Michener [1977].

C10 - **** - DEF(eigenvalue/eigenvector)
C20 - * - DEF(Characteristic polynomial/equation)
C25 - ** - DEF(algebraic multiplicity)
C28 - * - MP(Distinct eigenvalues are good)
C50 - ** - MP(Singular shifts)
C55 - DEF(the resolvent)
C60 - *** - DEF(eigenspace)
C65 - * - DEF(invariant subspace)
C70 - * - DEF(geometric multiplicity)
C73 - DEF(simple eigenvalue)
C80 - ** - MP(spectral invariance)
C110 - **** - DEF(UT form)
C150 - **** - DEF(diagonalizable)
C451 - CP(diagonal entries eigenvalues)

C250 - **** - MP(symmetric)
C451 - - CP(composition of eigenvalues)

6.2.2 Results-space

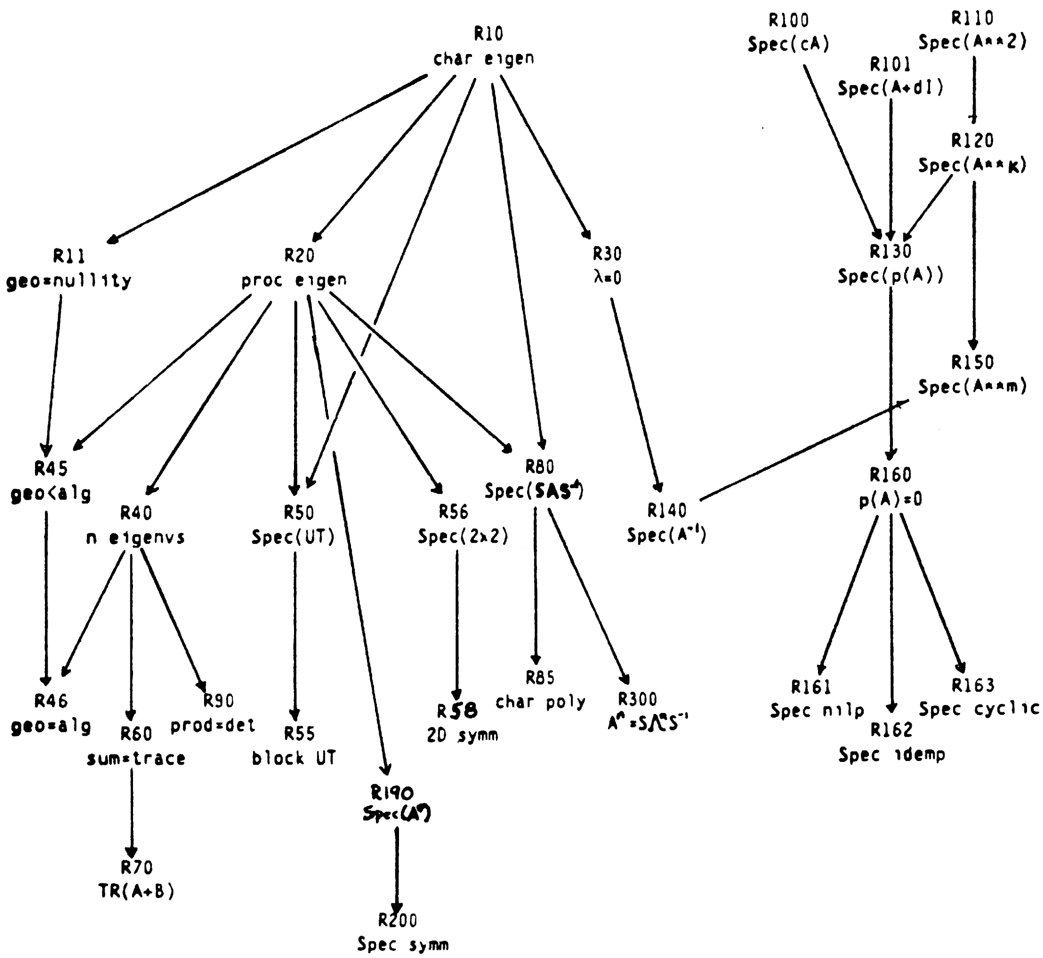
We won't say too much about the Results-space for EIGEN but rather refer the reader to its graph and the summary list.

The graph is fairly orderly, and except for the link between R140 and R150, it is almost disconnected into two major components. This indicates that the two major branches take their logical support from different sources. The successor results to item R10 rely on its procedural formulation of eigenvalue in terms of the characteristic polynomial; the stream of results on the spectra of matrices (i.e., the successors to R100, R101, R110) arrived at by algebraic operations on other matrices (for which spectral information is known), are proved directly from the declarative formulation of eigenvalue and do not make use of the procedural formulation proved in R10.

The following is a list of the result items (ID - RATING - CLASS - NAME) for the EIGEN domain:

R10 - *** - BASIC(characterization of eigenvalues)
R11 - TECH(geometric mult and nullity)
R20 - *** - KEY(procedural formulation of eigenvalue)
R30 - BASIC(0 as an eigenvalue)
R40 - * - BASIC(existence of n eigenvalues)
R45 - BASIC(geo mult $\leq$ alg mult)
R46 - TECH(when geo mult = alg mult)
R50 - ** - KEY(Spec of diagonal/triangular matrices)
R55 - TECH(Spec of block diagonal form)
R60 - * - BASIC(eigenvalues the trace)
R70 - BASIC(elementary properties of the trace)
R80 - *** - KEY(similarity invariance of eigenvalues)
R85 - ** - CULM(similarity invariance of characteristic polynomial)
R90 - * - BASIC(eigenvalues the determinant)
R100 - * - BASIC(scalar multiples of Spec)
R101 - * - BASIC(shift of Spec)
R110 - ** - KEY(square of Spec)
R120 - ** - KEY(positive powers and Spec)
R130 - *** - CULM(polynomials and Spec)
R140 - * - BASIC(inverse and Spec)
R150 - ** - CULM(general powers and Spec)
R160 - *** - CULM($p(A)=0 \implies p(\lambda)=0$)
R161 - TRANS(Spec of nilpotents)
R190 - KEY(transpose, conjugate and Spec)

Figure 4. The Results-graph for the EIGEN domain.



6.2.3 - Examples-space

The examples-graph is different from the both the concepts-graph and the results-graph in that it has several (six) starting nodes and five fairly distinct branches. The main feature of construction of examples in this domain is derivation by *increasing complication*.

For instance, we can start out with the simplest of all matrices, the matrix whose entries are all zero (E10), and change the diagonal entries to 1's, thereby getting the Identity matrix (E20). The 1's of the Identity can all be changed to another non-zero constant, and then one has the "scalar matrix", cI (E30). By changing the c's to other numbers, i.e., lifting the constraint that the diagonal entries be equal, one then gets the "diagonal" example (E40). By sprinkling entries either above or below the diagonal, one gets "upper triangular" and "lower triangular" examples (E50 and E51).

Although we did not continue much further along this particular branch of Examples-space in EIGEN, one could go on to construct more examples by adding more elements to E50 and E51 either with the constaint of symmetry, thereby building symmetric matrices, or without, thereby building general matrices. One could also change any of these matrices to matrices of functions -- specifically dependent on time, for instance -- by considering numbers to be constant functions and replacing them by polynomials or general functions of some class or just adding some non-constant terms. This would lead very easily to matrices that arise in differential equations and more general spectral analysis. Such considerations give rise to the examples of the convolution operator $*t$, E70, and then the Fredholm operator, E85.

The differential operator is a good source of examples not only in finite dimensional settings, but also in more general ones, where it is one of the few operators that one can work with directly. The simple differential equation

$$x'(t) = a\,x(t)$$

serves as a start-up example, E80 [Strang, pp.172-173]. In his book, Strang considers a coupled pair of such differential equations (see Chapter 3) and makes an analogy with the one dimensional case to introduce the ideas of eigenvalues in R^2 and more generally in a finite dimensional vector space.

Halmos carries analysis of the differential example, E60, further by varying its setting to generate examples E65 and E66. E65 is set within the ring of polynomials of degree less than n, and E66 is set within the span of n exponentials $\{exp\ c_i t\}$. These last two examples are used to sharpen one's appreciation of the setting in which one looks for the eigenvalues. E60 can be combined with the convolution example, E70, to produce E75, which is an example dealing with the commutator, AB - BA.

The next principal branch of the examples-graph deals with the Basic 16 example and its offshoots. The Basic 16 is the cluster of sixteen 2X2 matrices whose entries are 0's and 1's.

(It is simply considered as a set of sixteen elements, and not as a group or matrix algebra. If it were, it would lead to the matrix algebra over Z/2Z which is interesting in its own right -- e.g., in coding theory -- but which is not included here.)

In addition to including all sixteen matrices together in a cluster as E100, a few of the really interesting ones are examined in more detail and entered individually: E102, E103, E104, etc. E104 is the extremely important example of a deficient matrix: a matrix with repeated eigenvalues which cannot be diagonalized. (The reason is that the geometric multiplicity of the eigenvalue 1 is strictly less than its algebraic multiplicity; hence, the deficiency.) It is the simplest example of this phenomenon and is a paradigm for the kind of problems one encounters with repeated roots and non-symmetric matrices.

E106 and E107 are the projection operators onto the first and second coordinates repsectively. E103 is what we call the *counter-identity* matrix; it has interesting spectral properties. The n-dimensional counter identity, N_n, is example E135. E103 is a successor of N_2 generated by adding a minus sign. E120 simply pulls out the two eigenvectors of N_2. They are called the *diagonal vectors* (think of the unit square); one of them, the vector of all 1's, in R^n is the example E121; it arises in the study of circulant matrices, an example of which is N_2.

E116, the *full* or "Jacobi matrix" [Davis] is another important example. Its 3-dimensional counterpart is E117. The J matrices also have interesting spectra. Scalar multiples of such matrices enter in discussions of round-off error (see [Ortega]). They are an example of matrices with repeated roots which do not have problems of deficiency.

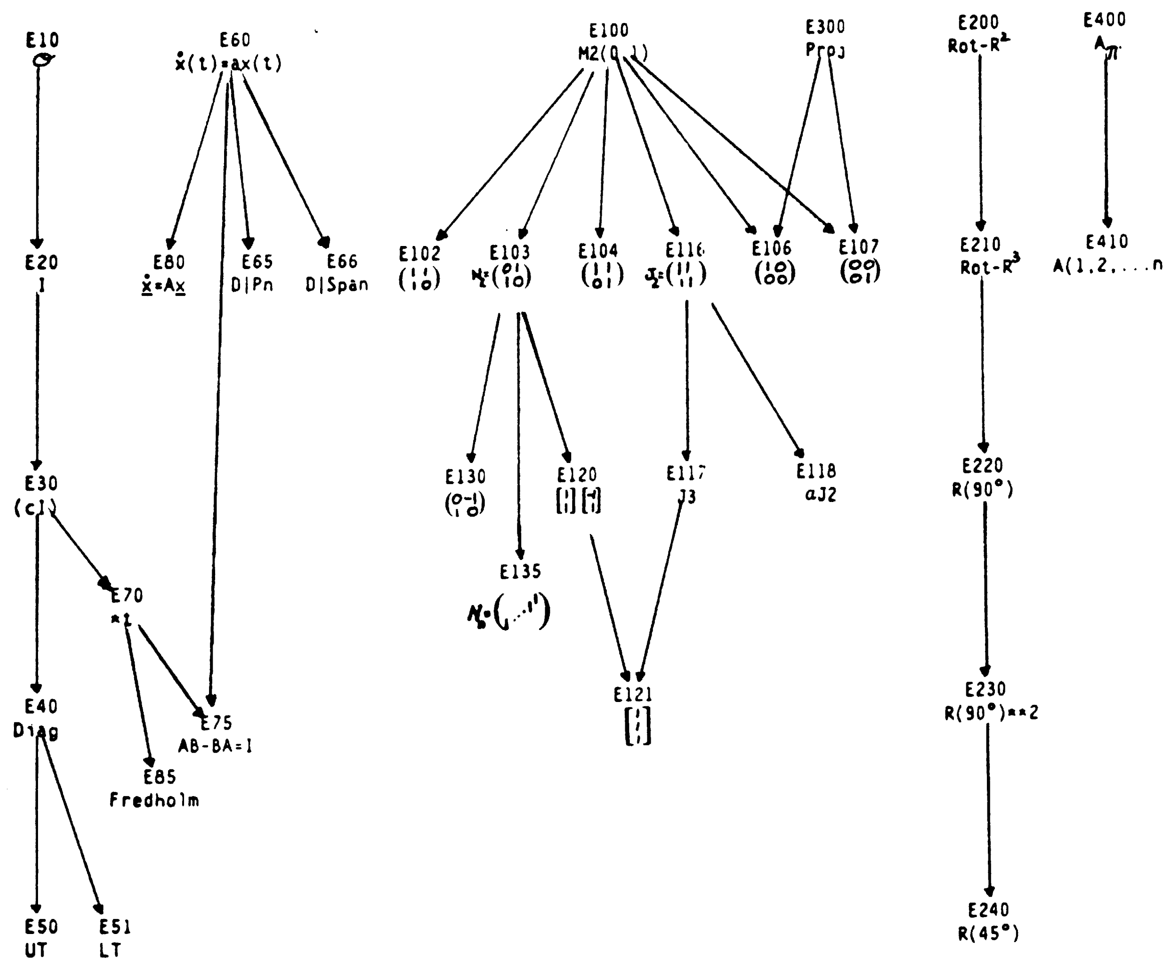
The fourth major branch of the examples-graph concerns rotation operators. Rotations provide additional easy entrance points to the theory of eigenvalues. For instance, the rotation in three-space, E210, the axis of rotation is an eigenvector corresponding to the eigenvalue 1, and the equatorial plane is a two dimensional eigenspace [Strang]; E210 is easily used as a start-up example. E200 is the general rotation through α degress in the plane described in terms of polar coordinates. E210 is the rotation operator in three space. E220 is the particular three dimensional rotation through 90^o; since the linear algebra is doable, this example is worked out in detail. E230 is E210 in the case of $\alpha=90^o$ iterated twice: i.e., E230 is the rotation in the plane through 180^o, an example which can be checked against one's common sense knowledge of the situation. E240 is another special rotation, that through 45^o; E240 comes up as the "square-root" of E220.

The last branch which is not well-developed grows from the permutation matrix A_π which is the matrix that corresponds to the permuation π. E410 is the particular specialization for $\pi = (123.....n)$, a pure circulation.

The examples-graph for EIGEN is shown in Figure 5. The following is a list of example items:

E10 - REF(zero operator)
E20 - ** - REF(identity matrix)
E30 - * - S-U(scalar matrix)
E40 - **** - MODEL(diagonal)
E50 - **** - MODEL(UT)
E60 - ** - S-U(differential operator)
E65 - CEG
E66 -
E70 - CEG(convolution *t)
E100 - **** - REF(Basic 16)
E102 - * -REF(2X2 Fibonacci generator)
E103 - * - REF(2X2 Counter Id, N_2)
E104 - * - MODEL(2X2 UT)
E106 - 2-dim projection
E116 - * - REF(2X2 full, J_2)
E200 - ** - S-U(2-D rotation)
E210 - ** - S-U(3-D rotation)
E220 - ** - REF(90^o rotation)
E230 - power of a rotation
E300 - * - REF(projection matrices)
E400 - ** - REF(permutation matrices)
E410 - * - REF(circulant matrix)

Figure 5. The Examples-graph for the EIGEN domain.



6.3 A Cluster from Calculus

In this section we sketch out a few pages from calculus that deal with Rolle's Theorem and the Mean Value Theorem. We use Thomas [Thomas 1972, Sections 4.7 and 4.8, pp., 129-134] as our source. Building a detailed knowledge base is left as an exercise to the reader.

6.3.1 Rolle's Theorem

Thomas starts the discussion of Rolle's Theorem by offering an example (Fig.36a) as "strong geometrical evidence" in support of the result. This example is nicely smooth and looks like a sine or two parabolas joined together. He also offers another example (Fig. 36b) with a point to indicate the necessity of requiring some degree of smoothness.

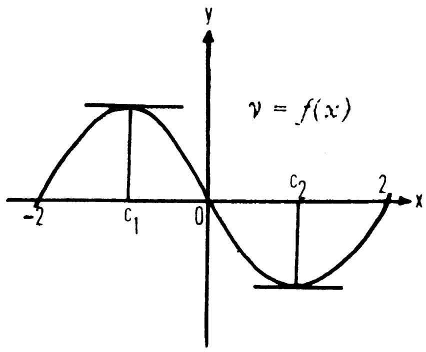


Fig. 36a

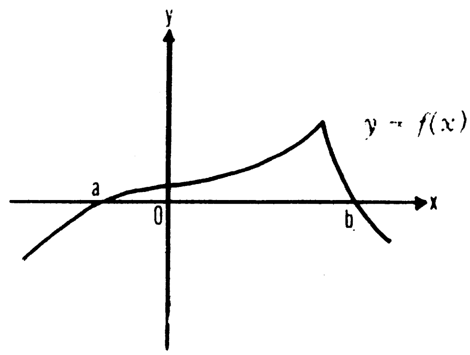


Fig. 36b

Then comes the theorem:

> **Rolle's Theorem.** *Let the function f be defined and continuous on the closed interval $[a,b]$ and differentiable in the open interval (a,b). Futhermore, let*
>
> $$fa) = f(a) = 0.$$
>
> *Then there is at least one number c between a and b where $f'(x)$ is zero; that is,*
>
> $$f'(c) = 0 \text{ for some } c \text{ in } (a,b).$$

The proof is done on two cases: (i) for f identically 0; (ii) for f not so. This could also be considered a wlog argument for the assumption of f not identically 0. Case (i) is trivial. Case (ii) is done by appealing to the result that on a closed interval (i.e., compact set), a continuous function achieves its max and min; this provides the point c. By using the result that the derivative must vanish at such critical points [Thomas, Theorem 1, p.119], the proof is completed.

Thus this proof uses two results as logical support: Theorem 15 [Thomas, Chapter 3, p.100] and Theorem 1 [Thomas, Chapter 4, p.119]. Theorem 15, itself, is stated without proof since its proof involves some deep properties of the real numbers and continuity.

More importantly, Thomas talks about Rolle's theorem in following *Remarks*. The first remark deals with the non-uniqueness of the point c. He refers to his first example to illustrate this. He also presents a polynomial, $x^3 - 4x$ on $(-\infty, \infty)$ to show how the theorem works.

Section 4.7 is concluded with Remark 2, actually a result stating conditions on when a function has a unique real root between a and b. This result is derived by combining Rolle's Theorem with the previous Theorem 15 (actually, the Intermediate Value Property of continuous functions).

Thus far we have acquired one result item (Rolle's Theorem) and two or three examples (the picture examples of his Fig.4.36a and b, and if we want to count it, the cubic polynomial). The item frame for Rolle's Theorem would appear something like the following:

ID CLASS Key RATING ** NAME Rolle's Theorem

STMNT SETTING R
SENT f defined and continuous on ...

DEMON-STRA-TION AUTHOR Thomas
MAIN-IDEA use points of vanishing derivative
PROOF Case 1. f identically 0 ...

PICTURE Fig 36a

REMARKS Caution: Nothing is guaranteed about uniqueness of the point c.

EXTRAS

PEDAGOGUES Thomas

IN-SPACE POINTERS BACK Theorem 15 (Chap. 3), Theorem 1 (Chap. 49
FORWARD

DUAL-SPACE POINTERS CONCEPTS differentiability, max/min
EXAMPLES Fig36a, Fig36b, cubic

6.3.2 The MVT

The next section concerns the Mean Value Theorem (MVT), which as he points out right at the beginning, is a generalization of Rolle's Theorem. (Thus a forward in-space pointer to the MVT would now be added to the representation for Rolle's Theorem.) The requirements on the function f are the same as for Rolle's. However, here he points out that the differentiability of the function at the endpoints a and b does not matter. For instance, at the endpoints the function can have a vertical tangent, as does his example:

$$f(x) = (a^2 - x^2)^{1/2} \text{ on } [-a,a].$$

Before giving the "analytic" proof he presents a geometrical paraphrase:

> "Geometrically, the Mean Value Theorem states that if the function f is continuous ...... and differentiable ...., then there is at least one number c in (a,b) where the tangent to the curve is parallel to the chord through A and B "

He argues the plausibility of this statement by referring to a general diagram (Fig.4.38) and asking the reader to imagine the chord moving "upward, parallel to its original position". He remarks that the analytic proof has its key idea from these geometric considerations.

The MVT is actually proved by an application of Rolle's Theorem. After the proof, Thomas then reiterates that the tangent-parallel-to-the-chord formulation is "a form that is easily recalled."

In the first post-dual example following this result, he elaborates about the c.

> "which is not very well defined....For a specific function f and specific values of a and b, however, the equation can be used to find one or more values of c.

His example is the standard reference example x^3 taken on the interval (-2, 2).

The second post-dual example is a reference for functions with points or cusps, $x^{2/3}$. He uses this example to show the sensitivity of the MVT to the hypothesis of differentiability. (Of course, this function is not diffenentiable at $x = 0$.)

The *Remark* that follows relates the MVT to instantaneous and average velocities.

Next come three corollaries, all easy consequences of the MVT. Since their statements are quite long, we only indicate them in an informal way:

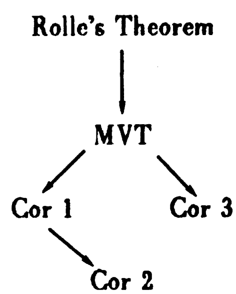
Corollary 1. $F' = 0 \implies F = constant$

Corollary 2. $F_1' = F_2'$ ===> $F_1 - F_2$ = constant

Corollary 3. $F' > 0$ ===> F increasing.
$F' < 0$ ===> F decreasing.

At the conclusion of these two sections, we have added a little to Results-space and Examples-space. The examples used in this section have contained a few new examples and several old reference examples (x^3 and $x^{2/3}$). Also note that the two sections we have discussed have contained no new concepts; they build deductively and illustratively on concepts from preceding sections. However, one should not be fooled by this seeming lack of concepts, since this whole discussion takes place against a background with a very extensive Concepts-space. The apparent lack of a Concepts-space is an artifact of our excision of these two sections from their context.

The following graph fragment, showing the results of these two sections, would be added onto the graph representing predecessor results:



6.4 Comments on Other Domains

Originally we had intended to illustrate the epistemology and representation by including some mini-domain for plane geometry, such as quadrilaterals. A few geometry textbooks were examined for their treatment of this area [Jacobs], [Beman].

While the Results- and Concepts-spaces for this domain grew at a steady rate into fairly interesting graphs, Examples-space was virtually non-existent. This lack of examples is no doubt due to several forces: (1) plane geometry as presented in high school level texts is designed to teach "abstract thinking" and as such is more of a stylized minuet of definition-theorem-proof than it is an excursion through "live" mathematics with all the inter-play between definition, conjecture, example, and theorem. (See [Lakatos] for a good example of this process; also, Polya's "alternation process" [MD]9.); (2) plane geometry has been around for a long time, and its conceptual and deductive houses are very much in streamlined order; (3) a lot of the examples are not very deep -- once you've seen one rhombus, you've seen them all -- or in other words, the diagrams are really model examples and these models really cover the possibilities.

It is interesting to note that just about all of the examples are presented pictures and almost all of them make their way into the text as diagrams for setting up the givens for a proof. If by any chance a student were not able to create an example for himself, especially its picture (e.g., a rhombus) and in addition he didn't look at the proof and its schematic diagram, then it is hard to see how he would ever learn what such a figure looks like. However, that's not too much chance of that since everyone draws diagrams (and hence examples) and besides the concepts are straight-forward and found everywhere in the real world.

Thus if one considers diagrams not to be bona fide examples (or even if one includes them), this domain is somewhat *examples-poor*. It is especially so in comparison with such *examples-rich* domains such as real analysis, eigenvalues and calculus. These observations relate to Lakatos' on "growing" and " theories. [Lakatos, p.].

It is interesting to consider what one can say about a theory by considering the richness of its Examples-space (and other spaces). Also, just what does it mean in regard to its development, *a la* Lakatos for instance, for a domain to be examples-poor. Are such domains in some sense mathematically dead or inactive? Are examples necessary for growth and development of the theory? Can theorems and concepts be discovered without them?

It is interesting to note that just about all of the examples are presented pictures and almost all of them make their way into the text as diagrams for setting up the givens for a proof. If by any chance a student were not able to create an example for himself, especially its picture (e.g., a rhombus) and in addition he didn't look at the proof and its schematic diagram, then it is hard to see how he would ever learn what such a figure looks like. However, that's not too much chance of that since everyone draws diagrams (and hence examples) and besides the concepts are straight-forward and found everywhere in the real world.

Thus if one considers diagrams not to be bona fide examples (or even if one includes them), this domain is somewhat examples-poor. It is especially so in comparison with such examples-rich domains such as real analysis, eigenvalues and calculus. These observations relate to Lakatos' on "growing" and "theories. [Lakatos, p.]

It is interesting to consider what one can say about a theory by considering the richness of its Examples-space (and other spaces). Also, just what does it mean in regard to its development, a la Lakatos for instance, for a domain to be examples-poor. Are such domains in some sense mathematically dead or inactive? Are examples necessary for growth and development of the theory? Can theorems and concepts be discovered without them?

Chapter 7. UNDERSTANDING MATHEMATICS

7.1 The Active Nature of the Understanding Process

Understanding mathematics is a very active process. While at first glance it may not seem so, especially in comparison with problem solving, it does involve significant effort on the part of the understander. To understand a theory, one must explore and manipulate it on many levels, from many angles, with facility and spontaneity. One must be able to travel freely through it, experiment with its items, survey its overall mathematical topography, shift the level of concern from nitty-gritty detail to broad overview and vice versa, and be able to ask questions. One gains understanding by examining relevant examples, perturbing settings and statements, and fiddling around (e.g., numerically and pictorially). To discover what makes an individual item or a whole theory tick, one must, in short, do quite a bit other than passively waiting for understanding to happen. Polya and Szego describe it in the Introduction to their famous analysis book:

> "One should try to understand everything: isolated facts by collating them with related facts, the newly discovered through its connection with the already assimilated, the unfamiliar by analogy with the accustomed, special results through generalization, general results by means of suitable specialization, complex situations by dissecting them into their constituent parts, and details by comprehending them within a total picture".

Understanding is a complementary process to problem solving. In many ways it is more difficult to describe than problem solving, since as Polya points out, it is a matter of "more or less and not yes or no" [Polya 1978]. That is to say, understanding has many levels and is never really totally finished. Actually, understanding, in our sense of building up a knowledge base with all its links and structures, can be taken together with problem solving expertise to comprise a larger view of understanding.

From an information processing point of view, there is a tremendous amount of activity that relates to the building of links and structures. Using the framework outlined in the previous chapters, we shall isolate and discuss aspects of the understanding process. We need not treat it as an opaque phenomenon that happens "as if by magic".

7.2 Deep Understanding

There are many senses and degrees of understanding. Polya abstracts four "levels" of understanding a rule from his readings of Spinoza [MD, p.134] (1) "mechanical" when one has memorized the rule and can apply it correctly; (2) "inductive" when one has tried out the rule in simple cases and is convinced that it works in these cases; (3) "rational" when one has accepted a demonstration; and (4) "intuitive" when one is convinced of its truth beyond a doubt.

Poincare also has some opinions on understanding. In particular, he points out the need for going beyond a mechanical level [Poincare 1929, p.240]:

> "What is it, to understand?...To understand the demonstration of a theorem, is that to examine sucessively each of the syllogisms composing it and to ascertain its correctness, its conformity to the rules of the game? Likewise, to understand a definition, is this merely to recognize that one already knows the meaning of all the terms employed....
>
> For some, yes; when they have done this, they will say: I understand. For the majority, no."

Clearly then, a deep understanding of a theory involves more than knowing just the details of theorems and proofs; it goes beyond simple in-space links. But what should we demand for full understanding? And how should we go about achieving it?

Having deep understanding of a body of mathematics has been likened to knowing one's way around a landscape. We continue with the quote of Polya and Szego:

> "There is a similarity between knowing one's way about a town and mastering a field of knowledge; from any given point one should be able to reach any other point. One is even better informed if one can immediately take the most convenient and quickest path from the one point to the other. If one is very well informed indeed, one can even execute special feats, for example, to carry out a journey by systematically avoiding certain paths which are customary...
>
> There is an analogy between the task of constructing a well-integrated body of knowledge from acquaintance with isolated truths and the building of a wall out of unhewn stones. One must turn each new insight and each new stone over and over, view it from all sides, attempt to join it on to the edifice at all possible points, until the new finds its suitable place in the already established, in such a way that the areas of contact will be as large as possible and the gaps as small as possible, until the whole forms one firm structure."

Thus if understanding is a matter of "more or less", then clearly deep understanding is a matter of "more". A richness of the knowledge base is needed for deep understanding.

Despite the lack of widely used, well-defined stages and criteria for understanding we should not be detered from trying to explicate the understanding process. In the next section we offer some questions to help make the process and levels of understanding more crisp and accessible.

7.3 Questions that Probe and Prompt Understanding

When one understands an individual result, example or concept item, one is obviously in command of much information about it. The following questions probe one's understanding of an individual item in the context of a mathematical theory. At the same time they present a general strategy for understanding. Being able to answer them is symptomatic of understanding an item in a thorough way. Being able to ask them indicates knowledge of *how to* learn and gain understanding. When one can answer these questions, we shall say that one *fully understands* an item.

The intent is not only to make explicit some of the ingredients and processes necessary in the the acquisition of understanding, but also to present them in such a way that the student can learn *how* to go about understanding. Thus the goal is similar to Polya's for problem solving [HTSI] for which his list of "How To Solve It Questions" is offered in the hope of aiding the problem solving process.

The questions are:

1. What is the statement of this item. The setting?

2. Do I understand the statement? Should I review or examine the ingredient concepts, especially the important ones and those to which I have previously not done justice?

3. What is a picture or diagram for this item?

4. Am I reasonably comfortable with this item's immediate predecessors? Are there any predecessors on which I should bone up? Or remember to come back to?

5. Do I know any (or even, all) of dual items for this item, such as counter-examples, model examples, reference examples, culminating results, basic results, etc.? Am I aware of the important ones? Should I peruse some of the others?

6. Can I say what is the gist of this item? Of its statement? Of its demonstration?

7. What is it good for? Why should I bother with it? What is its significance to the theory as a whole?

8. What is the main idea of its proof, construction or procedure? Are the details important? If so, can I summarize them?

9. Is there some way I can fiddle with this item? Perhaps check out a few test cases?

10. What happens if I perturb its statement? Does it generalize? Is it true in other settings? Can it be strengthened by dropping some hypotheses or adding some conclusions? If not, why not: can I cite a counter-example and can I pinpoint what goes wrong? If so, is the new demonstration similar or different from the original. Is it much harder? Should I just be aware that it exists, and forget about the details until I need them?

11. Can I see how this item fits in with the development of the theory as developed in the approach I am taking? What about other approaches. Is this item important or critical or is it simply a stepping stone or a peripheral embellishment?

12. Can I close my eyes and visualize or describe this item's connections to other items in the theory, to the theory as a whole, to other theories? Have I seen anything like it before?

Clearly this list of questions is rather long and one should not be attempt to answer all of them at once. But one should try to pick off as many questions as possible on an initial try, and if the item is important and worth the effort, come back to the list several times. Eventually through work directly with the item and indirectly with other items, one will flesh out answers to most of the questions. The last question is a keystone to understanding in a deep way and should be given a try from the very first exposure to an item. At first, the answer given will be very local, but later it will become more global and encompassing. It might take two or three passes over the material over several years time perhaps, to be able to expound upon these questions, but that is the fullness of understanding that a mathematician strives for in his work and a student should also set as his goal.

In teaching and learning experiences, we have found that the acquisition of full understanding is often a three pass process. On the *first pass*, i.e., on one's first exposure to a subject, which often occurs while one is taking a course, one tries simply (although it is not so simple to do) to become familiar with an item and its immediate associates (predecessors, successors, dual items [1]). One tries to learn the definitions, read through proofs and demonstrations perhaps checking them out on a step-by-step basis. This first phase is very much concerned with one item at a time; it is very local in outlook.

On the *second pass*, which often comes in reviewing a course, one tries to get a more overall feeling for the subject and the flow of its development. Minimally one tries to be able to recite definitions, examples, theorems and their demonstrations. One hopes to see what the essential assumptions and the culminating items are and know how to get to them. This second phase is concerned with items and their relations within their representation spaces and the theory as a whole; it is more global in outlook.

The *third pass* often comes after the course is over, perhaps on another exposure to the material through a different presentation or context, for instance, when listening to a series of lectures "for culture". One starts to see connections between several subjects. One recognizes that the *raison d'etre* of the subject is to address certain questions and that the whole development hinges on certain underlying ideas, axioms or examples; that the subject is very similar to another subject; that many of its items are shared by another subject and are in some sense "the same" as items in another subject. The third pass is thus involved with the theory in a global and trans-theoretic way.

7.4 Knowledge Involved in Understanding

Many of the answers and processes needed to find answers to these questions can be described in terms of our epistemology. Briefly put, the following information is involved in the answers:

1. the statement and setting of the item;

2. the concepts used in the statement especially those in the pre-concepts-dual;

3. a picture, diagram, or ikon for the item;

4. review of predecessor items; possible tagging of items on the basis of worth as items to be placed on an agenda if items to be examined in future;

5. the item's dual with emphasis on epistemlogoical classes;

[1] The could be said to be the "first order" dual items, i.e., $D(I)$, as opposed to second or higher order dual items, i.e., $D^2(I)$ or $D^n(I)$.

6. gist: paraphrase and/or synopsis of statement and demonstration with perhaps a skeletal outline;

7. significance involves look-ahead through the in-space successor and post-dual items with an eye for important items and epistemological classes.

8. overall structure of demonstration: main idea, plan and skeleton;

9. fiddle with variable elements in statement and/or picture;

10. perturb: look in more general setting; drop/add elements of statement; look up references; retrieve known counter-examples;

11. fits-in with successors and motivates post-dual items; depends on predecessors and is motivated by pre-dual items; comes after items in pedagogical trail; there are detours around it; it doesn't go anywhere, i.e., is on a short branch of its representation graph;

12. intra-space, inter-space, and trans-theoretic connections; investigation of *same-ness* relations through dual and analogy relations.

Thus to understand an item in a deep way, one ought to know about: (1) the item itself; (2) its intra-space relations to other items of the same type; (3) its inter-space relations to other items of different types; (4) dual relations to other items of like type; and (5) relations to items in other theories.

7.5 Infra, Intra, Inter and Trans

Having extracted the information needed to answer the questions we can re-group the knowledge into the following categories:

1. *INFRA-ITEM* knowledge: i.e., knowledge of the item itself, specifically of certain slots or elements from the item frame, such as:
   - statement
   - setting
   - picture/diagram

2. *INTRA-SPACE* knowledge: i.e., knowledge of the representation space of the item:
   - the predecessors
   - the successors
   - other items in the same space which are perturbations of the item;

3. *INTER-SPACE* knowledge:
   the dual space
   motivating (pre-) and motivated (post-) dual items;
   the three spaces together

4. *TRANS-THEORY* knowledge:
   items in other theories
   dual and analogy relations with items in other theories and other theories as a whole

5. *EXO-THEORETIC* knowledge:
   references (e.g., bibliographic) to perturbed items
   other expositions

Procedures to feret out such information are provided in the Grokker System [Michener 1977]. For example, some of the procedures that can be applied to an item *I* are:

1. Infra-item knowledge:
   FSTMNT(I)
   FSETTING(I)
   FPICT(I)
   FDEM(I)

2. Intra-space knowledge:
   BACK(I)
   FOR(I)

3. Inter-space knowledge:
   D(I)
   (<*D*)(I), (>*D*)(I)

5. Outside knowledge:
   FDEMO(I)
   FBIBLIO(I)
   FAPPLE(I)

Thus to fully understand an item, one must be able to "zoom" in and out on the item, i.e., shift level of concern among the infra, intra, inter and trans levels; travel around via infra and inter links; perturb items (i.e., solicit information on the perturbations if not actually establish the perturbations) [2]; and survey the overall topography of the individual spaces

[2] To follow the perturbative approach in the fullest sense, one would need interaction with a theorem prover -- human or automated -- and access to large libraries of mathematics -- computerized or not.

and all three spaces together; and link with other theories. Thus in achieving and possesssing full understanding, one establishes and exercises many links, as well as large quantitites of information.

7.6 Degrees of Understanding

Obviously, this is a lot to ask or invoke. Full understanding is very demanding and is clearly not appropriate for all the items in a theory. It does seem appropriate for the most important items: *** and **** items, all model and some reference examples, key and culminating theorems, all mega-principles and some counter-principles. On the other hand, technical and transitional results, some counter-examples and defintions should be treated much more lightly: perhaps with attention restricted to the infra-, intra- and a little of the inter-levels of knowledge.

Concentrating on questions 1-5 and their answers defines a much less demanding level of understanding. We shall call the understanding involved in being able to answer, 1-5 with question 5 modified to call only for pre-dual items, *minimal understanding.* Minimal understanding deals primarily with an item in a limited, local way: that is, only with the item and its immediate predecessors and pre-dual items. It involves only a "1-item" neighborhood of items. There is no global, or overall, survey of its connections. Minimal understanding involves a first-pass level of understanding: the kind of understanding one has when one can state and perhaps repeat the statement and demonstration but cannot say much more.

Most items deserve something between minimal and full treatment. One could define the appropriate level of understanding for an item based on its Michelin rating and its epistemoligcal class.

Such criteria could become part of a model of understanding used in programs that advise learning efforts [Michener 1977]. When the student has satsfied the criteria, one could then say that within the model he understands the item.

7.7 Understanding a Theory as a Whole

Understanding a theory as a whole is more than the sum of understanding its individual items. In addition to the knowledge required for understanding member items, it includes knowledge of the links within the theory and the links to other theory, i.e., of global, cohesive ties that bind the theory together and to other theories. Understanding a theory as a whole, like understanding an individual item, involves information about items and their connections. In addition, it has a perspective which always seeks to view the item in relation to the whole theory.

One can make the analogy between learning an item and learning a theory: to learn a theory, one "pops" up to a level where the "items" are now theories, and the relations between the items are now relations between theories. This is related to the Piagetian sense of "trans" [Sinclair 1978].

Briefly, understanding a theory as a whole involves:

1. knowledge of the epistemological classes: knowing which are the start-up, reference and model examples, the MP's, the CP's, the basic, key and culminating results: *epistemological knowledge.*

2. knowing the "pros" and "cons" of items: which items are good for what; which items are appropriate and when; how to use them; what their limitations are: *annotative knowledge.*

3. seeing the overall intra-space relations of the individual representation spaces; knowing routes and detours (e.g., "from this item I can get to that one"; "this string of items doesn't lead anywhere"; "the following is a quick and dirty way to derive item X"): knowledge of a *mapping nature.*

4. knowing the inter-space relations such as the items used in recurring dual relations; which items are the basis for striking dual relations; knowing which items are dual equivalent, or nearly so; knowing which items are strikingly similar in the dual sense but are not so within their own representation graphs: knowledge of *sameness* and *closeness*, especially in the sense of the dual idea.

5. abstracting and naming the "arrows", or intra- and inter-space relations, (e.g., Q-->R construction is called "completion" process).

6. recognizing dual and analogy links between items in other theories and theories as a whole: knowledge of *trans-theory links.*

7. recognizing clusters of items generalizing or sharing common features and perhaps eliminating common redundancies and elevating them to the "default", "common sense" or "foundation" knowledge.

7.8 Understanding Understanding Mathematics.

Understanding mathematics is a process that can be understood and to some extent taught. In our view of understanding, a good part of the process is concerned with building and enriching a knowledge base. This includes creating associations of many kinds as well as items. It also involves differentiating between various kinds of items according to their function in acquiring knowledge, familiarity, and expertise.

In summary, some of the ingredients of the process of understanding mathematics are:

1. Categorical knowledge of items and relations, general types such as the item/relation pairs of the three representation spaces and dual relations, as well as particular ones such as generalization and specialization;

2. General strategic or control knowledge such as: knowing to restrict the situation under consideration to the particular case of an example, such as a reference example (The "Restriction Principle"); in particular, restricting the situation under consideration to the case of an example of known generality, such as a model example, analysing how things work, and then lifting back up (The "Projection Principle"); knowing to fool around with examples, especially reference or models, when out of ideas; knowing to perturb statements and settings;

3. Meta-knowledge such as knowing to keep one's eyes open for items of special note such as models, references, MP's, etc.; and knowing that keeping track of links by mapping out one's knowledge base (at least thinking about trying to do this) can be a useful not only to keep track of what one knows but to build global understanding;

4. Epistemological knowledge -- knowing that certain items serve particular functions in understanding; and that some ideas and processes, such as the "group" idea [Bourbaki 1950] or the "divide and conquer" technique are very general and pervasive through all of mathematics.

5. Representational knowledge of knowing how to organize and keep track of what one knows such as through maps and networks of items and relations, and through representation schemes, such as frameworks for individual items.

Thus, to understand an item or a theory fully, one must be able to examine it at different levels of detail and from several points of view; follow infra-space and inter-space associations; perturb and fiddle with itmes; and survey the overall topography of the spaces individually and together; and link them with other theories. In short, to achieve a deep sense of understanding one must have established many links of all kinds.

7.9 UUM as a Teaching Methodology

The ideas presented here were used in a seminar with six MIT freshmen. The purpose of this seminar was two-fold: (1) to teach and explore the rich theory of eigenvalues (e.g., the perturbation and location of eigenvalue theorems such as found in Ortega's book [13]); and (2) to make young mathematicians aware of the ingredients and processes involved in understanding mathematics.

The epistemological and organizational ideas seemed natural to the students, especially in discussions in which the students worked out their ideas about keeping track of what they knew and wanted to know. They essentially asked for a representation that included examples, results and definitions, with orderings, and cross-space, i.e., dual, connections. [4] These ideas were also a source of homework problems. For instance, a standard type of problem in the seminar was:

List the dual items for a given item.

Another was:

Tell everything you can about this item.

After the discussion on representation [4] the students were asked as a homework assignment to map out the knowledge domain of the seminar according to our representation scheme; about a month later, they were asked to update their representations. In the seminar we all worked together to meld our representations. While there were some lively debates on how to weave an item into the representation, these sessions always seemed to benefit the students by making them aware of larger issues of how the subject hung together. Thus the organizational process, itself, proved very helpful for developing understanding.

---

[4] After about a month, the students wanted to review and catalogue what had thus far been covered in the seminar. At first, they attempted to list all the items in chronological order. Next, they split this list into two lists (definitions and theorems) and then, a third (examples); they tried to order these according to when items occurred. This, they found unsatisfactory since items came up more than once and chronology seemed to have very little to do with anything. Next, they re-ordered results according to what we here have called "logical support", and examples, by a mixture of chronology and increasing complexity; concepts remained in chronological order (which was essentially this author's pedagogical order). This author then told them about directed graphs and trees and with a little prompting, they adopted the three representation graphs of this paper. They were then happily proceeding to organize everything this way in three colors of chalk, when one of the students jumped up, grabbed another color chalk, and pounding his fist on the blackboard, said, "But that's not all there is: each of these results should be connected to some examples and definitions." And so entered the dual idea.

Another type of problem which they enjoyed involved comparing theorems that addressed a similar topic (e.g., the location of eigenvalues in the Gerschgorin Circle Theorem, Symmetric Perturbation Theorem, and Hoffman-Wielandt Theorem [Ortega, Chapter 3]):

Which theorem is easiest to use, and when?
Which provides the best results, and when?
Cook up at least three (2X2 or 3X3) examples to illustrate your answers.

Most students used reference examples (e.g., the identity, Basic 16) and model examples (e.g., diagonal, upper triangular) in their answers. Together we investigated more complicated matrices with less simple entries (e.g., non-symmetric matrices, matrices with entries of ϵ's and 10^{-1}'s, the Hilbert matrix; [Ortega] has many good examples).

In general, the students displayed a level of mathematical maturity that one would be happy to see in advanced students. They became excellent question askers and idea generators; discussions often left the areas of the author's expertise and entered areas where all were on "hands and knees" together. In short, they became *active*.

7.9.1 A Theorem Proving Anecdote

Even though the emphasis of this course was not on proving theorems but on understanding them, the following anecdote shows how natural some of the ideas of this paper were to them. One of the students, Ken, requested that we prove the Cayley-Hamilton Theorem (CHT) which states that every matrix A satisfies its own characteristic polynomial, $\det(A-\lambda I)=0$. The students agreed to try to find a proof, but they did not want to work out a purely computational proof involving manipulation of 2x2 and then 3x3 matrices with an induction argument for the general case. Also, we did not want to become involved in considerations of the "minimal polynomial" and its attendant algebra. The following is a nearly verbatim report of the dialogue that ensued when the students were asked to suggest a plan of attack:

JOHN *The theorem is certainly true for the identity matrix.*

DAVID *Check. Further if the CHT is true in general, it must be true for diagonal matrices. Right?*

ERM *Right.*

JOHN *That case is easy.*

DAVID *OK. So now we should be able to show it's true for diagonalizable matrices, by using the similarity transform $S^{-1}DS$, on diagonal matrices and* *hoping that the algebra goes away.*

KEN *So?*

DAVID *So, then we can get the general case by doing the same thing on upper triangular matrices and using the fact, i.e., the Jordan Normal Form Theorem which we haven't proved, but know about, and all believe, that all matrices are similar to upper triangular matrices.*

KEN *That sounds good to me.*

JOHN *Does all the algebra come out right?*

ERM *Let's try it and see.*

And so we followed David's plan which does indeed lead to a proof of the theorem once the upper triangular case is established (see [Strang, p. 224] for one way of doing this).

There are several noteworthy features about this episode: (1) the line of reasoning parallels *exactly* the direction of constructional derivation of one branch of the examples-graph we built (see the last chapter): Identity --> Diagonal --> Upper Triangular; (2) they strongly used reference and model examples (e.g., diagonal and upper triangular) of the eigenanalysis domain; (3) the whole interchange was completely spontaneous and took less than a minute. Also, in the actual working out of the details, they argued from the 3x3 case to the general *nxn* case. The rest of the seminar was truly amazed at the speed at which David formulated his plan, and also how pretty it was. David commented that it seemed the "obvious" thing to do. Ken chose to write about this theorem, its proof and the importance of examples as his term paper.

7.9.2 Some Comments on Problem Solving

During the semester, the students met to work on some selected problems in a one-on-one manner. The ground rules were that: these sessions were not tests; they could look up anything they wanted in our notes and references; they could always ask for suggestions and advice; there were no time constraints; and if possible, they would try to think out loud while they worked.

All the sessions were tape-recorded. The problems ranged in difficulty and style from standard questions with a stated goal, such as:

Show that the possible eigenvalues of an involution ($U^2=I$) are +1 and -1.

or:

> *Give a counter-example to show that interchanging rows of a matrix does not leave its eigenvalues unchanged.*

to more vaguely-posed problems, such as:

> *What can you say about the spectrum of a permutation matrix?*

Most all of the students handled the first question by using the declarative definition for "eigenvalue". All the students answered the second question by examining the reference collection of the "Basic 16" (the sixteen 2X2 matrices whose entries are 0's and 1's). Most students attacked the third question by examining the 2X2 cases to form a preliminary conjecture and then some of the 3X3 cases to test and refine it; not all started out this way, but those that tried to attack the problem through more general arguments found they could not get a handle on the problem and thus followed the heuristic of examining the two-dimensional case. To this author's delight they handled these problems with great poise and enthusiasm. They were, for the most part, completely undaunted by the fact that they had to decide what to do with them. As a bonus their answers were very complete.

7.10 Applications to Theorem Provers

The ideas we have developed to describe the understanding process have applications to programs which automate mathematical tasks, in particular the proving of theorems. Among researchers in the field of automatic and man-machine non-resolution theorem-proving, there is considerable interest in providing the theorem-proving programs with a knowledge base that includes heuristic methods and other domain-dependent knowledge. Many [Bledsoe 1975, 1977; Reiter 1973] have found that not only do their programs run more efficiently with the addition of such knowledge, but also that they behave in a way more similar to a working mathematician. Knowledge plus handles and ways to use it are fundamental. As Bledsoe says [Bledsoe 1975]:

> "the word *knowledge* is a key to much of this modern theorem proving. Somehow we want to use the knowledge accumulated by humans over the last few thousand years, to help direct the search for proofs.... So in a sense all of our concerns have to do with the storage and manipulation of knowledge."

It seems reasonable both to this author and some of the researchers [Bledsoe 1978] that ideas presented in this report can be of help.

In particular, in situations where a non-resolution theorem prover (NRTP) desires advice abpout how to proceed with its efforts to prove a theorem, knowing *how to understand* can be important. For instance, the NRTP could be advised to:

(1) Try out the proposed theorem in the special case of a reference or model example, and use this instantiation as evidence -- for or against -- the theorem;

(2) Custom tailor a model example to the specifics of the proposed theorem and examine how the thoerem "works" in this case, and then lift or bootstrap to the general proposed theorem; i.e., apply the Projection Principle.

(3) Find examples for the proposed theorem and consider other theorems that share these examples; in particular, see if the proposed theorem can be proved by methods used in the other theorems; i.e., use the dual idea to locate new data and then pull it over as an analogy;

(4) Look for MP's and CP's that might apply to the proposed theorem and see what they "say" about it;

(5) Find two or more items that are related as a chain of predecessors and successors and which have relevance to the proposed theorem and then try to abstract the procedural information inherent on the connecting arrows;

(6) Look for counter-examples in the collection of known reference and counter examples.

Some of these ideas have already been used by theorem provers, such as (1) and (2) in Gelernter's work.

Chapter 8. CONCLUSIONS

In this report we have talked about the structure and ingredients of mathematical knowledge as used by students and expert mathematicians and found in textbooks and less formal sources. We have developed a structure for representing mathematical knowledge and have singled out noteworthy classes of items in it. We have made explicit some of its non-formal aspects. This study produced a vocabulary and a framework by which to talk about mathematics and how it is understood.

The main idea was to distinguish several broad categores of knowledge, and structure them according to their natural morphisms. We also explored how the categories were related to each other through what we called the dual idea. The main classes together with the internal morphism of each defined three representation spaces for a mathematical theory which can be pictured by directed graphs, where the arrows reflect the sense of the relations: Results-space consisting of results organized according to the deductive morphism of logical support; Examples-space consisting of examples organized according to the constructional derivation of examples; and Concepts-space, organized according to pedagogical ordering. Settings-space is an additional category which we did not explore in great detail; the morphism here would be is-a.

Within each space of objects we also distinguished certain classes that shared similiar and noteworthy roles in learning and understanding a theory. We called these epistemological classes.

In addition to laying out an epistemology of examples, results, and concepts, we also considered some of the processes by which items are generated. We looked at the construction of several examples, briefly touched on some of the general ways in which concepts evolve, and examined the architecture of proof.

We illustrated these ideas in detail in three standard areas from the undergraduate college mathematics curriculum: calculus, linear algebra, and real analysis. Briefly we touched on plane geometry.

With this conceptual framework for mathematical knowledge, and some case studies of particular domains, we then discussed the understanding of mathematics. We tried to explicate the understanding process more crisply than it has been. We offered a group of "How To Understand It" questions which probe and prompt understanding. We discussed how being able to answer them is indicative of a deep understanding and that the information involved in answering them can be easily described in terms of our framework: items and types and also the various in-space and cross-space connections.

Briefly we reported on experience teaching these ideas in a classrooom situation.

With this report's conceptual framework and epistemology for mathematical knowledge, one is now in the position of using it to help explore some of the processes involved in doing mathematics and eventually to provide computational means to support, augment and mechanize such processes. Some of the epistemological homework that is a prerequisite to experimentation is completed and it is now appropriate to experiment.

There are two or three main directions for future work. One is to use these ideas to support programs that must keep track of past mathematical knowledge, such as theorem provers. It is felt that for theorem provers to prove some hard theorems, they are going to need to draw on a bank of past mathematics, including a rich stockpot of examples. The work reported here may prove to be of help with representation problems for such a knowledge base.

These ideas can also be applied to the design and support of interactive environments for experienced mathematicians and neophyte mathematics students. GS (the GROKKER SYSTEM) and GLA (GROKKER LEARNING ADVISOR) described in detail in Parts II and III of Michener [1977] are examples of the sort of systems we envision. GS is designed to help professional mathematicians retrieve and explore mathematical domains. GLA is an advisor program to be used in conjunction with GS to enable neophytes to work in the manner of experienced mathematicians and help them to learn how to understand, in short, to learn as some expert students do. Systems such as GS/GLA have a nice two-way relation with theorem provers: GS/GLA needs the prover to answer user queries and perturbations on the existing knowledge base, and the prover needs a GS-like system to manipulate the knowledge base.

Another project to which these ideas can be applied is the generation of examples to meet certain constraints, which we call *CEG* or *Constrained Example Generation* [Michener 1978]. This project can be pursued for many different purposes. Obviously, it ties in closely with work on improving theorem provers. It has a pedadgogical aspect in that it seeks to understand the example-building process and knowledge involved in it; such analysis and explication could make it more accessible to students. CEG research also is in the classic AI paradigm of trying to unmask the knowledge that experts bring to bear on a problem, structure and represent it, and then build a system to perform the tasks in order to test the model. Future work should pursue all three aspects of the CEG problem. A computational model for the CEG task should be tested not only for its sufficiency (can the model perform), sensitivity and necessity (what knowledge affects the outcomes and is critical), but also for the style in which it performs (do humans do CEG that way).

Thus to put it briefly, it is now appropriate to investigate the evolutionary aspects of mathematical knowledge. That is the next step towards the long range goal of developing a comprehensive theory of how one understands, learns, and does mathematics.

BIBLIOGRAPHY

A. The following mathematics books were used as primary references for this report.

Acton, F. S., *Numerical Methods that Work.* Harper and Row, New York, 1970.

Banks, H. T., Notes from Applied Math 211-212, Brown University, 1968-69.

Beman, W. W., *New Plane and Solid Geometry.* Ginn and Company, New York, 1900.

Bender, C.M., and S.A. Orszag, *Advanced Mathematical Methods for Scientists and Engineers.* Mc-Graw Hill, 1977.

Boravich, Z.I., and I.R. Shafarevich, *Number Theory.* Academic Press, New York, 1966.

Bourbaki, N., *Elements de Mathematique.* Hermann, Paris, 1947-77.

Dugundji, J., *Topology.* Allyn and Bacon, Boston, 1966.

Dunford, N., and J. T. Schwartz, *Linear Operators Part I: General Theory.* Interscience, New York, 1958.

Gelbaum, B. R., and J. Olmstead, *Counterexamples in Analysis.* Holden-Day, California, 1964.

Godement, R., *Theorie des Faisceaux.* Hermann, Paris, 1964.

Goffman, C., and G. Pedrick, *First Course in Functional Analysis.* Prentice-Hall, New Jersey, 1965.

Halmos, P. R., *Finite-Dimensional Vector Spaces.* D. Van Nostrand, New Jersey, 1942.

Hartley, B., and T. O. Hawkes, *Rings, Modules and Linear Algebra.* Chapman and Hall, London, 1970.

Herstein, I. N., *Topics in Algebra.* Blaisdell, New York, 1964.

Hewitt, E., and K. Stromberg, *Real and Abstract Analysis,* Springer Verlag, New York, 1965.

Hocking, J. G., and G.S. Young, *Topology.* Addison-Wesley, 1961.

Hoffman, K. M., *Analysis in Euclidean Space.* Prentice-Hall, New Jersey, 1975.

Ireland, K., and M.I. Rosen, *Elements of Number Theory.* Bogden and Quigley, New York, 1972.

Jacobs, H. J., *Geometry*. W. H. Freeman, California, 1974.

Jans, J. P., *Rings and Homology*. Holt, Rinehart and Winston, New York, 1964.

Kelley, J. L., *General Topology*. D. Van Nostrand, New Jersey, 1955.

Knopp, K., *Infinite Sequences and Series*. Dover Publications, New York, 1956.

Lang, S., *Algebra*. Addison-Wesley, New York, 1965.

Marcus, M., *Basic Theorems in Matrix Theory*. U. S. Government Printing Office, Washington D. C., 1960.

__________, *A Survey of Matrix Theory and Matrix Inequalities*. Allyn and Bacon, Boston, 1964.

Marcus, M., and H. Minc, *Introduction to Linear Algebra*. Macmillan, New York, 1965.

Nevanlinna, R., and V. Paatero, *Introduction to Complex Analysis*. Addison-Wesley, New York, 1969.

Ortega, J. M., *Numerical Analysis: A Second Course*. Academic Press, New York, 1972.

Royden H. L., *Real Analysis*. Second edition, Macmillan, New York, 1963.

Rudin, W., *Principles of Mathematical Analysis*. Second edition, McGraw-Hill, New York, 1964.

__________, *Real and Complex Analysis*. McGraw-Hill, New York, 1966.

Shilov, G. E., *An Introduction to the Theory of Linear Spaces*. Prentice-Hall, New Jersey, 1961.

Spivak, M., *Differential Geometry*. Vol. 2, Publish or Perish Press, Boston, 1972.

Steen, L., and J. Seebach, *Counterexamples in Topology*. Holt, Rinehart and Winston, New York, 1970.

Strang, G., *Linear Algebra and its Applications*. Academic Press, New York, 1976.

__________, and G. Fix, *An Analysis of the Finite Element Method*. Prentice-Hall, New Jersey, 1973.

Thomas, G. B., *Calculus and Analytic Geometry*. Fourth Edition, Addison-Wesley, Massachusetts, 1972.

B. The following works, written by mathematicians, discuss the nature of mathematics.

Banchoff, T., and C. M. Strauss, "Computer-Encouraged Serendipity in Pure Mathematics". IEEE Proceedings, Special Issue on Computer Grpahics, 1974.

Bourbaki, N., "The Architecture of Mathematics". American Mathematical Monthly 57 (1950), 221-232.

Davis, P. J., "Fidelity in Mathematical Discourse: Is one and one really two?" American Mathematical Monthly 79 (1972), 252-263.

____________, "Visual Geometry, Computer Graphics and Theorems of Perceived Type". *Proceedings of Symposia in Applied Mathematics*, Vol. 20, American Mathematical Society, Rhode Island, 1974, 113-127.

____________, and J. A. Anderson, *Non-Analytic Aspects of Mathematics and Their Implications for Research and Education*. Division of Applied Mathematics Memo, Brown University, 1977.

Freudenthal, H., *Mathematics as an Educational Task*. Reidel, Holland, 1973.

Griffiths, H. B., *Mathematics, Hedgehogs and Foxes, An Inaugural Lecture*. University of Southhampton, England, 1966.

Hadamard, J., *The Psychology of Invention in the Mathematical Field*. Dover, New York, 1954.

Halmos, P. R., "Innovation in Mathematics" in [Kline]

Jones, F. B., "The Moore Method". The American Mathematical Monthly, Vol. 84, No. 4, Mathematical Association of America, April 1977.

Kline, M. (ed), *Mathematics in the Modern World: Readings from Scientific American*. W. H. Freeman and Co., San Francisco, 1968.

Lakatos, I., *Proofs and Refutations*, British Journal for the Philosophy of Science, Vol 19, May 1963, No. 3. Also published by Cambridge University Press, London, 1976.

Michener, E. R., *Epistemology, Representation, Understanding and Interactive Exploration of*

Mathematical Theories. Doctoral dissertation, Department of Mathematics, MIT, 1977.

_________, *Constrained Example Generation.* Proposal to NSF-NIE, 1978.

Papert, S., *Teaching Children to be Mathematicians Versus Teaching About Mathematics.* MIT A.I. Memo No 249, 1971.

_________, *Teaching Children Thinking.* MIT A.I. Memo No. 244, 1971.

_________, Theory of Knowledge and Complexity. Lecture to IFIPS World Congress on Computers in Education, Amsterdam, The Netherlands, 1970.

Poincare, H., *Science and Hypothesis.* Dover, New York, 1952

Polya, G., *How To Solve It.* Second Edition, Princeton University Press, New Jersey, 1973.

_________, *Mathematics and Plausible Reasoning, Volumes I and II.* Second Edition, Princeton University Press, 1968.

_________, "Let us Teach Guessing". Etudes de Philosphie des Sciences, en hommage a Ferdinand Gonseth, Editions du Griffon, Switzerland, 1950.

_________, *Patterns in Plausible Inference.* Second Edition, Princeton University Press, 1968.

_________, *Personal communication.* 1978

Polya, G., and G. Szego, *Problems and Theorems in Analysis I.* Springer-Verlag, New York, 1972.

Schoenfeld, A. H., *INTEGRATION: Getting It All Together.* UMAP Module Nos. 203, 204, 205, Educational Development Center, Massachusetts, 1977.

_________, *Can Heuristics Be Taught?.* Proceedings of the Amherst Conference on Cognitive Processes, 1978.

Strauss, C. M., Real Time Computer Graphic Techniques in Geometry. *Proc. Conference on Influence of Computers in Mathematics and Education,* Missoula, Montana, 1973.

Thom, R., "Modern Mathematics: An educational and philosophic error?" American Scientist, 1971, 695-699.

Whitney, H., *Elementary Mathematics Curriculum.* Institute for Advanced Study, Princeton, New Jersey, 1976.

C. The following are references in artificial intelligence and computer science.

Ballentyne, A. M., *Some Notes on Computer Generation of Counterexamples in Toplogy.* The University of Texas at Austin Math. Dept. Memo ATP-24, 1975.

Bledsoe, W. W., *Non-Resolution Theorem Proving.* The University of Texas at Austin Math. Dept. Memo ATP-29, 1975.

_____________, *Discussions of Automatic Theorem Proving,* The University of Texas at Austin Math. Dept. Memo ATP-10, 1973.

_____________. *A Maximal Method for Set Variables in Automatic Theorem Proving.* The University of Texas at Austin Math. Dept. Memo ATP-33, 1977.

_____________, and P. Bruell, A Man-Machine Theorem-Proving System. Artificial Intelligence Journal, 5 (1974), 51-72.

_____________, and M. Tyson, *The UT interactive Theorem Prover.* The University of Texas at Austin Math. Dept. Memo ATP-17, 1975.

Bobrow, D. G., and A. Collins, *Representation and Understanding.* Academic Press, New York, 1973.

Brown, J. S., et al, *Artificial Intelligence and Learning Strategies* Report No. 3634, Bolt Beranek and Newman Inc., Massachusetts, 1977.

Bundy, A., "Analysing Mathematical Proofs (or Reading Between the Lines)". Fourth International Joint Conference on Artificial Intelligence, 1975

Bush, V., "As We May Think". Atlantic Monthly, 176 (1945), 101-108.

Carmody, et. al., "A hypertext editing system for the 360". *Proceedings, Conference in Computer Graphics.*, University of Illinois, 1969, 291-330.

Chester, D., "The Translation of Formal Proofs into English". Artificial Intelligence Journal, 7(1976), 261-278.

Feigenbaum, E. A., and J. Feldman, *Computers and Thought,* Mc-Graw Hill, New York, 1963.

Gelernter, H., "Realization of a Geometry Proving Machine" in [Feigenbaum and Feldman].

Kling, R.E., "A Paradigm for Reasoning by Analogy". Artificial Intelligence Journal, 2(1971), 147-178.

Lenat, D. B., *AM: An Artificial Intelligence Approach to Discovery in Mathematics as Heuristic Search.* Stanford A.I. Lab Memo AIM-286, Stanford University, 1976.

Michener, E. R., "Representing Mathematical Knowledge". Proceedings Second National Conference of the Canadian Society for Computational Studies of Intelligence, Toronto, July 1978.

Minsky, M. L., "A Framework for Representing Knowledge" in [Winston 1975].

Minsky, M. (ed), *Semantic Information Processing.* The MIT Press, Massachusetts, 1968.

__________, "Frames." in [Winston]

__________, and S. Papert, *Artificial Intelligence.* Condon Lectures, Oregon System of Higher Education, 1974.

Newell, A., and H. A. Simon, *Human Problem Solving.* Prentice-Hall, Inc., New Jersey, 1972.

Quillian, M. R., "Semantic Memory" in [Minsky 1968].

Reiter, R., "A Semantically Guided Deductive System for Automatic Theorem Proving". *Proceedings International Conference on Artificial Intelligence*, 1973, 41-46.

Slagle, J. R., "A Heuristic Program that Solves Symbolic Integration Problems in Freshman Calculus" in [Feigenbaum and Feldman].

van Dam, A. *An Experiment in Computer-Based Education Using Hypertext.* Division of Applied Math. Report, Brown University, Rhode Island, 1976.

Winston, P. H., *Artificial Intelligence.* Addison-Wesley, Massachusetts, 1977.

__________, *The Psychology of Computer Vision.* McGraw-Hill, New York, 1975a.

__________, "Learning Structural Descriptions from Examples" in [Winston 1975a], 1975.

D. The following are references from psychology.

Bruner, J. S., *The Relevance of Education.* Allen and Unwin, England, 1971.

Lindsay, P. H., and D. A. Norman, *Human Information Processsing.* Academic Press, New York, 1972.

Piaget, J., *Structuralism.* Harper and Row, New York, 1968.

Piaget, J., *Genetic Epistemology.* W. W. Norton, 1971.

Sinclair, M., *Personal communication.* 1978

E. Miscellaneous

Guthrie, A., *Alice's Restaurant,* 1967.

Heinlein, R. A., *Stranger in a Strange Land.* Avon Books, New York, 1961.

CS-TR Scanning Project
Document Control Form

Date : 3 / 21 / 96

Report # AI-TR-472

Each of the following should be identified by a checkmark:
Originating Department:

☒ Artificial Intellegence Laboratory (AI)
☐ Laboratory for Computer Science (LCS)

Document Type:

☒ Technical Report (TR) ☐ Technical Memo (TM)
☐ Other: ____________

Document Information

Number of pages: 138 (143-images)
Not to include DOD forms, printer intstructions, etc... original pages only.

Originals are:
☐ Single-sided or
☒ Double-sided

Intended to be printed as :
☐ Single-sided or
☒ Double-sided

Print type:
☐ Typewriter ☐ Offset Press ☐ Laser Print
☐ InkJet Printer ☐ Unknown ☒ Other: XEROX COPY

Check each if included with document:

☐ DOD Form ☒ Funding Agent Form ☒ Cover Page
☐ Spine ☐ Printers Notes ☐ Photo negatives
☐ Other: ____________

Page Data:

Blank Pages (by page number): FOLLOW ACK, PAGE, FORWARD, TABLE OF CONTENTS & PAGE#125

Photographs/Tonal Material (by page number): ____________

Other (note description/page number):

Description : Page Number:

Ⓐ IMAGE MAP: (1-138) UN#'ED TITLE PAGE, UN# FUNDING AGENT, UN# ABST. (2), UN# ACK, UN# BLANK, UN# FORWARD, UN# BLANK, UN# TABLES OF CONT (3), UN# BLK, 1-125, UN# BLANK,
(139-143) SCAN CONTROL, COVER, TRGT'S (3)

Scanning Agent Signoff: [signature]

Date Received: 3 / 21 / 96 Date Scanned: 4 / 19 / 96 Date Returned: 4 / 19 / 96

Scanning Agent Signature: Michael W. Cook

Rev 9/94 DS/LCS Document Control Form cstrform.vsd