An Analysis of Mathematical Notations: For Better or For Worse

Barry Biletch, Kathleen Kay, & Hongji Yu

November 8, 2015

Abstract

Mathematical notation is an essential tool for mathematics and sciences. However, the modern system contains a great number of variations and contingencies. This project sets out to explain such contingencies and provide a set of guidelines for good use of notation. We analyze current and historical mathematical notations, trace the developments of notations, and identify the various reasons that they fall in and out of favor. We device a theory to organize principles and analyze the interactions among them. Finally, we examine the validity of our model on several typical examples from modern mathematics and science.

1 Introduction

Mathematical notation is a symbolic representation of mathematics. Mathematical notation range from simple symbols, such as the numerical digits and arithmetic symbols, to more complex concepts and operations, including logical quantifiers and integration, and even to graphical representations of objects, such as Feynman diagrams and Penrose graphical notation. Although termed "mathematical", such notation is used widely in all disciplines of science and engineering. One may argue that mathematical notation is to modern sciences as the Latin alphabet is to English.

Just as there is good writing and bad writing, there are good forms of mathematical notation and bad ones. Good forms of notation make arguments readable and easy to understand, while bad ones render the text illegible or incomprehensible. Throughout the history of mathematics, myriads of new notation emerged, but most of them died out; however, the choice of one notation over another is no simple matter. Some notation has been given up simply because of historical or cultural reasons. Some went out of use because more concise alternatives arose. Some changed as the science itself evolved, and new notation demonstrated new ideas. Moreover, some concepts have more than one accepted notation, each preferred in different contexts.

Different methods of notating a concept can indicate semantically different perspectives. For instance, the Fourier transform of a function f is commonly expressed as $\hat{f}(\omega)$, $F(\omega)$, $\mathcal{F}[f](\omega)$, or $\mathcal{F}[f(x)]$. The former two indicate that the function and its transform are two related functions, while the latter two notations emphasize that the Fourier transform is an operation on a function, producing another function. $\mathcal{F}[f(x)]$ technically operates on an expression, which reduces mathematical purity at the expense of more easily allowing one to take the Fourier transform of an anonymous function. These interpretations, while equivalent, represent very different modes of thought.

The purpose of this work is to describe the criteria by which people choose mathematical notation. We wish to draw out principles that govern the usability of mathematical notation and assess their importance through experiments.

We begin by briefly reviewing the existing literature on this topic in chapter 2. Specialized treatment of this topic is rarer than we expected, so we mainly draw from a few comprehensive studies on mathematical notation, and refer to many other works that treat notation in specific fields or record history of some specific symbols. In chapter 3, we looked more carefully at a few authors who have attempted to describe what constitutes good notation or have criticized bad usage. Through this we try to find a rudimentary set of criteria that we come back to later. Chapter 4 looks at the evolution of mathematical notation in a few specific fields, in attempt to examine the validity of the principles developed in the previous chapter, and to discover new perspectives when the existing criteria turn out unsatisfactory. Next we revisit the criteria of good notation in chapter 5, proposing a new theory that evaluates notation not based on individual symbols but by virtue of the system of rules that generate the symbols. Finally in order to confirm the validity of our hypotheses, in chapter 6 we design a test that compares the usability of different forms of notation. An example of such a test is given in the appendix.

In this work, we will focus on the discussion of principles rather than producing a list or a comprehensive history of all mathematical notation. We will overlook the variations in notation that are completely stylistic or are the result of arbitrary conventions, but carefully examine those where there are scientifically or contextually relevant reasons to favor one over the other.

2 Literature Review

2.1 History

Exhaustive discussions focused on mathematical notation are rather rare. Florian Cajori published the most comprehensive study on the history of mathematical notations[11]. He organized his work by different fields of mathematics, including arithmetic and algebra, geometry, modern analysis, and logic, and concluded by a discussion of general principles. He wrote extensively and with great depth, covering a great volume of notations used during his time, most of which present themselves in usage today. However, mathematics and other sciences have advanced considerably in 20th century, making his work rather outdated. Nevertheless, his survey of previous scholars' discussions on good and bad notation still lends insight to our study. Other books titled history of mathematical notations can also be seen, such as Mazur's *Enlightening Symbols*[27]. Many fall into the category of popular science and focus more on elementary symbols in mathematics, such as digits and arithmetic operators, and therefore shed little value onto our discussion.

Works on the general history of mathematics often mentioned the history of mathematical notations. However, in many of these works, someone would transcribe the results found by early mathematicians into the modern notation for the sake of clarity. In his work on the history of mathematics[9], Cajori recounted the history of mathematics; he covered very much the same subjects as in his book on notation. Also, alongside the history of various concepts, he mentioned the invention of the notations used to represent them. Ball[2], Smith[33], and Miller[28] produced similar work from that period with varying coverage of topics and depth of discussion. Authors, including Boyer[7], Cooke[13], and Hofman[23], composed more recent works on these topics. However, even the latter works seldom go beyond the mathematics of the 19th century or the early 20th century. Except for early number systems, such writings only mentioned notations in passing, with the date and first user noted and while lacking a discussion on the reasons for adopting certain notations.

Besides the general histories of mathematics, some books on the history of a specific subfield of mathematics exist. Since these works have a more limited scope, they often have discussions with greater detail. Some have lent considerable insight into the reasons for adopting some notations. Boyer's volumes on calculus[6] and analytic geometry[5] provide good examples of this. At many points in his work, Boyer cited differences in notation used by early authors and those by modern conventions. Kleiner's work on the history of abstract algebra[25] also laid weight on the development of notation in early algebra.

2.2 General Discussions

Besides historical accounts, there are also many attempts to describe the inventory of modern mathematical notation, often for pedagogical purposes.

Scheinerman compiled a guide[31] to mathematical notation aimed at engineers and scientists. He covered mostly common symbols and notations seen in applied mathematics. Since it merely listed of all the notations and corresponding concepts, no discussion pertained to the notations' uses. However, such a record of prevalent mathematical notation still provides a useful reference. Similarly there are also many "mathematical handbooks" which often are collections of definitions and theorems in more applied branches of mathematics.[8][29] These books sometimes recorded variations in notation for those with more than one accepted form.

Additionally, some of the books that feature discussions on mathematical notation aim at teaching or describing the general style of mathematical literature, or more specifically, mathematical proofs. For example, a textbook by Bloch, *Proofs and Fundamentals*[4], "introduce[s] students to the formulation and writing of rigorous mathematical proofs". Methods of construction of mathematical proofs are introduced, and guidance on styles of presenting such proofs are given. Some concepts incorporated the comparison of different notations. Steenrod, Halmos, Schiffer, and Dieudonne each wrote about the style of mathematical writing, compiled into the collection of essays, "How to Write Mathematics"[35]. The contributors, being all well-known mathematicians, gave the compilation great value and their rich experience in writing mathematics bestowed it great credibility. All of the authors admitted that there is no uniform convention, and that it is difficult to write a guide that even working mathematicians agree on. They did, however, point out the importance of consistency in choosing symbols and the balance between symbols and words.

The Princeton Companion to Mathematics edited by Timothy Gowers is a comprehensive reference that introduces fundamental mathematical concepts, modern branches of mathematics, important theorems and problems, and well known mathematicians. Since the book is designed to be an encyclopedia, whenever a concept is introduced, the corresponding notation is also discussed. Invention of notations is also sometimes mentioned as contributions of certain mathematicians.[20]

2.3 Specific Usages

2.3.1 Elementary Geometry

In his book on history of mathematical notation, Cajori grouped symbols used in elementary (Euclidean) geometry into three types, pictographs for geometrical objects, ideographs for concepts and relations in geometry, and symbols from algebra. He then proceeded to trace the history of various symbols, some still in use, such as $\triangle$, $\angle$, some obsolete, such as the symbol for parallelogram and using $\Bumpeq$ for equivalence. For ideographs, usages derived from their geometrical meanings are also mentioned, such as using $\blacksquare$ for the algebraic operation of squaring a number. The convention for using letters in geometry was noted. Usages of $+$, $-$, $=$ in geometry were described and compared to their algebraic uses. Cajori also wrote about the conflict of ideology between symbolists and rhetoricians in elementary geometry, from which he drew the principle of moderation, or more specifically balancing the usage of symbols and natural language.[11]

Notation was not the focus of Boyer's *History of Analytic Geometry*, notation was not the focus, but is often discussed. Boyer focuses on the idea conveyed by a new notation, which is helpful to our studies. It was especially notable that in the book he emphasized the transition from geometrical notation, such as using two letters that represent the endpoints to denote a line segment, to analytic notation, where one uses letters to represent lengths, coordinates, and uses equations to represent geometrical objects.[5]

2.3.2 Advanced Geometry

The meaning of the word geometry to working mathematicians has changed greatly from 19th century to now. While it used to mean exclusively Euclidean geometry, now it generally refers to the study of objects on manifolds. The change has been gradual, and much of the notation was simply an extension of related symbols used in analysis and algebra. An account for the early development of differential geometry (extrinsic differential geometry) along with the notation can be found in Struik's two articles on the subject[36][37].

Einstein's theory of relativity relies heavily on the idea of tensors, where he employed the summation convention now named after him, which was originally developed by Schouten[32] to denote the Ricci calculus. This notation gained great popularity following Einstein's introduction into physics due to its conciseness in calculation.

Modern geometry as we see it now in the coordinate-free formulation owes its appearance to Élie Cartan. He introduced the exterior derivative, reintroduced the idea of exterior algebra which was initially invented by Grassman, and constructed the spine of today's formulation of geometry on a manifold[12].

2.3.3 Arithmetic and Algebra

In ancient history, mathematical notation was not a rigidly defined ideal. Different people, even within the same region and time period, used their own notations, usually pictures or words related to the concept. In ancient Egypt and Mesopotamia, for instance, addition and subtraction were sometimes represented by an image of legs facing towards and away from the operands, other times by the words "tab" and "lal," and the rest of the time by yet other representations [39].

The ancient Greeks also used nonstandardized notation, usually full words representative of the operation. However, circa 250 A.D., Diophantus invented a system that standardized the Greek world. This notation used $\overset{\circ}{M}$, ζ, Δ^{Υ}, K^{Υ}, $\Delta^{\Upsilon}\Delta$, and ΔK^{Υ} to represent the 0th through 5th powers of an unknown [39], which were concatenated with coefficients (also represented by letters, in a manner similar to Roman numerals) to form what we would today recognize as polynomials. Interestingly, this notation requires all negative terms to be separated from the positive ones (subtracting their sum from the positive terms), suggesting that there was not yet the concept of negative numbers; only that of subtraction of the corresponding positive number.

In early Indian mathematics, operations were represented by an abbreviation of the word (e.g. multiplication was represented by "gu," a shortening of "guna," meaning "multiplied") [39]. Unknowns quantities were not used until the 6th century, when they and their powers were again represented by abbreviations, similar to the Greek style. Curiously, while the first unknown was a shortening of "yāvat-tāvat," meaning "so much" or "how much," the others were all shortenings of colors.

Early Chinese notation differed extremely, both from contemporary notation and from anything that we might expect today. Functions of unknowns were expressed as a 2-dimensional arrangement of numbers [39]. This eliminated the need for explicit naming of the unknowns and for operations to even be written. While this notation sufficed for its initial purpose, it generalized extremely poorly to new concepts, which hindered Chinese mathematics as time progressed.

There is less information available on Islamic notation. However, the evidence points to the notation in use by the 12th to 15th centuries as being more engineered and less organic than other notations [39]. This notation includes proportions, square roots, quadratic equalities, and, by the 15th century, symbolic algebra. Of course, it also includes the numerals that, with only minor modifications, we use today.

2.3.4 Abstract Algebra

On notation used in abstract algebra, Florian Cajori's work only studied two main objects, determinants and vectors. The notation for determinants were traced back to seventeenth century, again to Leibniz. Cajori observed the variation in the letters and indices representing each entry, and the alignment and organization of the entries. He then studied the conceptual and notational relation between matrices and determinants. Notation for special determinants such as the Jacobian and Hessian were also noted. On notation for vectors, Cajori recorded the evolution from geometrical symbols to algebraic symbols. He also discussed the notation for operations on vectors and for vectorial operators. He then commented on later attempts at unification of notation in vector analysis, which by his time has reached a "deadlock", due to many unsettled disagreements between Hamilton's quaternions and Gibb's vector notation, as well as World War I. Finally, notation for tensors were mentioned briefly with a short introduction on Schouten's index notation and summation convention. Various symbols for Christoffel symbols were mentioned in passing. In this chapter, there were not many discussions on why some notations were more popular, possibly due to the relative novelty of the subject at that time. Cajori did, however, refer to various authors who wrote on the best notation for vectors, and posed several questions that must be answered in order to reach that end. [11]

In *A History of Abstract Algebra*, several notations were specifically noted, including Cauchy's introduction of the cyclic notation for permutations, Cayley's introduction of matrices, and Gauss' congruence notation. The benefits of these notations were not discussed in great detail, only the conceptual importance was noted of Cayley's notion that matrices themselves should be subject of study and should constitute a symbolic algebra.[25]

In modern textbooks on abstract algebra, many notations are introduced as analogues of familiar algebraic operations for real numbers. For examples, in group theory, both additive notation and multiplicative notation for groups are used. Multiplicative notation is used for general arguments due to its conciseness, while additive notation is used when the group is commutative or when denoting cosets. For rings and fields, fraction notation are also introduced.[21][19]

In abstract algebra, classifying objects of certain property is of great interest, and developing a method of naming them systematically is an accompanying issue. The Schönflies notation and Hermann-Mauguin notation for point and space groups[14], the names of finite dimensional representations of point groups[14][18], and the names for Lie groups and Lie algebras [18][22] are good examples of these notations. Since these notations are often invented systematically and used by the discoverer of the results, there is little dispute over them. Nevertheless there are coexisting conventions[14].

2.3.5 Logic

According to Cajori in his *A History of Mathematical Notations*, the earliest contributor to formal logic, and the pioneer in inventing notation for concepts in mathematical logic is again Leibniz. Even though logic was not treated as an independent field, many symbols for concepts that we recognize in mathematical logic were invented. Mixed among logical operators and quantifiers, the history of shorthands in arguments such as $\because$ and $\therefore$ are also noted. Cajori then proceeded to trace the work of logicians up to his time. Among them de Morgan, Boole, Frege, Peano, Moore, and Russell's contributions were discussed in detail. There has been great diversity in the choice of symbols throughout history, but the origins of each symbol in the modern notation can be traced back to specific authors. At the time of Cajori's writing, there was not yet a consensus on the proper notation for mathematical logic.

Edited by Jean van Heijenoort, *From Frege to Gödel: A Source Book in Mathematical Logic, 1879-1931* is a collection of individual writings and letters among logicians. From this collection one can see explicitly the notation used by each author, and read many discussions on symbols and formalization of mathematics in general.[38]

From Cajori and van Heijenoort, one can see that the concern of mathematicians of early 20th century was of a more general nature than inventing symbols. As famously attempted by David Hilbert, the program was to completely and consistently axiomatize all of mathematics.[34] Using only formal language was a prerequisite. Even though the task of proving the consistency of such a formal system has been shown impossible, completely expressing mathematics with unambiguous symbols is possible, if not practical.[30]

Later efforts developed advanced fields in logic such as proof theory and the study of axiomatic systems. However, no notably new notation was invented, and most symbols were derived from set theoretic conventions or already existing in earlier mathematical logic.[3]

2.3.6 Analysis

In Cajori's *A History of Mathematical Notations*, he presented an extensive discussion of notation in modern analysis. First he traced the origin and development of trigonometric notation, which include symbols for angles, sides of a triangle, the trigonometric ratios, and functions derived from the trigonometric functions, such as hyperbolic functions and inverse trigonometric functions. Next he opened a discussion on notation used in differential and integral calculus, first by noting Leibniz' great devotion and contribution to the subject of notation. A list of notations invented or used by Leibniz was given. The development of symbols for differentials were studied, in a both chronological and logical manner. Mathematicians before and after nineteenth century were studied separately, and the symbols for total differentiation and partial differentiation were each discussed in detail. The symbols for integrals were then reviewed, mostly attributed to Leibniz. Other symbols used in calculus such as the symbols for limits are mentioned. Besides calculus, Cajori also wrote about symbols in theory of functions. Both symbols for functions in general and those for special functions are noted. In the discussions of this chapter, he often cited supporters and dissenters of certain notations for the reasons why they are good or bad. At the end of the section on notation in calculus, he also made some general remarks on qualifications for a successful notation, which we will refer to later. [11]

The History of the Calculus and its Conceptual Development by Boyer is a conceptual history of calculus from early Greek mathematicians to the end of 19th century. Compared to Cajori, Boyer focused much more on the discovery and formalization of calculus. However, the notations used by various mathematicians were also described carefully, sometimes accompanied by a discussion of why one notation was or became popular. Similar to Cajori, Boyer also spent many pages on Leibniz' notation, accrediting him for many symbols that are generally accepted now, but also pointing out the difference in their meanings throughout the history. In his description of early mathematicians' works, Boyer was also very aware of misunderstanding that might be caused by transcribing early concepts into modern notation. Unlike Cajori, however, Boyer did not attempt to give a summary of the reasons why notations are adopted or abandoned. [6]

3 Why are Notations Good or Bad?

The most used notation in mathematics is prose. Historically, before we developed more specialized notations, mathematics was written in prose [42]. Even today, we use prose (to varying degrees) to logically connect the mathematically notated statements in proofs and other large mathematical works. As computers assume an increasingly prominent role in modern mathematics, it is constructive to consider the value in this approach. It may be easier to express and to understand more complicated concepts using prose, but this can also introduce ambiguity, which is antithetical to the principles of mathematics and especially problematic in today's age of computerized calculations and theorem provers.

Therefore, it is important to achieve a useful balance between the use of symbolic notation and prose. Historically, before symbolic notations were developed, mathematics was written exclusively in prose ("$x+x^2 = 1$" would be written as "an unknown plus its square is equal to one"). This is highly tedious, for both the writer and reader, and can mask patterns due to its excessive verbosity. As a result, mathematicians developed symbols to take the place of commonly used words and phrases: first numerals, then operations ($+$, $\cdot$, etc.) and algebraic variables. In 16th century, the $=$ sign was introduced, solidifying the concept of an equation as a mathematical object in its own right. This is approximately the level of abstraction used by most mathematicians today: equations are written fully symbolically, but connected logically with prose (e.g. "We know that $3x^2 + 9x - 12 = 0$, so, using the quadratic formula, can find that $x = \frac{-9\pm\sqrt{9^2-4\cdot 3\cdot -12}}{2\cdot 3} = -4, 1.$"). However, there is further notation available. In the 1880s, Peano introduced a symbolic notation for logical reasoning, which make it possible to reduce or even completely replace the connective prose. In fact, Whitehead and Russel's *Principia Mathematica* [40], published approximately 30 years later, is famous for doing exactly that: it is "probably. . . the most notation-intensive non-machine-generated piece of work that's ever been done" [42]. This has the advantage of being far more easily understood by computers (for the purposes of verification, etc.), but the disadvantage of being *less* easily understood by humans.

Thus the question is raised: is this disadvantage inherent to the use of notation for logical reasoning, or can it be overcome? One can easily make the assumption that upon the introduction of any new symbolic notation meant to replace prose, mathematicians were similarly confused. However, as these notations are assimilated over the years, we become able to wield them proficiently, even preferring their conciseness over their prosaic predecessors.

More importantly, given the current convention of writing mathematics in a blend of natural language and logical expressions, what makes an optimal configuration that is both concise and rigorirous, while being clear easy to understand by humans?

As we have seen from the brief literature review, despite the vast amout of writings that mention mathematical notation, not many people have systematically considered this last question. In this chapter, we shall look at a few serious attempts at answering it, and hope to draw a few general criteria that determines what is good notation.

3.1 Steven Wolfram

In a 2010 talk, Stephen Wolfram discussed the mutual influence between mathematics and its notations. [42]

Long before recorded history, numerical representations started with unary notations, such as tally marks. It is a very simple system that maps cleanly to the early concept of a number, so it is no surprise that the idea arose independently all over the world. However, while unary works well for counting and trivial arithmetic, it rapidly becomes impractical when trying to deal with larger quantities or more complicated mathematics. As a result, more complicated number systems emerged.

The two main classes of numerical systems beyond unary are positional notations (where the location of a digit within the representation affects its value) and value-based notations (where each symbol has a fixed value, and the value of a number is the sum of its symbols). Positional systems have many benefits, including the ability to more easily express very large numbers and (usually) simpler calculations. These advantages arise from their level of abstraction; however, this same abstraction makes them more difficult to understand: a 3 can mean different things based on where it is located. Most early civilizations were not yet advanced enough to understand this concept, so they arrived at value-based systems, which more closely model our language and how we think.[1] The notable exception to this is the Babylonians, and potentially their predecessors, the Sumerians, who used a base 60 positional system.

Wolfram also argues that notation has historically held back the development of mathematical ideas by preventing the extensions necessary to support these new ideas. Number systems using letters (including Greek

[1] Most, if not all, languages have separate words for one, ten, hundred, thousand, etc. Therefore, value-based systems, such as Roman numerals, which also have separate symbols for these values, are easier to understand.

and Roman numerals), for instance, impeded the development of algebra because they did not permit the use of letters as symbolic variables. Various systems evolved to work around this (see Section 2.3.3 for Diophantus' notation for polynomials), but they all failed to properly capture the concept in a way that is easy to work with, generalizes well, or exposes useful patterns.

3.2 Florian Cajori

At the end of his two volumes on the history of mathematical notations, Cajori made some summaries of his discoveries in the history of mathematical notations [11]. He traced the forms of symbols used throughout history, and first categorized the forms of symbols into single characters, which include ones abbreviated or derived from words and pictographic symbols, and compound symbols. He noted that the increased use of typographic machines has driven authors away from notations that have to be set in multiple lines. (This is an interesting observation, and we shall explore it in later chapters.) He also pointed out that most inventions of symbols are done by individuals. He then wrote that some symbols are merely shorthands used to compress the writing, and others are designed and organized to demonstrate "logical relationships". He asserted that superior notations should be adaptable "to changing viewpoint and varying needs".

On the selection and spread of symbols, Cajori pointed out that the adoption of mathematical symbols is usually brought about by groups of mathematicians, and the inventory of notations accepted now has a great variety of sources. Furthermore, all attempts at creating a grand system of notations for the entire science led to little or no success. According to his findings, the choice of symbols is often a result of social circumstances, rather than inherent values of the symbols, and it is often hindered by habits. Finally, he observed that mathematical notations cross language barriers faster than mathematical concepts, even more so with abbreviations and ideographs than with pictographs.

Cajori also noted that the selection of symbols is not a perfect process; mathematical symbolism is always in a "state of flux". Symbols once popular can go into disuse, and previously ignored symbols can gain popularity. There are also often symbols that have multiple meanings and multiple symbols that share a single meaning. We shall argue that this is due to both the latency between the invention of a notation and its acceptance or popularization, and the changing standard for optimal notations resulting from the development of the science itself.

In the end, Cajori criticized sharply that as a community, mathematicians do not have any effective cooperative mechanism that could smooth out the many conflicts and confusions in mathematical notations. Historically, notations have mostly evolved without any systematic method for developing or choosing the best, resulting in a very diverse and often contradictory system. In constrast, he took the example of the studies of electromagnetism and astronomy to show that such international collaboration is possible. Cajori did mention that near when he wrote his book, there was group effort in mathematics to unify the notations in vector analysis, theory of potential and elasticity, and acturial science. The former two were met with failure, partly due to World War I, but the latter was relatively successful. He proposed that introducing uniform notations should be done in two steps. First, international representatives should meet and reach agreement, and then individual mathematicians should be willing to adopt the results of such conferences. The previous failures, he concluded, are because one or both of these steps were not carried out.

Cajori's extensive survey is so far the only literature that deals exclusively with mathematical notations and is an important reference for our work. From his discussion, we can draw the following principles for good notations:

- Good notations should be concise.
- Good notations should be extendable when new ideas emerge.
- Good notations should be typographically easy to produce.

Rather surprisingly, these principles are not very selective. Cajori did not endeavor to develop these principles in his work. To him, it seems that most notations adopted in history are results of social circumstances and are largely arbitrary.

3.3 Norman Steenrod

In *How to Write Mathematics*, Norman Steenrod wrote about what he considers good practice in writing mathematics [35]. He took care to distinguish clearly the two components of mathematical writing: "the formal or logical structure consisting of definitions, theorems, and proofs, and the complementary informal or introductory material consisting of motivations, analogies, examples, and metamathematical explanations." The matter of choosing mathematical notations appears mostly in the formal part, which is what we will focus on.

Steenrod summarizes the criteria of a good organization of the formal structure of a mathematical work as “(1) length, (2) the quickness with which one obtains major or interesting results, (3) the simplicity of the start, and the gradualness of the approach to difficulties, (4) the quickness with which examples and intuitive materials can be developed, and (5) aesthetic satisfaction.” At face value, criteria (3) and (4) do not overlap considerably with the domain of notations, and we have agreed to ignore criterion (5). Upon closer examination, what Steenrod intended to establish has most to do with the organization of writing, so he deemed deciding what symbols to use “a minor problem”. However, from the perspective of organization of formal structure, he did point out that the author should use well-accepted notations, and use extensive global notations. He pointed out the advantages and disadvantages of such choices; that is, using fewer global notations reduces the length of the work and allows for compact expressions, but may also confuse a careless, scanning reader and is less robust against typographical errors. Deciding that the pros outweigh the cons, he also pointed out that accompanying mathematical notations with redundant explanations can reduce the burden of readers, and thus recommended that they be “strategically placed”.

Despite not paying close attention to notations, Steenrod still proposed some principles for good notations, which we shall summarize as:

- Use notations that reduce the length of expressions.
- Use notations that are well-known.
- Use notations consistently throughout a work.
- Accompany notations with moderate redundant explanations in a long work.

3.4 Paul Halmos

In Paul Halmos’s essay in *How to Write Mathematics*, he set off by describing the difficulty of making a guide to good mathematical writing, and resorted to presenting his own guidelines of writing. The essay is rather loosely organized, and develops fewer than twenty topics. Three sections are dedicated to notations, and more discussions are scattered among many of the topics.

The first section that explicitly concerns notations is “Think about the Alphabet”. Halmos recommended that before writing, one should consider first all the letters or symbols that one will use. He used a few examples to first suggest that a good choice of alphabet should be consistent and without conflicts. Avoiding conflicts means that a symbol should not carry different meanings in the same expression or even in the same article. Notations that cause implicit conflicts such as writing x_p^i for one type of object and x_i^p for another are also considered bad, because when the indices take numerical values, it becomes impossible to differentiate them. Consistency is more subtle, and perhaps also more stylistic. Halmos considers $ax + by$ or $a_1x_1 + a_2x_2$ more consistent than $ax_1 + bx_2$, from which a rule can be drawn that one should utilize as few ways of generating new symbols as possible; that is, labeling with indices, exploiting the alphabet, or using prime marks or dots.

The second suggestion that Halmos offers on symbols is to resist them. With a few examples, it appears that what he actually intended to say with this quite vauge statement is that one should avoid using superfluous symbols in arguments. For instance when saying, “On a compact space every real-valued continuous function f is bounded”, it is unnecessary to include the “f”. He did admit that in some cases, a dummy symbol helps the argument, but insisted that one should be careful referring to those statements. The general rule that we can draw from this is that mathematical notations should be used only when necessary.

The last guideline that is explicitly related to mathematical notations is to “use symbols correctly”. In this section, Halmos calls for attention to symbols that may be translated multiple ways into natural language. For example, $\in$ can be read both as the preposition “in” and the verb phrase “is in”, where only the latter is appropriate in the strict sense of the symbol. The same can be said about $\subset$and $\leq$, which are really predicates rather than connectives.

In other sections of the essay, Halmos also contemplated the balance between using symbols and using words to express a statement. He argued that even though logical symbols make statements short and concise, they also make statements difficult to construe. If the reader would need to translate the logical symbols to natural language before he could understand it, then the writer should use words in the first place.

Summarizing the points made by Halmos, good mathematical notation should satisfy the following:

- Be consistent.
- Avoid explicit and implicit conflicts.
- Avoid superfluous symbols.

Here, consistency has a multitude of meanings, which we shall expand and reformulate in the following discussions.

3.5 Summary

To summarize our findings through the readings in this chapter, good notation should satisfy a few criteria. As mentioned by all of the authors discussed above, consistency or unambiguity is an important principle. This principle manifests in two ways. First, each object or concept should only be represented by one symbol or one class of symbols. Any switch of notations if absolutely necessary should be also indicated explicitly. Second, each symbol should only correspond to one object. There should not be any ambiguity in what a symbol means given reasonable context. It is impossible to keep the meaning of generic symbols such as x unchanged throughout an article that is long enough, but within one proof, this guideline should always be followed.

The next criteria that many has mentioned is conciseness. Conciseness is realized by choosing the shortest form of expression without compromising consistency. This principle is also tied to another important factor that is the easiness or difficulty in producing a notation typographically. Since in mathematics and sciences, digital word-processing and typesetting is the norm, symbols that cannot be encoded and displayed easily should be not used.

Another important principle is following the convention. Modern mathematics has always been very formalized, with many traditions set by previous mathematicians. Therefore, when developing new notation, one must be aware that existing conventions in the academic community is very resilient to change. This is also why many attempts at unifying the notation in all field of mathematics have failed.

As we shall see in some case studies, these principles are rather primitive. In practice, some of them may conflict and one may face the dilemma of choosing which principle to give up. Another more fundamental problem is that these criteria apply primarily to individuals symbols, but since mathematical notation is a system of symbols, it would require a theory that addresses the system as a whole to really evaluate its usability.

4 Case Studies

4.1 Methodology

In the following sections, we shall look at the evolution of mathematical notation in particular fields. First we study notation in calculus, to demonstrate the criteria for good notation, and to show that different symbols need to be chosen in different contexts, and it would be counterproductive to try to unify them all under all circumstances. Next we take a brief look at how different notation affects the teaching and learning of mathematics. After that, we investigate the history of notation for vectors, focusing on how changes in notation are driven by evolution of mathematical concepts. Afterwards we show the advantages and disadvantages of three different notations for Maxwell's equations. With this example, we demonstrate that the shortest notation may not be the best, and further that the criteria developed in the previous chapter is not universal. Finally, we introduce the notation for space groups as an instance of notation that contains many diverse symbols, but can be generated by relatively few rules. This final case study will lead to a new perspective on mathematical notation.

4.2 Calculus

> Perhaps no mathematician has seen more clearly that LEIBNIZ the importance of good notations in mathematics. His symbols for differential and integral calculus were so well chosen that indeed one is tempted to ask in the words of GOETHE'S FAUST, "War es ein Gott, der diese Zeichen schrieb?" But this excellence was no divine inspiration, it was the result of patient and painstaking procedure. [10]

Leibniz held off on printing new notations until he was satisfied with them; he wanted to have good notation. This desire allowed for his notations to stand the test of time; most notably, Leibniz notation in calculus has survived to this day.

4.2.1 Derivatives

Today, small increments of a variable x are denoted by dx or ∂x or even Δx. That was not always true. Before Leibniz's accepted dx, Fermat used e for a small increment of x whereas Barrow used a for x and e for y. If a and e

still represented small increments today, confusion would occur with e, the base of the natural logarithm, and the common use of a as a coefficient.

Other than Leibniz's $\frac{dy}{dx}$ and Lagrange's $f'(x)$, two other forms for the derivative are f_x and Newton's $\dot{y}$. The former is a Lagrangian form for partial derivatives, whereas the latter is a time derivative used primarily in physics and mathematics.

Basic issues arise when using $f'(x)$ and $\dot{y}$ for higher order derivatives. When taking the fourth order or higher derivative, the standard prime mark is not used. The main notations are $f^{(n)}$ for the nth derivative and either $f^{(xi)}$, $f^{(XI)}$, or f^{XI} for the fourth. Newton's notation for time derivatives is not typically seen above second order. If it were, then the numerous dots above the function would be unsightly and difficult to count. Additionally, LaTeX does not have a simple way to denote a third time derivative in dot notation, suggesting that it is not common.

On the other hand, $\frac{d^n}{dx^n}$ is a much cleaner notation. In partial derivatives, a similar notation is used: $\frac{\partial}{\partial x}$. Between $\frac{\partial}{\partial x}$ and f_x, the former is more pleasing when taking many partial derivatives and when it is with respect to different variables. For instance, when writing $\frac{\partial^4 f(x,y)}{\partial x^3 \partial y}$ and f_{xxxy}, Leibniz's notation reads and writes more compactly. Using Lagrange's notation for even higher powered partial derivatives can become cumbersome.

These notations have their advantages and disadvantages. When one wants to denote the nth derivative in one-variable space, the Leibniz and Lagrange notations use a simple superscript n. For low order derivatives, especially time derivatives, Newton's dot notation and Lagrange's prime notation have the upper hand in terms of doing written calculations and with saving time. With Leibniz's notation, the object being differentiated is identified as well as the variable in respect to. This aids in separation of variables.

4.2.2 Integrals

Leibniz was the first to use $\int$ for integration. He preferred *calculus summatorius* with a long "s" over Bernoulli's "I" for *calculus integralis*. Interestingly, both men have ties with the terminology used today. Bernoulli's "integral" and Leibniz's "$\int$" are seen. [10] Furthermore, in a correspondence with Bernoulli, Leibniz wrote

$$d^m = \int^n \text{ when } n = -m. \text{ [11]}$$

This is not the conventional nomenclature for integration. The "n" is the n-th integral. The same concept today would display

$$d^{-2} = \int\int$$

although we do no use d^{-2}.

At one point of time, the integral was used similar to summation, of which the $\int^n$ was seen. A comparison between the two times is as follows,

$$\begin{gathered}\textstyle\sum_{x=1}^{m} x = \int x \\ \textstyle\sum_{x=1}^{m} x^2 = \int^2 x \\ \dots \\ \textstyle\sum_{x=1}^{m} x^n = \int^n x.\end{gathered}$$

The old notation lacks the maximum value to be added. The summation using sigma identifies the first and last number to be added.

4.3 Teaching and Learning of Mathematical Notations

When initially learning multiplication, children use "×". With the introduction of algebra, this "×" can be mistaken as the variable x, or vice versa. The use of "·" (or simply using parentheses) replaces "×", but "×" does not go away. Scientific notation uses "$\times 10^n$", and so does the cross product. The "·" becomes common practice in representing multiplication. With vectors, it is defined as the dot product. The initial use of "·" works the same way as with one-dimensional 'vectors', which avoids ambiguity. The dot product and parentheses present themselves in a clearer and less mistakable way than "×" does. Of course, they have minor disadvantages as well. Parentheses are used to denote vectors, points, and the object consumed in a function. The latter includes functions that have components which vary based on value (non-constants), so either the function symbol is mistaken for another meaning or the variable is not conventional (e.g. $\sigma(x)$ or $P(\cos\psi)$).

The dot notation takes on a different meaning when working with group theory. The dot notation is referred to as a binary operation.

In textbooks, notations and symbols are usually defined. Occasionally, the notation is changed so that its original symbol can be used to represent a different value. When teaching or learning using one of these books, confusion can arise. Additionally, notation is varied from field to field or book to book. One strong example of such matters is in *The Space Environment and Its Effects on Space Systems* by Vincent L. Pisacane. (It is even an AIAA education series book.) Stereotypically ϵ stands for eccentricity. We see energy being denoted by ϵ because 'E' is used for irradiance, with eccentricity changing to 'e'. Then, in the next chapter, 'E' goes back to being energy, with the different types having subscripts. (i.e. E_k for kinetic energy)

4.4 Notations for Vectors

The concept of vectors has great importance in today's mathematics, science and engineering. Since it appears in a variety of fields, the notation and even the concept itself also vary with the applications. Here we shall mainly consider Euclidean vectors in two and three dimensions, and Lorentzian vectors in four dimensions. More specifically, when we say Euclidean space, we mean the vector space $\mathbb{R}^n$ for $n \in \mathbb{N}$, or the affine space where the vector space acts as translations.

To represent a vector, the common notation used in type is $\mathbf{v}$, a lower case Latin letter in boldface. In some cases the boldface is omitted, or a right arrow above the letter is used, as in $\vec{v}$. In handwriting, because boldfaces are difficult to produce, writing an arrow above or simply writing a lower case letter is preferred. Sometimes in elementary geometry, vectors are represented in a similar way as line segments are written, where one uses capital Latin letters that represent the starting point and end point and write a right arrow above the two letters. For example, $\overrightarrow{PQ}$ represents a vector whose starting point is P and end point is Q.

When specifying a vector by its components, the most common practice is to choose the canonical basis in $\mathbb{R}^n$, and either write each component on a new line (column vectors), or writing them as a horizontal sequence, sometimes separated by commas (row vectors). For example:

$$\mathbf{v} = \begin{pmatrix} 3 \\ 2 \\ 0 \end{pmatrix} , \quad \mathbf{v} = \begin{pmatrix} 3 & 2 & 0 \end{pmatrix} , \quad \text{or} \quad \mathbf{v} = (3, 2, 0) . \tag{1}$$

Sometimes the column vectors are written with a pair of square brackets, and row vectors are also sometimes seen with chevrons.

Through history, there has been great dispute over what notations should be used to denote vectors. Furthermore, the meanings of these symbols have also been contested and altered. Some of these notations do not correspond to the concept that we defined above, and some aspects of the modern notion of vectors did not emerge until quite recently. Thus, we shall try to deliver a chronological account of the history of notations for vectors, and draw a few principles behind these changes.

Around when Issac Newton published his *Principia Mathematica*, the idea of describing a physical force by the magnitude and direction had become popular. The parallelogram law of addition of forces had also been discovered. However, at that time, the abstraction of vectors as objects that can be added and multiplied was far from developed. Vectors were treated as synonymous with directed line segments; therefore, the notations were also the same, which is often two capital letters that stand for the beginning and end points of the vector.

Later, roughly in the first half of 19th century, a more algebraic notion of vectors emerged from the geometrical interpretation of complex numbers. Many authors have worked on this topic, including Caspar Wessel, Carl Friedrich Gauss, and Jean Robert Argand.[11] During this period, there were two trends in the notations for vectors. One was the geometrical tradition that we described above. The other was an algebraic tradition, where a vector is treated as a number which can be added and multiplied. In the latter case, one would simply use a single lower case letter to denote the object. When components are explicitly written, various authors have found symbols for the imaginary unit. Wessel used ϵ. Euler first used i. Argand wrote $\sim$, which stood for a rotation by 90°. Even later, Cauchy used the notation a_α, where a is the modulus of the complex number, and α is the phase angle.

Many attempts have been made to generalize complex numbers to a "three dimensional number system". All such effort was met with failure until Hamilton developed a four dimensional algebra, called quaternions.[15] Hamilton used the notation $q = a + bi + cj + dk$. Here i, j, k are units of quaternions, satisfying $i^2 = j^2 = k^2 = ijk = -1$, and a, b, c, d are all real numbers. a was called the scalar part of the quaternion, and $bi+cj+dk$ was called the vector part of the quaternion. The ingenuity of using quaternions to represent vectors is that, if we have two quaternions whose scalar parts are 0, the scalar part of their product is the negative of the scalar product of the vectors, and the vector part is equivalent to the modern vector product or cross product. That is, suppose $q_1 = x_1 i + y_1 j + z_i k$, and $q_2 = x_2 i + y_2 j + z_2 k$. Then:

$$\begin{aligned} q_1 q_2 = &-(x_1 x_2 + y_1 y_2 + z_1 z_2) \\ &+ (y_1 z_2 - z_1 y_2)i + (z_1 x_2 - x_1 z_2)j + (x_1 y_2 - y_1 x_2)k\ . \end{aligned} \tag{2}$$

In fact, the modern notation of writing the three canonical Euclidean basis as i, j, and k stems from the quaternion units. Before Hamilton, Hermann Grassmann also attempted the problem.[25][15] He developed much of the later vector algebra, but in a very different formulation. Grassmann represented a point that we would use (x, y, z) to represent today as $p = xe_1 + ye_2 + ze_3$. Grassmann's notation resembles the alternative modern notation for basis vectors in $\mathbb{R}^3$. However this resemblance is somewhat deceptive. In Grassmann's program, the basis elements e_1, e_2, etc. were introduced with more general purpose in mind.

In the early systems of vector algebra, various products of vectors, as well as the corresponding notations, were also developed. For Hamilton, there is only one product between two vectors, which is simply the product of the quaternions. Thus, they were denoted in the same way as multiplication in algebra of real numbers; that is, if q_1 and q_2 are two quaternions, their product is q_1q_2. If there is a need to extract the scalar or vector part of the product, Hamilton would write $S_{q_1q_2}$ for the scalar part of the product, and $V_{q_1q_2}$ for the vector part. In Grassmann's case, an "internal product", which corresponds to the scalar product, and an "external product", which corresponds to the vector product, were developed for vectors. For the scalar product, Grassmann used the notation $a \times b$ at first, and then changed to $[u|v]$, while his vector product was $[uv]$.

After Hamilton and Grassmann's invention of vectors became known, many prominent scientists made use of the concept as well as the notations.[15] Most notably, Peter Tait used quaternions enthusiastically in his studies with application to physics. He often separated the scalar and vector part of the quaternion product, making his notations very similar to the modern convention. The gradient operator $\nabla = i\frac{d}{dx} + j\frac{d}{dy} + k\frac{d}{dz}$, which was initially introduced by Hamilton, was also discussed extensively in Tait's work. Just as other vectors, this operator was also treated as a quaternion. For example, $V\nabla a$ represented the same object as today's $\nabla \times \mathbf{a}$. On the other hand, despite praising the invention of quanternions, James Maxwell did not use the methods or notations in his works very much. As described in the previous section, Maxwell's treatise of electromagnitism was done mostly in Cartesian form, working separately with each component.

In 1880s, the modern system of vector analysis was created by Josiah Willard Gibbs and Oliver Heaviside independently.[15] Gibbs and Heaviside's vector analysis was a simplification of Hamilton and Grassmann's system, where the scalar and vector parts of the product were completely separated. Gibbs denoted the scalar product as $\alpha.\beta$, from which the modern notation is derived, and the vector product as $\alpha \times \beta$, the way we do today. These changes in notation are good for two reasons. The first is that giving up the quaternion and simply using a three component object to denote Euclidean vectors is more straightforward conceptually, since Euclidean vectors used in physics are most often 3-d objects. The second is that the dot and cross notation explicitly established these two products as two different binary operations. Because the scalar product is commutative but the vector product is anticommutative, denoting them both with the anticommutative quanternion product could cause confusion.

At the turn of the 20th century, many mathematicians and scientists used vector analysis. Because it was a relatively new field, there was little consensus on what notations should be used. Various organizations and individuals have called for communications and meetings to unify the notations. Ironically, in this process, there were even more notations suggested, such as using $a \times b$ for scalar products and $a \wedge b$ for vector products. After 1910, the modern system was finally adopted by most mathematicians and physicists, perhaps because it was very close to Gibbs' original design. However, even up to today, mathematicians have different conventions of denoting the gradient, curl, and divergence operator. Some use Gibbs' system, $\nabla, \nabla\times, \nabla\cdot$, while some use $\text{grad}, \text{curl}, \text{div}$.[9]

From the history described above, a few interesting conclusions can be drawn. In the early days, when vectors were treated as largely geometric objects, the notations were drawn completely from those used in elementary geometry. Moving into late 19th century, when the idea of vector algebra had emerged as the calculus of quaternions, mathematicians began using algebraic symbols. This reflects that vectors were then considered not only geometrical entities that usually cannot be added or multiplied, but also algebraic objects that can undergo more productive operations. However, this "over-algebraization" which led to the quaternion system also obscured the nature of Euclidean vectors, producing confusing notations that were inconsistent. Because a three dimensional vector can only be represented by a quaternion whose scalar component is zero, but the product produces a full quaternion and cannot represent a vector by itself, the product is not closed. This means that the quaternion calculus is not an algebra in the sense that real or complex numbers are. Mending this problem, the modern vector analysis was invented, and in a sense returned to a more geometrical notion of vectors. Therefore we see the usage of an arrow to mark a vector grow in later times.

4.5 Maxwell's Equations

In the 19th century, mathematics started to look more like what we see today. Many concepts had been formalized, and new notations had been invented to represent them. Maxwell's original representation of the famous equations named after him were written in components of the fields[26], which is extremely cumbersome in the eyes of today's mathematicians and physicists. Hamilton's studies in quaternions, Grassman's discoveries in algebra, and most importantly, Gibb's invention of vector notations as seen now, have all contributed to a more condensed form of Maxwell's equations[15].

Today, the most widely accepted form of Maxwell's equations contain four equations:

$$\nabla \cdot \mathbf{B} = 0 \tag{3}$$

$$\nabla \times \mathbf{E} + \frac{\partial \mathbf{B}}{\partial t} = 0 \tag{4}$$

$$\nabla \cdot \mathbf{E} = \frac{\rho}{\epsilon_0} \tag{5}$$

$$\nabla \times \mathbf{B} - \frac{1}{c^2}\frac{\partial \mathbf{E}}{\partial t} = \mu_0 \mathbf{J} \tag{6}$$

which is considerably fewer than Maxwell's original 20. In most textbooks, when these equations are presented together, they are written in the above form. In these equations, the symbol ∇ denotes the gradient operator $\nabla = i\frac{d}{dx} + j\frac{d}{dy} + k\frac{d}{dz}$, with $\nabla\cdot$ standing for the divergence, and $\nabla\times$ standing for the curl. Of the three vector quantities, **B** is the magnetic field, **E** is the electric field, and **J** is the current density. The scalar function, ρ is the charge density.

An obvious reason for favouring the vector notation is its conciseness. Another more subtle but equally, if not more, important reason is that the main subjects of these literature are the electric field and magnetic field, which in many contexts are considered as vector fields. Thus, it only makes sense if the dynamics of the fields are written in terms of vector quantities. On the other hand, we still often see special cases of Maxwell's equations written in the component form, for example, time harmonic fields in a wave guide, or when the system has some other symmetry. In those cases, the main subject of the equations, which depends heavily on the context and is no longer general, becomes the specific components.

In the 20th century, many great discoveries in mathematics and physics took place; one of which was Einstein's theory of relativity. An especially noteworthy by-product is the Einstein summation convention, which makes it possible to write Maxwell's equation as:

$$\partial_{[\alpha} F_{\beta\gamma]} = 0 \tag{7}$$

$$\partial_\alpha F^{\beta\alpha} = \mu_0 J^\beta \ , \tag{8}$$

or in terms of the potentials:

$$F_{\alpha\beta} = \partial_{[\alpha} A_{\beta]} \tag{9}$$

$$\partial_\alpha \partial^{[\beta} A^{\alpha]} = \mu_0 J^\beta \ . \tag{10}$$

The summation convention sums over all repeated indices, ranging from 0 to n, where n is the spatial dimensionality, often 3. $F_{\alpha\beta}$ is the electromagnetic tensor, incorporating the information of the electric and magnetic field in an antisymmetric second rank tensor. A_α is the vector potential, whose 0-th component corresponds to the electric potential, and the 1st to 3rd component correspond to the magenetic vector potential, which is often denoted as $\mathbf{A}$. J_α is the electromagnetic current, incorporating both the charge density and current density in a 4-vector.

A direct consequence of this type of new convention that suppresses summations signs is that proofs in tensor algebra are made considerably easier, and by explicitly differentiating between the upper and lower indices, the distinction between vectors and covectors is brought to light. For example, one can see from the notation that the vector fields J^β and A^α and the differential operator ∂_α have different transformation properties. First of all, this notation further reduces the number of equations from four to two, which again makes the expression more compact. Second, in this notation, the electric field and magenetic field are united in the same object, the electromagnetic field tensor. When considering (special) relativity, writing the electromagnetic field in a Lorentz invariant form is desired for consistency of the theory. This notation makes clear the distintion between vectors and covectors, which is very important in relativity, which developed from the study of electrodynamics of moving bodies descibed by Maxwell's equations. Third, this notation makes proofs that involve complicated vector operations considerably easier. Nevertheless, the vector notation is still the most popular in science and engineering, because the context does not usually call for such a compact notation when computation concerning only the electric field or magnetic field or their components are done. Furthermore, if the problem includes complicated boundary conditions, it becomes particularly difficult to write them down in the index or tensor notation, which makes the computation cumbersome and obscures the geometry of the system. A more simple reason why this notation is not used as often might be that it requires more advanced background knowledge to understand than the vector notation, which many people might not know of or find irrelevant to them.

Further developments in differential geometry provide another form of Maxwell's equations:

$$\mathrm{d}F = 0\,, \quad \text{and} \quad \mathrm{d}\star F = \mu_0 J\,, \tag{11}$$

or in terms of the potentials:

$$F = \mathrm{d}A\,, \quad \text{and} \quad \mathrm{d} \star \mathrm{d}A = \mu_0 J\,. \tag{12}$$

Here d is the exterior derivative that sends k-forms to $(k+1)$-forms. F is the electromagnetic field 2-form, A is the electromagnetic potential 1-form, and J is the current 1-form. Here F, A, and J here contain the same information as $F_{\alpha\beta}$, A_α, and J_α above.

This expression looks very similar to the tensor-index notation; however, there are some fundamental differences. First, the equations are written in a coordinate-free form, which makes them true in any space-time. This promotes the generality of the equations so that they are true even when gravitation is present, as in the general theory of relativity. Second, the equations in this form make it easy to deduce topological and geometrical properties of the field. The first equation means that the electromagnetic field tensor is a closed form, which can be used to proof the properties of line integrals or surface integrals of electromagnetic fields in regions with nontrivial topology. The second (nonhomogeneous) equation also incorporates the continuity equation for electromagnetic current, $\mathrm{d}J = 0$. Third, this equation can now be easily generalized to Yang-Mills fields, which are similar to the electromagnetic field but more general. Such a generalization is less transparent and elegant when the equations are written in terms of components, and is almost impossible to derive if one were to write them in the vector form. However, the power of differential geometry does not always simplify computation when one needs to know the solution to some boundary value problem down to each component. The same difficulty with the tensor notation is seen here. Because this notation requires an understanding of differential geometry to understand and use easily, the average student would not use this notation even though it saves space. The benefit and disadvantage of this notation again depends heavily on the context.

Having analyzed the three forms of the Maxwell's equations, we can already conclude some principles of choosing notations. First, the notation needs to be concise. Second, the notation needs to emphasize the objects that are most relevant to the discussion. Third, the notation needs to make computations easy. Further, the notation should be easy to generalize. These principles could be conflicting in practice, where the context decides which is more important.

Beyond the considerations that we have mentioned, there are other variations to these equations that we neglected. For example, different authors may choose to use a long left curly brace to indicate that (3), (4), (5), and

(6) constitute a set of closely related equations. Since this is purely a matter of style, we will not study these choices in notations.

The above discussion is far from exhaustive. For example, the issue of consistency did not come up, since we are only looking at one set of equations. Through the following discussions, we will try to construct a more complete theory by examining more different subjects.

4.6 Naming of Space Groups

Groups are important algebraic objects that emerge from the studies of symmetry. They are often constructed as transformations (in real space or some abstract space) that leave the symmetric object in consideration unchanged or invariant in some sense. In describing and classifying symmetric objects as in crystallography, molecular chemistry or particle physics, one needs to describe and classify the groups of symmetry. A result of such projects is the systems of notations for these groups. Here we are concerned with space groups that encode all possible discrete symmetry in three dimensions.

The space groups are constructed by combining the 32 crystallographic point groups with the 14 Bravais lattices. There are different methods of classifying both of these two classes of groups, thus also leading to different notations.

Crystallographic point groups contain symmetry operations that leaves a central point fixed. Due to the crystallographic restriction theorem[17], only 32 crystallographic point groups are compatible with discrete translation symmetry. There are two major conventions for denoting these groups. The first is the Schoenflies notation, in which point groups are classified into several families, and the symbols for them are derived from the first letter of the word that describes the family, and subscripted by the order and additional properties of the rotation axes. For example, C_n, with "C" for "cyclic" stands for n-fold rotation axes, C_{nh} has an additional horizontal plane of reflection, with "h" standing for "horizontal". S_{2n}, where "S" stands for "Spiegel", which is German for mirror, contains a $2n$-fold rotation-reflection axis. D_n stands for dihedral groups with an n-fold rotation axis with n twofold reflection axis. For groups with several higher order rotation axes, there are only three that are compatible with translation groups, they are represented by T for tetrahedral, O for octahedral, and I for icosahedral.

The second type of notation is the Hermann-Mauguin notation, or the international symbol. This notation enumerates all the unique rotation or rotoinversion axes along with the reflection planes. Uniqueness here means that one should not include axes or planes that can be generated by rotations about an axis already included or reflection about such a plane. A rotation axis is denoted by the order of the axis, and a rotoinversion axis has an extra macron. For example, a two-fold rotation axis would be 2, and a three-fold rotoinversion axis would be denoted by $\overline{3}$. A reflection plane that is not perpendicular to any rotation axis is denoted by the letter m. A reflection plane that is perpendicular to one of the axes is denoted as a fraction. For example if we have a three-fold rotation axis and a reflection plane perpendicular to it, then the notation is $\frac{3}{m}$. One thing to note is that for all odd n, the notation $\frac{n}{m}$ and $\overline{2n}$ express the same symmetry, in which case we always prefer the latter.

The Bravais lattices, which are all the possible lattices in three dimensions, can be classified in two ways. The first is by the point symmetry that each of them are compatible with, which leads to the seven lattice systems that chemists are familiar with, triclinic, monoclinic, cubic, etc. On the other hand, one can classify them purely by translation symmetry. See 1 for the seven types of translation symmetry.

Table 1: Classification of 3D lattices by translation symmetry.[1]

Letter	Position of extra lattice points
P	Primitive (no extra lattice points)
I	Body centered
F	Face centered
A	centered on A faces only
B	centered on B faces only
C	centered on C faces only
R	Rhombohedral

When the point symmetry and translation symmetry are combined, there are two other types of symmetry emerging. The first is screw symmetry, which is a rotation around an axis followed by a translation along the same axis. This is denoted by the degree of the rotation subscripted by the proportion of the lattice vector that it shifts by. For example, 3_2 is a 120° rotation followed by a translation of two-thirds of the lattice vector along this axis. The second extra symmetry is glide symmetry, which is a reflection about a plane, and a shift along an axis. A glide plan only exists when a reflection plane already exists, so it is denoted by replacing the original m by the letter a, b, or c, depending on which axis it shifts by. Additionally there is the n glide, which is a glide along half of the diagonal of a face, and the d glide, which is along a quarter of a face or the body diagonal. Thus $2/b$ would mean a reflection by the plane perpendicular to a two-fold axis followed by a glide along the b axis.

Finally, to denote the full space group of a system, the Hermann-Mauguin notation adds the letter denoting the lattice type in front of the symbols denoting the point symmetry. For example, a notation like $P222_1$ means the object has a primitive translational symmetry. It has three two-fold rotation axes, and there is also a screw symmetry along one of the axes. An absence of any higher order axes tells us that this is an orthorhombic crystal system. On the other hand, to represent space groups in Schoenflies notation has much less logic to it. The procedure is simply first listing all the space groups that correspond to the same point group, give them an arbitrary enumeration, and write that number as a superscript. For example, since the above space group has point group D_2, we shall list all the 9 space groups, which in Hermann-Mauguin notation are $P222$, $P222_1$, $P2_12_12$, $P2_12_12_1$, $C222_1$, $C222$, $F222$, $I222$, and $I2_12_12_1$, which makes the given group D_2^2.

Here we have described the process of writing down the name of a space group in detail, hoping to show that there is an underlying general procedure that can also be seen in other areas and applications of mathematics. What we have seen is that notations for a relatively large number of related but distinct objects can be generated by a relatively few number of rules. The rules for Schoenflies notation can be summarized as

1. Use the uppercase of the first letter of the name of the symmetry class as the main symbol;

2. Use the order of rotation symmetry and the additional reflection planes to subscript the main symbol;

3. Use the number of the given space group in the list of space groups corresponding to the same point group to superscript the main symbol;

and supplemented by a list of the symmetry classes, and all the space groups corresponding to all of the point groups. The rules for Hermann-Mauguin notation can be summarized as

1. The symbol is composed of at most three symbols for symmetry in three directions;

2. If a direction has an n-fold rotation symmetry, the symbol for that direction is n;

3. If a direction has an n-fold improper rotation symmetry, the symbol for that direction is $\overline{n}$;

4. If a direction has a plane of reflection perpendicular to it, the symbol is m;

5. If a direction has both an n-fold rotation and a reflection, the symbols is $\frac{n}{m}$;

6. When n is odd, the notation $\overline{2n}$ is favored over $\frac{n}{m}$;

7. A screw symmetry is denoted by subscripting the rotation symbol by the proportion of the lattice vector that it translates by;

8. A glide symmetry is denoted by switching the letter m for a, b, c, d, or n, depending on the translation;

9. The translation symmetry type is denoted by adding the letter representing the symmetry class to the symbols representing point symmetry;

and also supplemented by a list of translation symmetry classes, and a list of possible translations in glide symmetry.

Comparing these two systems of generating notations, we can immediately see that the criteria that we have discussed in chapter 3 do not apply very well here. For example, when a set of rules are given, one can be sure that all notations generated by them are consistent as long as the rules do not confuse or overcount objects. Conciseness also becomes a somewhat awkward notion, since one cannot adjust each symbol individually to make them short and hope to keep consistency with the rules. Thus we need new criteria that target not towards the individual symbols but towards such a system of generating notations.

5 Developing the Theory

In this chapter, we shall develop a new theory of evaluating the usability of notations base on the rules that generates the system of symbols, not the symbols themselves.

5.1 Criteria for Good Notation

To review our findings so far, we have drawn out a few criteria that distinguish good notations from bad notations, using as sources both the writings of previous authors who directly discussed the issue and the history of notations for certain subjects. Many of these principles are intuitive. For example, "conciseness", "unabiguity", and "consistent". However, we need to clarify the meaning of these words in context, because there could be many interpretations.

5.1.1 Conciseness

The first principle that we want to examine is conciseness. This is perhaps the most fundamental driving force for inventing new notations. Early mathematical notations where often either symbols created to denote quantities that would otherwise time consuming or difficult to represent, or contracted or abbreviated words that developed into symbols. Later inventions, such as the summation convention for tensors, were also fueled by similar intentions.

However in inventing or switching to a notation, the ultimate reason why such notations are used might be to pursue conciseness, even though it might seem so at first glance. There are two possible purposes that could be discerned from such an action. First, one may be genuinely looking for a shorter notation than what is readily available. This should be the case when the plus sign (originally the abbreviation for Latin *et*), the differential sign, and the existential quantifier $\exists$, were invented. Another possibility is that the mathematician finds the need to create a symbol to represent a new type of object. This happens when a new field of studies emerges and the mathematician sees the need to denote the new entities that might be composed of familiar objects. For example, even though when authors initially started using single letters to denote vectors, it was because denoting them by components takes up much more space, when linear algebra was invented, vectors are no longer just a list of numbers or points in $\mathbb{R}^n$, but more abstract objects, thus they are no longer denoted by their components in a certain basis. The reason for such choice is not pursuing conciseness, but emphasizing the subject of study. We will talk about that principle later.

5.1.2 Unambiguity and Consistency

The principle of unambiguity and consistency addresses two related problems that appear in mathematical writing. The first is using the same notation for multiple purposes, and the second is using different notation for the same concept. Here the word notation is a more general construct than individual symbols. It refers to the system of choosing the attributes of symbols so that each attribute represents a type of object. Such attributes may be the alphabet that a letter belongs to, or typographical variations of the characters, or indices that are attached to the symbol. In a piece of writing, one does not use a single symbol on its own, but construct each symbol by some preset rules. We argue that these rules are often the “real” notation system that this principle speaks to, rather than individual symbols that are generated from them.

Unambiguity means that in a notation system, one should only use a symbol or rule to represent one object or one class of objects. For example, in an article, the letter p should not be used to denote pressure and magnitude of linear momentum at the same time. As another example, when typographical distinction is not used, the letter d should not be used to denote a variable whose derivative is taken. A more subtle example is that when lower case Greek letters starting from α are chosen to be indices of a tensor, as in $F^{\alpha\beta}$, they should not be used also for coefficients that are multiplied to tensors, for example as in $\gamma F^{\alpha\beta} = G^\alpha H^\beta$, even though there is no single symbol that carries multiple meanings at the same time. In this context, if in an expression more than two indices are to be used, the next natural choice (here the “naturalness” is also governed by the implicit rules of generating symbols) is to use γ, but it is already taken by a coefficient, so one has to face the dilemma of choosing between using the same symbol for the different objects that do not appear in the same position, and violating the implicit rules of using Greek letters in alphabetical order.

Consistency means that an object should be only represented by one symbol or that a concept should correspond to only one rule of generating symbols. A simple example would be to not use both lower case and upper case letters to denote functions with the same domain and codomain, unless one needs to emphasize difference between two classes of functions. A more subtle example is to not use both α, β, γ and μ, ν, ρ as indices of tensors. This may confuse reader into thinking that these two sequences of letters represents different types of indices, which may have different ranges of values.

Unambiguity and consistency are related and are often violated at same time, if one fails to use good discretion. For example, one may choose c to stand for a constant, but only later to realize that c is also needed for speed of light in vacuum, so he decides to redefine c as speed of light, because the constant c is not used in the rest of the derivation, and when another constant is needed, he resorted to the capital C. This is a bad usage because even though the letter c was never used in the same expression for both its meanings, within the same article, making such transitions confuses a less careful reader, so it fails in unambiguity. Furthermore, there is also no consistency in the switch, because the rule of using c as the default constant was given up and a new rule was made.

5.1.3 Emphasizes the Subject of Study

What this criteria entails depends completely on the context of the treatise. Again we take Maxwell's equation as an example. The benefit of using the vector notation over the component notation is not only that it saves a lot of space, but also that it represents the true subject of study with a single symbol. If the component notation were used, the expression would definitely be much longer, and it would be less clear that the subject of study are the vector fields $\mathbf{E}$ and $\mathbf{B}$. There is also a more subtle implication of this principle. In a system of rules that generates notations, it is preferred to have the class of objects that is the subject of the study take the most simple, or "unmarked" form. For example, in multilinear algebra, if an article attempts to discuss vectors and tensors extensively, then one may choose to use regular lower case latin letters to denote vectors, instead of using boldface for every vector. Thus when a scalar is used, one may either distinguish them from vectors by choosing the letters from different parts of the alphabet, or use Greek letters.

5.1.4 Easy to Produce Typographically

Mathematical writing is sometimes difficult to produce in typography, because they may contain uncommon symbols or have many superscripts or subscripts. Before digital typography was invented, this was a contributing factor to the choice of some notations. In some older works, one can also find that symbols like the horizontal line in a radical symbol are not used as often, because they are more difficult to typeset. After digital typography was invented and became popular, this still persists to be an issue, because the encoding and fonts strongly restrict one's options. It is much more difficult to invent a new symbol today than two hundred years ago, because for a character to be represented correctly through a digital typesetting system, it must be included in the system encoding, and its glyph must be stored in the typeface. Even if one succeeds in inserting an invented symbol to the system, it is still very difficult to distribute it. All this makes fanciful inventions of symbols impossible. Moreover, one can no longer invent variations of letters as easily. In the popular word-processing softwares and typesetting systems such as Microsoft Word, Adobe InDesign, or LaTeX, there is only a fixed number of styles of letters that once can use. If one wishes to employ more than the roman, italic, calligraphic, blackletter, and blackboard variations of the alphabet, one has to devote a great amount of time into modifying the system.

5.1.5 Compatible with Existing Conventions

One difficulty of inventing new notation is to make sure it is compatible with existing conventions. One may attempt to invent a new, consistent system of mathematical notations for all of the active fields of mathematics, but no one has succeeded so far. There have been attempts in history, but none of the systems have survived as a whole. One reason why this would not work today is that there are simply too many sub-fields in mathematics, so an attempt to establish a general convention that does not have any ambiguity or inconsistency is extremely difficult, if not impossible, and if other sciences and engineering notations are included, matters would only be worse. Another reason is that the math and science community is quite resilient to changes. Sometimes supporters of rival notations both have a good argument, while sometimes it is simply a matter of habit. For example, generally the meaning of the Greek letters θ and ϕ in spherical coordinates is opposite among physicist and mathematicians. Physicists would use ϕ for the azimuthal angle and θ for the colatitude, and mathematicians the opposite. The former is supported by an ISO standard, while the latter corresponds to the notation for polar coordinates in a plane.

A good notation should consider existing conventions, so that a reader who is accustomed to reading work written in those conventions could understand it without difficulty. Thus it is not wise, for example, to use e for some variable and write the exponential function as exp, since it may cause confusion.

5.1.6 Difficulties

Having listed what criteria there are for good notations, the discussed seemed complete. However, there are many cases when the criteria for a good notation conflicts. For example, we have seen that even though conciseness is an important principle to follow, the shortest notation for Maxwell's equations is not the most popular one. Thus we need to develop a theory that assigns priorities to the criteria, so that in case of a conflict, one could make unconfused judgment.

More importantly, we need to look at notation not as a collection of individual symbols, but symbols that are organized in a system through a few rules. This means that we cannot apply the criteria to each symbol separately, but look at the underlying rules that generate them. This problem has already manifested as soon as we started considering consistency as a principle. One cannot talk of consistency when there is only one symbol in consideration. Instead, only a system can be determined to be consistent or inconstistent. For example, in our study of Hermann-Mauguin notation for point groups, one can always claim that since primitive lattices have not extra lattice points in the unit cell, and is the most simple translation symmetry, one could omitt the letter P for primitive lattices, so that each symbol would be shorter. However, this gives the point groups and space groups the same name when there are no screw or glide symmetry. Thus even though one could argue that 6_1 is a shorter notation than $P6_1$, and this does not make the notation ambiguous or superfluous, it is still bad, because we either have to make additional rules to differentiate the point group 6 and the space group $P6$, or suffer ambiguity.

5.2 An Analogue of a Linguistic Theory for Mathematical Notations

After seeing that the principles conflict, we must decide which principle prevails when all of the options violate the criteria to some degree. Here we propose a solution that draws from the optimality theory in linguistics.[24]

The motivation for this analogy is that language is considered by a school of scholars to be generated from a set of universal rules. The forms of expressions vary by the meaning one wants to express, and the number of possible articulations is infite. However, generative linguists believe that there is only a finite number of rules that generates all possible expressions. We find this mode of thought applicable to our subject of study. Here, we have some underlying concepts that need to be articulated in terms of mathematical symbols, and there is a set of rules that generates possible forms. A key question linguistics tries to answer is why are some forms "correct" or "natural", while other forms "incorrect" or "awkard", which is also the question we seek to answer for mathematical notation.

The linguistic model that we will emulate here is optimality theory. It is a linguistic model mainly applied in phonology, the branch of linguistics concerned with pronunciations, that attempts to explain by interactions of constraints why language takes its observed phonological form. There are three basic components of the theory. First, there is a set of generating rules that maps the underlying structure to multiple possible outputs. Second, there are many constraints, which may be violated by one or more of the candidates. Finally, there is a definition of optimality, by giving a method of evaluating the number of violations by each of the candidates and an ordering among the constraints. This model is applied to phonology and syntax, and has achieved significant success.

Similarly, in our theory of mathematical notation, we shall look at three components. First, a set of generating rules that maps the underlying concepts to multiple possible forms of notation. Second, a collection of constraints or criteria for good notation, drawn from previous discussions, which may be violated. Finally, the definition of optimality that evaluates the possible forms of expression by how badly they violate the criteria, also ordering the criteria by priorty.

Here we will first give an example taken from Wikipedia[41] to demonstrate how optimality theory operates, and then move on to developing an analogous theory for mathematical notations. The following procedure generates the observed form of the plural of the English word "cat".

Table 2: Analysis of the "cat+z > cats" by optimality theory

cat + z	*SS	Agree	Max	Dep	Ident
catiz				*	
catis				*	*
catz		*			
cat			*		
cats					*

Here the input is the underlying form of the plural of the word "cat": the word itself and the plural suffix "-z" concatenated. The possible variations of the plural suffix that we observe are listed in the left column. The examples of each form are (written phonetically, not necessarily orthographically correct): "-iz"/"dishes", "-is"/"bushes", "-z"/"dogs", "fish"/"fish", "-s"/"bats". The constraints or criteria are listed in the first row. "*SS" prevents sibilant-sibilant clusters, i.e., combinations such as "*dishs" without the intermediate vowel violate this constraint. "Agree" means adjacent obstruents must agree in voicing, so combinations such as "*ratz" with the unvoiced "t" and voiced "z" violate this constraint. "Max" means to maximize all input segments in the output, which prevents deletion, so outputs such as "fish + z>fish" in fact violates this constraint. "Dep" means that output segments should depend on the input segments, so outputs such as "dishes" where an intermediate vowel is added violates this constraint. Finally "Ident" means that the voicing of the output should stay identical to the input, which means that the voiced "-z" is favoured to the unvoiced "-s". Each violation of a constraint is marked by the five criteria listed here are in the order of dominance. The first two are "markedness" constraints, which are constraints that are imposed on output forms to rule out "unnatural" forms. The other three are "faithfullness" constraints, which make sure that the output form does not diverge far from the underlying form. In optimality theory, markedness constraints dominate faithfulness constraints.

5.2.1 Generating Rules and Constraints for Mathematical Notations

The first step to develop an optimality theory for mathematical notations is to formulate the rules that generate mathematical notations. A generating rule in such a system can either assign a class of symbol to a class of objects, or derive symbols from already assigned symbols when one needs to describe a related object. For example, in an article where one discusses vectors and tensors, a rule that generates the notations for components of a contravariant second rank tensor would be: the "body" of the symbols should be the same letter as the letter representing the tensor itself, that is, upper case Latin letters going in the sequence $T, S, F, G, \ldots$, and two lower case superscripts should be added to the letter, in the order $i, j, k, \ldots$.

We shall adopt for now the criteria in section 5.1 as constraints, and use the following example to demonstrate how optimality theory might work for mathematical notations. This example is described in words first, then we shall see a few examples of possible realizations in symbols.

Consider a proof that involves the following objects. First a function from a Euclidean space to a Euclidean space of the same dimension. A point in the domain of the function, and a neighbourhood containing the point. In this proof we need to invoke the fact that the function has nonsingular

Jacobian at this given point. We then consider the image of this point under this function. There needs to be two open balls, one in the domain of the function, centering at the point given above, the other in the codomain of the function, centering at the image of the given point. Then we take an arbitrary point in the second open ball. We want to construct a sequence of points in the first open ball, and argue that the sequence converges, and their image converges to the arbitrary point that we took in the second open ball.

To describe this again in symbols, we could say the following. Let $f : \mathbb{R}^n \mapsto \mathbb{R}^n$. Given $x_0 \in U \subset \mathbb{R}^n$, the Jacobian of the function $Df(x_0)$ is nonsingular. Consider the open balls $B(x_0, \epsilon)$, $B(f(x_0), \delta)$, and a point $y \in B(f(x_0), \delta)$. We then construct a Cauchy sequence $\{x_i\}$, and let its limit be $x = \lim_{i\to\infty} x_i$. Finally we need to prove that the sequence $\{f(x_i)\}$ converges to y.

Now we have a system of notations that presents many subtleties in choices of specific symbols. First of all, the given point x_0 can alternatively be represented as x, since it is a special point that we chose in the domain of the function. However, if this choice is made, we cannot use x to represent the limit of the sequence $\{x_i\}$. The notation for the Jacobian of the function also has other alternatives, such as J_f or simply J. We may also choose to use $(x_i)_{i=0}^{\infty}$ or (x_i) to denote the sequence.

Since there are many possible variations to this system of notations, we have tabulated a few and labeled them for further discussion.

The criteria that we use to evaluate the possible notations need to be modified from the ones given in 5.1, or at least reformulated more formally.

Analogous to the procedure in optimality theory, we will count the num-

Table 3: An example for possible system of notations

Objects and Concepts	Version 1	Version 2	Version 3	Version 4
function	f	f	f	f
Open subset of domain	U	A	U	Ω
Given point in domain	x_0	x	x	p
Jacobian matrix at point	$Df(x_0)$	$J\|_x$	$J_f\|_x$	$Df\|p$
First open ball	$B(x_0, \epsilon)$	$B_1 = B(x, \epsilon)$	$B(x, \epsilon)$	$B(p, \epsilon)$
Second open ball	$B(f(x_0), \delta)$	$B_2 = B(f(x), \delta)$	$B(f(x), \epsilon)$	$B(f(p), \delta)$
Point in codomain	y	y	y	Q
Sequence	$\{x_i\}$	$(x_i)_{i=1}^{\infty}$	(x_i)	$\{p_i\}$
Limit of sequence	x	x'	z	P

ber of violations of the criteria, and the best output will be the option that has the fewest violations, or in case there is a tie, the option that violates lower priority constraints.

The constraints are the following, ordered from highest priority to lowest:

1. Unambiguity: no two objects have exactly the same symbol.

2. Conventional: for every notation that is unconventional, one violation is counted.

3. Smallest number of rules: the number of rules that generates the system should be minimal, so if a system is not minimal, the difference in the number of rules between it and the minimal system is the number of violations against this constraint.

4. Shortest length: for any object, if a shorter representation is possible, one violation is counted.

The reason why the constraints are ordered in this way, with unambiguity at the top is that when notations are ambiguous, mathematical writing becomes incomprehensible or at best confusing. Using conventional notation is weighed over using the smallest number of rules to generate the symbols because in mathematics, conforming to existing habits is in fact more important than having consistent rules to generate symbols for a piece of writing to be well-received. Finally the shortest length rule was weighed not as important, because it is merely a constraint that rules out superfluous symbols and components of symbols, but not necessarily the major concern of writing.

Taking the example presented here, we can count the violations of principles as follows. First we consider version 1 in the table. The underlying implicit rules that generated this notation is

1. A function should be represented by lower case Latin letters starting from f.

2. An open set should be represented by upper case Latin letters starting from U.

3. The Jacobian or differential of a function f at x is expressed by $Df(x)$.

4. An open ball centered at x with radius ϵ should be represented by $B(x, \epsilon)$.

5. The radii of open balls should be represented by Greek letters in the sequence $\epsilon, \delta \dots$.

6. A sequence of points $x_0, x_1 \dots$ should be represented by $\{x_i\}$, where each term is represented by the same letter with indices, and the limit should be the same Latin letter.

7. A point represented by x with indices and other variations should have its image represented by y with the same indices and other variations.

This list of generating rules does not necessarily appear in the order of writing the work.

In comparison, we can work out the generating rules for the version 2.

1. A function should be represented by lower case Latin letters starting from f.

2. An open set should be represented by upper case Latin letters starting from A.

3. The Jacobian or differential of a function f at x is expressed by $J|_x$.

4. An Open ball centered at x with radius ϵ are be represented by $B(x, \epsilon)$.

5. The radii of open balls should be represented by Greek letters in the sequence $\delta, \epsilon \dots$.

6. The open balls required in the proof are given symbols B_1 and B_2.

7. The arbitrary point chosen in the codomain shall be represented by y.

8. A sequence of points $x_1, x_2 \dots$ should be represented by $(x_i)_{i=1}^{\infty}$, where each term is represented by the same letter with indices, and the limit should be the same Latin letter with a prime.

9. The first point of the sequence should be x.

10. A point represented by x with indices and other variations should have its image represented by y with the same indices and other variations, unless specified otherwise.

Now for both of the versions, we can count the violations of constraints. The first version violates the shortest length constraint twice because the using B_1 and B_2 for the open balls $B(x_0, \epsilon)$ and $B(f(x_0), \delta)$ is a shorter notation. The second version violates the shortest length constraint once, because its notation for sequences requires more symbols. The second system has three more rules than the first system, so it violates the smallest number of rules constraint twice. Additionally, the second version can also be considered to violate the conventional rule once, because usually one would want to have $f(x) = y$ instead of $f(x') = y$, but no symbol for $f(x)$. Thus overall, the first version has two violations, and the second version has four violations. This make the first version a better system of notation.

Now to compare version 3 and version 4, we first list the rules in these systems of notations. For version 3, we have

1. A function should be represented by lower case Latin letters starting from f.

2. An open set should be represented by upper case Latin letters starting from U.

3. The Jacobian or differential of a function f at x is expressed by $J_f|_x$.

4. An open ball centered at x with radius ϵ are be represented by $B(x, \epsilon)$.

5. The radii of open balls should be represented by Greek letters in the sequence $\epsilon, \delta \dots$.

6. The arbitrary point chosen in the codomain shall be denoted by y.

7. A sequence of points $x_1, x_2 \dots$ should be represented by (x_i), where each term is represented by the next unused Latin letter.

8. The first point of the sequence should be x.

9. A point represented by x with indices and other variations should have its image represented by y with the same indices and other variations, unless specifed otherwise.

The rules forming system 4 are

1. A function should be represented by lower case Latin letters starting from f.

2. An open set should be represented by upper case Latin letters starting from Ω.

3. The Jacobian or differential of a function f at x is expressed by $Df|_p$.

4. An open ball centered at p with radius ϵ are be represented by $B(p, \epsilon)$.

5. The radii of open balls should be represented by Greek letters in the sequence $\epsilon, \delta \ldots$.

6. A sequence of points $p_1, p_2 \ldots$ should be represented by $\{p_i\}$, where each term is represented by the same letter with indices, and the limit should be the same Latin letter in upper case.

7. The first point of the sequence should be p.

8. A point represented by p with indices and other variations should have its image represented by q with the same indices and other variations.

Both notation have symbols of the same length, and have the same level of typographic variations, so neither violates the conicseness constraint. Version 3 violates smallest number of rules constraint once. Version 4 violates the convention rule once, because even though using p and q for points in Euclidean space is conventional, using the upper case Latin letters P and Q for points is unusual. Now both versions have the same number of violations of constraints, but the version 4 violated a higher priority constraint than version 3. Therefore version 3 is a better notation.

6 Testing of Hypotheses

To evaluate quantitatively the value of each of our principles of good notation, we propose a series of experiments to determine the usability of various notations. By varying one aspect of a notation at a time, our aim is to quantify the effect of each aspect of a notation on its overall quality. We will measure two indicators of quality: ability to relay information and time required to read. Further, we will distinguish between levels of experience with each notation.

6.1 Selected Notations

There are clearly too many notations in use today to be able to conduct an exhaustive test; therefore, it is necessary to choose a small set of topics that collectively cover a wide range of notational differences. We have selected the following three topics:

6.1.1 Polynomials

Polynomials today can be expressed in standard form (e.g. $x^3 - 3x^2 + 4$) or in factored form (e.g. $(x-2)^2(x+1)$), each of which is more appropriate in different circumstances. Furthermore, there are historically used forms, most notably Diophantus' (see section 2.3.3), that have been used to represent polynomials. We intend to measure when conventional wisdom holds on which notation to use in which situations, and the advantages of modern notations over that of Diophantus.

6.1.2 Derivatives

There are several notations for derivatives, including $f'(x)$, $\frac{\partial f}{\partial x}$, and $D_x f(x)$, each with its own advantages and disadvantages. We intend to measure the effectiveness of each at conveying its intended meaning.

6.1.3 Vector and Tensor Notations

Einstein's notation reduces the tensor index notation, leaving out many operations as implied. This produces a much shorter and more compact notation. However, it may be that the implied operations place more cognative load on both the reader and writer, which would negate its benefits. We intend to measure whether or not this assumption is true.

6.1.4 Logical Connectives

The statements in a document, especially a proof, must be logically connected in some way. These relationships can be expressed in several ways, ranging from purely prose to soley symbolic notation (see the introduction to section 3), with arguments to be made for each side. Prose allows clearer expression of complicated thoughts, but symbolic notations allow these thoughts to be more rigorously defined and proven. An approach in the middle combines some of the advantages of both. We intend to measure the understandability of these different approaches.

6.2 Procedure

For each topic, we will identify a notation-heavy text addressing a subject within that field. We will then translate the instances of notation into each of the forms described above, producing an equivalent text for each notation (an example of these translations, applied to a proof of the chain rule, can be found in Appendix A). Next, a sample population will be selected and grouped by experience with the topic in general, as well as with the specific notations. These groups will then be randomly partitioned for each notation. Each subject will be given the version of the document corresponding to their partition and will be timed as they read it. Finally, a test will be administered to identify how much of the information contained in the document was conveyed.

6.3 Analysis

In performing this experiment, we will measure several statistical variables. The independent variables include the notation that is used and the subject's experience with the topic. The dependent variables are the time taken to process the document and the subject's measured understanding of the content. These are all meaningful: while the measured understanding is the most important measure of the quality of a notation, it cannot be ignored if, for instance, one notation takes twice as long to read as compared to another. Furthermore, it is important to account for the subjects' prior experience with the topic for two primary reasons. First, it presents a confounding factor that must be accounted for: otherwise, an experienced mathematician assigned an inferior notation may make the notation appear useful than it truly is, even though it is his experience, not the notation itself, causing the measured productivity. Therefore, we must normalize the data to account for this phenomenon. Second, the quality of a notation may be affected by the user's experience. It is not difficult to imagine a field where one notation more accurately reflects the underlying structure of the concepts, making it difficult for amateurs who do not yet fully understand the structure to use, while simultaneously being more useful for the professional who can use it to its fullest.

6.4 Goals

The purpose of these experiments is to correlate the differences in notations with their effectiveness at communicating information. This allows us to more rigorously and objectively determine the which characteristics of notation are truly important, and which are red herrings that merely appear to be significant.

7 Conclusion

In this work, we have studied the evolution of mathematical notation, and what constitutes good notation. Our purpose is to formulate a set of principles that helps one determine whether a notation is good by formal standards, rather than empirical or aesthetic judgments. To do this, we first looked at existing works concerning the topic, and summed up a few principles. These principles are shown to be insufficient when we look at certain developments in mathmeatical notation. This showed that a more sophisticated theory than a few universal principles is necessary.

Through the analogy between mathematical notation and language, we proposed a theory that resembles optimality theory in linguistics. Applying this theory to a practical example of mathematical writing, we showed that this theory does indeed make reasonable judgments on the usability of notations, at least on a small scale and at an elementary level. Considering that the results produced by our theory is also confirmed by intuition and heuristics, we believe that it is successful. To systematically test the theory, we also developed an experiment that surveys mathematicians about their intuition and emprical decisions.

For future studies on this topic, we suggest conducting the experiments to collect more data on what is considered good notation in mathematics academia. The results may challenge our theory and drive for more sophisticated and complete model, or even lead to new persepctives on mathematical notation, and eventually reveal new aspects of mathematical writing.

A Proof of Chain Rule in Different Notations

To accompany the experiment described in chapter 6, we present here an example of a set of documents that we will prepare for the experiment. The text is quoted from [16, p. 133]; the only changes are the removal of specific references to other sections of the book and changing the notation used. It is worth noting that the original source uses all three notations interchangably.

A.1 Outline of a Proof of the Chain Rule from [16] Using f' Notation

To *outline* a proof of the chain rule, suppose that we are given differentiable functions $y = f(u)$ and $u = g(x)$ and want to compute the derivative:

$$(f \circ g)'(x) = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(g(x + \Delta x)) - f(g(x))}{\Delta x} \tag{13}$$

The differential form of the chain rule[2] suggests the factorization

$$\frac{\Delta y}{\Delta x} = \frac{\Delta y}{\Delta u} \frac{\Delta u}{\Delta x} \tag{14}$$

where

$$\Delta u = g(x + \Delta x) - g(x) \text{ and } \Delta y = f(u + \Delta u) - f(u).$$

For x fixed, the factorization in Eq. (14) is valid if $g'(x) \neq 0$, because

$$g'(x) = \lim_{\Delta x \to 0} \frac{\Delta u}{\Delta x} \neq 0$$

implies that $\Delta u \neq 0$ if $\Delta x \neq 0$ is sufficiently small—for if so, then $\Delta u = (\Delta u / \Delta x) \cdot \Delta x$ is the product of nonzero numbers. But the fact that g is differentiable, and therefore continuous, at the point x implies that

$$\Delta u = g(x + \Delta x) - g(x) \to 0 \text{ as } \Delta x \to 0.$$

The product law of limits therefore gives

$$\begin{aligned} (f \circ g)'(x) = \lim_{\Delta x \to 0} \left(\frac{\Delta y}{\Delta u} \cdot \frac{\Delta u}{\Delta x} \right) &= \left(\lim_{\Delta u \to 0} \frac{\Delta y}{\Delta u} \right) \cdot \left(\lim_{\Delta x \to 0} \frac{\Delta u}{\Delta x} \right) \\ &= f'(u) \cdot g'(x) = f'(g(x)) \cdot g'(x). \end{aligned}$$

[2] Although we are using the $f'(x)$ notation, the original text is motivated by the differential form. As this is only tangentially referenced but of extreme relevance, we left it as is.

Thus we have shown that $(f \circ g)'(x) = f'(g(x)) \cdot g'(x)$ at any point x at which $g'(x) \neq 0$. But if $g'(x) = 0$, then it is entirely possible that Δu *is* zero for some or all nonzero values of Δx approaching zero—in which case the factorization in (14) is invalid. Our proof of the chain rule is therefore incomplete. In another section we give a proof that does not require the assumption that $g'(x) \neq 0$.

A.2 Outline of a Proof of the Chain Rule from [16] Using Differential Notation

To *outline* a proof of the chain rule, suppose that we are given differentiable functions $y = f(u)$ and $u = g(x)$ and want to compute the derivative:

$$\frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(g(x + \Delta x)) - f(g(x))}{\Delta x} \tag{15}$$

The differential form of the chain rule suggests the factorization

$$\frac{\Delta y}{\Delta x} = \frac{\Delta y}{\Delta u} \frac{\Delta u}{\Delta x} \tag{16}$$

where

$$\Delta u = g(x + \Delta x) - g(x) \text{ and } \Delta y = f(u + \Delta u) - f(u).$$

For x fixed, the factorization in Eq. (16) is valid if $\frac{du}{dx} \neq 0$, because

$$\frac{du}{dx} = \lim_{\Delta x \to 0} \frac{\Delta u}{\Delta x} \neq 0$$

implies that $\Delta u \neq 0$ if $\Delta x \neq 0$ is sufficiently small—for if so, then $\Delta u = (\Delta u / \Delta x) \cdot \Delta x$ is the product of nonzero numbers. But the fact that g is differentiable, and therefore continuous, at the point x implies that

$$\Delta u = g(x + \Delta x) - g(x) \to 0 \text{ as } \Delta x \to 0.$$

The product law of limits therefore gives

$$\frac{dy}{dx} = \lim_{\Delta x \to 0} \left(\frac{\Delta y}{\Delta u} \cdot \frac{\Delta u}{\Delta x} \right) = \left(\lim_{\Delta u \to 0} \frac{\Delta y}{\Delta u} \right) \cdot \left(\lim_{\Delta x \to 0} \frac{\Delta u}{\Delta x} \right) = \frac{dy}{du} \cdot \frac{du}{dx}.$$

Thus we have shown that $\frac{dy}{dx} = \frac{dy}{du} \cdot \frac{du}{dx}$ at any point x at which $\frac{du}{dx} \neq 0$. But if $\frac{du}{dx} = 0$, then it is entirely possible that Δu *is* zero for some or all nonzero values of Δx approaching zero—in which case the factorization in (16) is invalid. Our proof of the chain rule is therefore incomplete. In another section we give a proof that does not require the assumption that $\frac{du}{dx} \neq 0$.

A.3 Outline of a Proof of the Chain Rule from [16] Using Differential Operator Notation

To *outline* a proof of the chain rule, suppose that we are given differentiable functions $y = f(u)$ and $u = g(x)$ and want to compute the derivative:

$$D_x[f(g(x))] = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x} = \lim_{\Delta x \to 0} \frac{f(g(x + \Delta x)) - f(g(x))}{\Delta x} \tag{17}$$

The differential form of the chain rule[3] suggests the factorization

$$\frac{\Delta y}{\Delta x} = \frac{\Delta y}{\Delta u}\frac{\Delta u}{\Delta x} \tag{18}$$

where

$$\Delta u = g(x + \Delta x) - g(x) \text{ and } \Delta y = f(u + \Delta u) - f(u).$$

For x fixed, the factorization in Eq. (18) is valid if $D_x[g(x)] \neq 0$, because

$$D_x[g(x)] = \lim_{\Delta x \to 0} \frac{\Delta u}{\Delta x} \neq 0$$

implies that $\Delta u \neq 0$ if $\Delta x \neq 0$ is sufficiently small—for if so, then $\Delta u = (\Delta u/\Delta x) \cdot \Delta x$ is the product of nonzero numbers. But the fact that g is differentiable, and therefore continuous, at the point x implies that

$$\Delta u = g(x + \Delta x) - g(x) \to 0 \text{ as } \Delta x \to 0.$$

The product law of limits therefore gives

$$\begin{aligned} D_x[f(g(x))] = \lim_{\Delta x \to 0}(D_u[f(u)] \cdot D_x[g(x)]) &= \left(\lim_{\Delta u \to 0} \frac{\Delta y}{\Delta u}\right) \cdot \left(\lim_{\Delta x \to 0} \frac{\Delta u}{\Delta x}\right) \\ &= D_u[f(u)] \cdot D_x[g(x)]. \end{aligned}$$

Thus we have shown that $D_x[f(g(x))] = D_u[f(u)] \cdot D_x[g(x)]$ at any point x at which $D_x[g(x)] \neq 0$. But if $D_x[g(x)] = 0$, then it is entirely possible that Δu *is* zero for some or all nonzero values of Δx approaching zero—in which case the factorization in (18) is invalid. Our proof of the chain rule is therefore incomplete. In another section we give a proof that does not require the assumption that $D_x[g(x)] \neq 0$.

[3] See footnote 2.

References

[1] N. Ashcroft and N. Mermin. *Solid State Physics*. HRW international editions. Holt, Rinehart and Winston, 1976.

[2] W. Ball. *A Primer of the History of Mathematics*. Macmillian & Company, 1895.

[3] J. Barwise. *Handbook of Mathematical Logic*. Studies in Logic and the Foundations of Mathematics. Elsevier Science, 1982.

[4] E. Bloch. *Proofs and Fundamentals: A First Course in Abstract Mathematics*. Undergraduate Texts in Mathematics. Springer, 2011.

[5] C. Boyer. *History of Analytic Geometry*. Dover Books on Mathematics. Dover Publications, 2012.

[6] C. Boyer. *The History of the Calculus and Its Conceptual Development*. Dover Books on Mathematics. Dover Publications, 2012.

[7] C. Boyer and U. Merzbach. *A History of Mathematics*. Wiley, 2011.

[8] I. Bronshtein, K. Semendyayev, G. Musiol, and H. Mühlig. *Handbook of Mathematics*. Springer, 2007.

[9] F. Cajori. *A History of Mathematics*. Macmillan & Company, 1893.

[10] F. Cajori. Leibniz, the master-builder of mathematical notations. *Isis*, 7(3):412–429, 1925.

[11] F. Cajori. *A History of Mathematical Notations*. Dover Publications, 1928.

[12] S.-S. Chern, C. Chevalley, et al. Elie cartan and his mathematical work. *Bulletin of the American Mathematical Society*, 58(2):217–250, 1952.

[13] R. Cooke. *The History of Mathematics: A Brief Course*. Wiley, 2012.

[14] F. A. Cotton. *Chemical applications of group theory*. John Wiley & Sons, 2008.

[15] M. J. Crowe. *A history of vector analysis: The evolution of the idea of a vectorial system*. Courier Corporation, 1967.

[16] C. H. Edwards and D. E. Penney. *Calculus, Early Transcendentals*. Prentice Hall Upper Saddle River, NJ, 7 edition, 2007.

[17] S. Elliott and S. Elliott. *The Physics and Chemistry of Solids.* Wiley, 1998.

[18] W. Fulton and J. Harris. *Representation theory*, volume 129. Springer Science & Business Media, 1991.

[19] J. Gallian. *Contemporary abstract algebra.* Cengage Learning, 2009.

[20] T. Gowers, J. Barrow-Green, and I. Leader. *The Princeton Companion to Mathematics.* Princeton University Press. Princeton University Press, 2008.

[21] P. A. Grillet. *Abstract algebra*, volume 242. Springer Science & Business Media, 2007.

[22] B. C. Hall. *Lie groups, Lie algebras, and representations*, volume 222. Springer Science & Business Media, 2003.

[23] J. Hofman. *The History of Mathematics.* Kensington Publishing Corporation, 1957.

[24] R. Kager. *Optimality Theory.* Cambridge Textbooks in Linguistics. Cambridge University Press, 1999.

[25] I. Kleiner. *A History of Abstract Algebra.* Birkhäuser, 2007.

[26] J. C. Maxwell. A dynamical theory of the electromagnetic field. *Philosophical Transactions of the Royal Society of London*, pages 459–512, 1865.

[27] J. Mazur. *Enlightening Symbols: A Short History of Mathematical Notation and Its Hidden Powers.* Princeton University Press, 2014.

[28] G. Miller. *Historical Introduction to Mathematical Literature.* Cornell University Library historical math monographs. Macmillan, 1916.

[29] A. Polyanin and A. Manzhirov. *Handbook of Mathematics for Engineers and Scientists.* Taylor & Francis, 2006.

[30] C. Reid and H. Weyl. *Hilbert.* Springer-Verlag, 1970.

[31] E. Scheinerman and J. Scheinerman. *Mathematical Notation: A Guide for Engineers and Scientists.* CreateSpace, 2011.

[32] J. A. Schouten. Der ricci-kalkül. 1924.

[33] D. Smith. *History of Modern Mathematics*. Mathematical monographs. Wiley, 1896.

[34] E. Snapper. The three crises in mathematics: logicism, intuitionism and formalism. *Mathematics Magazine*, pages 207–216, 1979.

[35] N. E. Steenrod. *How to write mathematics*. American Mathematical Soc., 1981.

[36] D. J. Struik. Outline of a history of differential geometry: I. *Isis*, 19(1):92–120, 1933.

[37] D. J. Struik. Outline of a history of differential geometry (ii). *Isis*, 20(1):161–191, 1933.

[38] J. Van Heijenoort. *From Frege to Gödel: a source book in mathematical logic, 1879-1932*. Source books in the history of the sciences. Harvard University Press, 1967.

[39] J. Čižmár. The origins and development of mathematical notation (a historical outline). *Quaderni di ricerca in didattica*, 9:103–123, 2000.

[40] A. Whitehead and B. Russell. *Principia Mathematica*. Number vol. 2 in Principia Mathematica. University Press, 1912.

[41] Wikipedia. Optimality theory — Wikipedia, the free encyclopedia, 2015. [Online; accessed May-29-2015].

[42] S. Wolfram. Mathematical notation: Past and future, 2000.