Aesthetic Analysis of Proofs of the Binomial Theorem

Lawrence Neff Stout
Department of Mathematics and Computer Science
Illinois Wesleyan University
Bloomington, IL 61702-2900
lstout@iwu.edu

August 16, 1999

This paper explores aesthetics of mathematical proof. Certain important aspects of proofs are not relevant to aesthetics (validity, utility, exposition) but others are (immediacy, enlightenment, economy of means, establishment of connections, opening of mathematical vistas). Three different proofs of the binomial theorem are used as illustrations.

1 Introduction

Proof in mathematics has two central roles: it provides the definitive criterion for truth in the subject (an epistemological role) and it is the canvas for part of the aesthetic of mathematics.

In order to meet the demands of the epistemological role, a proof must follow the rules of deductive logic. Each statement in the proof must either be an axiom or definition or be known to be correct from a previous proof, or it must follow from earlier statements in the proof. Proofs are usually informal in that they do not fill in all of the steps, but rather depend on the mathematical knowledge of the reader to provide, if desired, all of the connections. Thus a proof depends on an intellectual tradition and a social context for satisfaction of its epistemological role.

That context will provide for an agreed upon notion of number, specification of the logical constructs allowed in the proof (usually classical predicate calculus, unless an explicit constructivist or intuitionist viewpoint is taken), notational conventions, and familiarity with other theorems which may be brought to bear. In particular, there is a need for knowledge of the proofs of those other results, so that hidden circularity is avoided.

But satisfaction with and appreciation of a proof does not end with determination of its validity. We ask for insight. A proof should not only tell us *that* a mathematical statement is true, but *why* it is true. We will find a proof more pleasing if it is elegant and efficient. A proof which shows how disparate parts of mathematics combine to give new results will be more satisfying than a proof which shows a result in a narrow context. Proofs illustrating the power of major theorems can either delight ("Wow, that was slick!") or disappoint ("Shooting a fly with a cannon") depending on whether the result seemed deserving of the tool. Some proofs provoke awe by their immediacy (Bhaskara's one word proof of the Pythagorean Theorem) and others by the element of surprise in how their pieces fit together (Euclid's proof of the Pythagorean Theorem).

In this paper I propose to consider several proofs of the the Binomial Theorem to see how aesthetic criteria can be applied to mathematical proofs. Since historically several slightly different related results have gone under that name, it is wise to specify exactly what we are proving.

Theorem 1.1 (Binomial Theorem) *For any natural number n and any numbers x and a,*

$$(x+a)^n = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) a^{n-k} x^k.$$

In order to make sense of the theorem we need to agree on some conventions. First, we define the binomial coefficients

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!}$$

using the convention that $0! = 1$ to cover the cases where either n, $n-k$, or k is 0.

We will also stipulate that $x^0 = 1$ and $a^0 = 1$. These are questionable if $x = 0$ or $a = 0$, so those should be dealt with as separate cases. Interpretation of the theorem in those cases gives either $a^n = a^n$ or $x^n = x^n$. If all of $n = 0$, $x = 0$, and $a = 0$ then we get the result $0^0 = 0^0$, which isn't particularly meaningful, but as long as we agree on what we mean by 0^0 we are forced to accept the result.

2 Three Proofs

The binomial theorem can be thought of as a solution for the problem of finding an expression for $(x+a)^n$ from one for $(x+a)^{n-1}$ or as a way to find the coefficients of $(x+a)^n$ directly. Solutions using what we call Pascal's triangle have a long history: Struik [8], p.21 gives references to books written in 1261 by Yang Hui and 1425 by Al-Kashi; Klein [4], p.272 notes that the result was known to thirteenth century Arabs and appears in a text by Stifel in 1544. Newton generalized the theorem to fractional and negative exponents in two letters to Henry Oldenberg in 1676, though he gave no proof.

2.1 Induction Proof

Many textbooks in algebra give the binomial theorem as an exercise in the use of mathematical induction. This can be thought of as a formalization of the technique for getting an expression for $(1+a)^n$ from one for $(1+a)^{n-1}$. The key calculation is in the following lemma, which forms the basis for Pascal's triangle.

Lemma 2.1 *For all* $1 \leq k \leq m$

$$\left(\begin{array}{c} m \\ k \end{array} \right) + \left(\begin{array}{c} m \\ k-1 \end{array} \right) = \left(\begin{array}{c} m+1 \\ k \end{array} \right)$$

Proof:

This is a direct calculation in which we add fractions and simplify:

$$\begin{aligned}
\left(\begin{array}{c} m \\ k \end{array} \right) + \left(\begin{array}{c} m \\ k-1 \end{array} \right) &= \frac{m!}{(m-k)!k!} + \frac{m!}{(m-k+1)!(k-1)!} \\
&= \frac{m!(m-k+1)!(k-1)! + m!(m-k)!k!}{(m-k)!k!(m-k+1)!(k-1)!} \\
&= \frac{m!(k-1)!(m-k)!(k+(m-k+1))}{(m-k)!k!(m-k+1)!(k-1)!} \\
&= \frac{m!(k+(m-k+1))}{k!(m-k+1)!} \\
&= \frac{m!(m+1)}{k!(m-k+1)!} \\
&= \frac{(m+1)!}{k!(m-k+1)!}
\end{aligned}$$

$$= \begin{pmatrix} m+1 \\ k \end{pmatrix}$$

∎

With this Lemma we can give a fairly quick induction proof.

PROOF:

We proceed by mathematical induction:

For the case $n = 0$ the theorem says

$$(x+a)^0 = \sum_{k=0}^{0} \begin{pmatrix} 0 \\ k \end{pmatrix} a^{0-k} x^k.$$

Now $(x+a)^0 = 1$ and

$$\sum_{k=0}^{0} \begin{pmatrix} 0 \\ k \end{pmatrix} a^{0-k} x^k = \begin{pmatrix} 0 \\ 0 \end{pmatrix} a^0 x^0 = 1.$$

Here we are using the conventions that

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

and that any number to the 0 power is 1. Given the artificiality of these assumptions, we may be happier if the base case for $n = 1$ is also given.

For the case $n = 1$ the theorem says

$$(x+a)^1 = \sum_{k=0}^{1} \begin{pmatrix} 1 \\ k \end{pmatrix} a^{1-k} x^k = \begin{pmatrix} 1 \\ 0 \end{pmatrix} a^1 x^0 + \begin{pmatrix} 1 \\ 1 \end{pmatrix} a^0 x^1.$$

This is equivalent to

$$(x+a) = \frac{1!}{1!0!} a + \frac{1!}{0!1!} x = a + x$$

which is true. Thus we have the base cases for our induction.

For the induction step we assume that

$$(x+a)^m = \sum_{k=0}^{m} \begin{pmatrix} m \\ k \end{pmatrix} a^{m-k} x^k$$

and show that

$$(x+a)^{m+1} = \sum_{k=0}^{m+1} \left(\begin{array}{c} m+1 \\ k \end{array} \right) a^{m+1-k} x^k.$$

This is a calculation using the Lemma

$$(x+a)^{m+1} = (x+a)^m (x+a) = \left(\sum_{k=0}^{m} \left(\begin{array}{c} m \\ k \end{array} \right) a^{m-k} x^k \right) (x+a)$$

$$= \sum_{k=0}^{m} \left(\begin{array}{c} m \\ k \end{array} \right) a^{m-k} x^{k+1} + \sum_{k=0}^{m} \left(\begin{array}{c} m \\ k \end{array} \right) a^{m-k+1} x^k$$

$$= \left(\begin{array}{c} m \\ 0 \end{array} \right) a^{m+1} x^0 + \sum_{k=1}^{m} \left(\left(\begin{array}{c} m \\ k \end{array} \right) + \left(\begin{array}{c} m \\ k-1 \end{array} \right) \right) a^{m-k+1} x^k + \left(\begin{array}{c} m+1 \\ m+1 \end{array} \right) a^0 x^{m+1}$$

$$= \sum_{k=0}^{m+1} \left(\begin{array}{c} m+1 \\ k \end{array} \right) a^{m+1-k} x^k$$

Completing the proof by induction. ∎

2.2 Combinatorial Proof

The combinatorial proof of the binomial theorem originates in Jacob Bernoulli's *Ars Conjectandi* published posthumously in 1713. See [2] p.383. It appears in many discrete mathematics texts.

PROOF:

We start by giving meaning to the binomial coefficient

$$\left(\begin{array}{c} n \\ k \end{array} \right) = \frac{n!}{k!(n-k)!}$$

as counting the number of unordered k-subsets of an n element set. This is done by first counting the ordered k-element strings with no repetitions: for the first element we have n choices; for the second, $n-1$; until we get to the k^{th} which has $n-k+1$ choices. Since these choices are made in succession, we multiply to get

$$n(n-1)\cdots(n-k+1) = \frac{n!}{(n-k)!}$$

such ordered k-tuples without repetition. Each k-element subset can be ordered in $k!$ different ways, so the count of ordered k-tuples is exactly $k!$ times too big for counting subsets. Thus the number of k element subsets of an n element set is

$$\frac{n!}{k!(n-k)!} = \left(\begin{array}{c} n \\ k \end{array} \right).$$

Next we observe that the process of multiplying out $(x+a)^n$ involves adding up 2^n terms each obtained by making a choice for each factor to use either the x or the a. The choices which result in k x's and $n-k$ a's each give a term of the form $a^{n-k}x^k$. There are $\left(\begin{array}{c} n \\ k \end{array} \right)$ distinct ways to choose the k element subset of factors from which to take the x. Thus the coefficient of $a^{n-k}x^k$ is $\left(\begin{array}{c} n \\ k \end{array} \right)$. This tells us that

$$(x+a)^n = \sum_{k=0}^{n} \left(\begin{array}{c} n \\ k \end{array} \right) a^{n-k}x^k$$

as desired. ∎

2.3 Derivation using Calculus

Newton's generalization of the binomial theorem gives rise to an infinite series. Careful consideration of differentiation inside the radius of convergence and uniqueness considerations from differential equations allow a proof (sketched, for instance, in Sallas-Hille [7], p. 679–curiously, most standard calculus books give this series at about page 670). If we restrict to natural number exponents, the convergence considerations are not necessary and a proof based on the differentiation of polynomials becomes possible. One needs to be careful not to use the binomial theorem in proving the power rule if one wants to use this proof or one will introduce a circularity.
PROOF:

We first note that since $(x-a)$ is a polynomial of degree 1, $(x+a)^n$ will be a polynomial of degree n and will thus be determined once we know what the coefficients of each of the $n+1$ possible powers of x are. For concreteness let us write

$$(x+a)^n = p(x) = \sum_{k=0}^{n} b_k x^k$$

and show how to determine the coefficients b_k.

Using the power rule and the chain rule for differentiation we observe that

$$\frac{d}{dx}(x+a)^n = n(x+a)^{n-1}$$

so that

$$(x-a)\frac{d}{dx}(x+a)^n = n(x+a)^n$$

with $(0+a)^n = a^n$. This gives a first order differential equation satisfied by $p(x) = (x+a)^n$, namely

$$(x+a)p'(x) = np(x)$$

with initial condition

$$p(0) = a^n.$$

We then determine what the coefficients b_k must be to satisfy this equation. The initial condition $p(0) = a^n$ tells us that $b_0 = a^n$. We can relate later coefficients to earlier ones using the differential equation:

$$p'(x) = \sum_{k=1}^{n} k b_k x^{k-1}$$

so

$$\begin{aligned}
(x+a)p'(x) &= \sum_{k=1}^{n} k b_k x^k + \sum_{k=1}^{n} a k b_k x^{k-1} \\
&= a b_1 + \sum_{k=1}^{n-1} (k b_k + a(k+1) b_{k+1}) x^k + n b_n x^n \\
&= \sum_{k=0}^{n} n b_k x^k
\end{aligned}$$

Since polynomials are equal when their coefficients are equal, this tells us that

$$\begin{aligned}
a\ b_1 &= n b_0 \\
(1\ b_1) + (a\ 2\ b_2) &= n\ b_1 \\
&\vdots \quad \vdots \\
(k b_k) + (a(k+1) b_{k+1}) &= n\ b_k \\
n\ b_n &= n\ b_n
\end{aligned}$$

Thus for $k = 1, \ldots, n-1$ we get

$$b_{k+1} = \frac{n-k}{(k+1)a} b_k.$$

Using the fact that $b_0 = a^n$ this gives us

$$\begin{aligned}
b_0 &= a^n \\
b_1 &= na^{n-1} \\
b_2 &= \frac{n(n-1)}{2} a^{n-2} = \left(\begin{array}{c} n \\ 2 \end{array} \right) a^{n-2} \\
b_3 &= \frac{n(n-1)(n-2)}{3 \cdot 2} a^{n-3} = \left(\begin{array}{c} n \\ 3 \end{array} \right) a^{n-3} \\
\vdots &\quad \vdots \\
b_k &= \frac{n(n-1)\cdots(n-k+1)}{k!} a^{n-k} = \left(\begin{array}{c} n \\ k \end{array} \right) a^{n-k}
\end{aligned}$$

which proves the theorem. ∎

3 Aesthetic principles in mathematical proof

What is it that makes a mathematical proof beautiful? While several authors have discussed beauty in mathematics, most of mainstream philosophy of mathematics deals with issues of ontology and epistemology rather than aesthetics. Philosophers are much more concerned with the nature of mathematical reality and the status of mathematical truths. The cumulative nature of mathematical truth (we don't revise the truth of previous results because rigorous proofs lead to a level of certainty not found in other disciplines) and the abstraction of mathematical objects make mathematics a special case in philosophical investigation. Mathematical theorems do not cease to be true, nor do proofs cease to be valid; they do, however, differ in their perceived significance over time. There are clearly fashions in style in mathematical proof and there are judgments made about what mathematics is interesting and thus worth the effort to understand and polish.

Davis and Hersh [3] give a chapter to aesthetic considerations – a short one, mostly noting the antiquity of the recognition of beauty in mathematics (quoting Aristotle) and the paucity of explanations of what that beauty consists of. They say

> Aesthetic judgment exists in mathematics, is of importance, can be cultivated, can be passed from generation to generation, from teacher to student, from author to reader. But there is very little formal description of what it is and how it operates. ...
>
> Attempts have been made to analyze mathematical aesthetics into components–alternation of tension and relief, realization of expectations, surprise upon perception of unexpected relationships and unities, sensuous visual pleasure, pleasure at the juxtaposition of freedom and constraint, and , of course, into the elements familiar from the arts, harmony, balance, contrast, etc. ... [3, p. 169]

Tymoczko's paper [9] noting that aesthetics as well as applications to science can provide a justification for mathematics, cites a need for criticism in mathematics. His discussion includes consideration of proof as both an art of composition and an art of performance, allowing for the refinement of proofs using earlier expositions as templates. Criticism has a role in teaching: "It can give rise to what critics in other arts call 'the canon': the body of lived proofs, the presentations still going on, that we want to teach to our students so that they can become gifted listeners of mathematics, sensitive critics able to judge new works as they appear."(p.73)

Borel's address [1] compares mathematics and painting. Both involve taking inspiration from either the real world (in mathematics, from applications and problems arising in applications; in art, from the subject of the painting) and an important role for abstraction. Edward Rothstein's *Emblems of the Mind* [6] gives an extended study of the similarities between music and mathematics, with concern for all of composition, technique, and aesthetics.

Gian-Carlo Rota [5] in his essay *The Phenomenology of Mathematical Beauty* stresses the variety of aspects of mathematics which can be considered beautiful (theorems, proofs, definitions, axiom systems) and notes that they need not go together: a beautiful theorem can have an ugly proof. He also notes the essential context sensitivity of judgments of mathematical beauty. By the end of the essay Rota concludes that discussion of beauty is a cop out and what mathematicians actually want is enlightenment:

> Mathematicians seldom explicitly acknowledge the phenomenon of enlightenment for at least two reasons. First, unlike truth, enlightenment is not easily formalized. Second, enlightenment admits degrees: some statements are more enlightening than others. Mathematicians dislike concepts admitting degrees, and will go to any length to deny the logical

> role of any such concept. Mathematical beauty is the expression mathematicians have invented in order to obliquely admit the phenomenon of enlightenment while avoiding acknowledgement of the fuzziness of this phenomenon. They say that a theorem is beautiful when they mean to say that it is enlightening.[5, p.132]

We can note first things which are not involved because they either relate to other mathematical issues or to results rather than proofs:

1. A proof must be logically correct to be a proof, so the truth of the result is a presupposition and is not relevant to the judgment of beauty.

2. Utility is not relevant since it is most often the result itself and not the proof that has utility.

3. For most proofs the beauty is not visual, but abstract.

4. While exposition matters, the beauty of a proof does not lie in the felicity of the wording. A poorly written beautiful proof can be rewritten to make the beauty more apparent, but an ugly proof will not be made beautiful through polished or poetic exposition. Rota [5, p.128] notes the distinction between elegance and beauty and notes that a beautiful proof can be presented both elegantly and inelegantly.

There are general criteria that we can use to judge the beauty of a proof. My object here is to discuss measures of beauty internal to mathematics, hence omitting many of the issues mentioned by Davis and Hersh as being more general aesthetic criteria:

1. A beautiful proof should make the result it proves immediately apparent.

2. It should explain (at least one aspect of) why the result not only *is* true but *should be* true.

3. It should be economical, using no more than is necessary for the result.

4. A beautiful proof often makes unexpected connections between seemingly disparate parts of mathematics.

5. A proof which suggests further development in the subject will be more pleasing than one which closes off the subject.

As with all aesthetic judgments, there is room for both cultural and individual variation in assessing the importance of different factors. What suggests further developments at one stage of the development of mathematics may not at another. What one mathematician perceives as disparate parts of mathematics may be so closely linked in another's mind that the surprise factor is absent.

Let us next attempt to apply these criteria to the proofs of the binomial theorem given earlier:

A beautiful proof should make the result it proves immediately apparent. What argument best makes a result immediately apparent depends a bit on the preparation of the beholder. A proof which is not understood will not produce the *aha!* reaction. Of the proofs given for the binomial theorem the induction proof and the proof using calculus extract the binomial theorem through calculations rather than giving a direct meaning to the coefficients. The combinatorial proof is somewhat more immediate, giving a single conceptual reason why the theorem is true. In general conceptual proofs are preferred to computational proofs, unless the computation involved is particularly elegant. (For instance the proof of Taylor's theorem which uses integration by parts to produce a telescoping series in which all but two terms vanish on rearrangement, while essentially computational, is elegant and provides the immediacy asked for in this criterion.)

It should explain why the result not only *is* true but *should be* true. Here the different proofs provide different aspects of why the binomial theorem should be true. The induction proof and the calculus proof show how the binomial theorem follows from well established machinery(calculation and differentiation). The counting argument uses a conceptual approach.

The inductive proof builds on the recursive nature of the definition of powers and shows how explicit calculation can tell how to get from one case to the next. Students at an early (concrete) stage of mathematical maturity find the calculation approach appealing, though induction is often difficult for them to grasp. Deep understanding of how induction and recursion are intertwined is needed for the induction proof to give the *should be true* reaction. For most mathematicians and students of mathematics induction proofs give little enlightenment and may be judged to be rather ugly because of that failure.

The other two proofs fare better on this criterion. The proof using calculus uses the central notion that knowing a first order differential equation and an initial condition should be enough to specify a differentiable function. In the details of how the

binomial coefficients follow from differentiation we gain insight into why the coefficient has the specified form. The counting argument also gives a clear reason why the binomial theorem should be true.

It should be economical, using no more than is necessary for the result. Of the three proofs, only the calculus proof looks like overkill. It uses much more machinery than the other two proofs. For economy it is hard to beat the induction proof. The calculation uses nothing more than basic algebra as does the induction step. The combinatorial proof also has an economy of means with only a minor side excursion into counting permutations..

A beautiful proof often makes unexpected connections between seemingly disparate parts of mathematics. Here the proof using calculus points out connections more distant than those suggested by either of the other two proofs. The induction proof stays firmly in one part of mathematics, suggesting few connections for the result. The combinatorial proof does make connections to counting technique, but the perceived distance between algebraic manipulation and counting is not large enough for the connection to be particularly surprising.

There is some conflict between the desire for a proof to have economy in means and to make connections with other parts of mathematics. Such unexpected connections may seem like side issues and diversions if other shorter more direct proofs are available. However, links to more distant parts of mathematics open up further development in a way that more self contained proofs do not.

A proof which suggests further development in the subject will be more pleasing than one which closes off the subject. A strong case can be made here for all three of the proofs.

The induction proof suggests the utility of recurrences. It also gives one of the most basic examples of an essential proof technique. As such it opens vistas on many parts of mathematics. The induction proof, however, gives little indication how to find new results about binomial coefficients, or how to generalize the binomial theorem to multinomials or fractional exponents. While the induction proof gives little guidance in how to develop generalizations in algebra, it does provide an important example of a key technique with wide application in mathematics.

The combinatorial proof suggests that algebraic identities can be proved by looking at the meaning of what counting methods they represent. A small cottage industry arises in combinatorics using similar ideas to prove identities using the binomial coef-

ficients. Turning the problem in another direction this proof shows how polynomials can be thought of as generating functions for counting problems. It is then a short and productive step to more general kinds of generating functions.

Within the calculus proof there is another idea which proves useful in combinatorics: thinking of $(x+a)^n$ as a function of x rather than as a formal expression in the symbols x and a. This opens the way for application of analytic techniques, useful for proving further identities by differentiation or integration. The calculus based proof I've given also illustrates how differential equations can be used to obtain coefficients in a Taylor series expansion for a solution.

In a sense the calculus proof of the binomial theorem is cheating on this criterion, since what I have given is the simplest case of the binomial series developed by Newton. Thus rather than having this proof suggest further developments to us, we have obtained the proof by specializing the further developments it leads to. In fact Newton gave the binomial series by generalizing from the form given by the binomial theorem and did not give a proof of the result.

References

[1] A. Borel. Mathematics: Art and science. *Mathematical Intelligencer*, 5(4):5–17, 1983. Translated by Kevin M. Lenzen.

[2] Robert Caliger, editor. *Classics of Mathematics*. Moore Publishing Co., Oak Park, IL, 1982.

[3] Philip J. Davis and Reuben Hersh. *The Mathematical Experience*, chapter The Aesthetic Component, pages 168–171. Houghton Mifflin, 1981.

[4] Morris Klein. *Mathematical Thought from Ancient to Modern Times*. Oxford University Press, 1972.

[5] Gian-Carlo Rota. *Indiscrete Thoughts*, chapter The Phenomenology of Mathematical Beauty, pages 121–133. Birkhäuser, 1997.

[6] Edward Rothstein. *Emblems of the Mind: The Inner Life of Music and Mathematics*. Times Books, 1995.

[7] S.L. Sallas and Einar Hille. *Calculus, One and Several Variables*. Wiley, 6 edition, 1990.

[8] D.J. Struik, editor. *A Sourcebook in Mathematics, 1200-1800*. Harvard University Press, Cambridge, 1969.

[9] Thomas Tymoczko. Value judgements in mathematics: Can we treat mathematics as an art? In Alvin White, editor, *Essays in Humanistic Mathematics*, number 32 in Notes, pages 67–78. MAA, 1993.