DESIGN

Creation of Artifacts in Society

Karl T. Ulrich
University of Pennsylvania

Published by the University of Pennsylvania.

Edition 1.0

PDF version formatted for 8 ½ by 11 printing
Other formats available at http://www.ulrichbook.org/

ISBN 978-0-9836487-0-3

Written, designed, illustrated, and photographed by Karl T. Ulrich, except where noted.

Cover Photo: Glen Mullaly (collection of spindle adapters for 45 rpm records)

DESIGN

creation of artifacts in society

Karl T. Ulrich

Contents

Preface

As a freshman entering MIT I intended to be a physician, but early in my first year I made new friends who were taking mechanical design courses. They were always carrying around bags of interesting components and displaying metal parts they had made in the machine shop. I was bitten early by the design bug and took all the courses I could on the subject. My identity as designer was solidified in 1979 as a winner of the MIT "2.70" design contest (now 2.007), an outcome that gave me near celebrity status in the hacker-designer crowd at MIT. I was fortunate to have as professors Ernesto Blanco, Woodie Flowers, David Jansson, Warren Seering, and others who were deeply committed to design education.

Karl Ulrich, age 19, winning the MIT "2.70" design contest. Source: MIT.

I did my doctoral work in the MIT Artificial Intelligence Lab, focusing on fundamentals of design theory and machine learning, and developed a whole new perspective on problem solving and design from Randy Davis, Marvin Minsky, Patrick Winston, and many other really interesting students and faculty in that lab. The AI Lab also had the best shop on campus, so after hours I designed and built cool stuff like a recumbent bicycle.

In the 25 years since I was a student at MIT, I've been lucky to lead a professional life that blends teaching design, doing design, and researching design, a luxury afforded by the culture of the Wharton School and the University of Pennsylvania.

My roots are in engineering design, and much of my professional life has been centered on product design. However, in the past 15 years, stints as an entrepreneur and a university administrator have broadened my conception of design to include the creation of services, businesses, and organizations. I intend for this book to be a synthesis of what I know about design based on the varied perspectives of teacher, researcher, and practicing designer.

Narberth
Pennsylvania
United States

ONE

Introduction to Design

Here are some of the human activities characterized as design:

Architectural design	Interior design
Automotive design	Landscape design
Business design	Lighting design
Ceramic and glass design	Machine design
Color design	Mechanical design
Communication design	News design
Engineering design	Packaging design
Environmental design	Product design
Experience design	Production design
Fashion design	Service design
Floral design	Software design
Furniture design	Sound design
Game design	System design
Garden design	Theatrical design
Graphic design	Type design
Industrial design	Urban design
Information design	User experience design
Instructional design	User interface design
Interaction design	Web design

The word *design* presents definitional challenges. Designers tend to view their own particular sphere of activity as the universe of the human activity of designing. For example, one of the twelve schools at the University of Pennsylvania is the School of Design. The school does comprise two clearly recognizable design activities—architecture and urban design—but also fine arts and historic preservation. At the same time, the trade journal *Design News*, with a subscription base of 170,000, focuses quite narrowly on engineering design, a domain not included in Penn's School of Design. I can't think of another human endeavor with such confusing intellectual jurisdictions.

Part of the problem is the English language. What we call *design* in English goes by several different words in other languages. For example, in German the words *konstruktion*, *bauart*, *entwurf*, *planung*, and *design* all refer to different activities that we call simply "design" in English. (See the appendix to this chapter for an overview of the etymology of the word and synonyms in other languages.)

Fortunately, at the level at which I treat design in this book, the activity of design is fundamentally similar across a wide variety of domains.

Design is conceiving and giving form to artifacts that solve problems.[1]

I use *artifact* in a broad and atypical sense to describe any product of intentional creation, including physical goods, services, software, graphics, buildings, landscapes, organizations, and processes. These artifacts can be categorized into *domains*, within which specialization of design methods can be useful.

Exhibits 1-1 through 1-8 are some examples of artifacts in different domains, all designed.[2] Each artifact was conceived and given form to solve a problem. The *form* for artifacts need not be geometric. For example, the computer program in Exhibit 1-1 takes the form of a nested list of symbols. The *problem* need not be a pressing societal need, but rather any perceived gap in a situation or experience. For example, the Insalata Caprese is a wonderful artifact, but hardly addresses a *problem* in the deepest sense of the word.

```
(define (smallest-divisor n)
  (find-divisor n 2))
(define (find-divisor n test-divisor)
  (cond ((> (square test-divisor) n) n)
        ((divides? test-divisor n) test-divisor)
        (else (find-divisor n (+ test-divisor 1)))))
(define (divides? a b)
  (= (remainder b a) 0))
```

Exhibit 1-1. A computer program to find the smallest divisor of an integer N, written in Scheme, a dialect of the programming language LISP. Source: Abelson and Sussman, 1996.

Exhibit 1-2. Insalata Caprese, allegedly originally from the island of Capri in the Campania region of Italy.

Exhibit 1-3. Connecting rods for an automotive engine. Source: LN Engineering.

Exhibit 1-4. The logo for Xootr brand scooters. Source: Lunar Design.

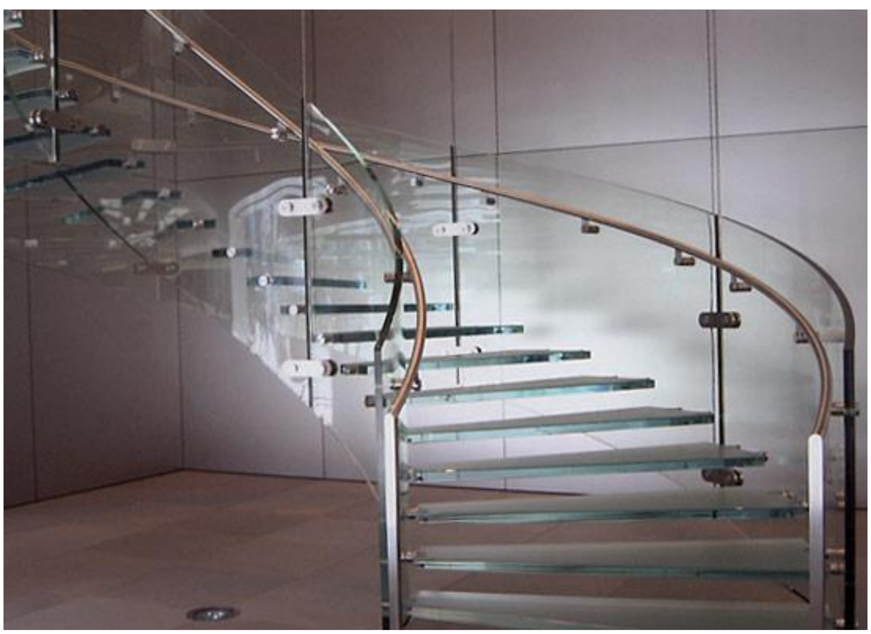


Exhibit 1-5. A glass staircase for the Apple Store in Osaka, Japan. Source: Koji Okumura

Exhibit 1-6. The Sony Cyber-shot digital camera. Source: Sony Corporation.

Exhibit 1-7. The Eclipse jet. Source: Eclipse Aviation.

Exhibit 1-8. Fisher Fine Arts library at the University of Pennsylvania. Designed by Frank Furness and completed in 1890. Source: Wikipedia.

Unifying Framework

From code to cameras and logos to libraries, design domains are highly diverse, and the tools and methods used by designers in these domains can be highly specialized. However, the activity of design across all domains can be usefully unified by a single framework.

I adopt an information processing view of design, largely consistent with that articulated by Herbert Simon in the 1960s (1996). From this perspective, design is part of a human problem-solving activity beginning with a perception of a *gap* in a user experience, leading to a *plan* for a new artifact, and resulting in the *production* of that artifact (Exhibit 1-9).[3] This problem-solving process includes both design and production of the artifact. Design transforms a gap into a plan, which might, for instance, be represented with drawings, computer models, recipes, or parameter values. Production transforms a plan into an artifact.

Note that the same word *design* is used in English as both noun and verb. The noun form may refer to both the activity of designing (e.g., Sammy is responsible for design of the Alpha 2000) and the plan that results from that activity (e.g., Sammy completed the design of the Alpha 2000).

In my model, the user is positioned at the start of the design process. The word *user*, while awkward, is a term of art in professional practice. Equally ugly synonyms include *customer*, *client*, *stakeholder*, and *consumer*. We can't even reliably substitute the term *human*, as users can be animals or aliens. (Instances of design for aliens are exceptional, of course, but consider that NASA's two Voyager spacecraft carry with them "golden records" designed in part for extraterrestrial users.)

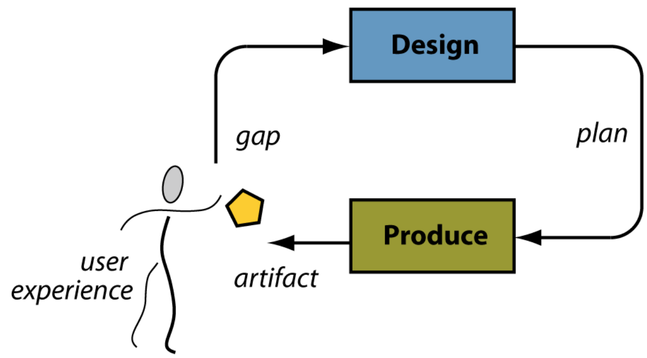


Exhibit 1-9. Design and production are the two activities that deliver artifacts to address gaps in the user experience.

Exhibit 1-10 further decomposes the design portion of the model into four steps. This is a codification of a process that may be implicit for many designers, yet these elements can be discerned in some form in most design efforts:

- **Sense gap.** Design begins with a perception of a gap in the user experience. Without a gap, there is no motive for design. The gap may be perceived by users themselves or by observers.
- **Define problem.** In effect, problem definition is the creation by the designer of an explanation of why the user experiences a gap. This diagnosis can be thought of as an identification of user needs that are not being met in the current state and/or the recognition of criteria for a high-quality solution. Problem definition is implicit in many design efforts, particularly when users are themselves designers, but is generally an explicit part of professional design efforts, expressed in the form of a *design brief*, customer needs list, or other document. Chapter 3 focuses on problem definition.
- **Explore alternatives.** Given a problem, designers almost always explore alternatives. (This step is sometimes called *search*.) I devote Chapter 4 to exploration.
- **Select plan.** Exploration typically exposes more than one solution, so design requires some sort of evaluation and selection from among alternatives. Some designers consider many alternatives simultaneously when selecting a plan. Others articulate, evaluate, and refine plans iteratively and select the first plan that is good enough.

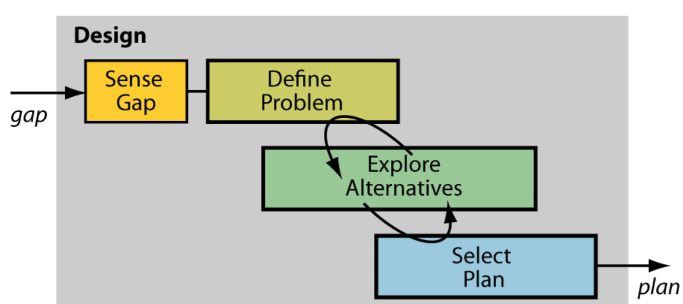


Exhibit 1-10. Design can be thought of as four information processing steps.

Exhibit 1-10 shows design as proceeding from left to right, from gap to plan. While this general flow is typical, the steps are rarely completed in a strict sequence. Exhibit 1-11 reflects the coding of a particular episode of design into three categories (roughly corresponding to my steps defining, exploring, and selecting) as a function of time (Günther 1996). In this instance, the designers jumped back and forth considerably among activities of the three types.

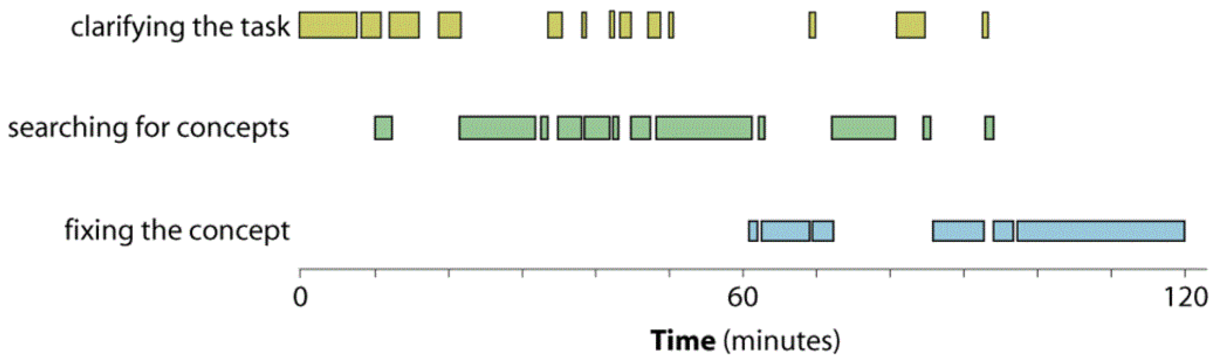


Exhibit 1-11. Actual allocation of design team time to three categories. Source: Günther 1996.

In addition to the iteration occurring among steps, the overall design process is typically executed multiple times, as the first artifact produced rarely results in a complete closing of the gap in the user experience. This iteration may occur across different time scales, ranging from high-frequency iterations by a single individual, perhaps over minutes or hours, to low-frequency iterations over multiple generations of artifacts within an entire society. For example, Rybczynski (2000) provides a detailed chronicle of the evolution of the screw and screwdriver as many iterations of problem solving over hundreds of years.

In the model I present here, design proceeds from gap to plan to artifact. In modern enterprises, the order is sometimes reversed. The designer sometimes begins with an artifact and searches for a gap that it might fill. For instance, DuPont discovered the compound polytetrafluoroethylene (i.e., Teflon) accidentally and then proceeded over many decades to find gaps that the artifact could address. This approach is typical of endeavors for which effective exploration methods are lacking—for example, pharmaceuticals and basic materials. The reverse sequence of design steps is sometimes called *technology push* because it begins with a solution rather than with a gap.[4]

What Is Good Design?

Design is difficult in that it absorbs substantial cognitive effort, typically requires multiple iterations, and rarely results in an optimal artifact, even in situations for which the notion of optimality can be defined. The few design domains that have been described by formal mathematical languages are, in the nomenclature of computational complexity, *NP-complete* search problems, meaning that the theoretically optimal solution cannot reliably be found.[5] Most design domains have not even been formalized, making the inherent complexity even greater and the prospect of optimality even more distant. However, users can generally still evaluate the quality of the outcome of the design process, and different artifacts designed to address the same gap can certainly exhibit markedly different levels of quality.

Design quality is derived from how well the artifact satisfies user needs, and thereby closes the perceptual gap in the user experience. The quality of an artifact is linked to at least these characteristics of the design process:

- How well did the designer diagnose the gap in the user experience? Is the problem as understood by the designer consistent with the causes of the gap experienced by the user? In simple terms, did the designer understand the problem?
- Has the scope for exploration been defined in a way that the space of possibilities includes high-quality solutions? In the nomenclature of cognitive psychology, has the design problem been framed in a way that allows for the discovery of high-quality solutions?
- Did the designer succeed in finding high-quality designs within the solution space that has been defined? Often this result depends on both the skill and knowledge of the designer and on the ease and accuracy with which the designer can forecast the quality of a design without actually having to produce it.

Of course, although not an attribute of the design process per se, the fidelity of production of the plan is also a determinant of user satisfaction.

In sum, did the designer understand the problem, frame it in a way that exploration could potentially lead to a good solution, find such a solution within the solution space, and deliver an artifact consistent with the plan?

Another way of thinking about design quality is to identify *defects* that can arise in the design process. For each element of the process, there is at least one potential defect: The designer may fail to accurately diagnose the gap in the user experience. The designer may frame the exploration problem in a way that excludes many high-quality designs. The designer may only be able to explore a limited portion of the solution space, finding only a few relatively lower-quality solutions. The artifact produced may not be an accurate embodiment of the plan.

Design Is Everything?

The marketing consultant Regis McKenna wrote a famous article in *Harvard Business Review* entitled "Marketing Is Everything" (1991). I know several designers whose blood boiled in response to this title. A common refrain among designers is that indeed *design* is everything (and certainly subsumes marketing). I'm sympathetic to this view, having observed a lot of dysfunctional managerial and political processes that would have been substantially improved by posing a challenge as a design problem and then applying the basic design process. (How often have you participated in a group effort for which no one had clearly articulated the problem, explored alternatives, or carefully selected a plan from the alternatives?)

However, a lot of human problem solving is not really design. The interactive, incremental, ongoing development and refinement of abilities that occurs between a coach and a performer doesn't quite strike me as design. Trading of financial instruments on Wall

Street does not involve much of what I think of as design. Construction of a building to faithfully execute a plan, even with the application of remarkable skill and craft, is hardly design.

The next chapter takes on directly the question of how design relates to human problem solving more generally. For now, let me just state that I believe that much of human problem solving would benefit from *more* design process, not less; however, I don't believe that "design thinking" addresses all challenges we face as individuals, managers, politicians, organizations, and institutions.

This Book

The central theme of this book is that a unifying framework informs the human activity of design across all domains. With few exceptions, each idea in this book applies to graphics, environments, products, software, services, machines, and buildings. I dream that the design process could be integral to the primary, secondary, and postsecondary education of all individuals in modern society. This book is an attempt to lay out some of the ideas that would form that education.

Earlier I alluded to the Nobel Prize–winning economist Herbert Simon and his information processing view of design. Simon was brilliant, and his book *Sciences of the Artificial* contains some beautiful ideas about design. In some ways it was the first serious intellectual treatment of design as a problem-solving activity across domains. But, despite all his merits, Simon didn't connect theory tightly to the practice of creating real artifacts. With this book I aim to marry deep concepts to the way real artifacts are created in society. I also hope to cover some of the big ideas that have been developed in the fifty years since Simon wrote about design.

This is a book about ideas. It is not a handbook for *doing* design. I am writing for three audiences. First, I am writing for designers with an interest in ideas about the design process. This isn't a huge population. I have spent my whole professional life working with the nuts and bolts of design, and I know that few designers have much patience for ideas like those in this book. One of the reasons they became designers was to *do* design, not *think about* design. Second, I am writing for those who do not think of themselves as professional designers, but who have an intellectual interest in design. This is a bigger group than the first, but clearly still not a mass audience. Third, this book is intended for university students and their instructors. There are very few design courses that are part of what might be considered general education in universities. This is unfortunate. However, there are a lot of courses on design or related to design in which one or more of the chapters in this book will be useful. For example, I use the chapters on aesthetics and variety in my product design course, which is intended to develop professional skills in those who want to design products.

This book assumes no specific disciplinary training. Economic principles are typically defined and any engineering concepts used are explained. The mathematics, though scarce, is basic. However, there is certainly an underlying tone to the book that arises from my own training and worldview as a structured thinker with education in engineering and computer science.

The book has eight more chapters:

2. Problem Solving and Design
3. Design Problem Definition
4. Exploration
5. Users, Experts, and Institutions in Design
6. The Architecture of Artifacts
7. Aesthetics in Design
8. Variety
9. Conclusion

Not all issues related to design are included here. For example, I don't write much about organizations, within which most real design gets done. I may address this and other topics in future editions. I believe good designers have a bias for action and learn through iteration. I created this book in that spirit.

Notes

[1] This definition draws on those proposed by at least two others. Edgar Kaufmann Jr., curator of the industrial design department at MOMA 1946–1948, wrote that "design is conceiving and giving form to objects used in everyday life" (Kaufmann 1970). Klaus Krippendorf and Reinhart Butter (1984) wrote, "Design is the conscious creation of forms to serve human needs."

[2] See the three-volume set *Phaidon Design Classics* for 999 "industrially manufactured objects of aesthetic value and timeless quality" (2006). Although they assume a more limited definition of design than I adopt in this book, the Phaidon Classics are nevertheless a fascinating collection of artifacts.

[3] Terwiesch (2007) provides a comprehensive discussion of product development as problem solving. Product development is a specific economic activity that includes design tasks.

[4] See Terwiesch and Ulrich (2009) for a more comprehensive treatment of various modes of innovation in industrial practice.

[5] *NP* means that the time required for an agent to find a solution increases with the size of the problem according to a relationship that is *not polynomial* (e.g., exponential, factorial, etc.). In other words, the problem "explodes" in magnitude in a way that finding a truly optimal solution is impossible in a reasonable amount of time, even with very fast computing.

References

Abelson, H., and Gerald Jay Sussman. 1996. *Structure and Interpretation of Computer Programs*. Cambridge, MA: MIT Press.

Günther, J., E. Frankenberger, and P. Auer. 1996. "Investigation of Individual and Team Design Processes in Mechanical Engineering." In *Analysing Design Activity*, eds. N. Cross et al. Chichester, UK: John Wiley and Sons Ltd.

Kaufmann, Edgar. 1970. *Introductions to Modern Design: What Is Modern Design & What Is Modern Interior Design*. New York: Museum of Modern Art Publications in Reprint.

Krippendorff, K., and Reinhart Butter. 1984. "Product Semantics: Exploring the Symbolic Qualities of Form." *Innovation* 3 (Spring): 4–9.

McKenna, Regis. 1991. "Marketing Is Everything." *Harvard Business Review* 69 (1): 65–79.

Phaidon Design Classics. 2006. Vols. 1–3. London: Phaidon Press.

Rybczynski, Witold. 2000. *One Good Turn: A Natural History of the Screwdriver and the Screw*. New York: Scribner.

Simon, Herbert A. 1996. *The Sciences of the Artificial*. Cambridge: MIT Press.

Simpson, John, and Edmund Weiner, eds. 1989. *The Oxford English Dictionary.* New York: Oxford University Press.

Terwiesch, Christian. 2007. "Product Development as a Problem-solving Process." In *Blackwell Handbook on Technology and Innovation Management*, ed. Scott Shane, 143–172. New York: Wiley-Blackwell.

Terwiesch, Christian, and K. T. Ulrich. 2009. *Innovation Tournaments: Creating and Selecting Exceptional Opportunities.* Boston, MA: Harvard Business Press.

Appendix: The Word *Design*

The word *design* comes to English via French from the Latin root *signum* and means literally to mark out. It was first used in English in the sense I use it in this book in the seventeenth century (OED 1989). By now, the word has assumed many meanings and covers a lot of territory in the English language. Exhibit 1-12 shows the words in several other languages that are used similarly to the way *design* is used in English. German has perhaps the most different terms for more precisely characterizing the different notions of design. Many of these words come from Latin roots, which are probably recognizable to most readers. Interestingly, the English word *design* is popular in other languages and has been adopted either exactly or phonetically (e.g., *dezain* in Japanese). In some of these languages, a word similar to *design* derived more directly from Latin and/or French has a different meaning. For example, in Italian, *disegnare* has the very narrow meaning "to draw," and either the English word *design* or the word *progettazione* (verb *progettare*) is used to refer to the activity of design; in French, the word *désigner* means to designate, not to design, and either *design*, *dessein*, or *conception* is used.

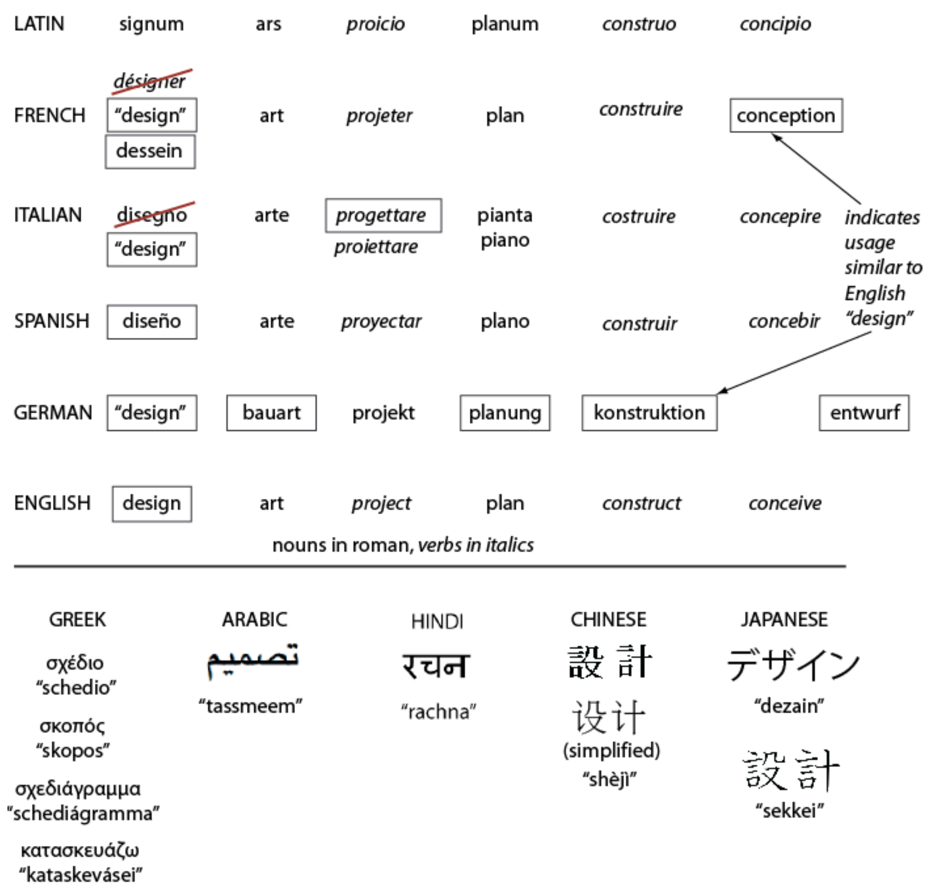


Exhibit 1-12. Words in several other languages used in a way similar to the English word *design*. The most similar terms are outlined with boxes.

TWO

Problem Solving and Design

Benjamin Franklin was an irrepressible problem solver, tackling challenges as diverse as fire prevention, higher education, and home heating. Yet I don't think of him as first and foremost a designer, perhaps because of some significant differences between problem solving and design. This chapter attempts to disentangle the real and perceived differences between design and problem solving and to elucidate both barriers and opportunities for the application of "design thinking" to problem solving more generally.

Exhibit 2-1 is a photograph of a pair of bifocals from Benjamin Franklin's time. Franklin is widely credited with the invention of bifocals, although there is some controversy about this attribution. In a letter to George Whatley in 1785 (Exhibit 2-2), he explains the difficulty of seeing both the food on his plate in front of him and the faces of his guests at the end of the table. He describes a way to address this difficulty by combining lenses in the now familiar bifocal configuration.

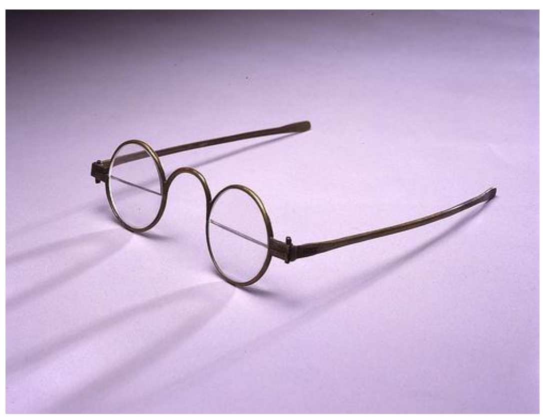


Exhibit 2-1. A pair of "Franklin-type" bifocals from the late eighteenth century. Source: The College of Optometrists (British Optical Association Museum), London.

Franklin's narrative (provided in the appendix) follows the design process I described in Chapter 1 and that is articulated in Exhibit 2-3. Franklin sensed a gap (vision out of focus), defined a problem (objects at different distances require different optical correction), searched for a solution, and then selected a plan (a lens formed from two halves, each with a different diopter).

This process is almost exactly the way I describe *problem solving* in a course I teach on the subject. Exhibit 2-4 shows how I articulate the problem-solving process to my students. An agent operating in the world senses a gap between the current state and some desired state.

The agent then defines a problem or problems, generates alternative solutions, selects an approach, and finally takes action by implementing the solution. In most cases the problem solver then assesses whether the gap has indeed been closed and, if not, the problem-solving process may be repeated iteratively.[1] This problem-solving process is almost exactly the design process in Exhibit 2-3. There is one conceptual difference and several practical distinctions. The conceptual difference between design and problem solving is the difference between *plans* and *outcomes*. The design process results in a plan for action, but not necessarily in a realization of that plan. One nice aspect of the way problem solving is typically taught and practiced is the relative emphasis on action, learning from that action, and improving on the initial solution as a result of that learning. Of course, many good designers are also complete problem solvers, remaining engaged in their challenges through the implementation, testing, and refinement of the artifacts they design. Because problem solving essentially includes the steps in the design process, problem solving is the more general human activity of which design is a critical element.

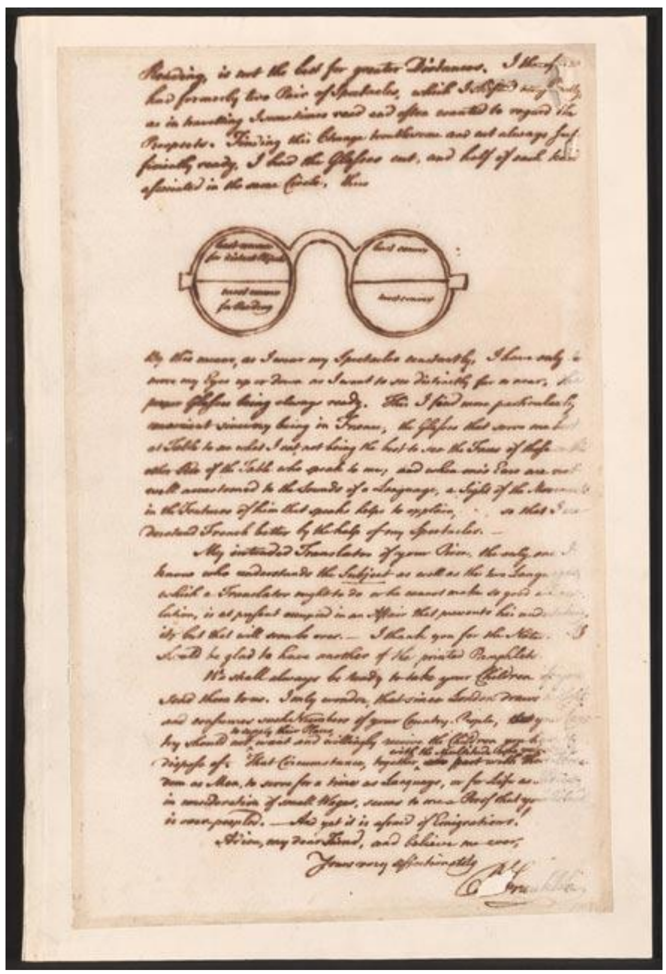
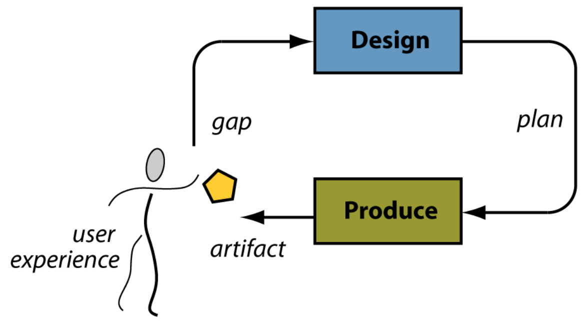


Exhibit 2-2. Letter to George Whatley from Benjamin Franklin describing the creation of bifocals to address the problem of vision correction for both near and far distances. The text of a portion of this letter is in the appendix. Source: United States Library of Congress.

There are other practical distinctions, though, beyond that of implementation and action. They derive from the relative emphasis of design on the creation of *new* artifacts and from the relative importance of time in some problem-solving challenges. These and other distinctions are clarified by a taxonomy of problem types, which includes design problems.



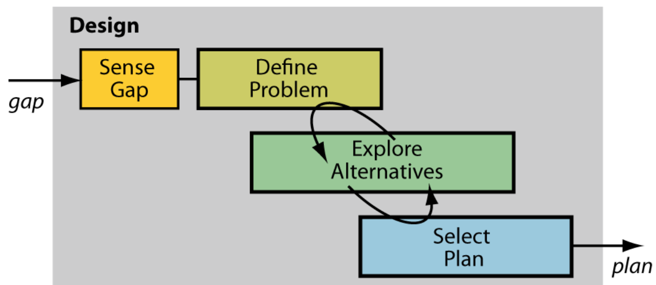


Exhibit 2-3. Design and production address gaps in the user experience. The design process can be thought of as four steps.

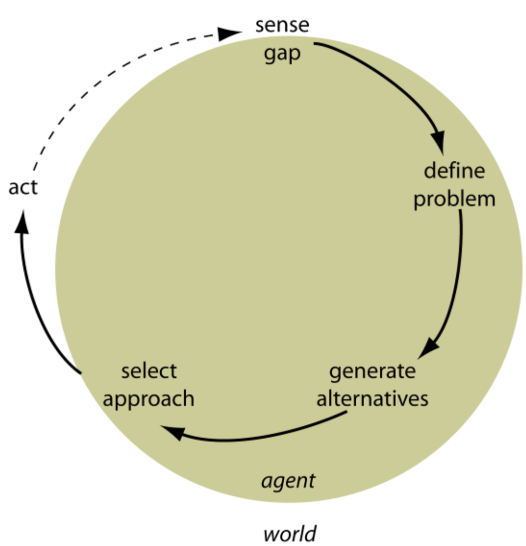


Exhibit 2-4. A generic problem-solving process. Problem solving addresses a gap between the state of the world and a desired state from the perspective of an agent.

Taxonomy of Problems

While pretty much all problem solving can be thought of as a process by which a gap in an agent's experience is closed, a taxonomy of problem types allows us to tighten the distinction between design problems and other types of problems. The categories in the taxonomy and their relationships are illustrated in Exhibit 2-5. The first distinction in the taxonomy is between problems for which there is an existing artifact or operating system and those for which there is no such artifact. This distinction separates all problems into two broad categories: design problems and system improvement problems. The other categories map either across or within these two divisions.

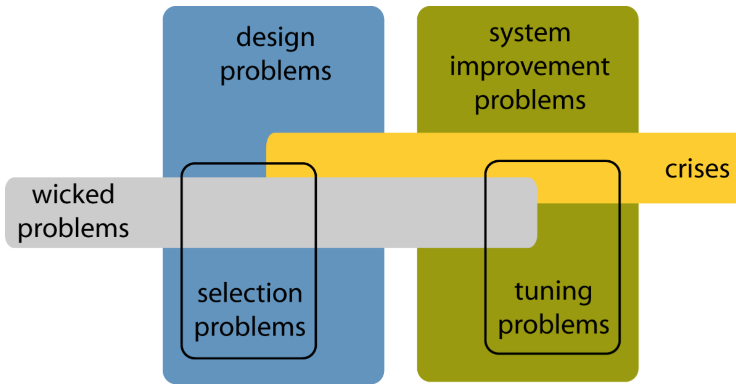


Exhibit 2-5. Six types of problems, one of which is design problems.

Design problems

The bulk of the book focuses on design problems. The hallmark of design problems is that the designer creates a plan for a *new artifact* in response to a gap. A central feature of design problem solving is the exploration of alternatives.

Selection problems

Selection problems are a subset of design problems in which the alternatives are already well articulated or relatively easy to discover. The central challenge is to select from among those clearly articulated alternatives. For example, when a firm needs to install a new accounting system, the problem solver can typically readily identify the available alternatives. These alternatives are the systems available on the market, as the firm would rarely create its own accounting system from scratch. The challenge is evaluating the alternatives and then selecting one. I include selection problems within the larger category of design problems because even with the most straightforward selection problems, the problem solver does have to at least articulate the alternatives, which is a form of exploration.

System improvement problems

Unlike design problems, system improvement problems concern modifications to existing artifacts or systems. The problem-solving process for system improvement problems typically involves the comparison of existing performance with some notion of ideal performance. Then, the problem solver focuses on exploring alternative approaches to improving performance. For example, the admissions process for business schools is a tricky undertaking requiring high levels of efficiency, fairness, and predictive accuracy. Most schools are continually attempting to improve the performance of the system. While creating an admissions process from scratch is clearly a design problem, improving an existing admissions process is qualitatively different. Some elements of difference, for example, are that improvement tends to comprise several incremental changes, often applied sequentially; the focus of problem solving is often *defect reduction*, which has a forensic quality to it; and system improvement typically benefits from a wealth of data from the existing system. None of these attributes is typical of design problems.

Tuning problems

A particular flavor of system improvement problems is *tuning problems*. Tuning problems are limited to incremental adjustments to parameters of an existing artifact. For example, consider the process for making plywood. A log is positioned on a machine (essentially a large lathe) that spins the log while a wide blade peels off a 2.5-meter-wide ribbon of wood veneer. That ribbon is subsequently cut into rectangular pieces, stacked into a sandwich with glue between the layers, and then squeezed in a heated press to cure the adhesive. Like most manufacturers, plywood makers are continually engaged in system improvement problems. One such problem is the tuning problem associated with the veneer-making process. The process parameters include, among others, rotational speed, blade shape, cutting angle, cutting pressure, and log moisture content. There are of course infinite possible combinations of these variables. The tuning problem is to find the combination that both achieves the best performance (wood utilization, consistency, surface finish, etc.) and delivers consistent results under varying conditions. A variety of methods have been developed for solving tuning problems. See particularly "optimal design" methods, which are appropriate in cases where mathematical models of the artifact exist (Papalambros and Wilde 2000) and experimental methods, which are appropriate for cases where analytical models are elusive (Ulrich and Eppinger 2011).

Crises

A crisis is simply a problem that must be solved quickly. In economic terms, the opportunity cost of time is very high for crises (e.g., a patient is bleeding, a company is failing, coal miners are trapped, public opinion is forming in the wake of an event). Crises can be design problems or system improvement problems. For example, when the crew of Apollo 13 said, "Houston, we have a problem,"[2] everyone soon knew that the problem had to be addressed quickly or the astronauts would die. The Apollo 13 crisis comprised, among others, a design problem—how to create an air filter from available materials (Exhibit 2-6)—as well as a system improvement problem—how to minimize the electrical current draw from the systems in the aircraft.

Wicked problems

Rittel and Webber (1973) defined a class of problems as *wicked*, kind of a catch-all term for problems that are extraordinarily hard to solve, and for which even clear definition is difficult. I like the term *wicked problem*, but have never felt it was defined with adequate precision. Here I use the term to refer to problems for which stakeholder objectives are fundamentally in conflict. Examples of such problems include territorial disputes in and around Jerusalem, global warming, public school reform, and terrorism. Like crises, wicked problems can be either design problems or system improvement problems.

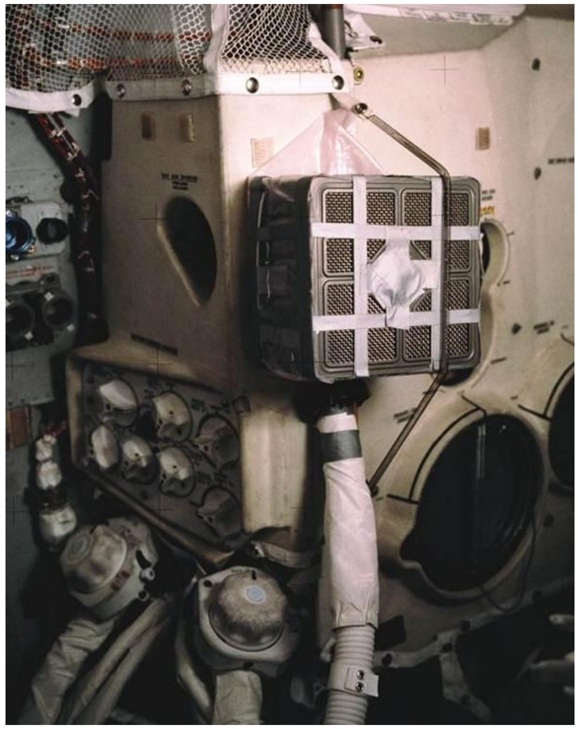


Exhibit 2-6. The air filter designed and built by the Apollo 13 crew from available materials when faced with a crisis. Source: NASA.

Deliberate Process, Importance, and Time

My view of problem solving and design is process oriented. My students often ask whether a deliberate process is always the best way to solve a problem.

I know of only two studies that have looked at the question of effectiveness of structured problem solving processes. Griffin (1997) studied the effectiveness of structured product development processes, a fairly close cousin of design processes (Terwiesch 2007). She found that firms that adopted structured development processes completed complex projects more quickly than those that did not.

Tyre and colleagues (1993) studied the effectiveness of structured problem-solving processes in addressing manufacturing problems at the Saturn division of General Motors. They found that the use of structured problem-solving processes (essentially the process articulated in Exhibit 2-4) was associated with both better solutions and faster completion.

This limited scientific evidence is consistent with my beliefs based on experience. At a minimum, structured processes act as checklists to ensure that no critical information processing task is omitted.

The question of whether a deliberate process is always warranted can be illuminated with some conceptual thinking. Exhibit 2-7 lays out two relevant dimensions. First, how important is getting the right solution? Second, how urgent is the problem? These two dimensions can be thought of in economic terms for the sake of relative quantitative comparisons. The first (horizontal) dimension can be thought of as the economic value of getting a near-optimal solution compared to the value of getting a typical solution.[3] For example, in branding a video recording and storage product, how much is the name TiVo worth compared to Replay? (My answer on that one is "millions.") Compare this to the value of figuring out exactly what to eat for breakfast one morning. (It's worth something to me, but not much.)

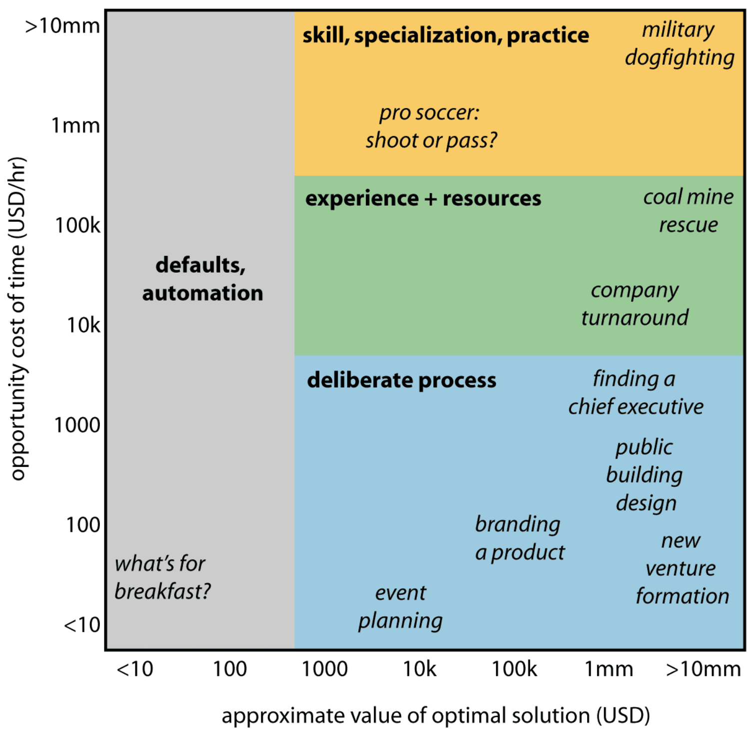


Exhibit 2-7. The use of deliberate process depends on the opportunity cost of time and the relative value of achieving the optimal outcome.

The second (vertical) dimension represents the opportunity cost of time. What is the cost of not having a solution? We can think about this in terms of cost per hour. For example, when formulating a solution for how to turn around a troubled company, one can think about the negative cash flow to be averted by that plan, which might be a million dollars per day (i.e., tens of thousands per hour).

The relative position of a problem on these two dimensions informs the question of whether to apply a deliberate problem-solving process, and if so, what type. For problems for which the value of an excellent solution is worth not much more than that of an average solution, there isn't much point to investing in problem solving, and so problem solvers may resort to just picking default solutions or to automated problem-solving techniques based on simple heuristics.

In the upper right portion of the exhibit, an excellent solution is valuable and time is extraordinarily precious. Examples of such settings would be action sports like soccer or hockey, or at the extreme, life-or-death military engagement, such as dogfighting between aircraft. In these settings, there is not time for even the most streamlined deliberate problem solving. Action must be reflexive and instantaneous. Humans cope with such situations only with extreme levels of skill, specialization, and practice, and they perform well only within a narrow range of problem types.

In the middle right portion of the exhibit, outcomes are still quite valuable, but time is a little less costly. Even in a life-or-death setting like a rescue in a collapsed coal mine, the problem solver has hours to deal with the crisis, not seconds. Some use of problem-solving process is warranted (e.g., the exploration of alternatives), but typically the problem solver will deploy a great deal of resources, perhaps pursuing several potentially redundant solutions at once. Typically, problem solvers responsible for addressing crises are deeply experienced, and so do not need to learn and adapt their problem-solving methods as they go.

Most of us spend most of our professional lives in the lower right portion of the exhibit. Problems are important, with the best solutions typically worth thousands or millions of dollars more than average solutions. And problems typically are not so urgent as to prevent us from devoting days, weeks, or months to their solution. We plan events. We brand products. We form new ventures. We design buildings. We staff organizations. For these problems, deliberate process is warranted. There is no reason not to carefully define the problem, explore alternatives, evaluate and select from the alternatives, and iteratively refine a solution.

Why the resistance to structured processes?

If a deliberate problem-solving process is warranted for a large fraction of the problems we face as professionals, why do humans so resist such processes? This resistance ranges from passive neglect to active loathing. I believe there are at least three reasons for the resistance. First, the application of structured processes is hard work, and most of us resist hard work when possible. Second, problem solvers rarely observe how well they might have done with the application of a structured process. That is, the opportunity cost of not applying a structured process is rarely obvious, and so the impetus for applying a process may not be well understood. The third reason is largely a conjecture on my part, but is interesting to think about.

I believe that for most of our evolutionary past, humans benefited from a bias for action. When faced with a decision of whether to flee or fight in the face of an enemy, those who reflected carefully on the problem and explored alternatives did not survive long enough to reproduce. This is an oversimplification, of course. One of the hallmarks of human behavior, going back tens of thousands of years, is that we use our brains to plan for the future. However, until the most recent few thousand years, this planning applied to small groups of people, perhaps over a time period of just one season. We were not prepared biologically for managerial life in a mass society. Today, we often make decisions as professionals that matter to thousands of people over time scales of many years. When faced with such decisions, deliberate processes are fully warranted, and yet our biological impulse may be to just act.

Design and Innovation

I wrote the bulk of this book at the same time I was coauthoring a book on innovation (Terwiesch and Ulrich 2009). Because of this confluence of activities, I was forced to reflect on the similarities and differences between *design* and *innovation*. Design and innovation are quite similar endeavors, but there are at least three distinctions.

First, innovation is typically broader in scope than design, and includes the entire set of activities that create a new match between a solution and a need. Design is often one of these activities, but so are market launch, the ramp-up of operations, and the management of regulatory issues.

Second, and related closely to the first distinction, innovation is often thought of as an economic activity whose basic unit of analysis is the innovation system, whereas design is an activity typically thought of at the level of a particular artifact or project. This is a tendency in practice more than a theoretical distinction. General managers in firms and other institutions worry about their innovation "pipelines" or "systems." When viewed as a system, innovation includes the more focused activity of design.

Third, design usually proceeds from the identification of a gap to the creation of a solution. Innovation can frequently proceed in the other direction, an approach sometimes called *technology push*. With technology push, an innovator begins with a new or existing solution and then searches for possible applications of that solution. This approach is typical of the pharmaceutical innovation process, in which a newly discovered chemical compound is screened for possible medically useful properties. Technology push also occurs frequently in innovation involving basic materials.

The Culture of Designers

Designers share some elements of common culture, even though diverse design domains typically possess idiosyncratic subcultures. Design culture sometimes clashes with, for example, the cultures of politicians, lawyers, and some managers. As I reflect on the unique aspects of design culture, I identify three key elements.

Optimism versus criticism

Designers are optimistic. They are accustomed to facing problems and solving them. This optimism contrasts with the culture of criticism one often finds in some other professions. For example, lawyers are trained to imagine the worst possible outcomes and protect against them. Designers are trained to imagine the best possible outcome that one might be able to create with a novel artifact. It is no surprise that these two groups of professionals often find themselves in a clash of cultures.

Prototyping and iteration

Good designers tend to have a bias for building, trying, and refining artifacts, rather than perfectly refining a theoretical plan. The design culture is one of prototyping and testing as much as it is one of conceptual exploration. This bias makes sense when faced with a high level of uncertainty and a lack of theory, as is often the case for design problems. The bias for action can be detrimental for problems in which data and analysis are powerful tools for finding solutions, as is the case for some problems in engineering and management.

Elegance

Designers tend to strive for elegance. It bothers most designers to create something sloppy even if it works. While elegance is an ill-defined concept, I think it tends to comprise originality, beauty, surprise, and an efficient use of resources. Many have tried to articulate what makes for good design. One effort I like is by Paul Graham (2004), a software entrepreneur, who argues that good design is, among other things, daring, timeless, slightly funny, and hard (but looks easy).

Nontraditional Design

This book is mostly focused on designed physical objects, although in Chapter 1 I offered a more general view of design and a more general notion of artifacts. I believe that most of the ideas in this book apply to the design of organizations, social systems, business models, and services as well as they do to the design of physical goods.

For example, consider the design of a business model. Exhibit 2-8 includes a template for essentially any business (Panel A). The template includes a customer acquisition process and a solution delivery process. An infinite number of possible business models can be created through exploration of the various alternatives for the elements of this generic model. For example, NetJets is a company that pioneered the commercialization of fractional jet ownership. Panel B in the exhibit shows the instantiation of the template with the key elements of the NetJets model. Panel C is a potential new business model that is an incremental perturbation of the existing NetJets model.

I believe that a structured process of exploration can be applied to the creation of new business models like that of NetJets as well as to the exploration of alternative models. This process is essentially similar to the way many effective designers explore alternatives for the design of physical objects and systems. Although most good designers of physical goods exhibit great discipline in exploring many alternatives, this discipline seems less well developed in the creation of businesses. In a course I teach in the MBA program at Wharton, I have tried to develop *design thinking* among business students faced with nontraditional design problems, and I believe this effort has been largely successful. An even further extension of design thinking to the creation of social systems and government policies seems quite promising.

Concluding Remarks

There is a lot of human problem solving that is not really design. However, I believe that much of human problem solving would benefit from *more* design process, not less. The hallmark of design is an exploration of alternatives and careful selection from among those alternatives, an approach that tends to make for good problem solving. I would also like to see greater diffusion of the culture of design, one of optimism, elegance, and a bias for action.

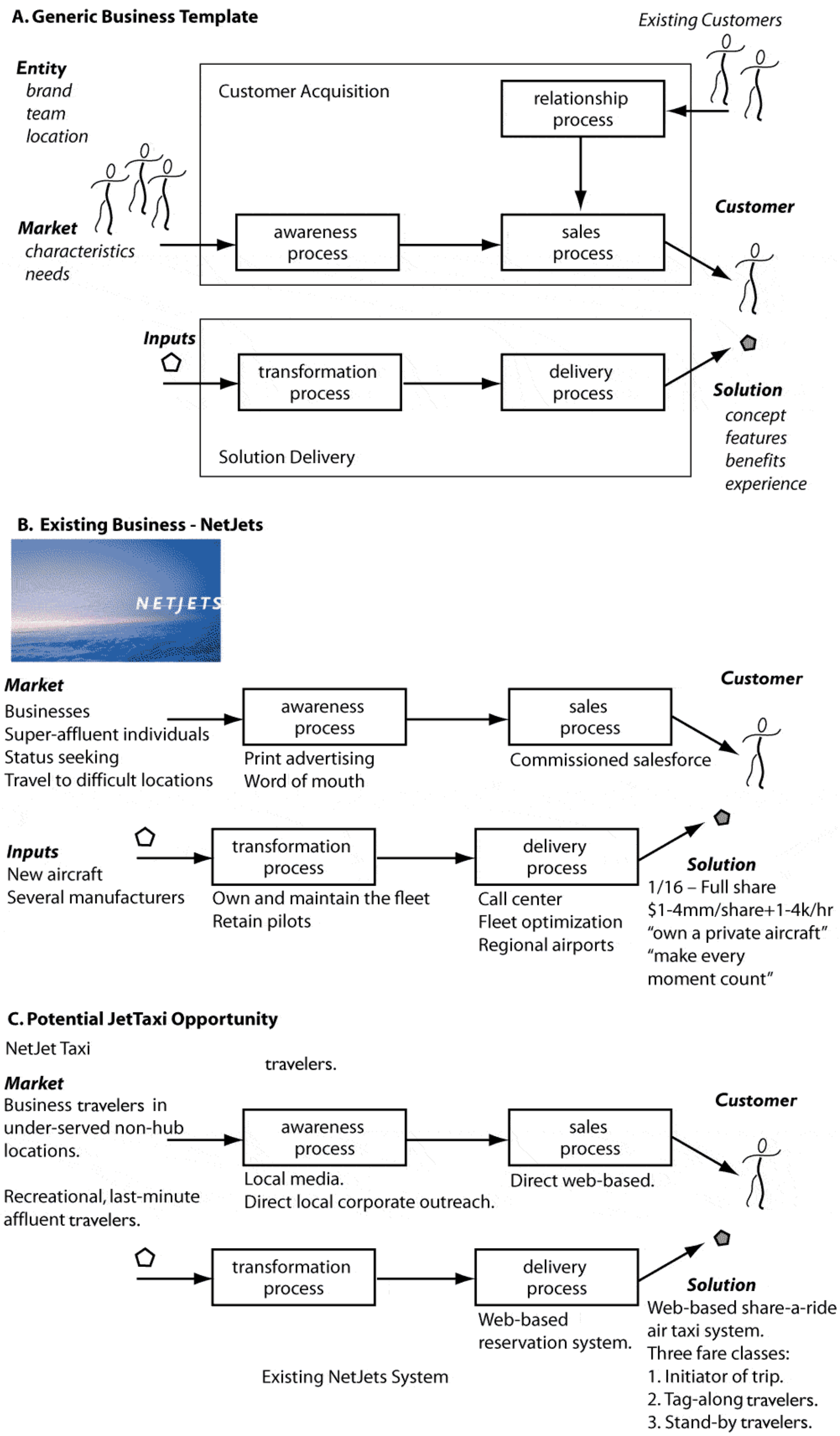


Exhibit 2-8. Design applied to business models.

Notes

[1] Karl Popper (1999) argued that "all life is problem solving" and that the basic elements of all problem solving are (1) recognizing the problem, (2) attempting alternative solutions, and (3) eliminating approaches that do not work.

[2] This quote isn't quite right. Astronaut James Lovell actually said, "Houston, we've had a problem," but the present tense sounds better.

[3] The notion of optimality is a bit loose here because most problems cannot be formalized in a way that optimality can really be defined. One way to think about the definition of the value of a near-optimal solution is to think about the probability distribution over the quality of solutions for a given problem. One might think of the horizontal axis in Exhibit 2-7 as the value of the standard deviation of this distribution.

References

Graham, Paul. 2004. "Taste for Makers." In *Hackers and Painters: Big Ideas from the Computer Age*, 130–145. Sebastopol, CA: O'Reilly Media. http://www.paulgraham.com/taste.html.

Griffin, Abbie. 1997. "The Effect of Project and Process Characteristics on Product Development Cycle Time." *Journal of Marketing Research* 34 (1): 24–35.

Papalambros, Panos Y., and Douglass J. Wilde. 2000. *Principles of Optimal Design: Modeling and Computation*. Cambridge: Cambridge University Press.

Popper, Karl. 1999. *All Life Is Problem Solving*. London: Routledge.

Rittel, H., and M. Webber. 1973. "Dilemmas in a General Theory of Planning." *Policy Sciences* 4: 155–169.

Terwiesch, Christian. 2007. "Product Development as a Problem-solving Process." In *Blackwell Handbook on Technology and Innovation Management*, ed. Scott Shane, 143–172. New York: Wiley-Blackwell.

Terwiesch, Christian, and K. T. Ulrich. 2009. *Innovation Tournaments: Creating and Selecting Exceptional Opportunities.* Boston, MA: Harvard Business Press.

Tyre, M. J., S. D. Eppinger, and Eva M. H. Csizinszky. 1993. "Systematic Versus Intuitive Problem Solving on the Shop Floor: Does It Matter?" Working Paper 3716, MIT Leaders for Manufacturing Program, July.

Ulrich, K. T., and Steven D. Eppinger. 2011. *Product Design and Development*. New York: Irwin/McGraw-Hill.

Appendix

Following is an excerpt from Benjamin Franklin's letter to George Whatley dated May 23, 1785. Whatley was a philanthropist and close friend of Franklin's.

> By Mr. Dollond's saying, that my double Spectacles can only serve particular Eyes, I doubt he has not been rightly informed of their Construction. I imagine it will be found pretty generally true, that the same Convexity of Glass, through which a Man sees clearest and best at the Distance proper for Reading, is not the best for greater

Distances. I therefore had formerly two Pair of Spectacles, which I shifted occasionally, as in travelling I sometimes read, and often wanted to regard the Prospects. Finding this Change troublesome, and not always sufficiently ready, I had the Glasses cut, and half of each kind associated in the same Circle, thus, [Franklin's sketch follows].

By this means, as I wear my Spectacles constantly, I have only to move my Eyes up or down, as I want to see distinctly far or near, the proper Glasses being always ready. This I find more particularly convenient since my being in France, the Glasses that serve me best at Table to see what I eat, not being the best to see the Faces of those on the other Side of the Table who speak to me; and when one's Ears are not well accustomed to the Sounds of a Language, a Sight of the Movements in the Features of him that speaks helps to explain; so that I understand French better by the help of my Spectacles.

THREE

Design Problem Definition

As I explained in the preceding chapters, I decompose the activity of design into four steps. This chapter focuses on *problem definition* and Chapter 4 focuses on *exploration,* the two middle steps, shown again in Exhibit 3-1.[1]

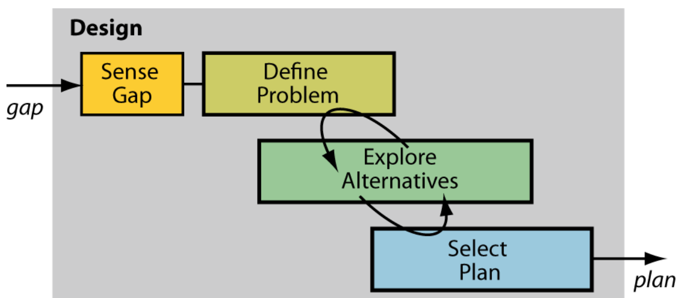


Exhibit 3-1. The design process can be thought of as four steps.

In a confusing mix of terminology, designers in different domains use different labels for problem definition. In architecture, problem definition is called *programming* (Duerk 1993). In industrial design it is *research*. In product design, problem definition is termed *identifying customer needs*. In engineering design, it is called *establishing specifications*.

Fundamentally, defining the problem is an articulation of what the designer sets out to accomplish, what gap the designer is attempting to close. The problem definition is the *what* but not the *how*. Conceptually, the definition of a design problem can be thought of as comprising two elements: (1) the basic function of the artifact and (2) the desirable qualities of an artifact that performs that function. For instance, Exhibit 3-2 is a new wood-fueled stove for use in developing regions of the world. The basic function of the artifact is *to provide heat for cooking*. The desirable qualities of the artifact include minimizing wood consumption, minimizing emission of pollutants, and providing stable support for cooking vessels. In this chapter I treat these two elements of product definition in turn, and then address the iterative nature of most problem definitions.

The Function of the Artifact

The function of an artifact is what it does as opposed to its structural characteristics (Crilly 2010). A theoretical ideal in design—one that avoids predetermining the solution—is to describe function in a way that does not imply a particular approach. Yet, describing function necessarily requires some such assumptions.

Problem Definition

Basic Function:
Provide heat for cooking

Desirable Qualities:
Minimizes wood consumption
Minimizes emissions of pollutants
Provides stable support for cooking vessels
Etc.

An artifact developed in response to a problem.

Exhibit 3-2. The wood-fueled Biolite cooking stove for developing regions of the world. The stove includes a fan for enhancing the flow of combustion air, which is powered by a thermoelectric device. Source: Biolite.

Problem hierarchies as a way to describe function

Theodore Levitt famously wrote that "people buy ¼-inch drill bits but need ¼-inch holes" (Levitt 1977). Of course, people don't really need ¼-inch holes either, but rather, they need, for instance, to fasten a book shelf to the wall. Indeed, any statement of a design problem, of the basic function of an artifact, reflects a decision about the level of abstraction at which the problem will be tackled, and assumptions about how the function will be addressed.

Exhibit 3-3 illustrates some possible levels of abstraction for a restatement of the Levitt problem "drive a 6 mm drill bit at 1000 rpm." The function can be stated more broadly by asking *why* as follows.

Q: Why do you want to drive a drill bit?

A: Because I want to make a hole.

Q: Why do you want to make a hole?

A: Because I want to fasten a book shelf to the wall.

and so forth…

The answers to a sequence of "why" questions form an abstraction hierarchy of design problem definitions. One such hierarchy is shown in Exhibit 3-3.

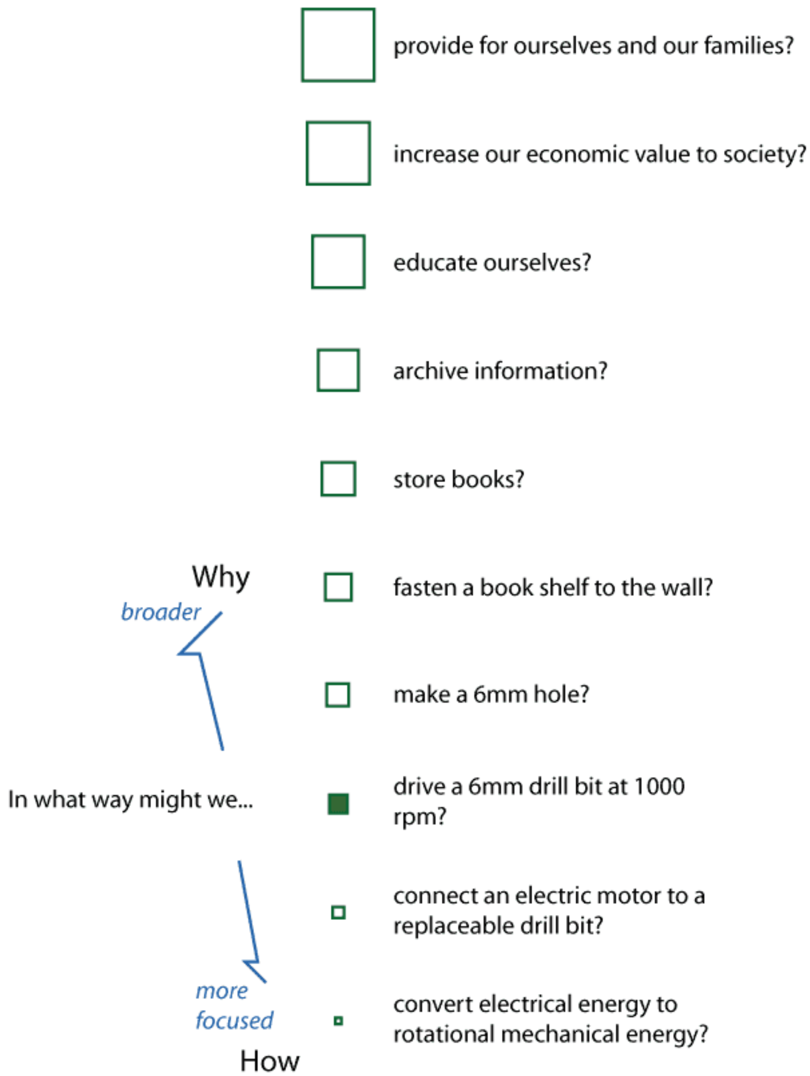


Exhibit 3-3. Problem hierarchy for the initial problem definition "In what way might we drive a 6 mm drill bit at 1000 rpm?"

The simple hierarchy of Exhibit 3-3 belies a more complex relationship among problems. There are typically multiple motives for solving a particular problem and there are multiple approaches that can be taken to solve it. Exhibit 3-4 illustrates a portion of a network of problem statements emanating from one of the more abstract problem statements in Exhibit 3-3— "In what way might we educate ourselves?" Articulating a design problem unavoidably reflects the designer's decision about the level of abstraction at which the design effort will focus.

Is there a correct level of abstraction? On the one hand, making design problems more abstract is good because doing so opens up additional avenues for exploration, broadening the range of solutions the designer considers. On the other hand, the more abstract the problem becomes, the less likely the designer will be able to significantly close the gap in the user experience. A power tool designer had best not abstract the design problem so much that he or she is trying to close gaps in higher education, a challenge for which the designer is likely ill suited. In the Toyota Production System, one of the techniques for problem solving is called the "five whys method," which is intended to get to the root cause of a problem (Ohno 1988). That method, of asking "why" five times, appears to be a good heuristic in design as well, forcing the designer to consider alternative definitions of the design problem, which may be somewhat more abstract than initially assumed.

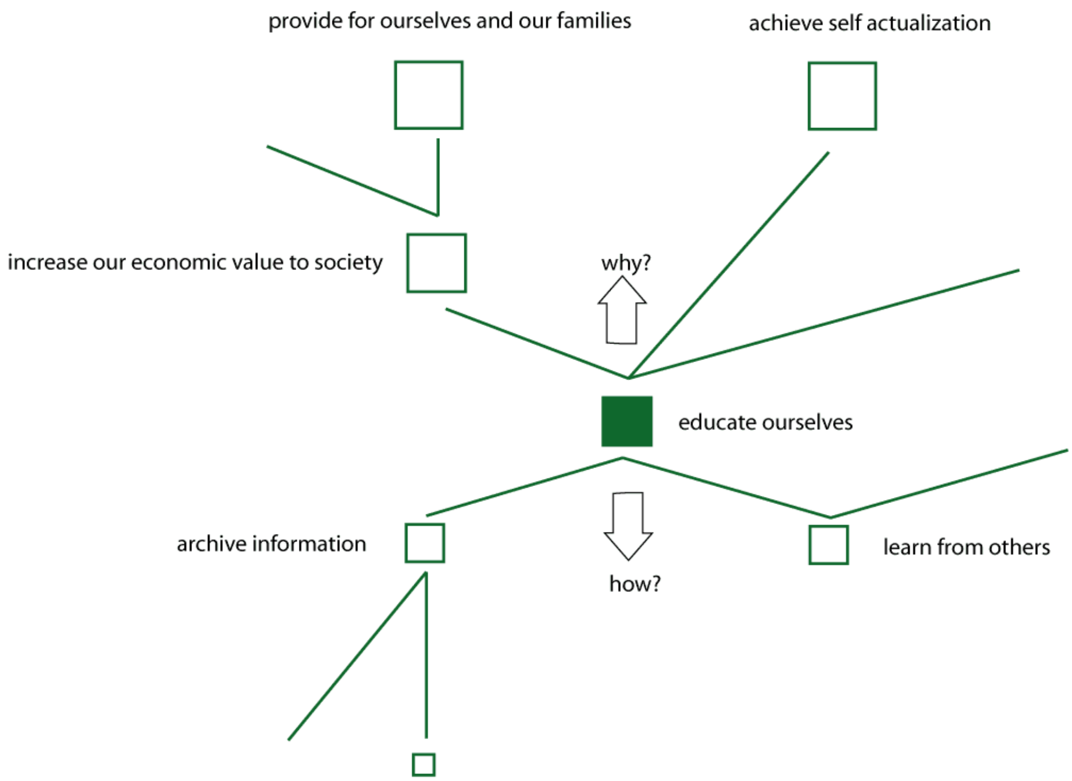


Exhibit 3-4. Problem network based on the question "In what way might we educate ourselves?"

Formal descriptions of function

In most cases, the basic function of artifacts is described using unstructured text, as in "drive a 6 mm bit at 1000 rpm." However in some domains, problems can be described more formally. For instance, in architecture, problem definition, or *programming*, often includes an adjacency diagram as shown in Exhibit 3-5. The most basic function of the facility for the Camden Community Center might be stated as "provide a physical space for programs that enhance the health and well-being of the San Jose community." An elaboration of that basic function specifies the types and sizes of spaces required and the desired relationships among those spaces. It does not contain any description of the form, position, orientation, or materials that would comprise the eventual building. In this sense it is still the *what* and not the *how*.

There have been several attempts in the design theory community to create formal languages for describing function (Finger and Dixon 1989), and there have been modest successes in narrow domains of application, such as electro- and fluid-mechanical systems and digital circuits (Mead and Conway 1980). There have also been efforts to create informal functional languages to facilitate the practice of design (Fowler 1990; Hubka and Eder 1988). These languages are sometimes used to create diagrams consisting of functional elements, expressed as linguistic terms like convert energy, connected by links indicating the exchange of signals, materials, forces, and energy. Some authors of informal functional languages provide a vocabulary of standard functional elements, while others rely on users to devise their own. Functional elements are sometimes called functional requirements or functives, and these diagrams have been variously called function diagrams, functional descriptions, and schematic descriptions (Pahl and Beitz 1984).

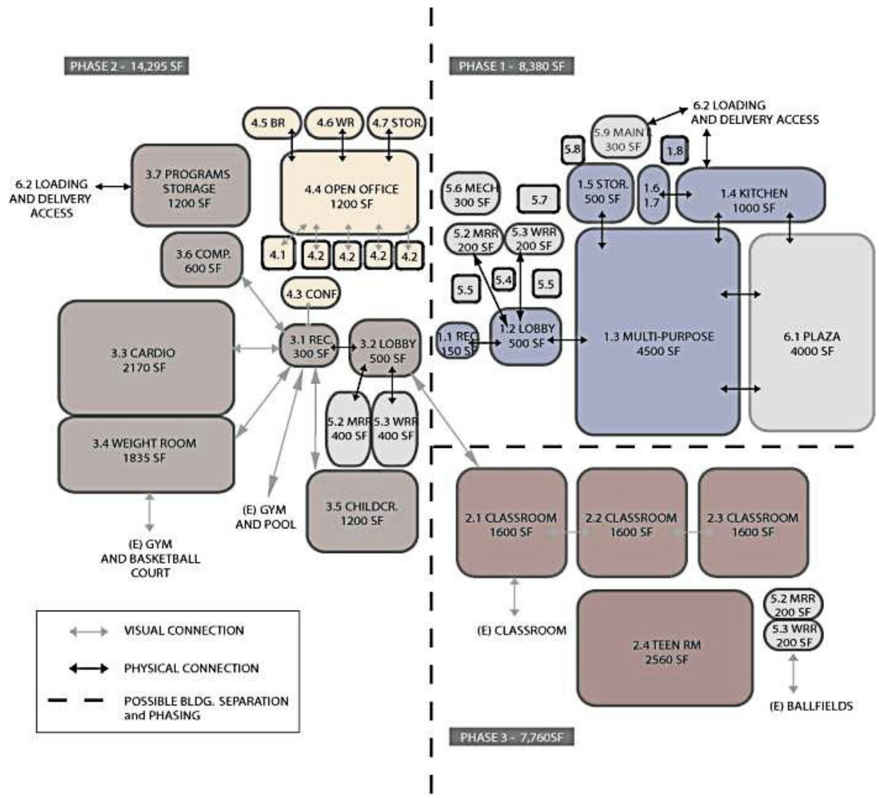


Exhibit 3-5. Bubble, or adjacency, diagram for Camden Community Center (San Jose, California). Source: Sabrina Phillips.

Desirable Qualities in the Artifact

The basic function of an artifact is rarely all the user cares about. For instance, even if a paperweight does its job of preventing paper from blowing about, the user probably also cares how it looks. In this case, while aesthetics is not a basic function of the paperweight, it is a quality of that artifact that must concern its designer.

Needs

In the field of product design, the desired qualities in an artifact are called *needs*, a term I'll adopt here. User needs are usually represented as a list of thirty to one hundred desired qualities of an artifact. That list is in essence a causal model of the relationship between artifact characteristics and user satisfaction. The list is an understanding by the designer of what the user cares about and what characteristics drive preference and satisfaction. If the artifact possesses those qualities, the user will be satisfied. Exhibit 3-6 is a list of 66 needs for a hand cart, derived from one-on-one interaction with potential users using the methodology of Ulrich and Eppinger (2011). In practice, needs are derived from both verbal interaction with potential users and from passive observational studies of potential users grappling with the basic problem the designer is trying to solve.

While needs are desired qualities of a solution, and ideally do not embody a preconceived design concept, they clearly reflect some assumptions about the direction of the solution. The qualities listed in Exhibit 3-6 suggest a human-powered, wheeled device. Many of the qualities in Exhibit 3-6 would be irrelevant for a different assumed design direction; for instance, a valet service that transported belongings on behalf of the user. Similarly, additional needs would almost certainly be important if the designer pursued a remote-controlled, electric-powered cart.

Stakeholders

When user-innovators create artifacts for themselves, the activity of defining the problem and articulating needs may never be formally conducted—needs remain implicit for the designer. When professional designers create artifacts for others, some process of understanding and documenting needs is almost always adopted in practice. When potential users are essentially aligned in their interests, they may be thought of as a *market segment* or *user community* and treated somewhat alike. However, in some cases, an artifact is intended to address the needs of a collection of stakeholders whose needs are not aligned. The same prison must serve both inmates and guards. The same school lunch must be tasty (for kids), healthy (for parents), and inexpensive (for school districts). Problem definition therefore benefits from a deliberate identification of stakeholder groups with interests in the resulting artifact, and from an articulation of the distinct needs of those stakeholder groups.

Do users really *need* most artifacts?

In the context of this chapter, *needs* are desired qualities of an artifact, rather than the attributes of the artifact that are needed in some fundamental sense. For instance, Exhibit 3-7 is the Hardee's Monster Big Burger, which features two 1/3-pound patties of beef, three slices of cheese, four strips of bacon, and mayonnaise on a buttered roll. No one *needs* this artifact, yet it clearly delivers satisfaction to individuals in the market segment targeted by Hardee's. (Incidentally, I am not arguing against the creation of the Hardee's Big Burger; I'm only observing that it is clearly not a healthy solution to meeting the basic need for everyday calories.)

Designers adopt a moral stance when creating artifacts. The dominant perspective in the practice of product design is of the designer as agent for a for-profit enterprise, and "needs" are those qualities that will lead to the greatest satisfaction for target consumers. Other perspectives are common and valid, including the perspective that designers have an ethical responsibility to create artifacts that are "good" for users. Of course, in doing so, designers are forced to make tricky judgments. Should Hardee's really not offer users the possibility of an occasional fatty, salty, and yummy (to some) experience? If not the Monster Big Burger, what about buttered popcorn, or premium ice cream?

Meeting the needs of mindful adults possessing full information can probably be reasonably justified by some designers, even if addressing those needs may detract from health or other desirable objectives. Designers can probably not reasonably justify addressing needs in ways that are manipulative or deceptive.

The Cart handles most terrain
- The Cart handles rough urban terrain
- The Cart cargo is retained over rough terrain
- The Cart works on grass
- The Cart works on sand
- The Cart works on off-road trails
- The Cart load remains stable when nudged or bumped
- The Cart can traverse steps
- The Cart can traverse curbs
- The Cart remains stable over cross-sloped terrain
- The Cart works in snow
- The Cart works in icy conditions

The Cart goes where I go
- The Cart works with all my travel modes
- The Cart can be used inside a grocery store
- The Cart works with my bike
- The Cart can be locked up on the street
- The Cart is allowed in fancy office buildings
- The Cart can be taken on Amtrak
- The Cart can be checked as luggage

The Cart navigates tight spots
- The Cart can be used in crowded urban spaces
- The Cart can be used on a crowded subway
- The Cart fits through narrow gaps—e.g., doorways, between file cabinets

The Cart makes transporting stuff a lot easier than carrying it
- The Cart requires minimal user effort
- The Cart carries all my stuff in one trip
- The Cart can be loaded quickly
- The Cart can be unloaded quickly
- The Cart (and stuff) can be easily loaded in the car
- The Cart (and stuff) can be easily unloaded from the car
- The Cart can be deployed in seconds
- The Cart can be deployed without instruction
- The Cart transports heavy stuff like file boxes
- The Cart can be conveniently lifted and moved when loaded

The Cart fits unobtrusively into my life
- The Cart consumes little of my living space when stored
- The Cart is affordable
- The Cart works well with my gear storage solution

The Cart is my single stuff-hauling solution
- The Cart handles stuff of different sizes and shapes
- The Cart transports a cooler
- The Cart can transport a longish object like a collapsible chair
- The Cart can be used as a baby jogger
- The Cart holds the gear for a family of four at the beach
- The Cart holds a 5-gallon water jug without spilling

The Cart evokes admiration from onlookers
- The Cart is not geeky
- The Cart is a rugged piece of gear, not a cheap gadget
- The Cart is practically invisible when not loaded with stuff
- The Cart is distinctive yet cool

The Cart is a mobile base of operations
- The Cart provides a temporary "table top" when outdoors
- The Cart accommodates little, easily lost items like pocketknives and flashlights
- The Cart identifies home base at the beach
- The Cart provides a temporary seat
- The Cart allows convenient access to all my stuff when loaded
- The Cart rests in a stable position

The Cart protects my stuff
- The Cart doesn't collect water
- The Cart protects my groceries from damage
- The Cart keeps my stuff dry in the rain
- The Cart keeps critters from my stuff when camping
- The Cart protects my stuff from dirt and mud on the ground

The Cart enhances rather than detracts from my safety on the streets

The Cart can be uniquely identified as mine

Exhibit 3-6. A list of needs for a personal hand cart. The needs in boldface are the primary needs, generalizations of the more detailed secondary needs.

Exhibit 3-7. The Hardee's Monster Big Burger. What need does this artifact address? Source: Hardee's.

Specifying and Quantifying Design Problems

Measurement is part of the religion of modern management. As annoying as attempts to measure everything can be, measurement has unambiguously led to dramatic and remarkable performance improvement in many human endeavors, including manufacturing, athletics, science, and medicine.

In design, quantification and measurement allow precision in defining problems. For instance, while a desirable quality of an automobile might be economical operation, that need can be made precise by specifying that the fuel economy exceed 20 kilometers per liter of fuel. In product design, these quantified needs are often called *specifications* (Ulrich and Eppinger 2011). In architecture, they are often called *performance requirements* (Duerk 1993).

Certainly attempts have been made to quantify pretty much every kind of need, including aesthetic responses and brand loyalty. While some of these attempts probably yield little benefit, quantification of needs in the form of metrics and values does allow for precise comparisons among artifacts and the use of the scientific method to refine and improve satisfaction.

In the academic field of marketing, the *ideal point model* leverages the notion of metrics and values. It holds that every user prefers a particular combination of performance values and that this combination is their ideal point. For instance, a mobile computing device display can be characterized by its area and resolution. A particular user will most likely prefer one combination of area and resolution. Market research techniques such as conjoint analysis can be used to estimate the ideal point for a target customer, and then these estimates can be used to characterize market segments and to decide on the target values for the specifications of the device.

Issues of translation dog the notion of an ideal point. Exhibit 3-8 illustrates this challenge conceptually. The user has an actual ideal point. However, the user expresses that ideal point with some error; he or she does not really know its precise location. Further error is introduced by the designer in attempting to understand the user's expression of needs. A final error is introduced in what the artifact actually delivers to the user. The resulting gap may be quite large for new categories of artifacts with which users and designers have little experience. In mature categories (e.g., cameras, computer monitors, automobiles), users and designers are much more likely to estimate ideal points accurately.

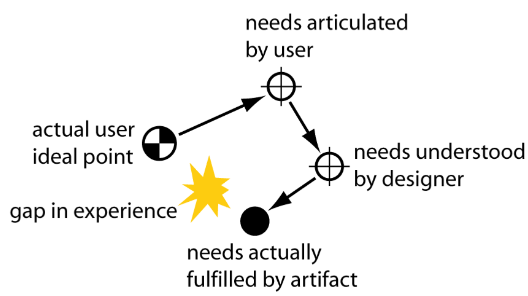


Exhibit 3-8. A user may have an ideal point, but errors in articulation, understanding, and translation may result in a persistent gap in the user experience.

Iterative Refinement and the Spiral Model

Recognizing the challenge of closing the gap in the user experience on the first try, Boehm (1984, 1986) articulated the Spiral Model of development. Although the Spiral Model was aimed at software development, the idea is valid in all design domains. Exhibit 3-9 illustrates the model. The designer engages in successive iterations of defining, exploring, building, and testing. During each iteration, a prototype is used to assess the extent to which the design problem has been properly defined and whether the designed artifact closes the gap in the user experience. The spiral aspect of the diagram arises from plotting cumulative investment in the design effort as the distance of the path from the origin as the designer cycles through the four steps. The Spiral Model was established in contrast to the Waterfall Model, in which design proceeds without iteration from one step to the next.

Boehm's Spiral Model is essentially a version of the Shewhart Cycle of *plan*, *do*, *check*, *act* (i.e., PDCA) which is a mainstay of the practice of quality improvement (Shewhart 1939). The Shewhart Cycle, in turn, is essentially derived from the scientific method, whose roots extend at least as far as Francis Bacon (*Novum Organum*, 1620).

Central to the Spiral Model is the use of prototypes. Prototypes are approximations of the intended artifact on one or more dimensions of interest. In this context they serve as instruments for measuring the extent to which the designer has understood the user's true needs and the extent to which the design actually addresses those needs.

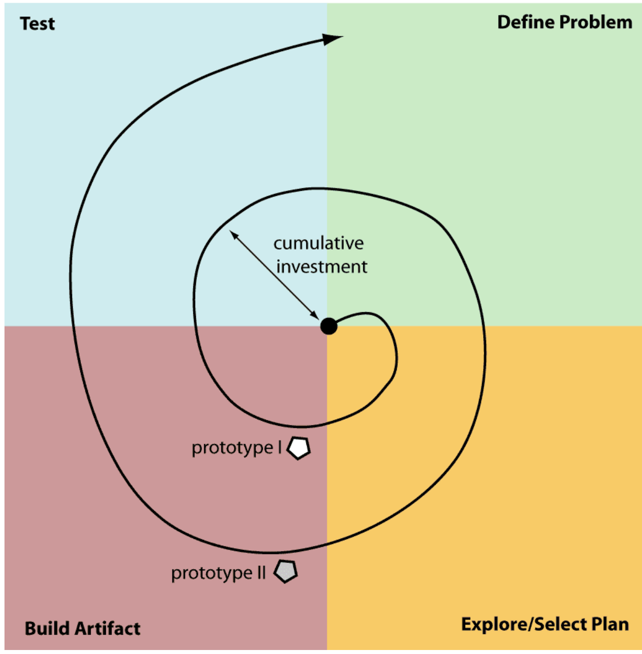


Exhibit 3-9. Spiral Model, adapted from Boehm (1986).

Concluding Remarks

A common defect in design is a failure to understand the gap the user is experiencing. By deliberately defining the design problem, this defect can be avoided. An additional defect is a failure to pose the design challenge broadly enough to allow the exploration and discovery of a wide range of potential solutions. The use of the five-whys method is one approach to balancing the benefits of a more abstract problem definition and the benefits of posing a tractable problem the designer is capable of addressing in a meaningful way.

Notes

[1] I do not address the problem of sensing gaps or of selecting plans, although I may do so in a future edition. Sensing gaps in a commercial setting is often called opportunity identification, and my book Innovation Tournaments (Terwiesch and Ulrich 2009) contains two chapters on that topic. Selecting a plan is a fairly straightforward activity once good alternatives have been identified. My book Product Design and Development (Ulrich and Eppinger 2011) contains a chapter called "Concept Selection," which describes some tools for effective selection.

References

Boehm Barry. 1986. "A Spiral Model of Software Development and Enhancement." ACM SIGSOFT *Software Engineering Notes*, 11(4):14–24.

Boehm, Barry W., Terence E. Gray, and Thomas Seewaldt. 1984. "Prototyping vs. Specifying: A Multi-Project Experiment." In *Proceedings of the 7th International Conference on Software Engineering*, Orlando, Florida, 473-484.

Crilly, Nathan. 2010. "The Roles That Artefacts Play: Technical, Social and Aesthetic Functions." *Design Studies* 31 (4): 311–344.

Duerk, Donna P. 1993. *Architectural Programming: Information Management for Design*. New York: John Wiley & Sons.

Finger, Susan, and John R. Dixon. 1989. "A Review of Research in Mechanical Engineering Design." *Research in Engineering Design* 1: 121–137.

Fowler, T. C. 1990. *Value Analysis in Design*. New York: Van Nostrand Reinhold.

Hubka, Vladimir, and W. Ernst Eder. 1988. *Theory of Technical Systems*. New York: Springer-Verlag.

Levitt, Theodore. 1977. "Marketing When Things Change." *Harvard Business Review*, November–December: 107–113.

Mead, Carver, and Lynn Conway. 1980. *Introduction to VLSI Systems*. Boston, MA: Addison-Wesley.

Ohno, Taiichi. 1988. *Toyota Production System: Beyond Large-Scale Production*. Portland, OR: Productivity Press.

Pahl, G., and Wolfgang Beitz. 1984. *Engineering Design*, trans. Ken Wallace, Lucienne Blessing, and Frank Bauert; ed. Ken Wallace. London: The Design Council.

Shewhart, Walter Andrew. 1939. *Statistical Method from the Viewpoint of Quality Control*. New York: Dover.

Terwiesch, Christian, and K.T. Ulrich. 2009. *Innovation Tournaments: Creating and Selecting Exceptional Opportunities.* Boston, MA: Harvard Business Press.

Ulrich, K. T., and Steven D. Eppinger. 2011. *Product Design and Development*. New York: Irwin/McGraw-Hill.

FOUR

Exploration

Exhibit 4-1 shows a few of Frank Lloyd Wright's sketches for a collection of cabins contemplated for a development at Lake Tahoe. Wright's sketches reveal a process of exploration of design alternatives, which is a hallmark of the activity of design. This chapter describes the essential elements of the exploration process. After explaining why exploration is necessary, I describe the concepts of *representation* and *abstraction*. I then discuss evaluation of design quality normally required to guide exploration. Next, I articulate the exploration strategies used most frequently to reduce the cognitive complexity of design problems. Finally, I connect these concepts to practice by touching on several examples.

Design Requires Exploration

Exploration inevitably involves consuming resources to develop and evaluate alternatives that will eventually be abandoned. We would of course prefer to avoid this wasted effort and just pick the right answer directly. Why do we need to explore, as opposed to determining the right answer analytically or with some other technique?

To illuminate the need for exploration, consider a counterexample, a design problem that can be solved without exploration: Design a *beam*—a structural element spanning some distance—to support an antenna on a space station. The antenna will be mounted in the center of the span and will apply inertial loads of up to 100 N perpendicular to the axis of the beam. The beam must not deflect more than 2 mm under that load in order to maintain signal quality and to limit vibration. The beam must span a 2 m wide opening and can be attached rigidly to both sides of the structure. The beam must be lightweight, but as inexpensive as possible. Assume that, like most other elements of the space station, the beam will be made of aluminum.

Because the beam design problem is simple, well defined, and highly constrained, and because it is informed by two centuries of development in the field of engineering science, it can be solved without exploration. The solution is a round aluminum tube, 2 m long and 43.5 mm in diameter, with a 1 mm wall thickness.[1]

However, let us make the problem slightly more realistic. Why constrain the solution to be aluminum? Why not allow titanium, or fiberglass, or steel, or carbon-fiber reinforced plastic? Why does the tube have to be a constant cross-section? Couldn't the wall be thinner in the middle? Given the attachment conditions at the ends of the gap, would a truss structure perhaps be more efficient than a tube? Could some additional structure be added to the antenna itself so that it could span 2 m instead of requiring a separate beam? With these simple questions, we have posed directions for exploration that would require a minimum of 32 different analyses, each with substantially different assumptions. Given a few moments of thought and a handful of questions, the design problem we could solve without exploration has been exploded into something that will require substantial exploration by even the most gifted designer. And even with complications, this is a relatively simple design problem.

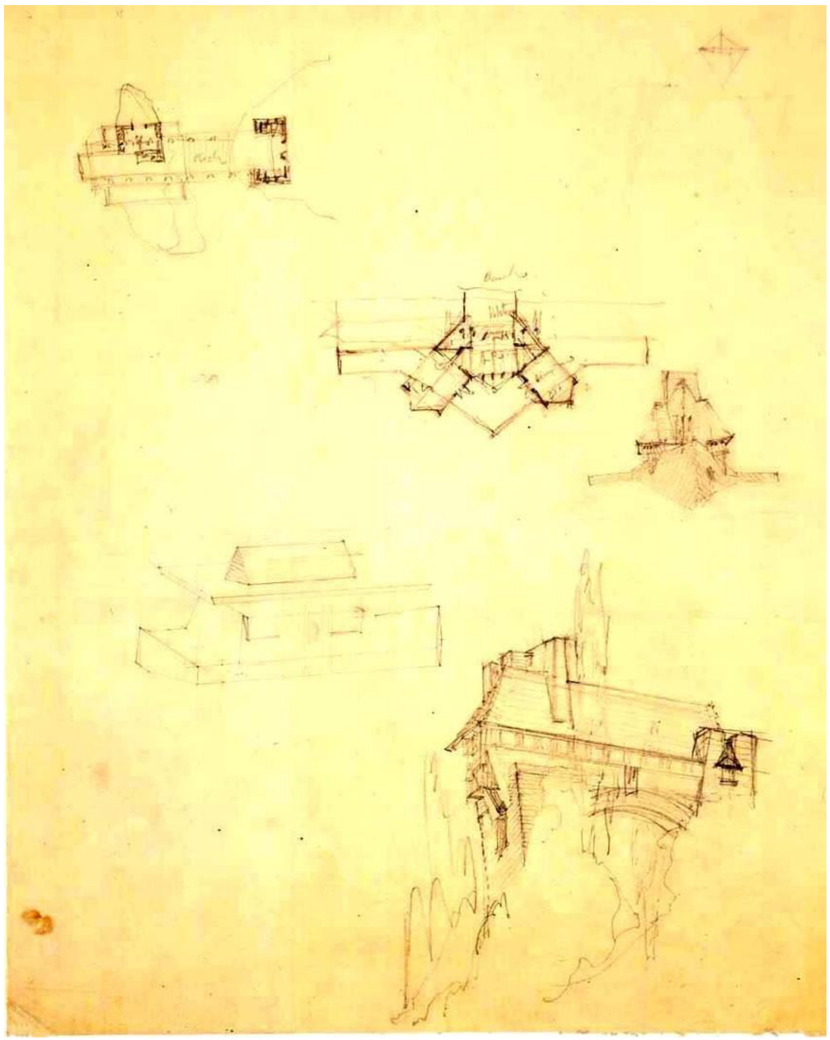


Exhibit 4-1. A page of sketches by Frank Lloyd Wright that explore the design space for a cabin at Lake Tahoe. Note the use of plan and perspective views to represent design concepts and the varying levels of abstraction employed in different sketches. Source: U.S. Library of Congress.

Practical design problems can rarely if ever be solved effectively without exploration. The problems simply cannot be fully formalized, there are too many discrete alternatives to consider, and the mathematical complexity would be overwhelming even if the problems could be formalized. And yet humans manage to design artifacts. The design strategy employed essentially universally by skilled designers is to explore a space of possibilities using a collection of heuristics that reduce the complexity of the task and to rely on knowledge to direct the exploration.

Representation and Abstraction

A representation is a language for describing designs using symbols. Consider Exhibit 4-2. The sketch on the right uses 13 line segments to denote the important edges of the geometry of a shed. Humans in general, and most designers specifically, are quite adept at interpreting such sketches as a representation of a geometric form.[2]

Most design involves a symbolic representation of artifacts. The alternative would be to explore directly in the physical world with the actual construction materials of artifacts. For example, to design a shed without the use of a symbolic representation, one would just start building. When the design proved unsatisfactory, the designer would either abandon the partially completed shed and start a new one or tear down portions of the shed and replace them with an alternative. The direct approach is quite rare in most domains because the cost in time and materials of manipulating the world directly is quite high. The designer can move much more quickly and with much less expense with pencil and paper, or with cardboard, glue, and a razor knife. (Curiously, design without representation may actually be the best approach in a few rare instances. For example, the details of dry-laid stone walls are largely designed by direct manipulation of the stones. This is because representing the detailed geometry of each stone would be more tedious than just looking at the array of stones on the ground and trying a few alternatives.)

Representation requires abstraction. An abstraction is a limited description of an artifact. A real shed can be described with essentially infinite detail. Imagine, for example, describing the precise geometry of the surface of each shingle on the shed. With detail comes complexity, which increases the cognitive burden of design. To manage cognitive complexity, designers employ representations that are abstract, encoding only the essential information about a possible artifact. Suppressing the details of materials, finishes, colors, trim, decoration, and adjacent plantings makes the shed design problem more straightforward. Good abstractions suppress details that have little relevance for the central design decisions at hand.

In addition to reducing the complexity of the design space by focusing attention on the key design decisions, representations are used to record design alternatives in *external memory*. Humans do not have the cognitive ability to store and recall the dozens or hundreds of alternatives typically explored during the design process. In contrast, paper, digital files, and physical models are quite effective storage devices for that task.

Representation and abstraction are important for exploration in nonphysical domains as well. Services and computer programs are often designed using flowcharts. Advertisements are designed using storyboards. Songs are designed with musical notation.

Most representations used in design exploration are informal, meaning that neither the syntax nor the semantics are defined precisely. Representations used to communicate a design for the purposes of fabrication are typically more formal (e.g., digital solid models or architectural drawings), but even these representations are not typically formal in a mathematical sense. To aid in illustrating some of the key ideas in this chapter, I employ a relatively formal representation and abstraction I call *shed world*.

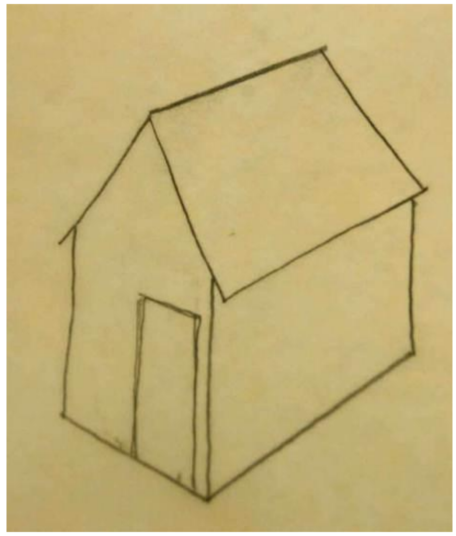


Exhibit 4-2. A real shed and a shed abstraction.

Shed world

Imagine a design problem posed by the need for a garden shed. The basic function of the shed is to shelter outdoor belongings. The user needs to store two trash cans, three bicycles, a wheelbarrow, and a stack of lawn chairs. The shed needs to fit on the edge of a terrace and harmonize aesthetically with the house. It needs to protect the contents from the weather in northern latitudes.

Shed world, shown in Exhibit 4-3, is a representation for describing sheds. Shed world is a particularly simple formal representation in which a shed is described by a quintuple: wall height, floor plan aspect ratio, roof type, roof orientation, and roof pitch. These five elements fully describe a shed in this formalism. A *tuple* is a simple form of a *design grammar*, a set of rules for constructing "legal" designs in a design space. A more complex grammar might allow for floor plans that are Ts or Ls or Hs or might allow for roofs that have nonconstant pitches. However, for our purposes, the simple shed grammar consisting of a quintuple is sufficiently complex.

To get a sense of the complexity of this design space, assume that designs are allowed to assume only the discrete options shown in Exhibit 4-3. These discrete choices result in 640 distinct sheds (4 heights x 5 aspect ratios x 4 roof types x 2 roof orientations x 4 roof pitches). Five of these designs are shown in the exhibit. Of course, adding other attributes like door location, window placement, or siding type increases the size of the design space geometrically, and if we allowed the attributes to take on continuous values (e.g., arbitrary aspect ratios instead of discrete choices), then there would be infinite possibilities.

Even in the highly stylized shed world, the complexity is daunting. It would be tedious to consider every alternative. And with more complex design problems, doing so is more than tedious, it is impossible.

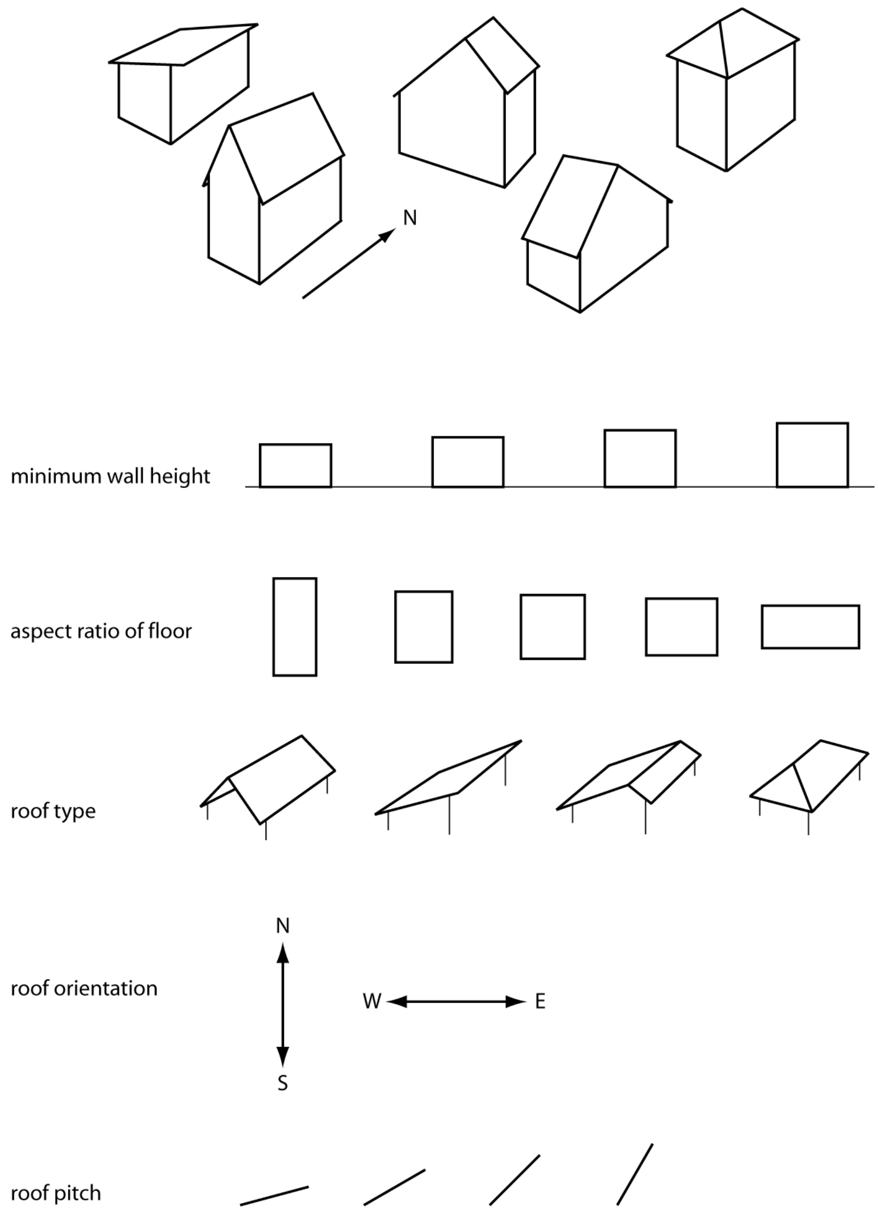


Exhibit 4-3. A formal representation of the design space for a shed.

Evaluation

Design requires exploration, and so the process of design must include evaluation of the alternatives considered; otherwise the selection of a design would be arbitrary. Of course, the designer could literally build and test every artifact contemplated. However, evaluation is much more efficient when based on an abstract representation of those artifacts.

In almost all cases of design by humans, the first evaluation of a design is a cognitive response of the designer to a sketch or other representation of the design. These evaluations are holistic judgments based on highly abstract descriptions. These judgments are efficient, but because they are formed rapidly based on limited information, they are plagued by uncertainty.

Subsequent evaluations of more refined designs may be more analytical, and may be based on a decomposition of the overall quality of an artifact into several dimensions or attributes. For the garden shed, the quality attributes might include space efficiency, aesthetics, cost, and ease of access to the contents. For a fruit salad, the attributes might include appearance, flavor, texture, and shelf life. For an airplane, the attributes might include fuel efficiency, payload, cruising speed, and minimum runway length. A rich history of academic research and industrial practice has shown that useful predictions of user preference can be made by first evaluating alternatives with respect to individual attributes and then aggregating those evaluations into a single overall measure of utility or preference (Keeney and Raiffa 1976).

As the designer narrows the possibilities under consideration to just a few, prototypes may be built and tested to refine and validate earlier judgments.[3]

Exploration Strategies

Armed with a representation and a way to evaluate design alternatives, the designer still faces daunting complexity in the exploration task. In this section, I outline four strategies commonly used to manage the complexity of exploration: hierarchical decisions, parallel exploration and selection, causal relationships, and existing artifacts.[4]

Hierarchical decisions

In shed world, the array of possibilities can be reduced substantially simply by fixing one of the design variables. For example, one might decide that the shed will have a rectangular floor plan with the long side facing the terrace. Assuming the designer is working from the discrete alternatives for the variables illustrated in Exhibit 4-3, this single decision reduces the number of alternatives from 640 to 128, a factor of five.

Of course, an arbitrary fixing of a design variable introduces the risk of having excluded an excellent design from consideration. But these decisions need not be arbitrary. Ideally, the designer makes a decision that substantially reduces the complexity of the problem and that can be made with high reliability without committing to decisions for the remaining design variables.

Subsequent design decisions can then proceed sequentially. Given the rectangular aspect ratio, the designer may decide that the ridge of the roof will be oriented the long way on the building. Having specified a roof orientation and aspect ratio, the designer may then decide that the roof will be a conventional gable-end peaked roof. Given those choices, the designer may decide that the roof pitch will be 45 degrees (or "twelve twelve" in roofing terminology, referring to a vertical rise of 12 inches over a horizontal run of 12 inches). Finally, the designer may commit to a 2-meter wall height. This process of sequential decision making and the resulting path through the design space is illustrated in Exhibit 4-4.

By considering design decisions hierarchically, exploration becomes a process of choosing which fork in the road to take as each decision is encountered. Typically, a sequential decision strategy is a heuristic approach—it is a rule of thumb that does not guarantee that the best alternative is found on the first pass. One cannot typically know that there is not a better design down some path that was not taken. As a result, most designers will explore several paths, may backtrack, and may explore several different sequences of decisions. Nevertheless, a collection of promising designs can usually be generated relatively efficiently by considering decisions hierarchically.

Parallel exploration and selection

A sequence of design decisions forms a trajectory of exploration in the design space. In Exhibit 4-4 one such trajectory is shown for the decisions explicit in shed world. However, to finalize the design of the shed, we would need to locate a door and possibly a window or two. We would need to choose materials and finishes. We would need to specify trim details and the characteristics of the foundation. For most design problems, many such detailed design considerations consume a great deal of effort. These details can rarely transform a poor initial concept into a high-quality artifact. No amount of cedar siding and polished brass hardware will transform a bad floor plan with an ugly roof into a nice shed.

We can exploit differences in relative importance and cost of design decisions in the exploration process. By arranging design decisions in order of decreasing importance and in order of increasing effort, the designer can focus on the high-impact, low-cost decisions first, and defer the high-cost, low-impact decisions for later. We then can divide the design decisions into a selection phase and a development phase (Sommer and Loch 2004).

In the selection phase, several trajectories are pursued in parallel, but only as far as necessary to make an assessment of the likely quality of an artifact that would result from pursuing the trajectory fully. The designer in effect walks down a path only far enough to get a sense of how the landscape looks in that direction. By exploring several alternative paths in a preliminary way, the designer avoids wasting resources refining a design concept that will ultimately prove unsatisfactory.

The multiple trajectories of parallel exploration can be pursued by several independent designers as part of a design team or possibly even in a tournament format among competing designers (Terwiesch and Ulrich 2009). Alternatively, several trajectories may be pursued in a preliminary way sequentially by a single designer and then compared simultaneously.

By ordering design decisions carefully and by pursuing several trajectories in a preliminary way in parallel, the designer first selects a promising design direction before committing the resources required to fully refine the design. In doing so the designer avoids a pitfall common among novices, which is to focus on a single design direction initially, investing substantial resources in a concept that will ultimately prove to be disappointing.

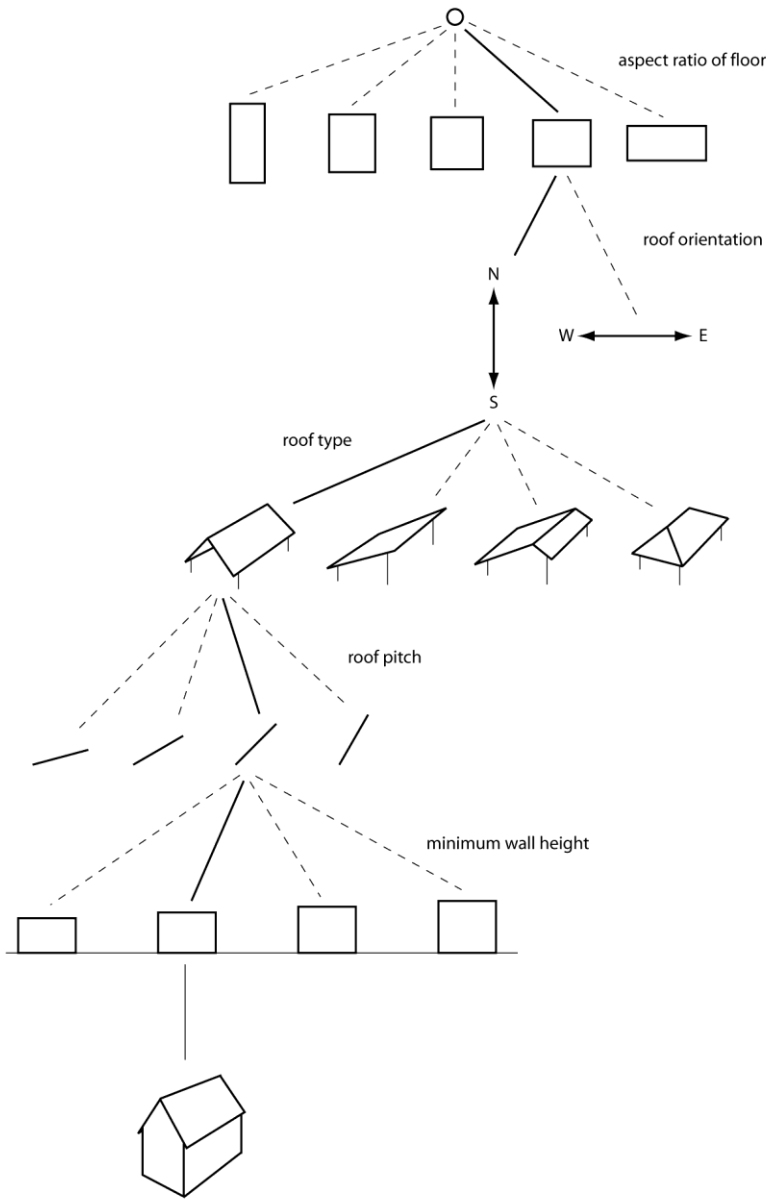


Exhibit 4-4. A partially instantiated tree structure representing hierarchical design decisions.

Causal relationships

Ideally, decisions about design variables are not made randomly. Rather, the designer benefits from knowledge about the causal relationship between a particular value of a design variable and the ultimate quality of the artifact. For example, if a shed will be built off site and transported by truck, the freight costs will be lowest if the shed can be placed on a trailer and can travel normally on roadways. In the United States, this requires that the shed and trailer be less than 14 feet high, which implies that the roof be less than 11 feet high. As a result, we know by simple geometry that for a peaked roof, the shed height is equal to the wall height plus the tangent of the roof angle times half the width of the shed. This knowledge allows us to eliminate from consideration a combination of a peaked roof and high walls and to constrain the roof pitch to be less than 45 degrees for the aspect ratios with wider walls. This kind of knowledge of the causal relationships among design variables and the ultimate quality of the artifact allows for entire regions of the design space to be eliminated from consideration.

Causal relationships need not be mathematically precise or even valid under all conditions. Rather, they can be heuristics that allow for more promising designs to be generated efficiently. For example, one heuristic is that to harmonize with a Victorian house style, the roof should be a gable-end or hipped roof with a pitch of at least 9/12. This is not universally valid, but works for the vast majority of situations. Another heuristic is to use the *golden ratio* (~1.6) for the ratio of the length to the width of the floor. Again, the causal relationship is not universally valid, but provides heuristic guidance that often leads to superior solutions.

If one were to apply all three of these examples of causal relationships to the shed design problem, there would remain only 12 alternatives, few enough that every one of them could be sketched or modeled, and evaluated. Shed world with these causal relationships applied is illustrated in Exhibit 4-5.

Causal relationships are learned through experience, and sometimes are codified and taught. Design in domains for which such relationships have not been learned, discovered, or developed is very, very difficult.[5] Knowledge of these relationships is one of the key factors that distinguish novices from experts as they approach design problems.

Existing artifacts

A fourth strategy for managing the daunting complexity of exploration in design is to exploit existing artifacts. Existing artifacts are jewels for designers. Someone else has expended the resources required to build and test the artifact. Existing artifacts are known landmarks in the design space that can be readily evaluated. By considering the solutions that others have designed to address a similar problem, one can start the exploration process with substantial knowledge. Indeed, if an existing artifact is close to being acceptable, it can become a starting point for incremental modification and learning. Exhibit 4-6 shows a few existing sheds. As a shed designer I could immediately make some useful inferences. (For instance, I discover that I prefer peaked roofs with steep pitches and substantially rectangular floor plans, and I discover many interesting possibilities for window and door placement, and for materials and finishes.)

A generalization of successful existing designs is a *template,* a pattern for designs that has proven successful in the past. Goldenberg and Mazursky (2002) provide compelling evidence for the power of a relatively few templates for guiding the creation of high-quality artifacts in the domains of product design and advertising. They have shown that these templates can be taught to professionals and used efficiently to create new product and advertisement concepts.

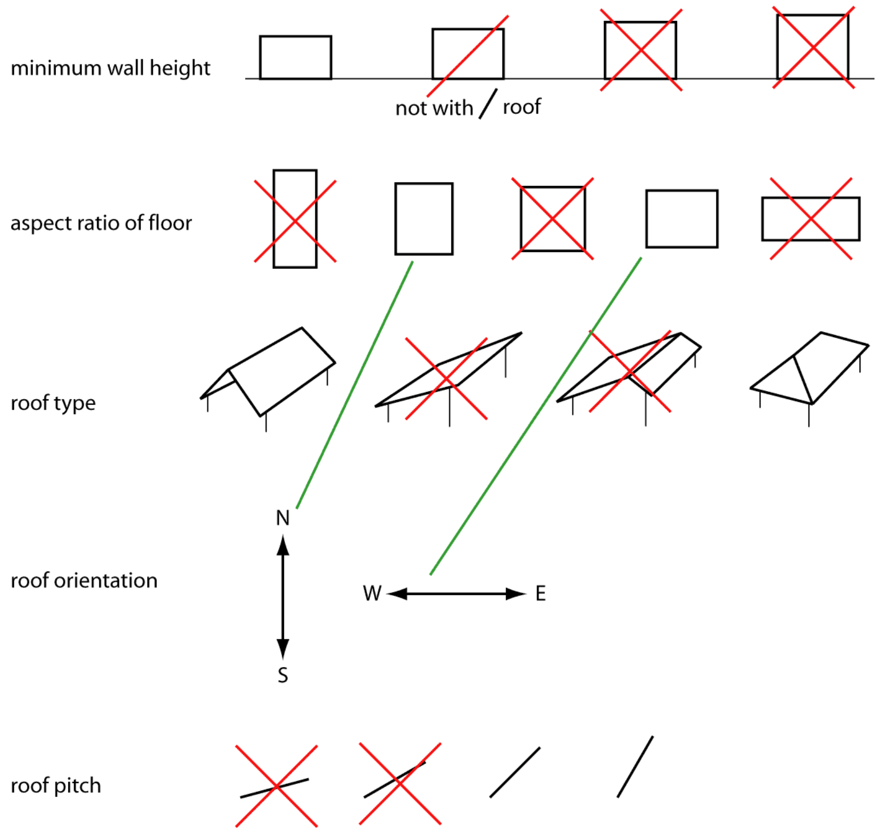


Exhibit 4-5. The design space, pruned through the use of causal relationships.

Shed World and the Real World

Shed world, or really any representation of a design space, is not the *real world* for at least two reasons. First, shed world is an abstraction of the space of possible artifacts that focuses on only a small subset of the attributes of real artifacts. Shed world does not capture the interesting contrast between trim painted baby blue and the weathered shingles on the shed in Exhibit 4-2. Shed world does not treat door and window placement. Shed world does not consider roofing materials. Shed world does not capture the treatment of the soffits and rafter tails on the roof. The real world is infinitely complex, and so any symbolic representation must necessarily omit certain attributes of artifacts. A good representation is one that suppresses detail that is irrelevant to the task of exploring the space of possible designs, yet makes explicit those attributes that have a large impact on the quality of an eventual artifact produced from the design.

Second, and perhaps more significant, shed world constrains exploration to the boundaries of the grammar; to the limits of the expression of the representation. Exhibit 4-7 is a collection of sheds that cannot be discovered through exploration in shed world. Limited expressiveness is the other edge of the sword of representation: Representations allow for efficient exploration by limiting the space of possibilities, but they also exclude many possible design alternatives. In practice, designers can overcome the limits of expressiveness by exploring designs using several alternative representations, in essence exploring under several different sets of constraints and abstractions.

By using shed world as the central example in this chapter, I hope I have not overemphasized the importance of representation in design. Most designers do not think explicitly about representation, and work perfectly comfortably without thinking about the symbol systems they employ. Most designers employ several informal representations when designing, sometimes nearly simultaneously, as evidenced by the Wright sketch at the beginning of the chapter. The theoretical concept of representation is useful, I believe, for better understanding the task of designing. However, I am not prescribing the use of formal representations as a tool or technique for practicing design.

Exhibit 4-6. A collection of existing sheds, each one representing a known point in the design space.

Other Examples

I have illustrated the key concepts of the chapter with the problem of designing a shed, because the domain is simple and easy to understand. However, I do not wish to leave the impression that these ideas apply only to the design of buildings. Following are a few other examples of design domains, associated representations, and exploration strategies.

Internet domain names

Naming problems are a highly structured form of design problem. Generating designs for Internet domain names is a fairly common problem in professional life. Domain names must of course be unique, in that they must map to a single Internet protocol numerical address. The design problem is to find a name that is available and that satisfies some other criteria. Common criteria for product and company names are that they be memorable, easy to spell, short, and evoke positive associations. Domain names may consist of only 38 possible characters (a-z, 0-9, -, +) and must end with a top-level domain (e.g., .com, .net, .edu, etc.). These rules are design grammar for the domain. If we assume that a practical domain name has 15 or fewer characters, then there are more than 38^{15} possible names for each type of top-level domain.[6] This is about enough to give a unique name to each grain of sand on earth.[7] Given this vast design space, finding a unique name is not typically a problem (e.g., xutq++012ayq858.net is highly likely to be available). The problem is that the space is rather sparsely populated with names that are in some sense *good*. Exploration can proceed fairly exhaustively for domains up to about three letters long, at which point the designer really has to begin invoking some brutally efficient heuristics to limit the possibilities considered.

Exhibit 4-7. An eclectic collection of sheds not represented by shed world. (Various sources.)

Exhibit 4-8 shows the later stages of exploration for a name for a teaching aid that I designed with my colleague Christian Terwiesch. The device is a catapult that launches table tennis balls and that can be adjusted in order to run experiments on the launching process. The names in the exhibit are the best of more than a thousand alternatives that were generated by a group of my students. Note the use of heuristics for generating alternatives. For example, a very common heuristic is to create compound names composed of two words (e.g., "flingthing"). Another heuristic is to construct an arbitrary string of characters that can be easily pronounced (e.g., "fooz"). A third heuristic is to take fragments of two words that have meaning in the domain of interest and graft them together (e.g., "catapong"). These names are much better than random strings of letters, and provide the designer with an efficient way to explore the space. A second idea illustrated by this example is that of selectionism. A large number of parallel trajectories were compared in tournament fashion, with successive rounds of filtering to arrive at a good solution. The name we finally selected was *xpult* and the domain is xpult.com. We were quite pleased to find a unique evocative name just five characters long, even though there are 38^5 five-character names out there.

An important insight is that if one of the most highly structured design domains imaginable (Internet domain names) is essentially infinite in scope, imagine the vastness of less structured domains such as architecture, graphics, industrial design, software, cooking, or engineering design.

Initial Concepts	Best Ten	Best Three	Final Name
AstroPong	Catapong	Catapong	Xpult
Catapong	Catapulooza	Varipult	
Catapulooza	Experipult	Xpult	
Experipult	FlingThing		
FlingThing	Funpult		
Fooz	Hurlicane		
Funpult	PingFling		
Hurlicane	Slingcat		
Hurlitzer	Varipult		
LearningLever	Xpult		
PennPong			
Physazz			
PingFling			
Pongit			
Slingcat			
Swish			
TheCatapult			
Varipult			
Xpult			

Exhibit 4-8. Exploration of alternatives for Internet domain names (and a product name) for an experimental catapult used as a training aid.

Utility knives

Exhibit 4-9 illustrates exploration for the domain of utility knives, in this case in response to a design problem posed by the company Henkel. The designer explored many alternatives in a preliminary way, 24 of which are illustrated on the left. Three promising alternatives are shown on the right with greater resolution of detail. Some of the variables evident in the designer's implicit representation of the problem are handle width, "beak curvature," grip padding placement, blade/handle interface, and blade replacement mechanism.

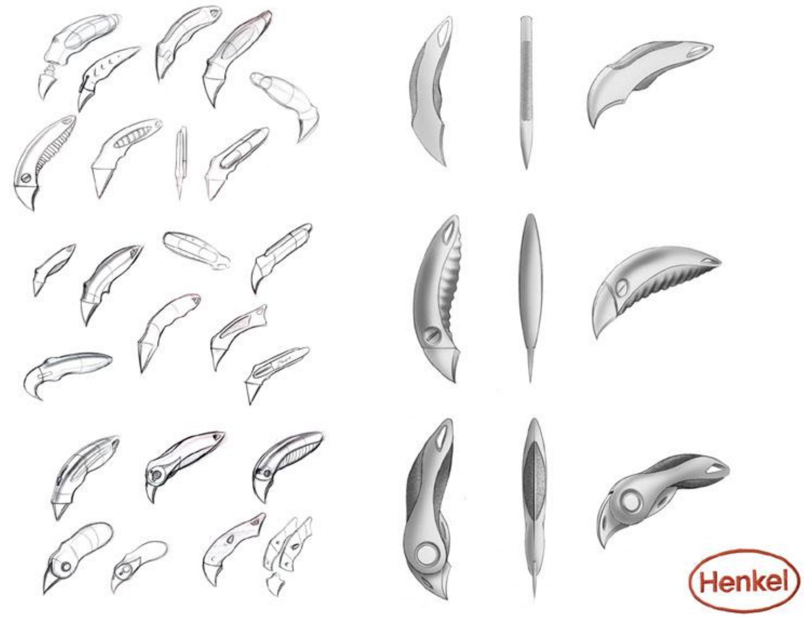


Exhibit 4-9. Results of exploration in the domain of utility knives. On the left are results of some preliminary exploration and on the right are the three most promising alternatives. Source: Apollo Paul Paredes.

Italian pasta dishes

In my experience, if one orders a pasta dish at a restaurant in Italy some distance from the obvious tourist destinations, it will be wonderful nearly every time. Many of these pasta dishes seem very simple, yet they represent highly successful artifacts in a design space that is incredibly vast. Consider the representation of pasta dishes shown in Exhibit 4-10. The pasta itself can be produced in infinite variety. (There is even pasta in the shape of a bicycle for the cycling fanatic.) Even if dishes are restricted to the few hundred readily available pasta types, adding the design variables associated with the sauce explodes the design problem into millions of possibilities. (This is without considering the variables associated with relative proportions of ingredients.) Designing a new pasta dish benefits from several of the exploration strategies introduced in this chapter. For example, we might address the problem hierarchically, perhaps first deciding the base for the sauce and deferring until last the shape of the pasta. We might invoke causal relationships, like the heuristic that tomato, garlic, and olive oil often combine harmoniously; or that vegetables with subtle flavors typically do not stand up to the strength of tomato-based sauces. We might use existing designs as starting points—say, beginning with a carbonara sauce (egg, pecorino cheese, pancetta, olive oil, and garlic)—and incrementally modifying it to be a meatless design, such as by substituting caramelized onions for pancetta.

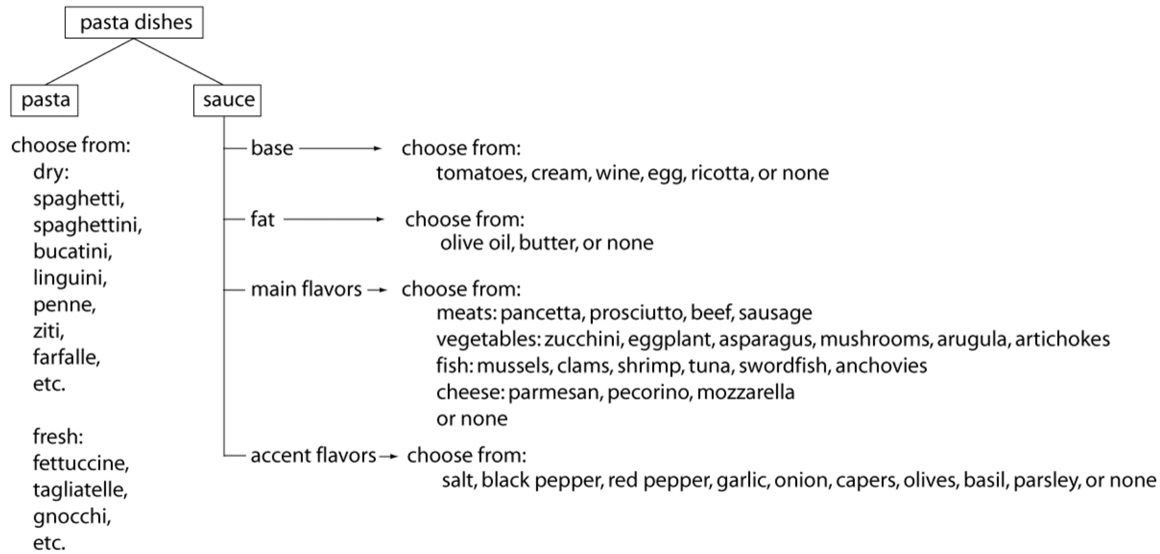


Exhibit 4-10. A representation of a design space for pasta dishes.

Logos

Terrapass is a company that sells environmental offsets for automobiles (www.terrapass.com). Shortly after the company was formed, a team of three graphic designers explored options for a logo for the company. Some of the exploration is shown in Exhibit 4-11. The process clearly proceeded hierarchically, with initial concepts articulated in black and white and then the more promising concepts developed further and finally detailed in color and with type. The team explored quite broadly initially and discovered a region of the design space they called the "yin yang arrows" (the two designs near the lower right corner of the first set), which everyone really liked. This region was explored further (the middle set of designs) and finally refined with color and detail in the final design on the right.

Concluding Remarks

An attempt to provide a theoretical framework for exploration in design raises at least two interesting questions. First, if exploration can be characterized formally, can it be automated? Second, what is the relationship between theory and practice; do practicing designers think of exploration as I've described it?

Automation

Over the past few decades, researchers have attempted to automate certain design tasks. By and large the most successful efforts have been confined to facilitating the description of designs (e.g., with solid modeling via computers), visualizing designs with computer graphics and rapid prototyping, and/or estimating the performance of artifacts. There has been very little progress in truly automating the exploration process. I believe that the biggest barrier to this endeavor is automatically estimating the quality of an artifact based on a partially completed design. I'm not optimistic about the prospects for full automation of the exploration process. However, I see great potential for further development of tools for allowing designers to more rapidly generate alternatives, visualize designs, and evaluate designs without having to build and test prototypes.

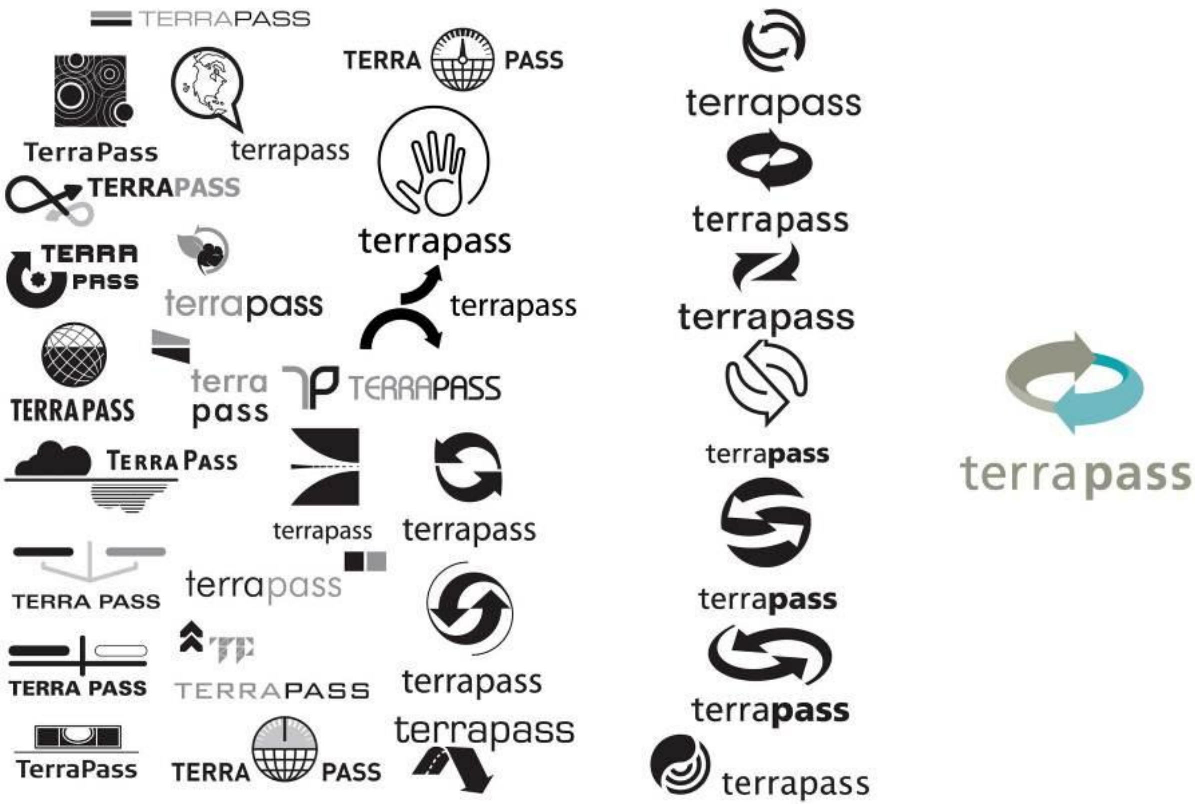


Exhibit 4-11. Exploration of alternatives for Terrapass logo. The seven logos toward the right resulted from further exploration in the region of "yin-yang arrows" discovered during initial exploration. Source: Lunar Design Inc.

But do designers do it this way?

I don't imagine the chef Thomas Keller will read this book and begin developing a pasta grammar for his exquisite restaurant The French Laundry. Indeed, very few practicing designers became experts at design by learning the theoretical foundations of exploration as outlined here. Let me make two comments on this reality. First, the fact that practitioners are not aware of the theoretical underpinnings of a task does not mean those underpinnings are not valid. Design is a complex information processing task. There is no way to avoid the inherent complexity of the task, although expert designers have developed many powerful techniques for avoiding blind search. Just because designers do not typically think of their tasks in formal terms does not mean that those tasks can be tackled without somehow confronting the basic tradeoffs and challenges inherent to exploration. In fact, I believe that most good designers learn the strategies I have described here as well as others, even if they cannot articulate them explicitly.

My second response is perhaps more controversial. Much of design education and almost all of design practice is atheoretical. I believe that theory can inform practice in design. In many domains, expertise is acquired through painstaking trial and error, often under the guidance of a seasoned expert. I believe that a robust theory of exploration can lead to more efficient learning of design expertise and a more thorough exploration of design alternatives in practice. Indeed, this belief was one of the motives for writing this book. I may be wrong in this belief, and so I leave it as a conjecture that remains to be validated.

Notes

[1] The beam is a round tube because we know that a round tube is the most weight-efficient structure for supporting loads that could come from any direction. The equation for the deflection of a tube rigidly supported on both ends with a load F applied in the middle is $\delta = FL^3/192EI$, where L is the length of the span, E is the modulus of elasticity of the material, and I is the moment of inertia of the beam cross section. (We know this thanks to at least Galileo, da Vinci, Euler, and Bernoulli.) The moment of inertia is calculated as $I = \pi(D^4 - d^4)/64$. We know that the minimum thickness of the wall is 1 mm to allow inexpensive joining techniques and to prevent buckling (i.e., $D - d \geq 0.002$ m), and we know that the lightest possible structure will be a tube with the minimum possible wall thickness. We can plug in values for δ, F, L, and E and solve for D. The resulting design is a 2 m long aluminum tube, 43.5 mm in diameter, with a wall thickness of 1 mm. Thus, we can solve a design problem while avoiding wasting effort on exploration.

[2] Winston (1992) provides a clear and detailed discussion of representation and search in his book on artificial intelligence (AI). Design is connected in many deep and important ways to AI and the Winston book provides a good introduction to the core concepts.

[3] See Ulrich and Eppinger (2011), Chapter 12, for a thorough discussion of prototypes in product design.

[4] Herbert Simon (1996) pioneered the view of design essentially embodied in this chapter, articulating the concepts of representation, complexity, and search. I deliberately avoid the term *search* in this book, preferring instead *exploration*. The term *search* tends to offend practicing designers. For many, it implies weak methods unguided by expertise. This is not the sense in which Simon and other early researchers intended it, but I find the word *exploration* more descriptive of the activity anyway, so I adopt it here.

[5] Fleming and Sorenson (2004) have done a fascinating study of the patent literature in which they show that science serves to guide search in complex design domains.

[6] There are more than this because domain names need not be 15 characters long, but this figure gives a sense of the essentially infinite scope of the design space.

[7] If you must know, poke around the Internet and you'll probably find estimates for the number of grains of sand on earth to be about $10^{22} - 10^{25}$. Note that 38^{15} is about 10^{23}.

References

Fleming, Lee, and Olav Sorenson. 2004. "Science as a Map in Technological Search." *Strategic Management Journal* 25: 909–928.

Goldenberg, Jacob, and David Mazursky. 2002. *Creativity in Product Innovation*. Cambridge, UK: Cambridge University Press.

Keeney, R. L., and H. Raiffa. 1976. *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. New York: Wiley.

Simon, H. A. 1996. *The Sciences of the Artificial*. Cambridge, MA: MIT Press.

Sommer, Svenja C., and Christoph H. Loch. 2004. "Selectionism and Learning in Projects with Complexity and Unforeseeable Uncertainty." *Management Science* 50 (10): 1334–1347.

Terwiesch, Christian, and K. T. Ulrich. 2009. *Innovation Tournaments: Creating and Selecting Exceptional Opportunities.* Boston, MA: Harvard Business Press.

Ulrich, K. T., and S. D. Eppinger. 2011. *Product Design and Development*. New York: Irwin/McGraw-Hill.

Winston, P. H. 1992. "Semantic Nets and Description Matching," 15–46; "Nets and Basic Search," 63–80. In *Artificial Intelligence*. Reading, MA: Addison-Wesley.

FIVE

Users, Experts, and Institutions in Design

The first act of design was almost certainly *user design*, in that the first artifact was given form by the user rather than by a third-party designer. Perhaps this first user designer contemplated frustration with a task nearly 3 million years ago, formed a plan to address the frustration, and then fashioned an artifact, possibly fracturing a river cobble to form a scraping tool. A clear distinction between expert designers and user designers emerged at some point, possibly first in architecture. Certainly by the time ancient Egyptians were creating pyramids, the roles of experts and users in design were separated. The activity of design appears to have become increasingly professional and institutionalized over the next few thousand years. By the nineteenth century, as the Industrial Revolution developed in full, expert designers with specific technical training assumed distinct professional roles, both because of the comparative advantage of expertise and because institutions, usually companies, were formed to exploit the benefits of mass production.

Although a separation between users and designers has increased in many domains over the past several thousand years, the practice of design by users is emerging again in current society. This chapter addresses the role of the user in design, with particular emphasis on design by users, and considers how experts and institutions interact with users to deliver artifacts in modern society. In this chapter, I articulate three modes of engagement in design: by users, experts, and institutions. Then, I outline the drivers of the selection of these modes in society. Finally, I discuss how emerging technologies and practices are enabling shifts in how these modes are applied.

Design Modes

Design is conceiving and giving form to artifacts that solve problems. Exhibit 5-1 reiterates the model of design I adopt for this book. Design is an information processing activity through which a plan for an artifact is created to address a gap in the user experience. A production activity transforms that plan into the artifact itself.

I have described the design process without characterizing the agents who do the design other than referring to them as *designers*. For the purposes of this chapter, I distinguish between *users* and *experts*. Users are the individuals experiencing the perceived gap between the current state and the goal state. They are essentially always a party to the design process.[1] Other terms for users include *customers*, *consumers, clients*, and *stakeholders*, although these terms evoke a more specific commercial context than I intend. Experts are individuals who have acquired skills and capabilities that allow them to perform most design tasks more efficiently and at a higher level of quality than novices. In some cases an expert may also be a user, as for instance when an expert employed as a designer by a cookware company is also an avid cook.

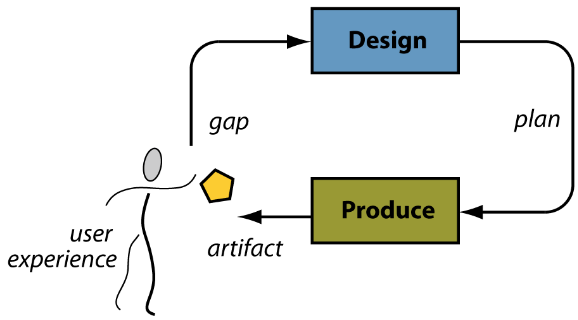


Exhibit 5-1. Model of the design process.

I make an additional distinction about the institutional context of design. Design may be performed for a particular user or for multiple users who share similar needs. When design is performed for a collection of users (e.g., a *market segment*), usually by providing artifacts produced from a common plan, some institution is required to coordinate the design and production of the artifact. These institutions are most typically firms, but may also be governments, clubs, religious organizations, universities, professional societies, user groups, or neighborhood associations.

I divide the modes of design into three categories—*user design*, *custom design*, and *common design*—according to the roles played by users, experts, and institutions. These modes are illustrated in Exhibit 5-2.

- **User design** consists of a single user designing for his or her own needs. Because the resulting plan is produced for a single individual, and therefore in low quantity, a flexible production process is required to deliver the artifact. Flexible processes are those that can produce different artifacts without incurring high fixed costs for each variant. In many cases, such flexible production processes are craft processes in which skilled people create artifacts with general-purpose tools, as is typically the case for unique furniture or unique buildings. An example of a flexible production process enabled by technology is digital printing.
- **Custom design** is also paired with flexible production of a unique artifact. However, an expert creates a plan on behalf of a particular user. In most cases, the user contracts with the expert for this service, as is the case when hiring an architect to design a unique house or engaging a machinery designer to design a unique piece of factory equipment.
- **Common design** differs from custom design and user design in that artifacts based on a common plan are delivered to a collection of users. Because this common artifact is in a relatively large quantity, it may be produced by mass production methods, processes that typically incur substantial fixed costs for each variant of the product, but relatively low marginal costs of producing additional units. Common design involves an institution of some kind, usually a firm, that assesses the gaps in a

set of users, creates a common plan for addressing those gaps and delivers an artifact based on a common design to those users.

This taxonomy focuses on differences in the way *design* is performed, and I do not distinguish between flexible production by users and flexible production by experts. Mass production because of its very nature must be performed by an institution of some kind as it serves a collection of users with a common artifact.

These categories are intended to be exhaustive and mutually exclusive relative to the variables identified here. However, all three modes may exist simultaneously to serve different individuals within the same community of users or market.

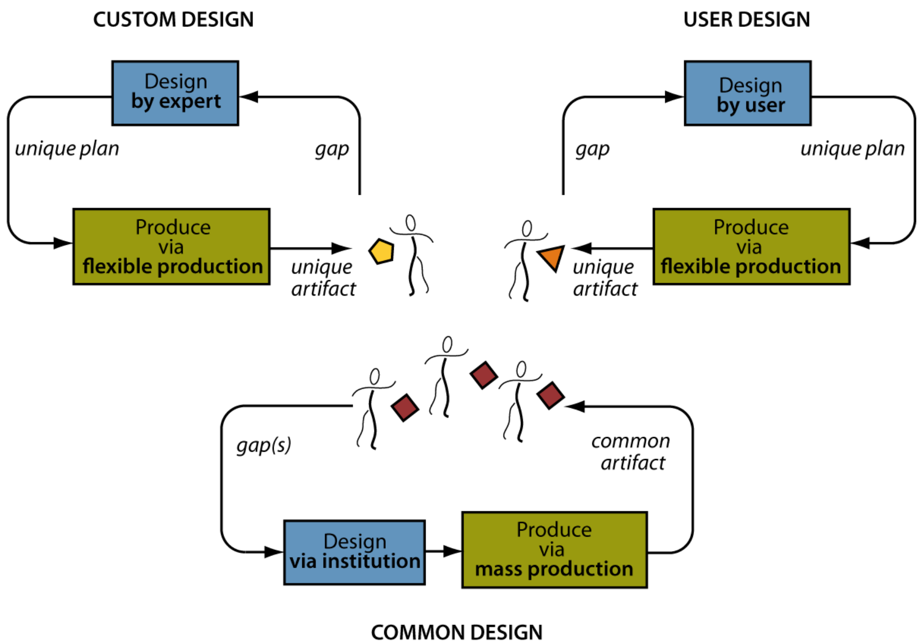


Exhibit 5-2. Three modes of design that may be exhibited within a community of users.

Drivers of Mode Choice

Assuming that historically the first design mode was user design, why did the other modes evolve and why do they exist? What are their relative advantages? What drives the choice of mode in a particular setting?

Economies of scale in production lead to common design

A very large fraction of the economic value in retail trade in current society flows through just a few very large distribution channels (e.g., Walmart, Target, Home Depot, Carrefour). Most products in these channels are produced in high volume (e.g., 10,000 to 10 million units/year) for a mass market. This is because for these products, mass production offers a crushing advantage in satisfying user needs at low cost. This advantage arises because of economies of scale in design and production. Creating 10,000 pairs of identical shoes can be a hundred times less expensive on a per-unit basis than creating only one pair of unique shoes. Very few consumers have distinct enough needs to be willing to pay a hundredfold premium for shoes made uniquely for them. Thus, the cost structure of most design and production processes provides a compelling motive for clustering similar groups of users and addressing their needs with a common design.

A common design requires an institution of some kind, because to achieve commonality, users must be grouped, the gaps in their experiences assessed, and a common artifact designed and produced for them. In sum, economies of scale lead to mass production; mass production requires a common design; a common design requires an institution. For this mode, user design is generally not possible. To the extent that design is performed by a single individual, or even by a team, the remaining individuals whose needs are addressed by the common artifact will not be designers. Instead their experience will be assessed vicariously by others in the common design mode.

Custom design versus user design

Design is performed for a single user when that user's needs are unique enough, given likely economies of scale in design and production, that a unique artifact is preferred to a common artifact (Lancaster 1990). This case arises frequently in architecture (custom homes, buildings, landscapes), food, software, and graphics. This mode is also exhibited occasionally in furniture, apparel, sporting goods, and tools. It is exhibited rarely in home appliances, automobiles, aircraft, medical devices, or computers, domains for which the economies of scale present nearly insurmountable barriers to unique artifacts, even for the very wealthy.[2] The design of a unique artifact in this context may be performed either by the user or by an expert on behalf of that user, leading to the two modes in the upper half of Exhibit 5-2.

All other things equal, design professionals develop expertise that allows them to perform design tasks better than novices (Ericsson 1996). Given that most users will be novices, experts will outperform users in most design tasks. However, costs are incurred in engaging an expert, and so the expert design mode will be selected only when the advantages of expertise outweigh its costs. These costs can be thought of as *direct costs* paid to the expert and as *transaction costs* associated with retaining the expert. Direct costs are straightforward: Most experts will be paid for their services. Transaction costs are more subtle; they are incurred in defining a design problem and in evaluating alternative solutions.

On first reflection, a user would appear to have an advantage over an expert in defining a design problem, in diagnosing the gap in his or her own experience. I believe that this is sometimes true, but not necessarily so. Experts by definition have encountered similar design problems many times before and will likely have observed empirical regularities in user needs. Experts typically also deploy techniques for probing user needs, such as interviews and observational methods (Ulrich and Eppinger 2011). In many cases user needs are *latent*, in that they cannot be spontaneously articulated by users, but if these needs are satisfied, the gap in the user experience is addressed. Of course, a risk of expertise is that it frames the designer's diagnosis of the problem. An architect may define a gap in the communication patterns within an R&D organization as a problem relating to the built environment, whereas a management consultant may define the same gap as a problem of organizational structure. These challenges in defining design problems may be manifest as financial costs and/or the costs associated with different levels of quality of the resulting artifact.

Exploration almost never results in a single plan, but rather exposes several alternatives that are promising enough for serious consideration. Evaluation of alternatives typically occurs "on paper" before an artifact is produced. Once an artifact has been produced, there is almost always an evaluation through testing by the user. Users are clearly best at assessing, through their own experience, whether an artifact actually closes the sensed gap in their experience. While experts may productively observe patterns in behavior, ultimately the user is the frame of reference for the gap in the first place, and is the only agent who can conclude that the gap has been addressed. However, users are typically ill equipped to forecast the extent to which a design alternative, represented abstractly, will meet their needs. Because they do not work daily with design representations, most users are not skilled at visualizing an artifact or a mental simulation of the artifact's function, and are not alert for common pitfalls for a category of artifact.

Given these characteristics of transaction costs, users are actually likely to have an advantage over experts when design alternatives can be readily generated and when plans can be accurately evaluated quickly and at low cost, as when realistic prototypes can be produced readily. In such environments, the user can achieve high-quality design through rapid iteration and learning. Expert design in the same context can incur high transaction costs because of the switching back and forth between search by the expert and evaluation by the user. In this situation, the more efficient search by an expert may be outweighed by the reduced transaction costs of user design.

An additional driver of user design is the utility (or disutility) some users derive from solving their own problems. To the extent that there is a psychological benefit derived from the process of design ("I designed it myself!"), a user may be willing to accept a lower-quality outcome even at the same cost of expert design (Franke et al. 2010).

For completeness, let me comment on an additional form of transaction costs, that emphasized in *transaction cost economics* (TCE). The TCE paradigm has been influential in thinking about industrial organization and so should be mentioned here. Consistent with the view articulated in this chapter, TCE would predict a bias for user design in the face of high transaction costs. However, the transaction costs contemplated in TCE are those associated with *asset specificity*. When a contracting relationship between a user and an expert requires a speculative investment in assets (e.g., knowledge and expertise) that are highly specific to a particular relationship between a user and an expert, both the user and the expert face a loss in bargaining power. This is because the asset that has been developed may be used only for the specific relationship. Under these conditions, TCE predicts that the user will prefer not to contract with another party, but will instead perform design for him- or herself. For a discussion of the theory of transaction cost economics and the related literature, see Ulrich and Ellison (2005). The problem with invoking TCE in this context is that most design is a "one-off" effort, and so when contracting with an expert, a user typically assumes that all transaction costs, including investments in specific assets, will be paid as part of the engagement. The expert would rarely, if ever, invest in specific assets without factoring those investments into the contract for design services. Terwiesch and Loch (2004) discuss some of these contracting and pricing issues in the context of customized artifacts.

Synergies among modes

All three modes of design can and typically do exist in the same community and for the same category of artifact. Some people engage in user design. Some people engage in custom design. Everyone participates in common design, at least through their consumption and use of artifacts.

A commonly occurring pattern of innovation is for a new artifact to emerge through user design and then to be adopted, often with some refinement, as part of a common design effort. This process of appropriation and improvement may take place over many years and even generations. This pattern of innovation has been documented in detail by von Hippel (1988). However, the migration from a unique design to a common design need not originate in user design. An essentially similar pattern involves the migration from expert design of a unique artifact for a single user to common design by an institution for a collection of users. In either case, an individual user uncovers a set of user needs and a design that addresses those needs. This design is subsequently exploited by an institution to deliver a common artifact.

Hybrid modes

An artifact may be the result of more than one mode of design if it comprises more than one element. For example, a common component may be used in combination with a custom component. Or, one or more attributes of a component may be customized, with the rest standardized. This approach is sometimes called a *platform strategy* and is closely related to the notion of *mass customization*. By adopting this strategy, a producer may be able to offer a user a unique design while exploiting the economies of scale associated with the standard elements of the product. Randall, Terwiesch, and Ulrich (2005) provide a detailed discussion of user design for customized products.

Enabling Processes and Technologies

Mode choice in design is strongly influenced by changes in design and production processes and technologies. New technologies and processes have emerged in the past few decades that are changing the way design modes are adopted in practice.

Templates

The challenge of exploration is dramatically simplified if a *template* is adopted. A template in this context is a fixed architecture for an artifact within which alternative elements may be placed (Ulrich 1995). For example, iPrint is a web-based system by which users may design printed items such as business cards, stationery, and party invitations (Exhibit 5-3). Each of several types of items is represented with a standard template. Within that template, choices may be made of typeface, type size, colors, position of graphic elements, paper, and textual content. By constraining search to a selection of elements within a fixed template, the design problem is bounded sufficiently that many users are able to find satisfying solutions without retaining an expert. Digital printing technology is sufficiently flexible that unique artifacts may be produced in relatively low volume (50–1,000 units) at reasonable cost.

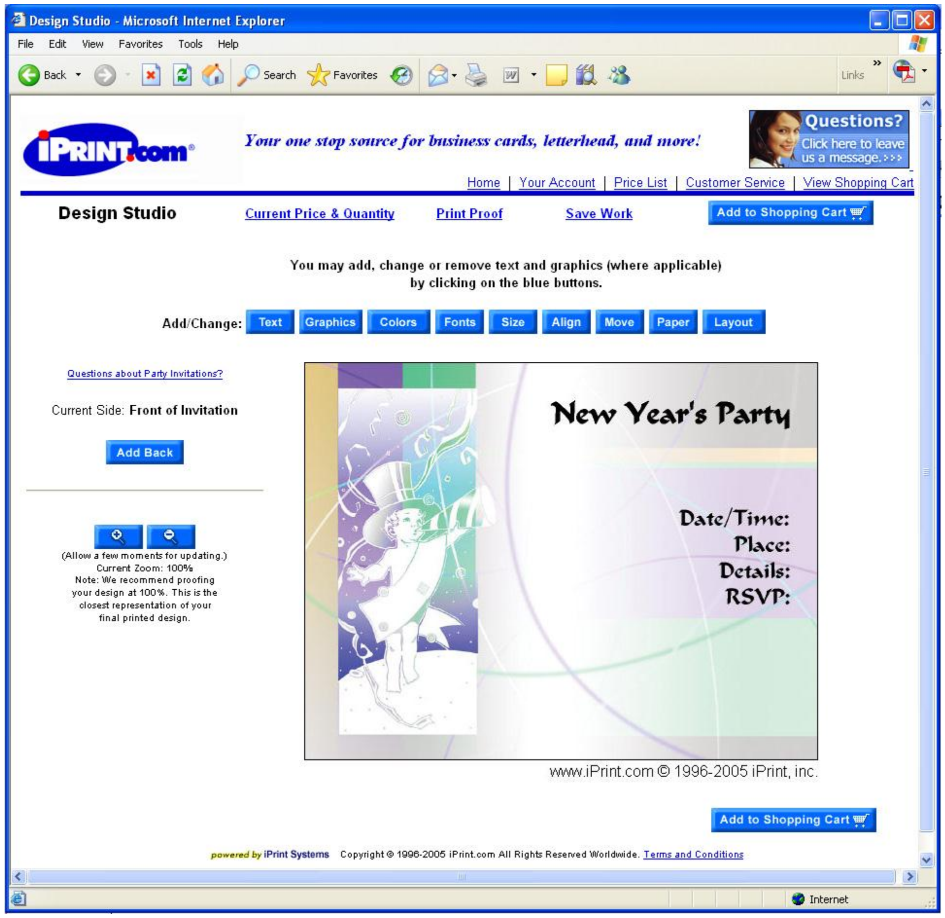


Exhibit 5-3. Web-based interface for user design of a New Year's party invitation based on a template.

Design grammars

A design grammar is a set of rules defining "valid" designs, including a definition of the elements of the design and the rules by which they may be configured. (A template is a very restrictive type of grammar in which the alternative selections of elements must always be configured in the same way.)

Stiny (1978) developed a design grammar for several domains in architecture, including Queen Anne style houses. Exhibit 5-4 is an example of several instances of valid Queen Anne houses within Stiny's grammar, each showing a different valid porch configuration for a single main house plan.

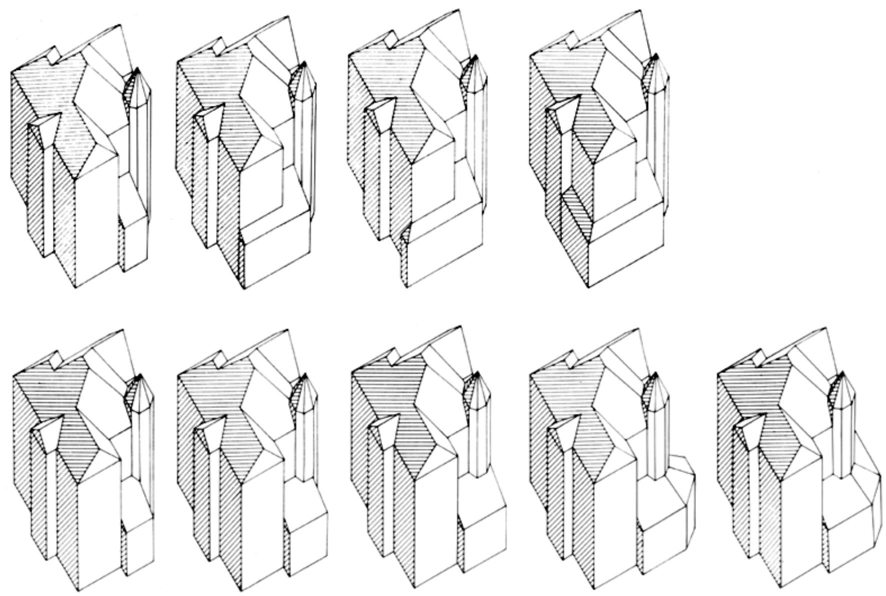


Exhibit 5-4. A few instances of a "Queen Anne" design composed within the Queen Anne grammar. Source: Pion Ltd, London (Flemming U, 1987, "More than the sum of parts: the grammar of Queen Anne houses" Environment and Planning B: Planning and Design 14(3) 323 – 350)

A grammar defines a universe of valid designs. While it may enable efficient exploration, it also restricts the space of possibilities to the scope of the grammar. Consider the designs of Frank Gehry such as the MIT Stata Center (Exhibit 5-5). In the late twentieth century, Gehry's work appeared fresh precisely because it deviated from existing grammars, possibly the way the Queen Anne style appeared fresh in the late nineteenth century. Interestingly, over his career Gehry has designed enough buildings that one can start to imagine a formal grammar defining a valid "Gehry style."

Grammars have been developed and used for VLSI circuit design, for computer system design, and for chemical process design. Formal grammars have otherwise rarely been used in design practice. However, the development and use of such grammars offers the prospect of making exploration more tractable for novices, or even computers.[3]

Automation of exploration

If a design domain can be formalized through a design grammar, then the prospect of automating exploration emerges. A second requirement for automating exploration is that a formal evaluation function (or *objective function*, in the language of optimization) be articulated. Without some way of automatically estimating the quality of a design, automating exploration is unlikely. For highly structured design problems, such as creating a customized personal computer to meet the needs of an individual, automation of exploration is currently feasible (Randall et al. 2005). Additional problems are likely to be addressed by automation in the future. If exploration were more highly automated, the value of expertise in exploration might be diminished, making user design more attractive.

Exhibit 5-5. The Stata Center at MIT, designed by Frank Gehry. Source: Wikipedia, original upload 3 August 2004 by Finlay McWalter.

Rapid prototyping

Most design efforts require the designer to forecast the extent to which a contemplated alternative will satisfy the needs of the target user. A forecast is required when the cost of producing the artifact, even in prototype form, is relatively high. Rapid prototyping technologies, which might more appropriately be called *inexpensive* prototyping technologies, allow the designer to produce relatively more prototypes for actual testing and can therefore reduce the importance of accurate forecasting of design quality. In the hands of a novice designer, the act of testing many prototypes can substitute to some extent for expertise in exploration and evaluation of designs and thereby enable user design where custom design or common design was previously the norm.

Exhibit 5-6 shows several chess pieces made directly from computer models using the *selective laser sintering* (SLS) process. The cost and time required to produce physical models of complex geometric forms like these have fallen by at least a factor of ten relative to conventional prototyping technologies (in this case, carving by hand), enabling more frequent evaluation of physical prototypes as opposed to requiring the designer to completely refine the form of an object before committing to an expensive and time-consuming prototyping process.

Flexible production

Flexible production is a means of producing artifacts with relatively low fixed costs per variant of the artifact. For example, laser printing of documents is quite flexible, allowing ten different documents to be printed at about the same cost as ten copies of the same document. Computer-controlled laser cutting machines allow arbitrary trajectories to be cut in plywood, sheet metal, and plastic sheet, with essentially no setup cost. To the extent that an artifact can be produced by flexible production means, unique artifacts can be produced for individual users at reasonable cost. Flexible production technologies therefore enable custom design and user design. Exhibit 5-7 shows a web-based design interface that creates instructions for a computer-controlled milling machine, which can be used to flexibly produce three-dimensional shapes as shown. CNC milling is a material removal process incurring only modest fixed costs per variant of the artifact and therefore enabling relatively low-volume production. As flexible production processes become increasingly available, both user design and custom design become more feasible.

Exhibit 5-6. Chess pieces fabricated using the selective laser sintering (SLS) process, a rapid prototyping technology. Source: http://www.kinzoku.co.jp/image/zoukei_p3_b.jpg.

Tournaments

Tournaments in design have increased in popularity with the advent of mass media channels, but have probably been used by institutions for a long time. In a tournament, many individuals or teams submit plans or prototypes, which are typically evaluated by experts, sometimes with panels of users, and sometimes through testing (Terwiesch and Ulrich 2009). Some tournaments are intended to be primarily design mechanisms for a producer or user. Examples of these competitions are QVC's product road show, which visits ten cities in the United States each year to screen new products, and the U.S. government agency DARPA's Grand Challenge autonomous robotic vehicle competition. Other tournaments are intended primarily to deliver entertainment to an audience. An example of this type of competition is *Million Dollar Idea*, a televised competition in which a winner is granted $1 million to commercialize his or her invention. Tournaments exploit large numbers of parallel searches by individuals, sometimes collecting design alternatives from thousands of entrants. This strategy can be particularly powerful when seeking new ideas for products in that a raw plan, perhaps only partially developed, can be selected from the efforts of many individuals and then refined professionally through common design by an institution. In this way, tournaments are a way of harnessing the value of independent exploration by user designers with the cost advantages of common design. Tournaments may also exploit a tendency by entrants to overestimate the probability of success, possibly resulting in more design effort per unit of investment by the tournament sponsor than could be achieved by other means.

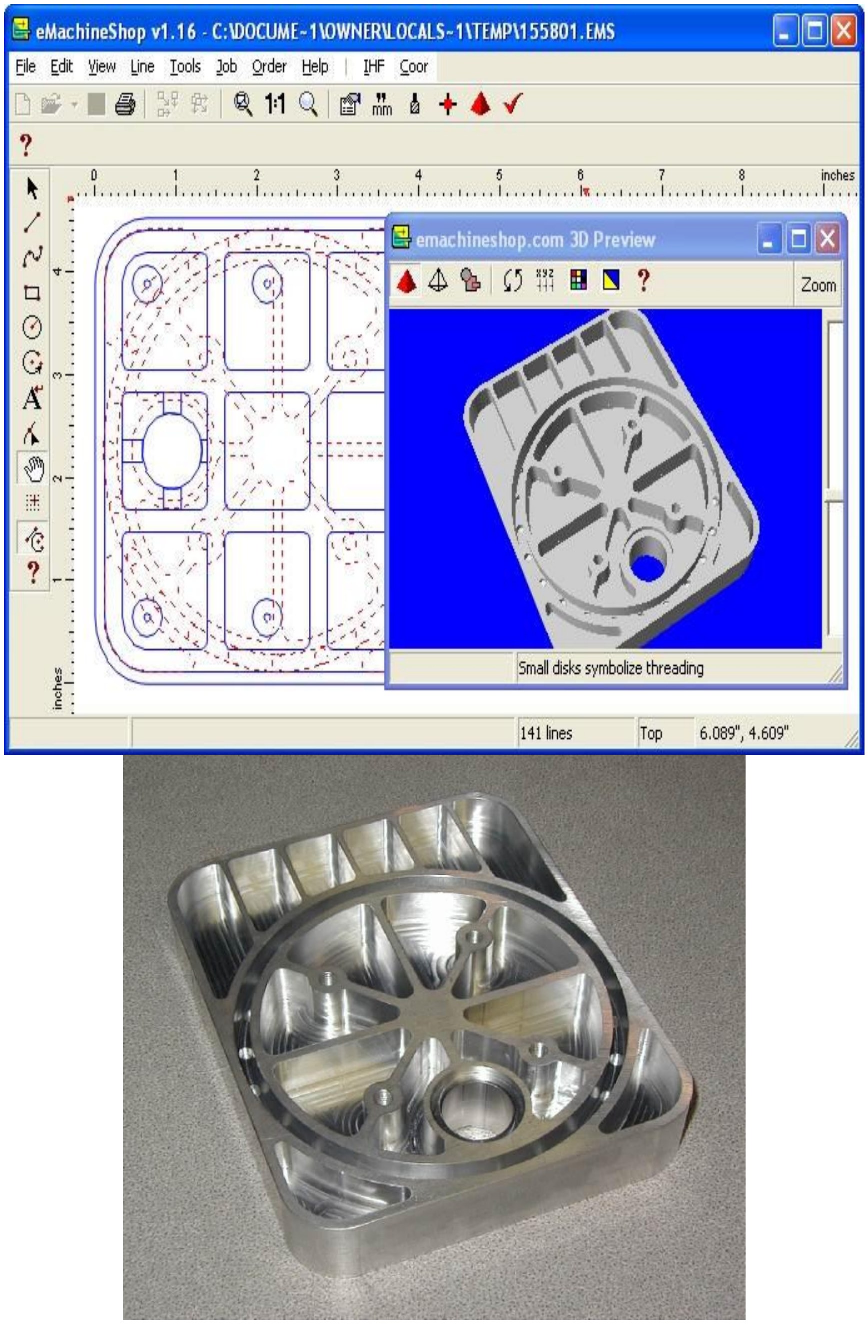


Exhibit 5-7. Aluminum part flexibly produced by a CNC milling machine. A web-based design program can be used to create instructions for the milling machine. Source: emachineshop.com.

Open source

The practice of *open source* arose in the software engineering community and comprises, at a minimum, the free publication of the "source code" for an artifact. For software, the source code is the program instructions in human-readable form, typically as they were written by the designer. For documents, the source code is the text, in readable, editable form. For a physical good, the source code might include geometric information, materials specifications, control algorithms, and/or process specifications.

The rationale for open source is that some users will sense opportunities for improvement in an artifact and will themselves make those improvements (Terwiesch and Ulrich 2009). Several open-source communities have developed and are active, the most famous being the Linux computer operating system. Most of these communities have some mechanism for evaluating and ratifying potential improvements submitted by members of the user community. Remarkably, some open-source artifacts evolve with almost no managerial oversight. For example, the Wikipedia encyclopedia is open source, and can be modified by anyone in the world with access to an Internet browser. Open-source communities need not be firms, but they are nevertheless institutions that enable the common design mode.

Design kits

Design kits are tools to facilitate the design process, often provided at no charge by firms seeking to produce the unique artifacts of designers, or who otherwise benefit from active design communities. Producers of specialized semiconductor devices will sometimes provide designers with "breadboard" systems incorporating the devices to enable experimentation and trial, and in the hopes that these devices will be used in a new artifact. Design kits reduce the fixed costs of designing a unique artifact and so enable expert design and user design.

User groups

User groups are sets of users with communication mechanisms to facilitate the exchange of information relative to a class of artifact. These mechanisms are increasingly electronic, typically implemented via the Internet. User groups are often structured around issues or questions, sometimes called discussion threads, although some user groups have formal administrative elements such as managers and committees. User groups enable user design by allowing plans from one user to be communicated to another with similar needs. User groups can also facilitate common design by allowing users to share information about gaps, coordinate plans, and even test prototypes.

An example of a user community is flashkit.com, a community of designers using the Macromedia Flash multimedia programming language. As of this writing, this community had about 500,000 members. In this case, a primary beneficiary of the user group is the firm Macromedia.

Concluding Remarks

This chapter articulates the modes of design adopted by users, experts, and institutions in creating new artifacts. User design is a tantalizing prospect by which users create unique artifacts to address their own needs. Yet, expert design and common design remain prevalent modes. The choice of a particular mode is driven by the comparative advantage of experts, by economies of scale in design and production, and by the transaction costs of engaging experts, features that remain the foundations of modern economic life. However, emergent processes and technologies such as rapid prototyping and design grammars can alter the economics of mode choice.

Notes

[1] An exception is perhaps a *design study* done in isolation by a professional designer, but even in this case the designer typically contemplates a virtual user. Design without a user seems to me to be more individual art rather than true design.

[2] Some artifacts can be decomposed into a platform and derivatives, with the platform a common artifact and the derivative a unique artifact. In a subsequent section, we discuss hybrid modes of design, which can arise in such cases.

[3] Goldenberg and Mazursky (1999) make a compelling argument that what they call “templates” (actually closer to a grammar in my nomenclature) can be used to characterize successful designs for advertisements and new product concepts.

References

Ericsson, K. Anders. 1996. “The Acquisition of Expert Performance: An Introduction to Some of the Issues.” In *The Road to Excellence: The Acquisition of Expert Performance in the Arts and Sciences, Sports, and Games*, ed. K. Anders Ericsson, 1–50. Mahwah, NJ: Erlbaum.

Franke, Nikolaus, Martin Schreier, and Ulrike Kaiser. 2010. “The ‘I Designed It Myself’ Effect in Mass Customization.” *Management Science* 56 (1): 125–140.

Goldenberg, Jacob, David Mazursky, and Sorin Solomon. 1999. “Creative Sparks.” *Science* 285 (5433): 1495–1496.

Lancaster, Kelvin. 1990. “The Economics of Product Variety: A Survey.” *Marketing Science* 9 (3): 189–206.

Randall, Taylor, Christian Terwiesch, and K. T. Ulrich. 2005. “User Design of Customized Products.” *Marketing Science* 26 (2): 268–280.

Stiny, G., and W. J. Mitchell. 1978. “The Palladian Grammar.” *Environment and Planning B* 5 (1): 5–18.

Terwiesch, Christian, and Christoph H. Loch. 2004. “Collaborative Prototyping and the Pricing of Custom Designed Products.” *Management Science* 50 (2): 145–158.

Terwiesch, Christian, and K. T. Ulrich. 2009. *Innovation Tournaments: Creating and Selecting Exceptional Opportunities.* Boston, MA: Harvard Business Press.

Ulrich, K. T. 1995. The Role of Product Architecture in the Manufacturing Firm. *Research Policy* 24 (3): 419–440.

Ulrich, K. T., and David J. Ellison. 2005. “Beyond Make-Buy: Internalization and Integration of Design and Production.” *Production and Operations Management* 14 (3): 315–330.

Ulrich, K. T., and Steven D. Eppinger. 2011. *Product Design and Development*. New York: Irwin/McGraw-Hill.

von Hippel, Eric. 1988. *The Sources of Innovation*. Oxford: Oxford University Press.

SIX

The Architecture of Artifacts

Architecture most commonly refers to the art or science of creating edifices. However, in this chapter I use the term to refer to the organizational structure of an artifact, and more specifically to the arrangement of its *function* and *structure*. This is the sense in which we speak of computer architecture or the architecture of an automobile. For physical artifacts, structure is composed of physical components. For software, services, or other intangible artifacts, structure comprises the intangible building blocks—routines, processes, code—used to assemble the artifact.

I define the architecture of an artifact more precisely as (1) the arrangement of functional elements; (2) the mapping from functional elements to components; and (3) the specification of the interfaces among interacting components. These notions can become abstract quite quickly, so in this section I explain the key points using the example of a simple vehicle trailer. After illustrating the idea of the architecture of an artifact, I articulate the implications of architecture for issues that matter to designers, producers, and users of artifacts.[1]

The arrangement of functional elements

The function of an artifact is what it *does* as opposed to its structural characteristics (Crilly 2010). There have been several attempts in the design theory community to create formal languages for describing function (Finger and Dixon 1989), and there have been modest successes in narrow domains of application such as electro- and fluid-mechanical systems and digital circuits (Mead and Conway 1980). There have also been efforts to create informal functional languages to facilitate the practice of design (Fowler 1990; Hubka and Eder 1988). These languages are sometimes used to create diagrams consisting of functional elements, expressed as linguistic terms like *convert energy*, connected by links indicating the exchange of signals, materials, forces, and energy. Some authors of informal functional languages provide a vocabulary of standard functional elements, while others rely on users to devise their own. Functional elements are sometimes called *functional requirements* or *functives*, and these diagrams have variously been called *function diagrams*, *functional descriptions*, and *schematic descriptions* (Pahl and Beitz 1984). Here I will call the arrangement of functional elements and their interconnections a *function diagram*. An example of a function diagram for a trailer is shown in Exhibit 6-1.

Function diagrams can be created at different levels of abstraction. At the most general level, the function diagram for a trailer might consist of a single functional element: expand cargo capacity. At a more detailed level, the function diagram could be specified as consisting of the collection of functional elements shown in Exhibit 6-1, that is, connect to vehicle, protect cargo from weather, minimize air drag, support cargo loads, suspend trailer structure, and transfer loads to road.

As they are expressed in more detail, function diagrams embody more assumptions about the physical working principles on which the artifact is based. For example, "expand cargo capacity" does not assume the trailer will be a device towed over the road (the trailer could be a lighter-than-air device), although the more detailed function diagram shown in Exhibit 6-1 does embody this assumption. For this reason, two products that at the most general level do the same thing may have different function diagrams when described at a more detailed level. While most functional elements involve the exchange of signals, information, materials, forces, and energy, some elements do not interact with other functional elements. An example of such an element might be "harmonize aesthetically with vehicle."[2]

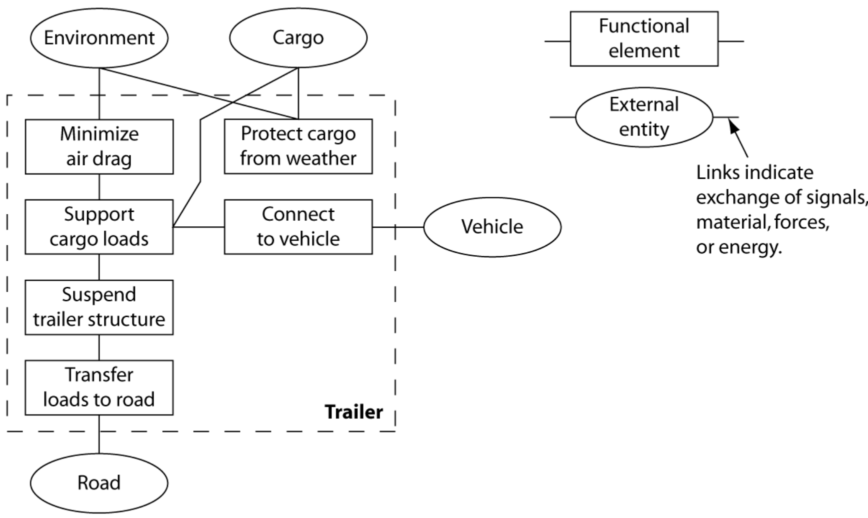


Exhibit 6-1. A function diagram for a vehicle trailer.

The mapping from functional elements to components

The second part of the artifact architecture is the mapping from functional elements to components. An artifact consists of one or more components. For clarity, I define a component as a separable part or subassembly. However, for many of the arguments in the chapter, a component can be thought of as any distinct region of the artifact, allowing the inclusion of, for example, a software subroutine in the definition of a component. Similarly, distinct regions of an integrated circuit, although not actually separate physical parts, could be thought of as components. Components implement the functional elements of the product. The mapping between functional elements and components may be one-to-one, many-to-one, or one-to-many. Two different trailer designs and their associated mappings of functional elements to components are shown in Exhibits 6-2 and 6-3.

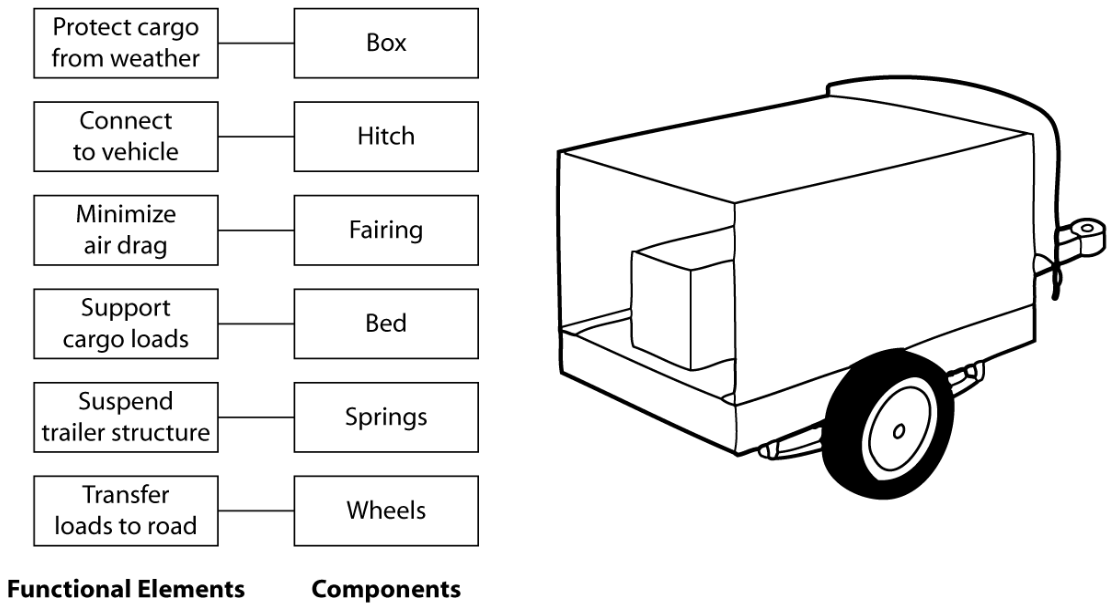


Exhibit 6-2. An example of a one-to-one mapping between functional elements and components and an associated trailer.

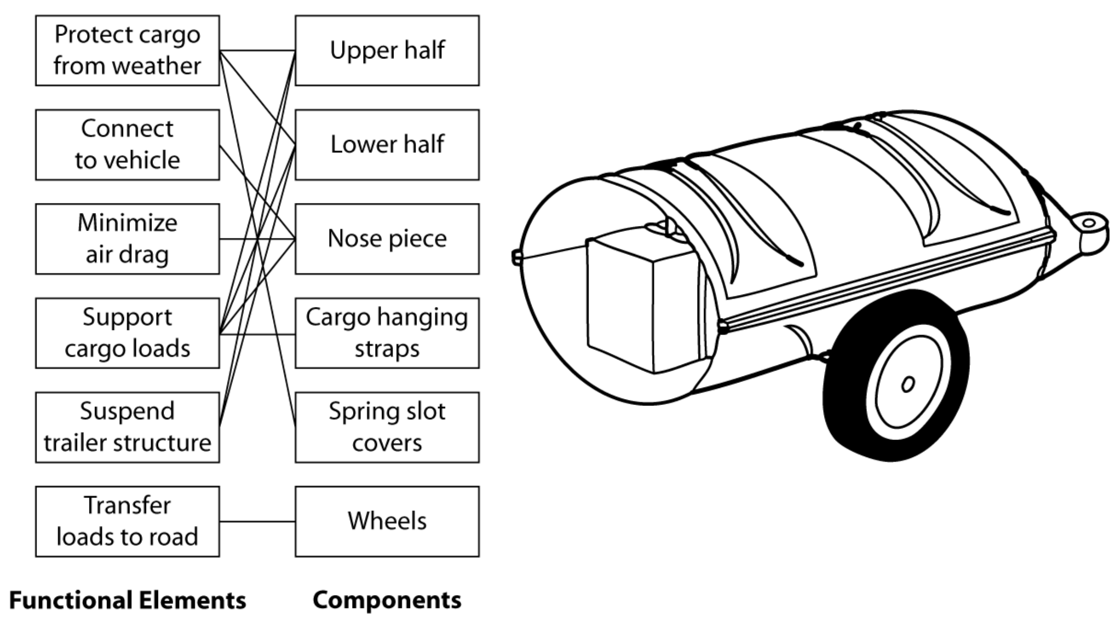


Exhibit 6-3. An example of a complex mapping between functional elements and components and an associated trailer.

The specification of the interfaces between interacting components

By definition, interacting components are connected by some interface. Interfaces may involve geometric connections between two components, as with a gear on a shaft, or may involve noncontact interactions, as with the infrared communication link between a remote control and a television set. An interface specification defines the protocol for the primary interactions across the interfaces, and the mating geometry in cases where there is a geometric connection.

For example, one of the interfaces for the trailer shown in Exhibit 6-2 is between the box and the bed. The specification of the interface includes the dimensions of the contact surfaces between the two components, the positions and sizes of the bolt holes, and the maximum force the interface is expected to sustain. Note that interfaces may be specified to adhere to a standard protocol. Examples of protocols that have been standardized across many different manufacturers' products are USB (universal serial bus), tire/rim standards for automobiles, a 3.5-millimeter audio jack for headphones, a garden hose connection thread, and a "ball-type" trailer hitch. Manufacturers sometimes choose to adopt a common protocol for interfaces used within their own product line, even though the interface may not adhere to an external standard.

A Typology of Architectures

A typology of architectures provides a vocabulary for discussing the implications of the choice of architecture for the user and producer. The first distinction in the typology is between a *modular* architecture and an *integral* architecture. A modular architecture includes a one-to-one mapping from functional elements in the function diagram to the components, and specifies decoupled interfaces between components. An integral architecture includes a complex (not one-to-one) mapping from functional elements to components and/or coupled interfaces between components.

Types of mappings from functional elements to components

The two trailers in Exhibits 6-2 and 6-3 illustrate two extreme examples of mappings from functional elements to components. One trailer embodies a one-to-one mapping between functional elements and components. Assuming that the component interfaces are decoupled (more on this later), this trailer has a modular architecture. In the field of software engineering, the notion of module cohesion or strength is similar to the one-to-one mapping of functional elements to components (Schach 1990). The other trailer embodies a mapping in which several functional elements are each implemented by more than one component, and in which several components each implement more than one functional element (a complex mapping). This trailer has an integral architecture. The phenomenon of a single component implementing several functional elements is called *function sharing* in the design theory community and is described in detail by Ulrich and Seering (1990). To some extent, whether or not functional elements map to more than one component depends on the level of detail at which the components and functional elements are considered. For example, if every washer, screw, and filament of wire is considered a component, then each functional element will map to many components. To more precisely define what a one-to-one mapping between functional elements and components means, consider an artifact disassembled to the level of individual piece parts. (This level of disassembly is sometimes called the *iota* level.) In general, many possible subassemblies[3] could be created from these iota parts. If there is a partitioning of the set of iota parts into subassemblies such that there is a one-to-one mapping between these subassemblies and functional elements, then the artifact exhibits the one-to-one mapping characteristic of a modular architecture.

Interface coupling

In addition to one-to-one mappings, modular architectures include *decoupled* component interfaces. Two components are coupled if a change made to one component requires a change to the other component in order for the overall artifact to work correctly. Two physical components connected by an interface are almost always coupled to some extent; there is almost always a change that can be made to one component that will require a change to the other component. (For example, arbitrarily increasing the operating temperature of one component by 1000°C will require a change to nearly any imaginable neighboring component.) However, in practical terms, coupling is relevant only to changes that modify the component in some useful way. (See Schach 1990 for a detailed discussion of the different types of coupling encountered in software.) Exhibit 6-4 illustrates an example of an interface between two components, the bed and the box from the trailer in Exhibit 6-2. The coupled interface embodies a dependency between the thickness of the bed and the vertical gap in the box connection slot. The decoupled interface involves no such dependency. For the coupled interface, when the thickness of the bed must be changed to accommodate a change in the cargo load rating, the box must change as well. Although the example in Exhibit 6-4 is geometric, coupling may also be based on other physical phenomena, such as heat or magnetism.

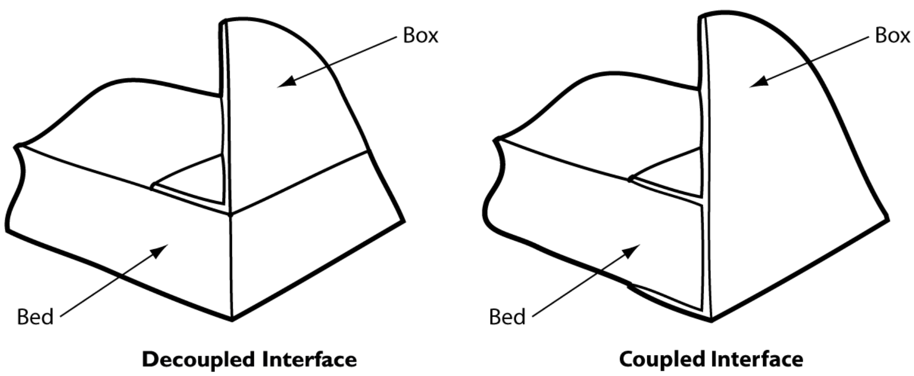


Exhibit 6-4. Two hypothetical designs for the interface between the box and bed of a trailer, one decoupled, the other coupled

Types of modular architectures

I divide modular architectures into three subtypes: slot, bus, and sectional. Because each of the three subtypes is modular, each embodies a one-to-one mapping between functional elements and components, and the component interfaces are decoupled; the differences among these subtypes lie in the way the component interactions are organized.

Slot

Each of the interfaces between components in a slot architecture is of a different type from the others, so that the various components in the artifact cannot be interchanged. An automobile radio is an example of a component in a slot architecture. The radio implements exactly one function and is decoupled from surrounding components, but its interface is different from any of the other components in the vehicle (e.g., radios and speedometers have different types of interfaces to the instrument panel).

Bus

In a bus architecture, there is a common bus to which the other physical components connect via the same type of interface. A common example of a component in a bus architecture is an expansion card for a personal computer. Nonelectronic products can also be built around a bus architecture. Track lighting, shelving systems with rails, and adjustable roof racks for automobiles all embody a bus architecture. The bus is not necessarily linear; I also include components connected by a multidimensional network in the bus subtype.

Sectional

In a sectional architecture, all interfaces are of the same type and there is no single element to which all the other components attach. The assembly is built up by connecting the components to each other via identical interfaces. Many piping systems adhere to a sectional architecture, as do sectional sofas, office partitions, and some computer systems.

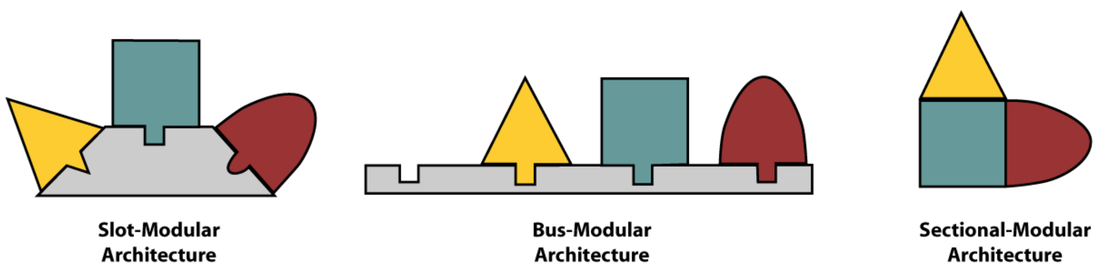


Exhibit 6-5. Three types of modular architecture. Source: Adapted from Ulrich and Eppinger 2011.

Examples

The next several exhibits illustrate this typology for the trailer example, for a desk, and for a variety of other artifacts. I intend for the typology to provide a vocabulary for describing different artifact architectures. The types shown are idealized; most real products exhibit some combination of the characteristics of several types. Products may also exhibit characteristics of different types depending on whether one observes the artifact at the level of the overall final assembly or at the level of individual parts and subassemblies.

A producer can design and manufacture artifacts without ever explicitly creating an architecture or even a function diagram. In the domains of software and electronic systems, the idea of a function diagram (labeled as a schematic, flowchart, etc.) is prevalent in industrial practice. However, the notion of a function diagram has only recently been disseminated in many mechanical domains. If an architecture is explicitly established during the development process, this step usually occurs during the system-level design phase of the process after the basic technological working principles have been established, but before the design of components and subsystems has begun.

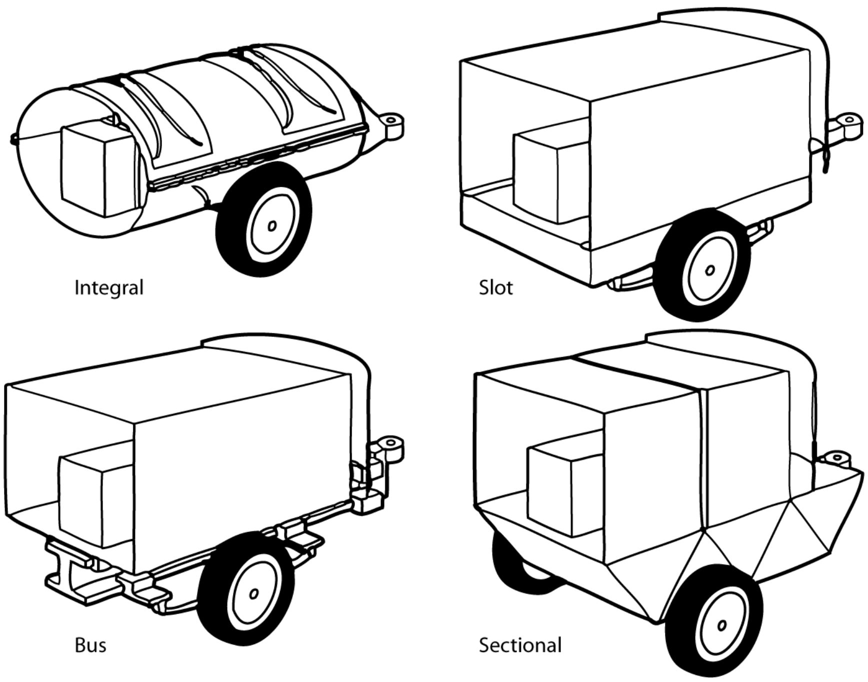


Exhibit 6-6. Four stylized trailers representing four different architectural choices.

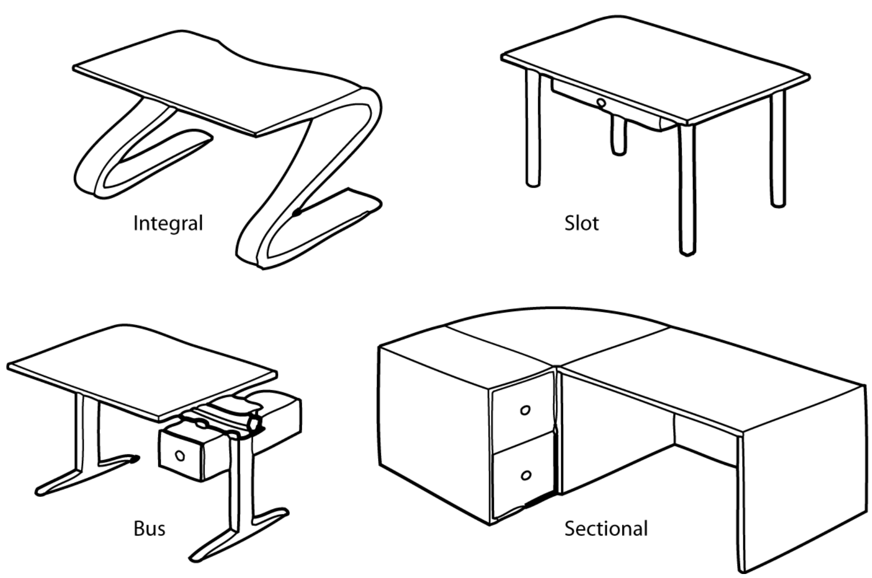


Exhibit 6-7. Four stylized desks representing four different architectural choices.

The examples in the exhibits suggest that designers possess substantial latitude in choosing an architecture, although the architecture of many existing products may be less the result of deliberate choice and more that of incremental evolution. Several scholars have prescribed a modular architecture as ideal. For example, Alexander (1964) presents an "optimal" design methodology, ensuring a lack of coupling between components.[4] I maintain that while artifact architecture is extremely important, no single architecture is optimal in all cases. The balance of the chapter discusses the potential linkages between the architecture of the artifact and a set of issues of technical, economic, and managerial importance. Recognizing and understanding these linkages is a prerequisite to the effective choice of architecture for a particular product.

Exhibit 6-8. A bus-modular architecture for a knife. Any tool from the set of possible tools can be added as a new slice to the "sandwich," a form of bus architecture. Source: Wenger.

Exhibit 6-9. Shimano pioneered the integration of the controls for shifting and braking with a gripping location for the rider's hands. This is an integral architecture, arising both from a complex mapping from functional elements to components and from coupled interfaces. Source: Shimano.

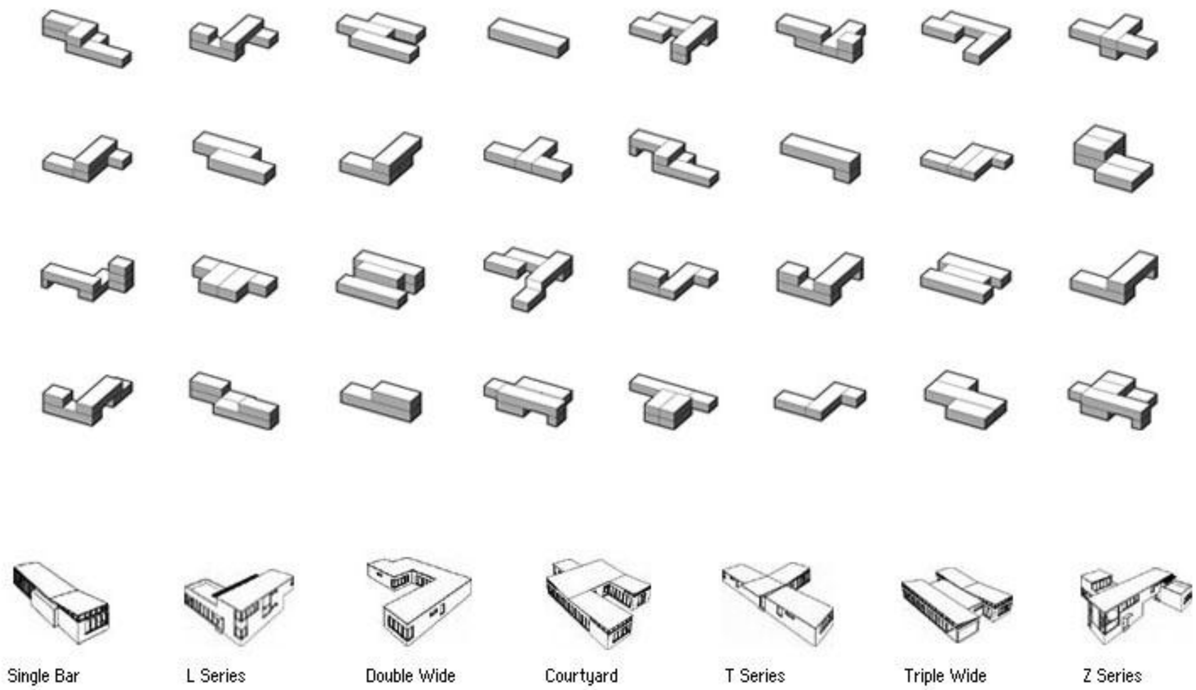


Exhibit 6-10. Some of the alternative forms of housing that can be created from a sectional-modular architecture. Source: Resolution 4 Architecture (http://www.re4a.com).

Artifact Change

This section focuses on two types of artifact change: change to a particular artifact over its life cycle (e.g., replacing a worn tire) and change to a product or model over successive generations (e.g., substituting the next-generation suspension system in the whole product line). The next two sections treat two closely related concepts: variety and standardization.

Architecture determines how the artifact can be changed

The minimum change that can be made to an artifact is a change to one component. The architecture of the artifact determines which functional elements of the artifact will be influenced by a change to a particular component, and which components must be changed to achieve a desired change to a functional element. At one extreme, modular architectures allow each functional element of the artifact to be changed independently by changing only the corresponding component. At the other extreme, fully integral architectures require changes to every component to effect change in any single functional element. The architecture of an artifact is therefore closely linked to the ease with which a change to an artifact can be implemented. Here I consider how this linkage manifests itself in implementing change within the life of a particular artifact and in implementing change over several generations.

Change within the life of a particular artifact

Artifacts frequently undergo some change during their lives. Some of the motives for this change are:

- **Upgrade.** As technological capabilities or user needs evolve, some artifacts can accommodate this evolution through upgrades. Examples include changing the processor board in a printer and replacing a pump in a cooling system with a more powerful model.
- **Add-ons.** Many artifacts are sold by a manufacturer as a basic unit to which the user adds components, often produced by third parties, as needed. This type of change is common in the personal computer industry (e.g., the addition of third-party mass storage devices to a basic computer). See Langlois and Robertson (1992) for a thorough description of several such cases.
- **Adaptation.** Some long-lived artifacts many be used in several different use environments, requiring adaptation. For example, machine tools may have to be converted from 220V to 440V power. Engines may have to be converted from a gasoline to a propane fuel supply.
- **Wear.** Physical features of an artifact may deteriorate with use, necessitating replacement of the worn components to extend its useful life. For example, one can replace vehicle tires, most rotational bearings, many appliance motors, and dull blades in nondisposable razors.
- **Consumption.** Some artifacts consume materials that are replaceable. For example, copiers and printers frequently contain toner cartridges, glue guns use glue sticks, torches contain gas cartridges, and watches are powered by batteries.
- **Flexibility in use.** Some artifacts can be configured by the user to exhibit different capabilities. For example, many cameras can be used with different lens and flash options, some boats can be used with several awning options, and some fishing rods accommodate several rod-reel configurations.

In each of these cases, changes to the artifact are most easily accommodated through modular architectures. The modular architecture allows the required changes that are typically associated with the artifact's function to be localized to the minimum possible number of components.

Although consumption and wear are frequently accommodated through a modular design with replaceable parts, another popular strategy is to dramatically lower the cost of the entire artifact, often through an integral architecture, such that the entire object can be discarded or recycled. For example, disposable razors, cameras, and cigarette lighters have all been commercially successful products, and disposable pens dominate the marketplace. Later, I explain how integral architectures can allow for a lower-cost artifact under certain conditions.

Change across generations of artifacts

When a new model of an existing artifact is introduced, the artifact almost always embodies some functional change relative to the previous version. (In relatively rare cases, a producer changes only the name of the artifact.) The architecture of the artifact has profound implications for a producer's ability to implement this product change. For artifacts with a modular architecture, desired changes to a functional element can be localized to one component. Artifacts with integral architectures require changes to several components in order to implement changes to the artifact's function. The observation helps to explain industrial practice in the area of generational change.

For example, the original Sony Walkman architecture allowed the cassette tape transport mechanism to be reused in many successive models, while the enclosure parts could be easily changed for each new model (Sanderson and Uzumeri 1995). *Virtual design* is a term Sanderson and Uzumeri use for this superposition of several product cycles involving changes to only a few components onto the longer life cycle of a technological platform.

This virtual design is enabled by the modular artifact architecture exhibited by the Walkman at the level of major subassemblies. In some settings, a firm introduces a product, gauges the market response, then develops and launches an incrementally improved product extremely quickly. A modular architecture is essential to being able to quickly change the artifact in this way. The benefits of a modular architecture for exploring a market and fine-tuning an artifact are also described in Langlois and Robertson (1992). Nobeoka and Cusumano (1997), in summarizing several previous studies of the world automobile industry, identify project scope—the percentage of unique components a manufacturer designs from scratch in-house—as a key variable relating to product development performance. The architecture of the product, and the degree of modularity in particular, dictate how much project scope will be required to achieve a particular level of functional change. Change to an artifact is not always confined to activities by a single manufacturer. In some markets, such as home entertainment, users create virtual products by assembling collections of products provided by diverse manufacturers. Modularity at the level of the entire system, when combined with standard interfaces, allows for the virtual artifact to evolve and change through independent actions by individual manufacturers (Langlois and Robertson 1992; Fine 1998).

Artifact Variety

For the purposes of this chapter, I define *variety* as the assortment of artifacts that a production system provides to society. (Chapter 8 is a comprehensive treatment of the subject of variety.) High variety can be produced by any system at some cost. For example, an auto manufacturer could create different fender shapes for each individual vehicle by creating different sets of stamping dies, each of which would be used only once. Such a system is technically feasible, but prohibitively expensive. The challenge is to create the desired variety economically.

The ability of a system to economically produce variety is frequently credited to production flexibility. When viewed at the level of the entire production system, this is a tautology—if a system is economically producing variety it is to some extent flexible. However, flexibility is often equated with the flexibility of the individual processes in the production system (e.g., computer-controlled milling machines), or with flexible assembly systems (e.g., programmable electronic chip insertion equipment). In this context, a flexible production process incurs small fixed costs for each output variant (e.g., low tooling costs) and small changeover costs between output variants (e.g., low setup times). This notion of flexibility is consistent with Upton's definition (1994, 73): "the ability to change or adapt with little effort, time, or penalty." I argue that much of a production system's ability to create variety resides not with the flexibility of the processes in the system, but with the architecture of the artifact the system produces. This section shows how both the flexibility of the production process and the artifact architecture interact to contribute to the ability to economically create variety.

Variety is meaningful to users only if the functionality of the artifact varies in some way.[5] This variation may be in terms of the set of functional elements implemented by the artifact (Does the trailer protect the cargo from the environment at all?), or in terms of the specific performance characteristics of the artifact relative to a particular functional element (Is the environmental protection normal or heavy-duty?). Consider the trailer example and a firm that produces trailers for its customers. Assume customers' needs can be neatly divided in the following ways. Some customers want to minimize air drag, some do not. Two types of vehicle connection and three alternatives for the type of environmental protection are desired. Three alternatives are also desired for both the structural load rating and for the ride quality of the suspension system.[6] Under these assumptions, if variety incurred no cost, the firm would offer 108 distinct trailers to the marketplace ($2 \times 2 \times 3 \times 3 \times 3 = 108$). If the firm uses the modular architecture shown in Exhibit 6-2, all of the 108 different trailers can be created from a total of only 12 different types of components: a single type of fairing (which is either included with the trailer or not), two types of hitches, three types of boxes, three types of beds, three types of spring assemblies, and one type of wheel assembly. Because each functional element maps to exactly one physical component, and because the interfaces are decoupled, the variety can be created by forming 108 combinations from a set of 12 component building blocks. I was not the first to observe that variety can be created by combinations of building blocks. In fact, this combinatorial approach to variety is part of a five-step technique called (somewhat confusingly) Variety Reduction Program (Suzue and Kohdate 1990). Nevins and Whitney (1989) also give several examples of such combinatorial assembly of artifact variants, and Pine (1992) popularized the notion of *mass customization*. The modularity of the artifact allows the variety to be created at final assembly, the last stage of the production process. Some firms are even delaying a portion of the final assembly until the artifact has moved through the distribution system and is ready to be shipped to a customer. This strategy has been called *postponement* (Lee and Tang 1997). If the firm wishes to offer all 108 variants and uses the integral artifact architecture shown in Exhibit 6-3, 73 different types of components will be required: 27 types of upper halves, 27 types of lower halves, 12 types of nose pieces, three types of cargo hanging straps, three types of spring slot covers, and one type of wheel assembly. Because in many instances each component implements several functional elements, there must be as many types of each component as there are desired combinations of the functional elements it implements. For example, to provide all of the different desired combinations of the two vehicle connection types, the two types of drag reduction, and the three load ratings, 12 distinct types of nose pieces will be required because the nose piece contributes to all three of the functional elements associated with the options.

Variety and flexibility

At first glance, producing 108 varieties of the integral design appears to be far less economical than for the modular design. In fact, the flexibility of the production process is an additional factor in determining the basic economics of producing variety. If the trailer components can be economically produced only in large lot sizes because of the large setup times required for the process equipment, or if each type of component requires large tooling investments, then in fact the integral design will be very expensive to produce with high variety. High variety under these conditions would require some combination of large inventory costs, large setup costs, or large tooling costs.[7] However, if the integral trailer components could be produced economically in small lots (e.g., setup costs are low) and without tooling investments, then variety could be offered for the integral design.

For example, consider the following production system for the integral trailer. The upper and lower halves are made by a computer-controlled rolling machine followed by a computer-controlled laser cutting machine. Plates of arbitrary thickness and material can be rolled to arbitrary diameters (within certain limits), and slots for the springs can be cut along arbitrary trajectories—all with small setup times, no tooling investment, and rapid processing times. The nose piece is created by laser cutting, computer-controlled rolling, and automated welding. The six components are then assembled manually. Because of the flexibility of the upper half, lower half, and nose piece production processes, the required component types can be produced as they are needed in arbitrary combinations, and then assembled into the required trailer types. Such process flexibility allows economical high-variety production of an artifact with an integral architecture.

Flexible production process hardware can also have an impact on the production of the modular design. Using inflexible processes requiring expensive tooling and large lot sizes, the 12 different components required to assemble the 108 different product variants would be held in inventory ready for final assembly. Alternatively, the components for the modular design could be produced with flexible production equipment, eliminating the need for the inventories and tooling expense. With a modular architecture, variety can be achieved with or without flexible component production equipment. In relative terms, to economically produce high variety with an integral architecture, the component production process must be flexible. This argument assumes in all cases that the final assembly process itself is somewhat flexible; that is, different combinations of components can be easily assembled to create the final product variety. This assumption is usually valid for products assembled manually, but some assembly systems, particularly high-volume automated assembly equipment, violate this assumption. For these systems, the flexibility of the final assembly process is also a key driver of the ability of the firm to offer product variety.

Infinite variety

Many flexible production processes can be programmed to produce an infinite variety of components. For example, a computer-controlled laser cutting system can cut along an arbitrarily specified trajectory. This flexibility allows systems incorporating these processes to create artifacts that can be infinitely varied with respect to one or more properties. This ability to continuously vary the properties of components by a flexible process provides a subtle distinction between the variety that can be created by assembling artifacts from a finite set of component alternatives, and the variety that can be created by flexible component production processes. Assembly from finite component choices is fundamentally a "set operation," in that it allows sets to be formed from discrete alternatives. Continuously variable process equipment can implement arbitrary mathematical relationships among component characteristics. For example, the laser cutting machine could be programmed to cut along a curve parameterized as a function of a set of other characteristics, such as expected climate of the use environment, the types of loads the trailer will carry, and the road quality in the customer's geographical region. Note that the ability to arbitrarily vary component characteristics can be achieved for both integral and modular architectures if components are fabricated with programmable processes.

A summary of the effect of architecture and component process flexibility on the resulting performance characteristics of the production system is shown in Exhibit 6-11.

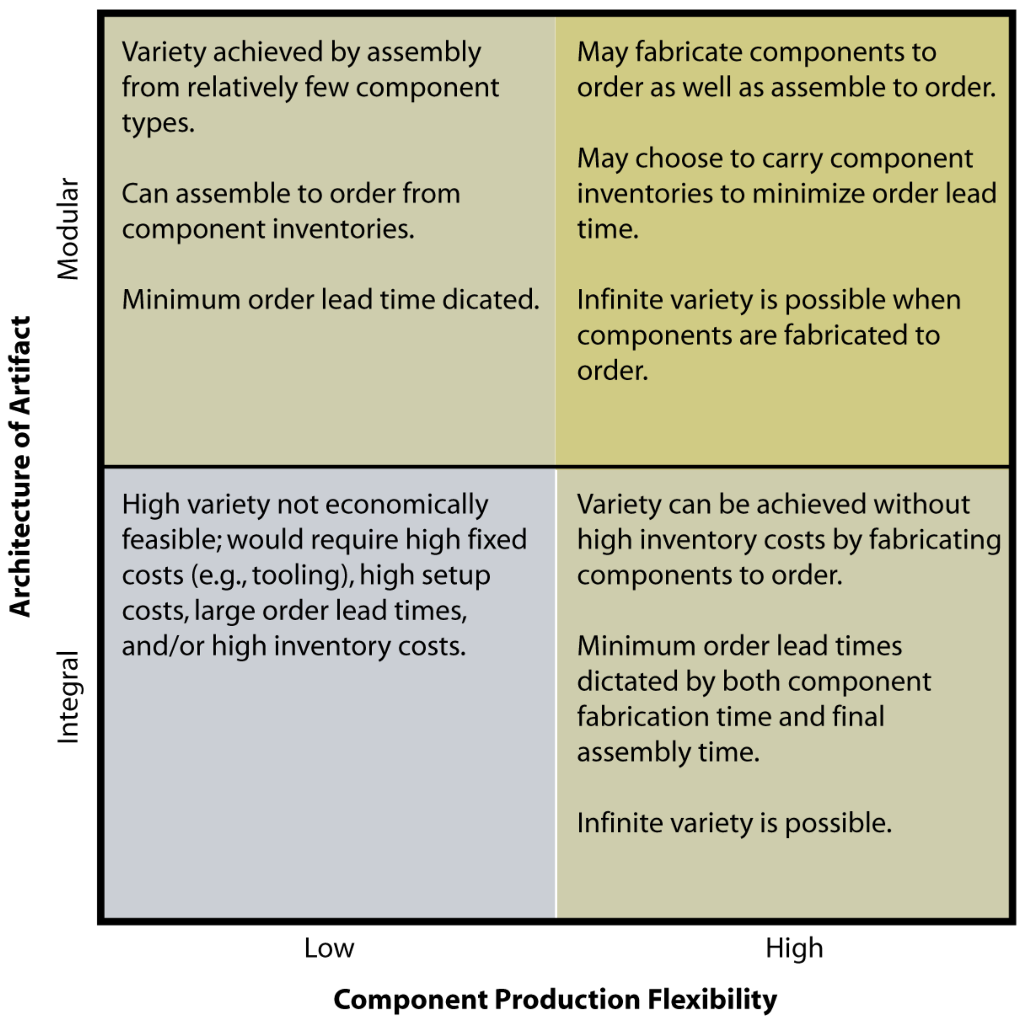


Exhibit 6-11. The relationship between component production process flexibility, the architecture of an artifact, and the ability to deliver variety.

Component Standardization

Component standardization is the use of the same component in multiple versions of an artifact and is closely linked to variety. Common standardized components include tires, batteries, bearings, motors, lightbulbs, resistors, and fasteners. Component standardization occurs both within a single entity (e.g., Quad4 engines at General Motors) and across multiple entities (e.g., Timken roller bearings at Ford, General Motors, and Daimler). I call the first case *internal standardization* and the second case *external standardization*. For internal standardization, components may be designed and manufactured within the entity or provided by suppliers. For external standardization, components are typically designed and manufactured by suppliers.

A modular architecture makes standardization possible

Standardization can arise only when (a) a component implements commonly useful functions and (b) the interface to the component is identical across more than one different product. Otherwise, a component would either not be useful in more than one application or would not match the interface of more than one application. A modular architecture increases the likelihood that a component will be commonly useful. When the mapping from functional elements to components is one-to-one, each component implements one and only one function. Such components are therefore useful in any other applications where their associated functions occur. Components of an artifact exhibiting an integral architecture would potentially be useful only in other artifacts containing the exact combination of functional elements, or parts of functional elements, implemented by the component. A modular architecture also enables component interfaces to be identical across several products. Interfaces in modular architectures are decoupled—that is, a particular component will not have to change when surrounding components are changed. Therefore, different sets of surrounding components, such as might occur in different applications, do not require different component interfaces. When interfaces are decoupled, an interface standard can be adopted and the same component can be used in a variety of settings.

What are the implications of standardization?

Component standardization, whether external or internal, has implications for the producer in the areas of cost, performance, and development. Under most circumstances a standard component is less expensive than a component designed and built for use in only one artifact. This lower cost is possible primarily because the standard component will be produced in higher volume, allowing greater economies of scale and more learning. Higher component volume may also attract several competitors who exert price pressure on one another. However, there are some circumstances under which the use of a standard component may incur higher unit costs than the use of a special component. Sometimes, in an effort to standardize, firms will use a component with excess capability for a particular application. For example, a standard enclosure may be slightly larger than necessary in a particular application, or a standard power supply may provide slightly more power than is strictly necessary in a particular application. In these cases, firms may choose to adopt the standard components even if their unit cost is higher than that of a component more closely matched to the application. This standardization may be justifiable because of the economic savings from reduced complexity in, for example, purchasing, inventory management, quality control, or field service.

Standard components, in general, exhibit higher performance (for a given cost) than unique designs. This performance advantage arises from the learning and experience the component supplier is able to accumulate. However, standardization may act as an inertial force preventing firms from adopting a better component technology because of compatibility issues in the installed base of products.

The use of standard components can lower the complexity, cost, and lead time of design and development. An existing standard component represents a known entity and therefore can reduce the number of uncertain issues the development team must cope with. An existing standard component also requires no development resources and so can lower both the cost and, if the component development would have been on the project's critical path, the lead time of a project.

Artifact Performance

I define *performance* as how well the artifact implements its functional elements. Typical performance characteristics are speed, efficiency, life, and noise. Performance, as defined here, excludes economic performance, except to the extent that it arises from noneconomic dimensions of performance, because economic performance is also highly dependent on the firm's production, service, sales, and marketing activities.

All physical artifacts occupy space, exhibit some shape, and are composed of materials with mass and other physical properties. Performance characteristics tied closely to the size and mass of an artifact typically are compromised by modular architectures. To minimize size, mass, and variable cost, designers adopt integral architectures. Nonphysical artifacts like software may exhibit performance characteristics somewhat analogous to those related to size and mass—for example, memory requirements or lines of code.

For most physical artifacts, several key performance characteristics are closely related to size and shape and/or to mass. For example, acceleration relates to mass, aerodynamic drag relates to size and shape, and, in the trailer example, vehicle fuel efficiency relates to size and shape as well as to mass. In most cases, increasing overall performance involves decreasing size and mass. (In relatively rare cases, increasing performance involves *increasing* size and mass; improving the holding power of a boat anchor or increasing the passenger comfort of an automobile may be such cases.)

Three design strategies are frequently employed to minimize mass or size: function sharing, geometric nesting, and part integration. *Function sharing* is a design strategy in which redundant physical properties of components are eliminated through the mapping of more than one functional element to a single component (Ulrich and Seering 1990). For example, a conventional motorcycle contains a steel tubular frame distinct from the engine and transmission. In contrast, several high-performance motorcycles contain no distinct frame. Rather, the cast aluminum transmission and motor casing acts as the structure for the motorcycle. For example, consider the BMW R1100S motorcycle shown in Exhibit 6-12. The motorcycle designers adopted function sharing as a means of exploiting the fact that the transmission and motor case had incidental structural properties that were redundant to the structural properties of the conventional frame. Through function sharing, the designers minimized the mass of the frame/motor/transmission system. In exploiting the secondary structural properties of the motor and transmission case, the designers mapped more than one functional element to a single component and thereby created an integral architecture.

Geometric nesting is a design strategy for efficient use of space and material and involves the interleaving and arrangement of components such that they occupy the minimum volume possible or, in some cases, occupy a volume with a particular desired shape. For example, the wheel, suspension, fender, and brake system of a modern automobile are arranged in a way that barely allows clearance for wheel travel; they are tightly nested. An unfortunate consequence of nesting is the coupling of the interfaces between components, the other hallmark of an integral architecture. For example, in an automobile the brake system cooling is tightly coupled to the shape of the wheel well, the wheel covers, and the fenders. A slight change to the shape of the wheel cover can require substantial changes to the brake disc design. Similarly, the road and wind noise from the wheels is coupled in a complex way to the shape of the wheel well and fender. Thus, a desire for increased global performance in the area of drag and aesthetics leads to a design strategy of geometric nesting. This design strategy causes components to be coupled, thereby sacrificing the modularity of the architecture.

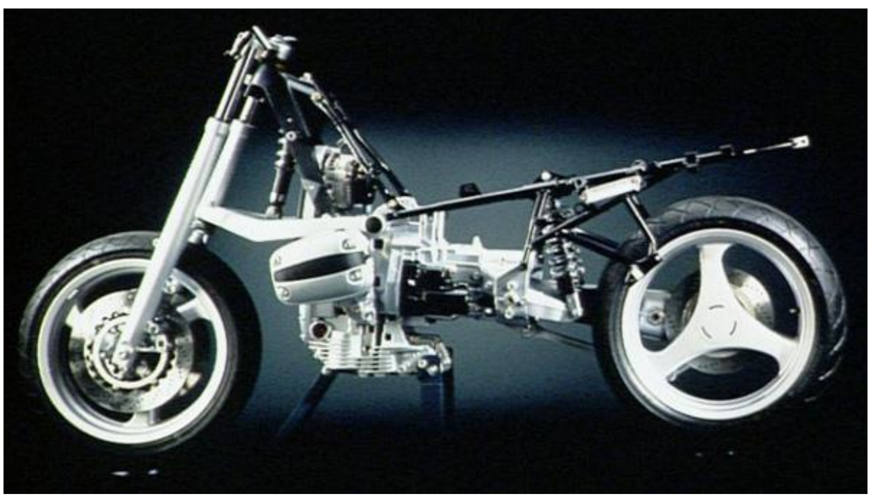


Exhibit 6-12. The BMW R1100S motorcycle includes a transmission component that not only transmits power from the engine to the rear wheel, but also acts as a key structural element for the frame and suspension. Source: BMW.

A similar argument applies to the *part integration* that is a common strategy in design for manufacturing and a common motive for integral architectures (Ulrich and Eppinger 2011). Part integration, or the combination of multiple parts into one contiguous part, minimizes the use of material and space associated with component interfaces, and may improve geometric precision, but compromises the one-to-one mapping from functional elements to components.

Integral architectures and variable cost

Minimizing size and mass is also part of a strategy for minimizing unit production costs for high-volume products, because as production volumes increase, materials costs become more and more significant. This explains why integral architectures are sometimes employed to achieve very low unit costs, such as are required for disposable products like ballpoint pens, razors, and single-use cameras.

The examples in this section illustrate extreme conditions. Most artifacts will embody hybrid modular-integral architectures. For example, although the high-performance motorcycle may exhibit little modularity in the architecture of the engine, transmission, and frame, the architecture of the ignition system may be quite modular (e.g., spark plug, wiring, coil, etc.). The designers of the motorcycle have avoided modularity only where the performance penalties are most severe.

Note that what may be considered a component of one artifact is itself the end product for the supplier of that component (whether the supplier is internal or external). As a result, the component itself may be designed with a highly integral architecture, but then may be used in a highly modular way as part of a larger system. For example, tires exhibit a highly integral architecture, but may be used as a component in a trailer with a highly modular architecture.

Management of the Design and Development Process

At a basic level the design and development process for complex artifacts can be viewed as consisting of four phases: concept development, system-level design, detailed design, and testing and refinement (Ulrich and Eppinger 2011). The architecture of the product has implications for the effectiveness of approaches to the three development phases following concept development. The following sections discuss these three phases and Exhibit 6-13 summarizes the differences in effective approaches for modular and integral architectures.

System-level design

A modular architecture requires relatively more emphasis on this phase of development than does an integral architecture. For the modular architecture, the focus of system-level design and planning is to carefully define component interfaces, specifying the associated standards and protocols. Performance targets and acceptance criteria are set for each component, corresponding to the particular functional element implemented by the component. Component design is frequently assigned to specialists, either internal or external to the enterprise. The development team leader can be viewed as a "heavyweight system architect." For the integral architecture, system-level design absorbs relatively less effort. The focus is on establishing clear targets for the performance of the overall system and on dividing the system into a relatively small number of integrated subsystems. These subsystems are frequently assigned to multidisciplinary teams who will share the responsibility for designing the components that make up the subsystem. The leader of these teams can be viewed as a "heavyweight system integrator."

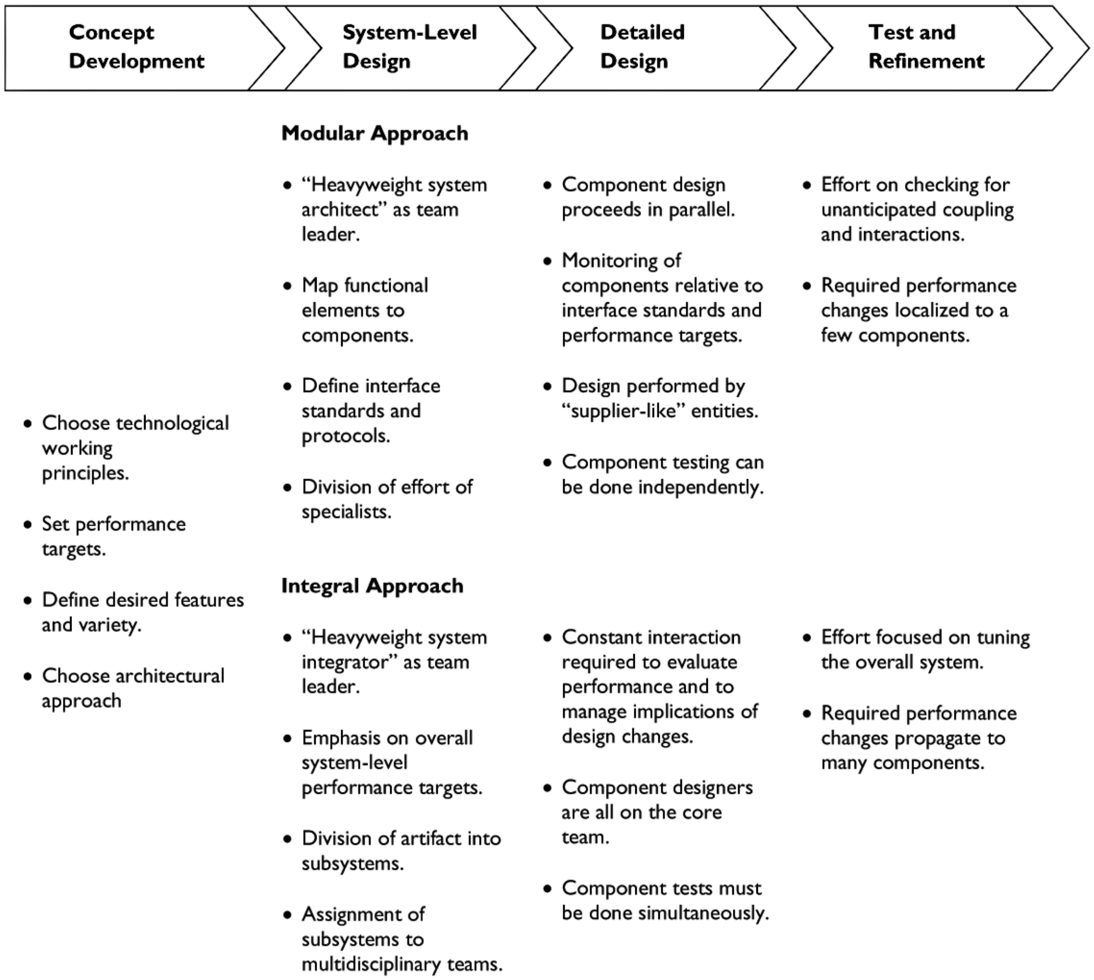


Exhibit 6-13. The design and development process benefits from different approaches, depending on the architectural choices made during the concept development phase.

Detailed design

For the modular architecture, detailed design of each component can proceed almost independently and in parallel. Management of the detailed design process consists of monitoring the progress of each individual component design activity relative to the component performance targets and interface specifications. The component design teams are "supplier-like" in that interaction is structured and relatively infrequent. Testing of each component can be performed independently and clear objectives define completion of each component design activity. For the integral architecture, component designers all form a "core team" and interact continually to analyze performance of the subsystem to which their component belongs and to manage changes required because of component interface coupling. Whether the components meet their performance targets depends on their interaction and not on whether they meet some prespecified criteria. Testing of components cannot be completed in isolation; subsystems of components must be assembled and tested as a whole.

Test and refinement

For the modular artifact, testing and refinement is a checking activity. The tests are intended to detect unanticipated interactions among the components. These interactions are viewed as "bugs" and their resolution is usually localized to changes to one or two components. For the integral artifact, testing and refinement is a tuning activity. If the artifact performance must be altered in some way, changes are likely to be required to many components. Relatively more time will be spent in this phase than for the modular artifact.

Organizational Implications

There are at least three organizational issues tied to a choice of architectural approach: skills and capabilities, management complexity, and the ability to innovate. Highly modular designs allow institutions to divide their design and development organizations into specialized groups, each with a narrow focus (Sanchez and Mahoney 1996). This organizational structure may also extend to the supplier network. If the function of a component can be precisely specified and the interface between the component and the rest of the artifact is fully characterized, then the design and production of that component can be assigned to a separate entity. Such specialization may facilitate the development of deep expertise relative to a particular functional element and its associated component (Fixson and Park 2008).

Required project management skills are different for different architectures. Modular architectures may require better systems engineering and planning skills, while integral architectures may require better coordination and integration skills.

Organizations with a long history of a particular architectural approach are likely to have developed the associated skills and capabilities. A modular architecture enables a bureaucratic approach to organizing and managing development. This approach, for relatively well understood technologies, allows the complexity of the development process to be dramatically reduced and may allow for better exploitation of supplier capabilities. For some domains the benefits of reduced complexity and enhanced supplier involvement may drive the choice of the architecture for at least portions of the artifact; software development is one such domain. In most cases the system-level performance penalties of a modular architecture are dwarfed by the benefits of a reduction in project management complexity. A potential negative implication of a modular architecture is the risk of creating organizational barriers to architectural innovation. These barriers appear to be unfortunate side effects of focus and specialization. This problem has been identified by Henderson and Clark (1990) in the photolithography industry and may in fact be of concern in many other industries as well.

How to Establish the Architecture

Dozens of issues are linked to the architecture of an artifact. The net effect is a complex set of relations among many areas of concern. While there are currently no deterministic approaches to choosing an optimal architecture, the process can be guided. In most cases the choice will not be between a completely modular or completely integral architecture, but rather will be focused on which functional elements should be treated in a modular way and which should be treated in an integral way. Listed here are questions the designer can ask in order to raise the important issues and to guide the development of an appropriate architecture. These questions are best posed during the concept development phase of the design and development process. These questions also serve as a summary of the linkages between architecture and the areas of managerial concern described in this chapter.

Artifact change

- Which functional elements are likely to require upgrade?
- Are third-party add-ons desirable?
- Which functional elements may have to be adapted to new use environments over the life of the product?
- Which functional elements will involve wear or consumption?
- Where will flexibility in configuration be useful to the user?
- Which functional elements can remain identical for future models of the product?
- Which functional elements must change rapidly to respond to market or technological dynamics?
- Which variants of the artifact are desirable to best match variation in user preferences?

Artifact variety

- What level of flexibility of the component process is available or easily obtained?
- How much advantage does minimizing order lead time for custom designs provide?

Component standardization

- Are existing components available internally or externally for any of the functional elements of the artifact?
- What are the cost implications of sharing a component with another version of the artifact?
- Where can adopting a standard component reduce development time or complexity of project management?

Artifact performance

- Which performance characteristics are closely linked to size, mass, and shape? Does high performance with respect to these characteristics require an integral architecture?

Design and development management

- How much focus and specialization is present in the organization and in the supplier network?

- Is the artifact inherently large and complex?
- Is the development team geographically dispersed?
- Are barriers to architectural innovation developing in the organization because of specialization?
- Has the organization demonstrated an ability to change in structure and style?

Concluding Remarks

In this chapter, I define the concept of the architecture of an artifact as the arrangement of its function and structure. The architecture of an artifact has deep implications for several issues of technical, managerial, economic, and organizational importance. Several of these implications are shown in Exhibit 6-14 for each of four types of architecture within the typology introduced here. Although the architecture of an artifact often evolves in an ad hoc way, it can also be deliberately chosen as part of the conceptual and system-level design process. The careful choice of architecture allows designers to achieve several objectives beyond the direct satisfaction of user needs.

<table>
<tr><th></th><th>Integral</th><th>Modular-Slot</th><th>Modular-Bus</th><th>Modular-Sectional</th></tr>
<tr><td rowspan="3">Definition</td><td>Complex mapping functional elements to components.</td><td colspan="3">• One-to-one mapping between functional elements and components.
• Interfaces between components are not coupled.</td></tr>
<tr><td>And/or component interfaces coupled.</td><td>Component interfaces all different.</td><td colspan="2">Component interfaces all the same.</td></tr>
<tr><td></td><td></td><td>Single component (bus) links other components.</td><td></td></tr>
<tr><td rowspan="5">Examples</td><td>Automobile body.</td><td>Truck body/frame.</td><td></td><td></td></tr>
<tr><td>Neon sign/lighting.</td><td>Lamp with bulb and shade.</td><td>Track lighting.</td><td></td></tr>
<tr><td></td><td></td><td>Shelf brackets and rails.</td><td>Stackable shelving units.</td></tr>
<tr><td>“Boom box” stereo.</td><td>Consumer component stereo.</td><td>Professional rack-mounted audio equipment.</td><td></td></tr>
<tr><td>Tanker ship (hull in particular).</td><td>Tractor-trailer.</td><td></td><td>Freight train.</td></tr>
<tr><td>Artifact Change</td><td>Change in functionality requires change to many components.</td><td colspan="3">Functional changes can be made to a product in the field.
Manufacturers can change the function of subsequent models by changing a single component.</td></tr>
<tr><td rowspan="2">Artifact Variety</td><td rowspan="2">Variety not feasible without flexible component production processes.</td><td colspan="3">Artifacts can be assembled in a combinatorial fashion from a relatively small set of component building blocks to create variety.
Variety possible even without flexible component production.</td></tr>
<tr><td colspan="2">Variety confined to the choices of components within a predefined overall structure.</td><td>Variety in overall structure of the artifact possible (e.g., Lego blocks, piping).</td></tr>
<tr><td>Component Standardization</td><td></td><td colspan="3">Components can be standardized across a product line.
Firms can use standard components provided by suppliers.
Interfaces may adhere to an industry standard.</td></tr>
<tr><td rowspan="2">Artifact Performance</td><td rowspan="2">May exhibit higher performance for holistic characteristics like drag, noise, and aesthetics.</td><td colspan="3">May facilitate local performance.
Decoupling interfaces may require additional mass and space.
One-to-one mapping prevents function sharing—the simultaneous implementation of more than one functional element by a single component.</td></tr>
<tr><td></td><td colspan="2">Standardized interfaces may result in additional redundancy and physical “overhead.”</td></tr>
<tr><td>Development Management</td><td>Requires tight coordination of design tasks.</td><td colspan="3">Design tasks can be cleanly separated, thus allowing the tasks to be completed in parallel.
Specialization and division of labor possible.
Architectural innovation may be difficult.
Requires the top-down creation of a global artifact architecture.</td></tr>
</table>

Exhibit 6-14. Summary of the key implications of architectural choice on issues of technical, managerial, economic, and organizational importance.

Notes

[1] Most of the ideas in this chapter first appeared in Ulrich and Tung 1991 and Ulrich 1995.

[2] In Chapter 3 I distinguish between overall function and the subordinate qualities of an artifact (i.e., user needs). Here, I consider all of the desirable attributes of an artifact as *functional elements*.

[3] A subassembly is a collection of components that (1) can be assembled into a unit and (2) can be subsequently treated as a single component during further assembly of the product.

[4] Alexander went on to do remarkable work in design theory, but somewhat dismissed his earliest attempts at formal approaches.

[5] Functionality, in this context, is used broadly to mean any attribute of the artifact from which the user derives a benefit, and so would include, for example, styling or color changes.

[6] Assume for the purpose of the example that the type of suspension and the load rating are independent choices. In practice, these two functional elements may in fact be related.

[7] Inventory costs and setup costs can be traded off against one another; inventory can be minimized by using small lot sizes, but this leads to high setup costs.

References

Alexander, Christopher. 1964. *Notes on the Synthesis of Form*. Cambridge, MA: Harvard University Press.

Crilly, Nathan. 2010. "The Roles That Artefacts Play: Technical, Social and Aesthetic Functions." *Design Studies* 31 (4): 311–344.

Fine, Charles H. 1998. *Clockspeed: Winning Industry Control in the Age of Temporary Advantage*. New York: Perseus Books.

Finger, Susan, and John R. Dixon. 1989. "A Review of Research in Mechanical Engineering Design." *Research in Engineering Design* 1: 121–137.

Fixson, Sebastian K., and Jin-Kyu Park. 2008. "The Power of Integrality: Linkages Between Product Architecture, Innovation, and Industry Structure." *Research Policy*. 37 (8): 1296–1316.

Fowler, T. C. 1990. *Value Analysis in Design*. New York: Van Nostrand Reinhold.

Henderson, Rebecca M., and Kim B. Clark. 1990. "Architectural Innovation: The Reconfiguration of Existing Artifact Technologies and the Failure of Established Firms." *Administrative Science Quarterly* 35: 9–30.

Hubka, Vladimir, and W. Ernst Eder. 1988. *Theory of Technical Systems*. New York: Springer-Verlag.

Langlois, R. N., and Paul L. Robertson. 1992. "Networks and Innovation in a Modular System: Lessons from the Microcomputer and Stereo Components Industries." *Research Policy* 21: 297–313.

Lee, H. L., and Christopher S. Tang. 1997. “Modeling the Costs and Benefits of Delayed Product Differentiation.” *Management Science* 43 (1): 40–53.

Mead, Carver, and Lynn Conway. 1980. *Introduction to VLSI Systems*. Boston, MA: Addison-Wesley.

Nevins, James L., and Daniel E. Whitney. 1989. *Concurrent Design of Products and Processes: A Strategy for the Next Generation in Manufacturing*. New York: McGraw-Hill.

Nobeoka, K., and M. A. Cusumano. 1997. “Multiproject Strategy and Sales Growth: The Benefits of Rapid Design Transfer in New Product Development.” *Strategic Management Journal* 18 (3): 169–186.

Pahl, G., and Wolfgang Beitz. 1984. *Engineering Design*, ed. Ken Wallace. London: The Design Council.

Pine II, B. Joseph. 1992. *Mass Customization: The New Frontier in Business Competition*. Cambridge, MA: Harvard Business School Press.

Sanchez, R., and Joseph T. Mahoney. 1996. “Modularity, Flexibility, and Knowledge Management in Product and Organization Design.” *Strategic Management Journal* 17: 63–76.

Sanderson, Susan, and Mustafa Uzumeri. 1995. “Managing Product Families: The Case of the Sony Walkman.” *Research Policy* 24 (5): 761–782.

Schach, Stephen R. 1990. *Software Engineering*. Homewood, IL: Aksen Associates/Irwin.

Suzue, Toshio, and Akira Kohdate. 1990. *Variety Reduction Program: A Production Strategy for Product Diversification*. Cambridge, MA: Productivity Press.

Ulrich, K. T. 1995. “The Role of Product Architecture in the Manufacturing Firm.” *Research Policy* 24: 419–440.

Ulrich, K. T., and S. D. Eppinger. 2011. *Product Design and Development*. New York: Irwin/McGraw-Hill.

Ulrich, K. T., and Warren P. Seering. 1990. “Function Sharing in Mechanical Design.” *Design Studies* 11 (4): 223–234.

Ulrich, K. T., and Karen Tung. 1991. “Fundamentals of Product Modularity.” In *Proceedings of ASME Winter Annual Meeting Symposium on Design and Manufacturing Integration*, Atlanta: 73–79.

Upton, David M. 1994. “The Management of Manufacturing Flexibility.” *California Management Review* 36 (2): 72–89.

SEVEN

Aesthetics in Design

The aesthetic response to an artifact is the immediate feelings evoked when experiencing that artifact via the sensory system. I consider aesthetic responses to be different from other judgments in at least three ways. Aesthetic response is *rapid*, usually within seconds of exposure to the artifact. Aesthetic response is *involuntary*, requiring little if any expenditure of cognitive effort. Aesthetic response is an *aggregate assessment* biased either positively (e.g., beauty or attraction) or negatively (e.g., ugliness or repulsion) and not a nuanced multidimensional evaluation.

By contrast, consider the response to a new mutual fund. While the financial service may be quite appealing and preferred over other alternatives, this assessment of preference is likely the result of a deliberate analytical process over an extended time period and will probably include a balancing of elements of like and dislike. The response to the fund takes significant time, requires effort, and it is multidimensional, and so for my purposes is not an aesthetic response.

An aesthetic response is most frequently stimulated by visual information, largely because the vision system provides data more immediately and at higher rates than do the other senses. Nevertheless, aesthetic responses can be stimulated via senses other than vision. For example, consider the varied responses to the sound of a recording of Aretha Franklin; the feel of a warm whirlpool; the taste of a chocolate truffle; the smell of spoiled meat; the acceleration of a roller coaster in a sharp turn.

We typically think of the aesthetic qualities of an artifact as distinct from its function. Two different hammers might perform the task of driving nails equally well and yet they may evoke different aesthetic responses in the user. Why, then, does aesthetics matter in design?

Let me cite three reasons, giving a preview of a theory of aesthetics to follow. All other things equal, most users will prefer a beautiful artifact to an ugly one, even in highly functional domains such as scientific instruments. Thus, beauty can be thought of as "just another attribute" in a user's evaluation of preference, alongside durability, ease of use, cost, and safety. In this respect, the aesthetic quality of an artifact is an important factor in providing a satisfying user experience, the prime motive for design.

Second, the aesthetic response to an artifact is usually the *first* response to the artifact. First impressions matter, and overcoming an initial aesthetic repulsion is a substantial challenge for the designer, better avoided in the first place.

Third, beauty may serve as a signal for unobservable attributes of quality, much as a brand does for products and services. In such cases, beauty itself is less important than what else the observer may infer from an exhibition of beauty.

So far I have avoided the question of why one artifact may be perceived as more beautiful than another. This question has been posed more generally for centuries by philosophers attempting to explain beauty across the domains of art, literature, music, landscapes, architecture, and the human body. Eighteenth-century philosophers David Hume and Immanuel Kant (Hume 1757; Gracyk 2003) wrote about aesthetics and engaged the fundamental question of the extent to which aesthetic quality is absolute and universal or dependent on context. Although the philosophy and psychology of aesthetic judgments is more nuanced today, this basic tension between universal standards and relative assessment remains prominent. I believe that the most compelling theory of universal aesthetic judgment derives from evolutionary psychology, and I review that perspective here. I then discuss the perspective that aesthetic judgments are derived from specific human experience and cultural context. After providing a brief review of these two perspectives, I synthesize them into the beginnings of a theory of aesthetics for design.

Evolutionary aesthetics

Most significant human adaptations evolved over the past 100,000 generations (2 to 3 million years) and so haven't changed much since the dawn of modern civilization. We live in a modern world, but are equipped with a stone-age mind. The evolutionary perspective is that aesthetic responses must be judgment adaptations that provided reproductive advantage in our ancient past (Thornhill 2003).

The classic example of evolutionary aesthetics is that humans on average find symmetry attractive in potential mates. And in fact, even today, facial symmetry is correlated with reproductive health, so it is plausible that rapidly detecting and being attracted to facial symmetry is an aesthetic judgment adaptation that could have led to relatively higher reproductive success (Thornhill and Gangestad 1993). Evolutionary aesthetics also convincingly explains a wide range of other responses, including an aversion to slithering snakelike objects and a preference for landscapes that provide protection and vantage points. A central tenet of evolutionary aesthetics is that adaptations are shared by essentially the entire species; thus to the extent that an adaptation explains an aesthetic response, it does so universally. (See Dutton 2003 for a nice summary of the key ideas in evolutionary aesthetics.)

On balance, I find quite compelling the idea that we possess many specific adaptations for quickly assessing attractive and repulsive properties of the physical world and that some of these adaptations are likely to be relevant to aesthetic judgments of artifacts. However, the evolutionary perspective cannot yet explain a great many of the interesting characteristics of aesthetic responses exhibited in society today.

Cultural aesthetics

The evidence is overwhelming that many aesthetic judgments differ widely across time and across cultures. As a result, anthropologists and psychologists have sought cultural explanations for aesthetic judgments.

The cultural perspective on aesthetics posits that the ideas prevalent in a social environment influence the aesthetic preferences of individuals within that environment. Therefore, when the environment differs, so do the aesthetic preferences.

One manifestation of cultural phenomena is the emergence of *schools* of design or *design movements*. Perhaps the most influential school of industrial design was the Bauhaus, formed by Walter Gropius in Germany in 1919 (Girard 2003). The central tenet of the Bauhaus was that good design arises from the seamless integration of art and craft. Gropius articulated a set of design principles including “organically creating objects according to their own inherent laws, without any embellishment or romantic flourishes.” One of the most famous designers to emerge from the Bauhaus was Marcel Breuer, whose bookcase from 1931 is shown in Exhibit 7-1. Although the Bauhaus survived less than 15 years, the aesthetic style of functional minimalism is still today broadly influential.

Exhibit 7-1. Bookcase c1931 by Marcel Breuer, a student and teacher at the Bauhaus school. Source: http://classicdesignshop.com/lang-en/accessories-bookcase/386-marcel-breuer-design-furniture-bookcase-s44.htm.

The Memphis movement was formed in 1981 as a consortium of Italian designers led by Ettore Sottsass. The movement was essentially a reaction against modernism, which was to a large extent an outgrowth of the Bauhaus. The Memphis designers produced whimsical, colorful, and even illogical artifacts. An example of Sottsass’s work within Memphis, another bookcase, is shown in Exhibit 7-2.

Exhibit 7-2. The Carlton bookcase c1981 by Ettore Sottsass, the founder of the Memphis group.
Source: http://boijmans.medialab.nl/en/work/V 258 (KN&V).

A theory of aesthetics that seeks to explain the aesthetic appeal of both the Bauhaus and Memphis bookcases seems likely to require cultural insights, in addition to the evolutionary perspective. Despite their apparent differences, the evolutionary and cultural perspectives are not mutually exclusive explanations for aesthetics. In fact, they can be harmonized in a relatively straightforward way as follows.

All aesthetic judgments are implemented by a biological information processing system made up of a collection of evolutionary adaptations. Some fundamental elements of this system are largely invariant across humankind regardless of education, culture, or experience. However, many mechanisms, even if invariant across the species, operate on *symbols* and not on minimally processed sensory inputs, and the values of the symbols on which the mechanisms operate may vary widely (Crilly et al. 2004). Also, many mechanisms are developed, or at least tuned, in a particular individual based on learning and experience.

For example, psychological mechanisms for determining status, prestige, and rank appear to be quite universal, but operate on symbols whose values depend on context. In one setting the symbols associated with status may be derived from body piercing and in another from a large automobile. Although, at this time, the explanatory power of evolutionary aesthetics is relatively weak for settings in which an aesthetic response is highly dependent on social environment, learning, and culture, by recognizing that psychological mechanisms may produce very different aesthetic responses depending on context, both the evolutionary and the cultural theories of aesthetics can be useful and harmonious.

A Theory of Aesthetics in Design

Despite the ambitious section heading, let me state clearly from the outset that I do not have a fully formed and comprehensive theory of aesthetics in design. Nevertheless, I offer some fragments of a theory, which I do think are useful in providing insights and in guiding practice.

The theory comprises these elements:

- The phenomena we lump together into *aesthetic response* are actually the result of many different psychological mechanisms.
- These mechanisms operate on basic sensory inputs and on symbols derived from these inputs and from memory.
- The psychological mechanisms that we consider aesthetic operate very rapidly and may be superseded by a more deliberate formation of preference based on cognitive analysis over longer time periods.
- Some important and significant aesthetic responses are vestigial adaptations for detecting physical features that were useful in an evolutionary sense.
- Other important and significant aesthetic responses are adaptations that operate on symbols derived from learning, experience, and cultural context.

Consider Exhibit 7-3, which is a schematic representation of the theory. We perceive an artifact through a sensory interface. Many psychological processes operate simultaneously, making inferences about the attributes of the object. Some are extremely rapid, detecting light and motion, for example. Others play out over a second or longer, like those detecting shape, symmetry, gloss, and temperature. Psychological processes continue to operate and may invoke symbols from memory. Finally, aesthetic responses may give rise to deliberate analytical thought, which may persist for minutes or longer. An overall preference may be formed within a fraction of a second, but this preference may change as additional information is processed. An initial positive impression may wane, or an initial aversion may turn positive.

Within this theory a sharp distinction between an aesthetic response and an analytical response is a somewhat arbitrary conceptual convenience. The boundary between aesthetics and analytics cannot be sharply drawn. However, I do think that judgments that play out over a few seconds feel qualitatively different from those that may play out in minutes, and certainly from those that operate intermittently over hours and days.

This theory also lets us distinguish between responses that are likely to be universal and those that are likely to be highly dependent on symbols determined from learning, experience, and culture. The most immediate responses are those that are derived from the information processing mechanisms closely tied into the sensory system. Those mechanisms that rely on retrieving symbols from memory are likely to require more time.

Within this overarching theory, let me make five propositions that I think can be useful in explaining aesthetics in design and in guiding practice. Certainly these propositions are incomplete and are yet to be validated empirically. With this disclaimer, here they are.

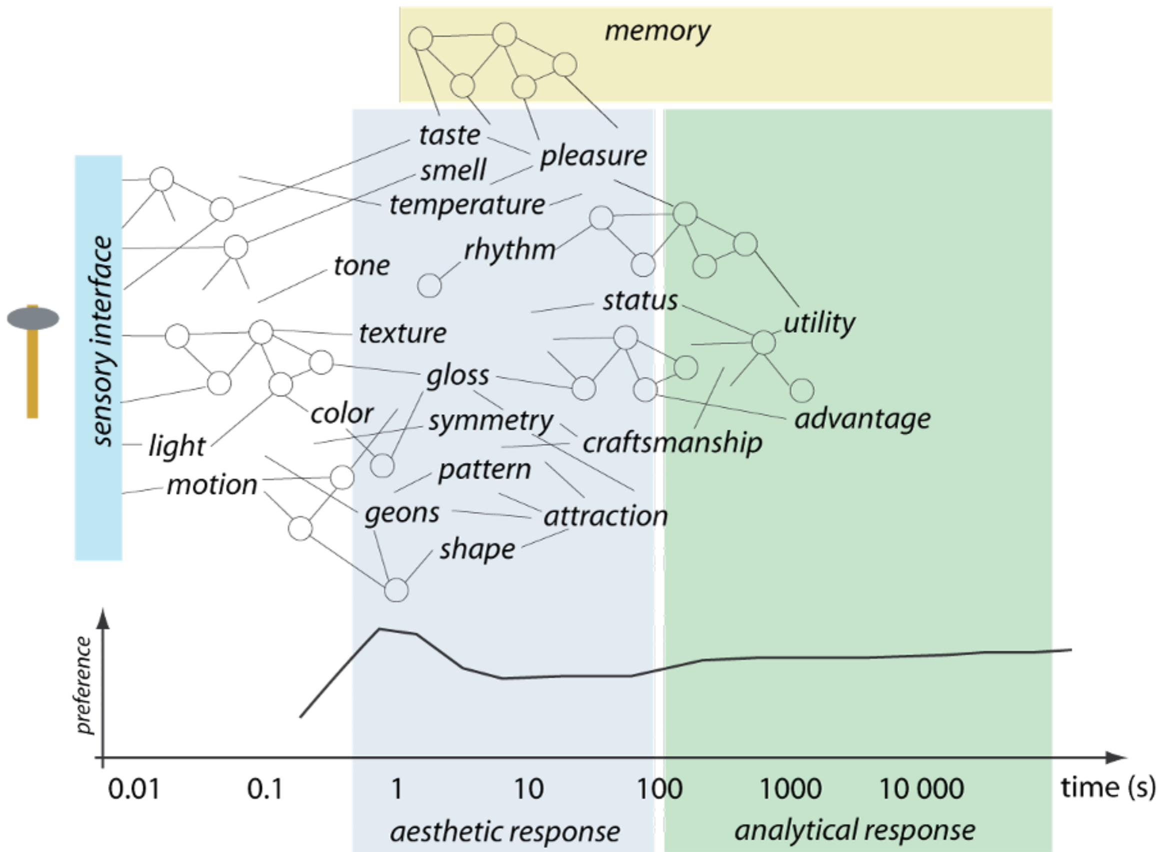


Exhibit 7-3. Schematic illustration of human cognitive response to an artifact (e.g., a hammer) with a hypothetical trajectory of preference as a function of time for a particular individual. Attributes of the object, represented by nodes and labels, are inferred over time based on the sensory inputs, memory, and other attributes.

First impressions matter

Aesthetic responses are immediate and involuntary and they result in the development of preferences. I conjecture that aesthetic responses influence subsequent analytical determination of preference. Specifically, a positive aesthetic response is more likely to lead to a positive ultimate preference than if the initial aesthetic response were negative.[1] Such a phenomenon could be exhibited for at least three reasons. First, and obviously, beauty itself is by definition preferred and so given similar analytical preferences, the beautiful artifact should still be preferred over the ugly artifact. Second, and more subtly, an initially positive aesthetic response may result in a greater chance of further analysis and exploration by the user. A negative aesthetic response may dissuade the user from ever learning more about the artifact and therefore reduces the chance that an ugly, but otherwise preferred, artifact will ever be fully evaluated. Third, I suspect that aesthetic preferences are "sticky." That is, positive aesthetic judgments create a positive bias that persists even in the face of mounting negative analytical evidence. Conversely, negative aesthetic judgments persist even when further analysis reveals highly positive attributes.[2]

The first-impressions proposition could be tested experimentally by providing information about artifacts to human subjects in different sequences and testing whether information relative to aesthetic judgment (e.g., appearance) has a stronger influence on preference when it is presented first than when it is presented after information relative to analytical judgments.[3]

Vestigial adaptations contribute to first impressions

There were no cell phones in our evolutionary past, and yet when we see a cell phone, our stone-age sensory system and aesthetic adaptations are involuntarily invoked. We are not able to command our retinas and visual cortex to evaluate a cell phone differently than it would a stone hand ax. I propose that for most modern artifacts, our most immediate aesthetic responses are vestigial; that is, they are the result of adaptations that were useful in our evolutionary past. However, when applied to modern artifacts, these adaptations do not today confer reproductive advantage. If true, this phenomenon does not make the aesthetic response any less real or any less powerful in determining ultimate preference, so understanding these vestigial adaptations may be usefully exploited in creating artifacts that are attractive.

As far as I know, there are no comprehensive catalogs of vestigial aesthetic adaptations. However, a few adaptations have been clearly articulated and fewer still have been convincingly established empirically (Voland and Grammer 2003). Here I describe two: gloss and cuteness.

Before I provide these examples, let me emphasize what I am *not* claiming. By arguing that there are fundamental vestigial aesthetic adaptations, I am not arguing that these adaptations are always paramount in determining aesthetic preferences. My theory posits that there are hundreds of information processing mechanisms that determine aesthetic response, and that some of these operate on symbols drawn from memory. An immediate vestigial response based on fundamental physical attributes of the artifact such as shape or surface finish could be quickly superseded by a response derived from what those attributes mean to the observer symbolically.

Exhibit 7-4 is a consumer electronic device, the iPod portable music player, created by product designers at Apple Computer. Most people find it attractive. Many explanations are possible, but one element of its attraction is that it is *glossy*. How could the surface finish of an engineered component invoke a vestigial aesthetic response? Coss and coworkers have argued that our brains are hardwired to love reflective surfaces because the only reflective material on the savanna in the Pleistocene was *water*, and water was a scarce and highly valuable substance (Coss and Moore 1990). They further showed that infants will pick up and lick glossy objects more frequently than the same forms with matte surfaces. To me, it is highly plausible that humans possess psychological mechanisms for detecting and rewarding the detection of glossy surfaces, and that these mechanisms are quite fundamental.

Exhibit 7-4. We like glossy objects, perhaps because of hard-wired attraction to water. Source: Apple Inc.

Exhibit 7-5 is a early Volkswagen Beetle automobile. Is there anyone who doesn't immediately find this car cute? How can a car be cute? Why do we like cute inanimate objects? We don't need much imagination to create a theory of cuteness. Babies exhibit certain physical features such as forward-facing eyes and rounded heads that are attractive to adults, who can provide resources and protection for the young. The cute phenomenon could have plausibly evolved to provide reproductive advantage to humans. So powerful are cute features in invoking attraction that our psychological mechanisms are tricked into oohing and ahhing over collections of sheet metal that resemble babies.

Exhibit 7-5. Why is this car cute? Source: Volkswagen Group.

Physics can be aesthetic

I believe that humans possess fast and effective *physics computers*. We are remarkably good at estimating trajectories, predicting imbalance, and sensing strength and rigidity of structures. One can easily imagine why such mechanisms would have been useful in an evolutionary sense. Consider Exhibit 7-6, which shows a walkway over Grand Canyon West. How attractive do you find this walkway? Personally, I want to turn and run back to the minivan. My physics computer does not understand tricky high-strength steel cantilevered structures, and its immediate reaction is that this is an artifact to be avoided.

This is an interesting example of where an initial aesthetic revulsion might be superseded by a higher-order preference. If I thought about the walkway for a few minutes, I would probably conclude that thousands of people had safely walked on it and that the chances of it falling down as I walked on it were pretty slight, probably less than the chances of being hit by a tour bus while crossing the parking lot. At that point, I might actually be attracted to doing something that stimulates my danger avoidance system, an opportunity I don't have very often as a university professor. Nevertheless, I think designers benefit from understanding that humans are likely to be attracted to things that appear safe and stable, and that this perception is based on the physics of pretty ordinary objects made of materials like tree branches and rocks.

Aesthetic features are honest signals of quality

Signals are essential elements of our means of making sense of the world.[4] We use signals to detect whether someone is bored with a joke, to decide whether to stop at a roadside restaurant, to choose a sofa for the living room. The concept of an *honest signal* arises in both evolutionary biology and in economics, and I believe plays a key role in aesthetics. An honest signal is one that is unlikely to be faked by the signaler and therefore can be relied on by the receiver of the signal (Bird and Smith 2005). In nature, the vertical jumping of a gazelle when encountering a lion is an honest signal that the gazelle is fit and can outrun the lion. This is mutually beneficial because the animals can effectively skip the expense of a contest with a predetermined outcome; the gazelle doesn't actually have to run and the lion doesn't actually have to chase. In economic life, agents develop behaviors in response to incentives, and signaling is an important element of this behavior. Spence (1973) showed that an overinvestment in education, say by attending a challenging university, is like the gazelle's leap. The action is a signal of ability that can be relied upon by an employer. Nelson showed that under certain conditions, advertising by a manufacturer can be viewed as an honest signal of product quality (1974).

In order to be *honest*, a signal must be difficult or costly to fake. In economic terms, it must provide more net benefit (benefit minus cost) for a more fit signaler than it does for a less fit signaler. Under these conditions, it is in the fit signaler's interest to provide the signal and the receiver can therefore rely on the signal as a true indicator of fitness.

Mithen (2003) has done a fascinating study of ancient hand axes, possibly the first aesthetic artifacts. Apparently, our ancestors developed an aesthetic preference for highly symmetric, carefully crafted stone hand axes. The leading theory of this aesthetic preference is that beautiful hand axes were honest signals of male fitness. A male who could be directly observed to craft a beautiful hand ax was one who (1) had access to scarce resources like obsidian, (2) had excellent strength, dexterity, and fine motor skills, and (3) could afford to sit and make axes for hours at a time and still survive. The signal is honest in that it is less costly for a fit fabricator to make axes than a less fit fabricator, and so the expenditure of effort to fabricate aesthetic hand axes can be relied on as a signal of fitness.

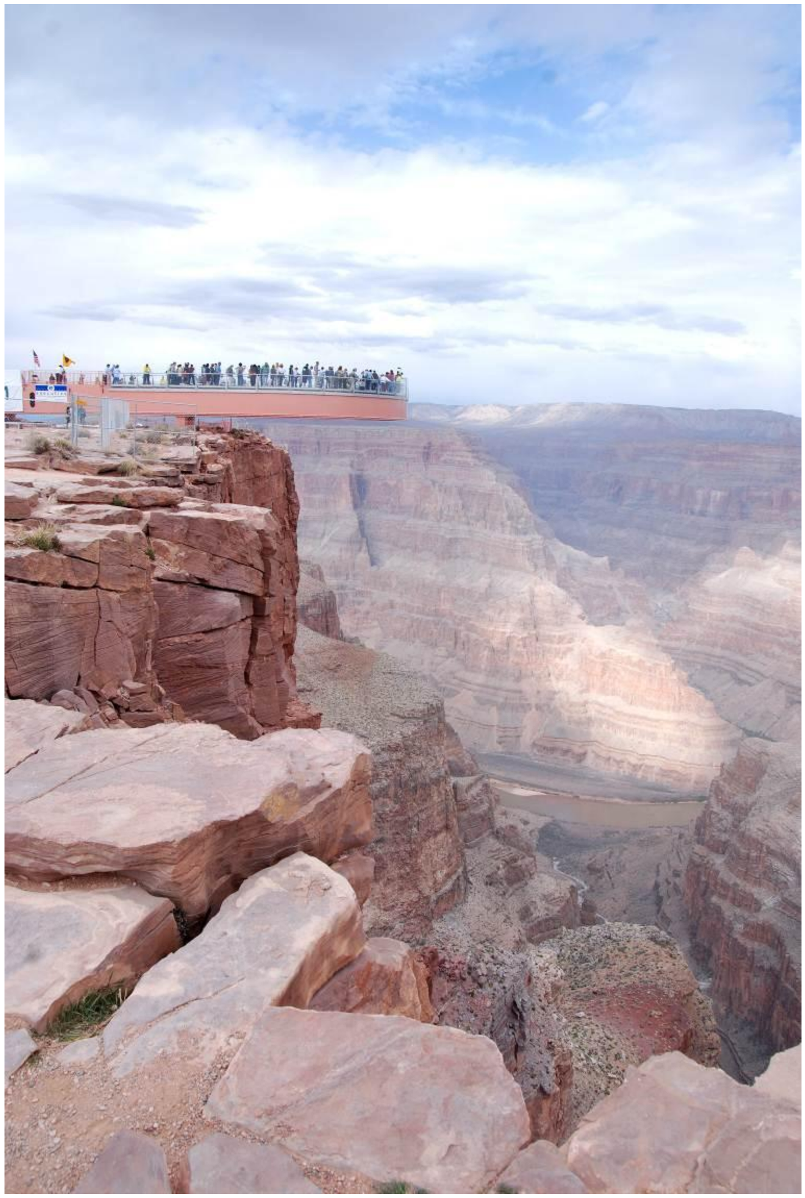


Exhibit 7-6. The cantilevered walkway over Grand Canyon West. What is your aesthetic response? What is your mental physics processor telling you? Source: Best American Destinations, Grand Canyon West / Hualapai Tribe.

In an analogous way, deliberate investment in designing aesthetically pleasing artifacts can be used by producers and consumers as an honest signal of the quality of the artifact. The key idea is that designing beautiful artifacts is costly for a producer. If an artifact is beautiful, it is unlikely it got that way by accident or by trivial imitation. Rather, a designer devoted care and attention to the forms, surfaces, and details of the artifact. In a profit-maximizing setting, the producer who stands to benefit the most from this investment is the one who produces goods that are preferred upon closer inspection and that will deliver long-term satisfaction to the user. In this way, the producer of better products benefits more from positive aesthetics than does the producer of lower-quality products. Thus the development of aesthetic features of artifacts satisfies the requirements of honest signaling.

Artifacts have symbolic value in social systems

Teenagers seem able within seconds to size up a fashion accessory and determine whether or not it is attractive. The aesthetics of fashion are highly dynamic, so it is hard to argue that some intrinsic physical properties of fashion accessories directly determine aesthetic preference. Rather, fashion artifacts must stimulate and invoke symbols in memory that determine the aesthetic response. I am not ambitious enough to try to explain fully such mechanisms, but let me conjecture how one such mechanism might work.

Exhibit 7-7 shows the hip-hop artist 50 Cent wearing huge jewel-studded items of jewelry known (as I write this anyway) as “bling.” My teenage son has a strongly positive aesthetic response to bling. Personally, I don’t get it. Indeed, the fact that I don’t get it may be a key reason my son likes it. A simple set of symbolic relationships seem highly predictive of his aesthetic response: An artifact whose physical attributes (1) invoke an association with a group a teenager admires and (2) invoke a disassociation with the parents will be attractive to the teenager.

Lest I dismiss this response as youthful folly, an almost identical mechanism explains in part why I am attracted to Patagonia brand apparel. I aspire to the dirt-bag, free-spirited culture associated with the brand, and wish to disassociate myself from the Ralph Lauren set. This is such a primitive symbolic aesthetic response that it persists despite the logical analysis that the more accurate association of Patagonia would be with middle-aged affluent professionals. True dirt-bag nomads buy their fleece at Walmart or Goodwill.

Exhibit 7-7. The hip-hop artist 50 Cent wearing his bling. Photographed by Bobin James. Source: www.khachaak.com.

It is easy to imagine other symbolic relationships that could explain aesthetic responses. Most of these relationships operate on symbols whose values are themselves dynamic. A few relatively straightforward relationships could give rise to phenomena that appear complex and dynamic, such as fashion in current society. For instance, Pesendorfer (1995) develops a simple economic model in which the latest fashion apparel is used as a signal in a dating game.

Creating Beautiful Artifacts

Even assuming you are persuaded by my proposed theory of aesthetics of artifacts, I have provided no prescriptions for how one might actually design beautiful artifacts.

We can certainly imagine a design process that can create beautiful artifacts, although perhaps not efficiently. Such a process requires only that we can generate alternatives and that we can evaluate the beauty of those alternatives. In Chapter 4, I discuss exploration in detail, but no great intellectual leap of faith is required to imagine a way to generate alternatives. One could engage a variety of different designers with different approaches, each of whom would generate different designs. One can also imagine a simple, even if costly, approach to evaluation. We could build prototypes of the alternatives, present them to the target user population, and observe which are preferred by the users. In fact, at the macro level of an entire industry or design domain, this is the process by which artifacts may become more attractive over time.

However, an unguided process of generating alternatives and evaluating them through testing in a user population is inefficient. Given a theory of aesthetics, a designer should be able to develop and apply heuristics based on causal relationships in the theory, resulting in the generation of more successful alternatives and a reduced requirement for testing. A sample heuristic is that all else equal, humans assume "normal physics" in evaluating objects, so chairs, tables, and other structural objects are more likely to be attractive if their forms appear to be stable, solid, and strong.

I believe that interested researchers could develop a more complete theory of aesthetics in design. With such a theory, I believe that useful design heuristics could be developed that would be highly effective in educating designers and in guiding practice. About thirty years ago, the architect Christopher Alexander and his collaborators wrote a brilliant book, *A Pattern Language* (1977), which is essentially a collection of heuristics for designing the built environment, some of which are based on thoughtful and careful observation of how humans respond to their buildings and outdoor spaces. Alexander's heuristics are surprisingly easy to apply, and have attracted a passionate following among some designers. For example, this is Alexander's heuristic (or "pattern") 159: "When they have a choice, people will always gravitate to those rooms which have light on two sides, and leave the rooms which are lit only from one side unused and empty" (Alexander et al. 1977: 747).

He goes on to articulate the theory underlying this heuristic, which is in part that light from two sides provides the optimal illumination of other people for detecting subtle expressions and movements, making the rooms conducive to understanding social exchanges.

While expert architects may over a long career develop strong intuition about natural lighting, a heuristic like Pattern 159 is highly useful in guiding a novice.

Practical aesthetics

As an academic, I am optimistic and intrigued by the prospect for a comprehensive theory of aesthetics, which might then be followed by the development of useful heuristics for design. However, as a designer, I know that we are probably decades away from that goal. As a practical matter, heuristics for aesthetics are likely to be of limited use. Rather, we will continue to rely on designers who possess skills, largely tacit, for creating beautiful artifacts.

Design spaces are *rugged*, meaning that incremental iterative improvement of a design is unlikely to result in finding a great solution. Better solutions are likely to be found in territory distant from the starting point. In such environments, we know that parallel exploration using divergent approaches is likely to result in better outcomes. As a result, competitions, simultaneous efforts by members of a design team, and the application of distinct methods for creating alternatives, are likely to be useful exploration strategies.

A substantial problem for designing artifacts that are strongly preferred overall is that the people who are skilled at designing beautiful artifacts may not be those skilled at designing artifacts to achieve other, more purely functional, objectives. One need only spend a few hours in an industrial design studio and then in an engineering lab to realize that the cognitive processes, social systems, and skills and capabilities of these two populations are nearly disjoint.

Nevertheless, when abstracted, the design process is the same. Designers consider a gap, explore alternatives, evaluate alternatives, and iterate. An organizational challenge is to coordinate the exploration and evaluation of alternatives with contributions from individuals who are very different in order to arrive at a design that stands out on many dimensions.

Concluding Remarks

The aesthetic response to an artifact is the immediate feelings evoked when experiencing that artifact via the sensory system. The aesthetic quality of an artifact is important in determining a user's eventual preferences. Theoretical foundations for aesthetics in design are emerging, even if still preliminary and speculative. A theory of aesthetics in design may eventually inform practice, leading to more efficient and reliable creation of attractive artifacts.

Notes

[1] Coates (2003) provides a nice discussion of a version of this idea in his work on "liking and disliking" products.

[2] In psychology this stickiness in preferences is a well-known property of human decision making called the *confirmation bias*.

[3] Carlson et al. (2006) did a similar experiment to test the effect of presenting (nonaesthetic) information about products in different order.

[4] Meaning in design is closely linked to aesthetics. The broader issue of what artifacts mean and how they communicate meaning is the focus of the intellectual area of design semantics (Krippendorff 2006), and has been treated in the marketing community as well (Solomon 1983; Kirmani and Rao 2000; Veryzer and Hutchinson 1998).

References

Alexander, Christopher, Sara Ishikawa, and Murray Silverstein. 1977. *A Pattern Language: Towns, Buildings, Construction*. New York: Oxford University Press.

Bird, Rebecca B., and Eric Alden Smith. 2005. "Signaling Theory, Strategic Interaction, and Symbolic Capital." *Current Anthropology* 46 (2): 221–248.

Carlson, Kurt A., Margaret G. Meloy, and J. Edward Russo. 2006. "Leader-Driven Primacy: Using Attribute Order to Affect Consumer Choice." *Journal of Consumer Research* 32 (March): 513–518.

Coates, Del. 2003. *Watches Tell More Than Time: Product Design, Information, and the Quest for Elegance*. New York: McGraw-Hill.

Coss, R. G., and M. Moore. 1990. "All That Glistens: Water Connotations in Surface Finishes." *Ecological Psychology* 2 (4): 367–380.

Crilly, Nathan, James Moultrie, and P. John Clarkson. 2004. "Seeing Things: Consumer Response to the Visual Domain in Product Design." *Design Studies* 25 (6): 547–577.

Dutton, Denis. 2003. "Aesthetics and Evolutionary Psychology." In *The Oxford Handbook for Aesthetics*, ed. Jerrold Levinson. New York: Oxford University Press.

Girard, Xavier. 2003. *Bauhaus.* New York: Assouline.

Gracyk, Ted. 2003. "Hume's Aesthetics." *The Stanford Encyclopedia of Philosophy*, ed. Edward N. Zalta. http://plato.stanford.edu/archives/win2003/entries/hume-aesthetics.

Hume, David. 1757. "Of the Standard of Taste." In *Four Dissertations.* http://www.mnstate.edu/gracyk/courses/phil%20of%20art/hume%20on%20taste.htm.

Kirmani, Amna, and Akshay R. Rao. 2000. "No Pain, No Gain: A Critical Review of the Literature on Signaling Unobservable Product Quality." *Journal of Marketing* 64 (2): 66–79.

Krippendorf, Klaus. 2006. *The Semantic Turn: A New Foundation for Design*. Boca Raton, FL: CRC Press.

Mithen, Steven. 2003. “Handaxes: The First Aesthetic Artefacts.” In *Evolutionary Aesthetics*, eds. Eckart Voland and Karl Grammer, 261–278. Berlin: Springer-Verlag.

Nelson, Phillip. 1974. “Advertising as Information.” *Journal of Political Economy* 82 (4): 729–754.

Pesendorfer, W. 1995. “Design Innovation and Fashion Cycles.” *American Economic Review* 85 (4): 771–792.

Solomon, Michael R. 1983. “The Role of Products as Social Stimuli: A Symbolic Interactionism Perspective.” *Journal of Consumer Research* 10 (3): 319–329.

Spence, Michael. 1973. “Job Market Signaling.” *Quarterly Journal of Economics* 87 (3): 355–374.

Thornhill, Randy. 2003. “Darwinian Aesthetics Informs Traditional Aesthetics.” In *Evolutionary Aesthetics, e*ds. Eckart Voland and Karl Grammer. Berlin: Springer Verlag.

Thornhill, Randy, and Steven W. Gangestad. 1993. “Human Facial Beauty: Averageness, Symmetry, and Parasite Resistance.” *Human Nature* 4 (3): 237–269.

Veryzer, Robert W., and J. W. Hutchinson. 1998. “The Influence of Unity and Prototypicality on Aesthetic Responses to New Product Designs.” *Journal of Consumer Research* 24 (4): 374–394.

Voland, Eckart, and Karl Grammer, eds. 2003. *Evolutionary Aesthetics*. Berlin: Springer-Verlag.

EIGHT

Variety

Exhibit 8-1 shows a portion of the production of the Model T for one day in 1913 at Ford's Highland Park Factory. Henry Ford supposedly said of the Model T, "You can buy it in any color, as long as it's black." In fact, before 1913 the Model T was available in red, gray, green, and blue. For the thirteen years following 1913, indeed black was the only color. Then, in the last two years of its product life, the Model T was available in 11 colors. Ford's design decision relative to paint colors was the response of a producer to economic factors of both supply and demand. In this chapter, I articulate those factors and use them to explore the use of variety in the design and production of artifacts.

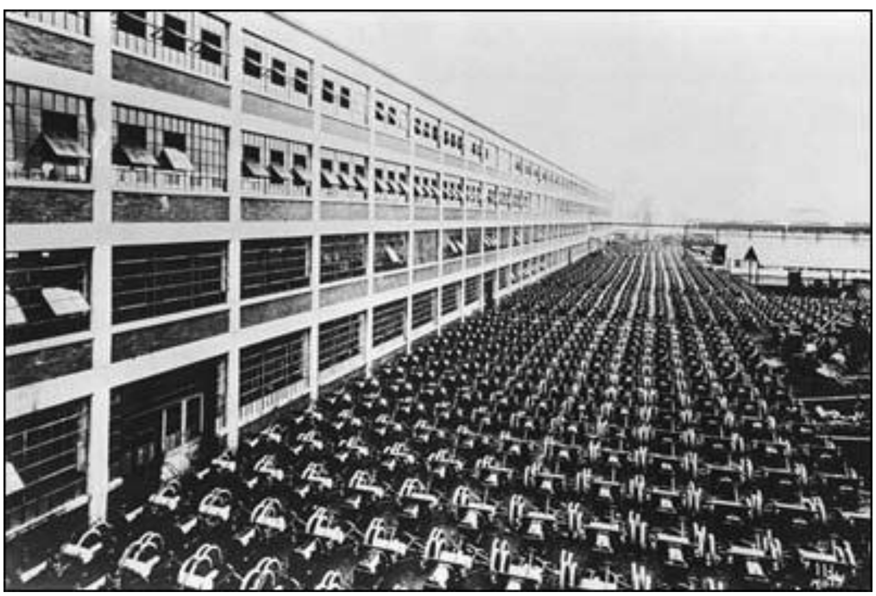


Exhibit 8-1. A single day's production of the Model T at the Ford Highland Park factory in August 1913. Source: Henry Ford Museum.

I use *variety* to refer to the assortment of artifacts that differ with respect to one or more attributes. I focus principally on the variety within a product category available simultaneously from the producers in a marketplace, although variety can also be thought of in terms of the frequency and extent to which a producer changes the artifacts it offers over time. Consider three examples of variety. Exhibit 8-2 shows several T-shirts (the *category*) that differ in their size (an *attribute*). Exhibit 8-3 shows several soft drinks that differ in their formulation. Exhibit 8-4 shows several bicycle cranks that differ in geometry, material properties, and surface finish.

In this chapter, I start by defining three types of variety. I then explain the economic motives for variety and the costs associated with variety. I conclude by providing a framework for designing variety—determining the type and level of variety for a family of artifacts.

Exhibit 8-2. The Hanes Beefy-T shirt is available in several different sizes for a given style, color, and quality level (shown here from S to XXXL). This is an example of *fit* variety.

Exhibit 8-3. A dozen of the many variants of Coke. This is an example of *taste* variety.

Exhibit 8-4. Shimano bicycle cranks are available in several different quality levels for a given size and application type. These four artifacts are all 170 mm 53/39-tooth cranks sold by Shimano at prices from \$60 (bottom) to \$350 (top). This is an example of *quality* variety.

Types of Variety

I categorize variety into three types: *fit*, *taste*, and *quality*. These categories are defined by the way a user's evaluation of an artifact changes as a function of changes in an attribute. Exhibit 8-5 illustrates how a single hypothetical user might value a T-shirt as a function of changes in three different attributes of a shirt. The first attribute is the circumference of the shirt, an element of its size. If the shirt were much too small to wear, it would be useful only as a dust rag. This hypothetical user values the shirt the most if it is 1100 millimeters in circumference, a perfect fit. The user can get by with a shirt a little too small or a little too big, but the value of the shirt falls off steeply as the fit gets too tight or too sloppy. The basic shape of this function characterizes a *fit* attribute. Note that a fit attribute need not refer literally to geometric fit. Rather, fit attributes are those for which the user's preference exhibits a single strong peak for a single value of the attribute, with satisfaction falling off substantially as the artifact diverges from this value. For example, for a software application, a fit attribute might be the operating system with which the application is compatible. For a bicycle crank, a fit attribute might be whether the crank is designed for mountain biking or road biking. Fit attributes are typically easy to measure, characterize, and forecast for the designer and producer, and relatively easy for the user to assess.

The second attribute is the T-shirt's color. For this attribute, the user may have a preference for blues, but also like greens, and the value function may exhibit a lot of peaks and valleys. For this example, color is a taste attribute, an attribute for which the user may have a complex, multimodal response. Preferences for taste attributes are typically much less sharply defined than for fit attributes, and the user may accept as substitutes artifacts with very different values of a taste attribute. I intend *taste* in a broad sense, and not only in the literal sense of flavor. For example, for the bicycle crank, a taste attribute might be the finish on the aluminum surfaces—whether polished or matte.

The third attribute shown is the T-shirt's durability as measured by the number of washing cycles the shirt can withstand before significant degradation. As expected, the user prefers increased durability. One thing that would prevent a user from preferring the most durable shirt might be the shirt's price. In the crank example, the prices vary by a factor of six as quality increases. Note that for most users, satisfaction increases at a decreasing rate as quality approaches a high level. Indeed, one must be a sophisticated cyclist to even detect which crank is considered the highest quality, and I doubt most cyclists could feel any kind of difference in the performance of these cranks.

For completeness and to avoid confusion, note that economists typically divide variety into two categories: *horizontal* and *vertical*. Horizontal variety is essentially what I call taste and fit variety; vertical variety is quality variety. The terms horizontal and vertical variety are not very descriptive, so I prefer to use fit, taste, and quality, which I find both more memorable and more useful conceptually.[1]

Motives for Variety

Variety is the result of decisions made by the producers of artifacts. Producers respond to seven basic economic motives for variety.[2]

Heterogeneous user preferences

Each individual user of an artifact values its attributes differently. In a commercial setting, a user is willing to pay the highest price for an artifact whose fit and taste attributes are at that user's *ideal point*. All else equal, to maximize user satisfaction, a producer would offer an artifact at the ideal point of each potential user. Hotelling (1929) wrote a beautiful paper that provides the seminal conceptual framework for thinking about consumer preferences and variety.

Variation in user experience

Some but not all users seek variety in their experiences over time (Kahn 1995), preferring different breakfast cereals on different days or different hotels on subsequent visits to a city. In a setting in which users seek variety for the intrinsic value of its diversity, producers will offer variety.

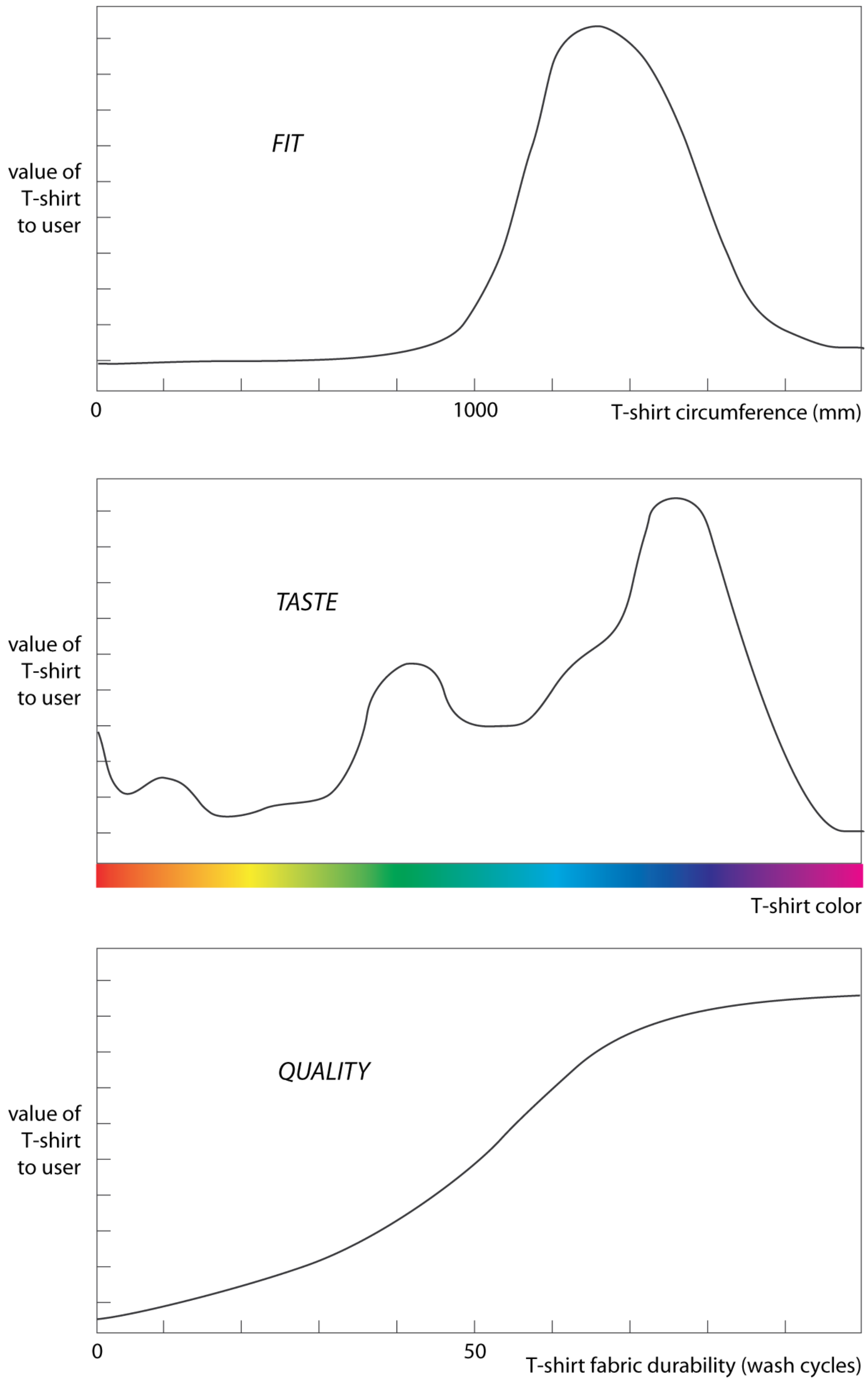


Exhibit 8-5. Illustration of how a hypothetical user's evaluation of a T-shirt might change as a function of changes in fit, taste, and quality attributes.

Sole source to customer

There are costs in time, effort, and money to procure goods and services from multiple sources, and so from the standpoint of convenience, a customer prefers to purchase from a single provider. Of course, the provider reaps benefits from being the sole source as well, including diminished price competition and higher volume of sales. These pressures may lead a producer to offer additional variety, some of which may not be profitable when viewed in isolation, in order to reduce the number of sources of goods and services a customer deals with.

Price discrimination

Different customers exhibit different levels of willingness to pay for quality attributes. Assuming that the profit margins as a percent of price do not diminish with higher prices, a producer would prefer that a high willingness-to-pay consumer buys a higher-quality, higher-price product than a lower willingness-to-pay consumer. This phenomenon leads producers to offer different quality levels of artifacts, often with fairly slight differences in their attributes, but at significantly different prices.

Niche saturation

Existing producers have an incentive to inhibit rivals from entering their markets. An existing product in a niche deters entry by a second firm. As a result, incumbent firms may offer products in small niches, even when the marginal benefit of doing so is not positive, in order to prevent a new entrant from gaining a toehold. Schmalensee (1978) provides a comprehensive discussion of the literature and a theoretical treatment of this phenomenon.

Avoiding price competition

Have you ever tried to find the best price on a new mattress? For a consumer, it's an exercise in frustration. The same producer will offer similar but not identical models through different retailers. At Acme Mattresses, one finds the SoftSleep Excel 2150 and at Beta Mattresses one finds the SoftSleep Delux B150. These mattresses may differ in terms of quilting pattern, number of ties on the springs, and which specific foam is used. However, discerning which is actually preferred is essentially impossible. This use of variety inhibits the consumer's efforts to directly compare prices, allowing Acme and Beta to avoid direct price competition, to offer lowest price guarantees with impunity, and therefore to charge higher prices.

Channel shelf space

Shopping can be a cognitively challenging task. When faced with a shelf of toothpaste options, few consumers will carefully evaluate each alternative, comparing features and benefits. In fact, there is a certain element of randomness to the purchase decision, and so almost anything on the shelf will garner some sales. In fact, holding all other factors constant, sales volume is remarkably proportional to the shelf space allocated to the product. Imagine a shelf in which there are two brands of toothpaste, say Colgate and Crest. Given the shelf-space phenomenon, the producer that adds a second variant, say Minty Crest, will have two thirds of the shelf and, all other things equal, will garner two-thirds of the sales. This action will of course lead to an "arms race" of variety. In fact, Crest toothpaste can be purchased today in about a hundred different formulations (even counting them all exactly is tricky), a figure that does not include variety in packaging and size.

Exhibit 8-6. The toothpaste aisle.

Costs of Variety

The economic motives for variety would quickly push producers to offer infinite variety if there were no associated costs. Indeed, variety incurs two basic types of costs: *reduced scale* and *consumer search costs*.

Reduced scale

Variety erodes scale for producers, and given the ubiquity of economies of scale, will therefore increase production costs. Holding total production quantity constant, if a producer substitutes two similar variants of a product for a single product, total costs will rise. Consider the specific example of the Xootr Mg scooter (a product I designed with my brother Nathan and Jeff Salazar, an industrial designer at Lunar Design). Exhibit 8-7 shows the product, whose central structural element, or *deck*, is a die-cast magnesium part. When we contemplated developing the Mg scooter, we considered offering two versions of the product, one with a wide deck and one with a narrow deck. The different decks represent fit and taste variety; different customers prefer different shapes and sizes.

The die-cast deck is produced by a very large press that brings together two halves of a die (or mold) into which molten magnesium is injected. When the part has cooled and the magnesium solidified, the die is opened and the part is ejected. The process is magnificent in that once the machine is set up, a precise and nearly finished part can be produced once per minute indefinitely, with a batch of about five hundred requiring only a single shift of production.

If the part were produced in two versions, then most costs would increase, including the costs of designing and testing the two versions, the costs of the dies to make two different parts, and the costs of supporting the production and sale of two variants. Even assuming that the two decks use approximately the same amount of magnesium, the unit production costs would also increase with two versions of the product, because the machine would have to be set up and adjusted for two different batches of parts instead of for just one batch. Exhibit 8-9 summarizes the cost comparison.

The costs for the scooter are idiosyncratic to this setting and to this production process. However, virtually all producers of artifacts face economies of scale in their production and delivery processes. Holding all else equal, when variety is increased, the volume per variant is decreased, and therefore the total costs of production increase.

Exhibit 8-7. The Xootr Mg scooter with a die-cast magnesium deck (left) along with computer models of designs contemplated for the deck. The two shapes appeal to different users for reasons of style, comfort, and kicking efficiency.

Consumer search cost

The second cost of increased variety is increased cognitive load on the consumer. When a dinner menu has only one item on it, choosing what to eat is easy. As the number of options increases, the likelihood increases that one of the choices will be pleasant, but the consumer must also invest more and more cognitive effort in identifying relevant alternatives and in making a selection. I call this consumer effort *search cost*. At some point, the increase in search cost may exceed the increase in value derived from additional variety. As variety reaches very high levels, the selection problem may become so painful that the consumer may actually prefer forgoing the product altogether to avoid the agony of the selection process (Iyengar 2010).

As producers develop an increasing ability to offer variety, due to enhanced process flexibility, there is a temptation to offer more variety than can be usefully absorbed by the consumer. With my colleagues Christian Terwiesch and Taylor Randall, I have explored methods for easing the cognitive burden of choosing from among many alternatives using decision support technologies (Randall, Terwiesch, and Ulrich 2007). These methods may serve to diminish the relationship between variety and search cost for consumers.

Exhibit 8-8. Production of the Mg deck requires a die (or mold) as shown in the upper two panels. The raw castings as they come from the die are shown in the lower two panels. The lower right panel shows a production batch of about 500 pieces.

Societal Perspective

Consider the amusing discussion by *Fast Company* magazine of the differences in four of Coca-Cola's offerings (Exhibit 8-10). Coke Zero is a diet cola with no calories and is sold alongside Diet Coke, another diet cola with no calories. The company calls Coke Zero "a new kind of beverage that features real cola taste and nothing else." How critical is it that consumers are now able to enjoy Coke Zero in addition to Diet Coke? Even if the Coca-Cola Company is economically rational in offering a dozen formulations of a diet cola, this action somehow seems wasteful and wrong from a societal perspective.

	Costs of 25,000 decks of one shape	Costs of two shapes with 12,500 of each
Design and testing costs	12,000	16,000
Tooling costs (e.g., dies and fixtures)	40,000	70,000
Material costs and processing costs	675,000 (27.00 per deck)	725,000 (29.00 per deck)
Purchasing, logistics, and inventory costs	6,000	10,000
Marketing communications (e.g., photography, brochures, website)	4,000	5,000
Total costs	**737,000**	**826,000**

Exhibit 8-9. Comparison of costs for 25,000 scooter decks over five years in one versus two variants. Approximate costs in US$.

A moral judgment about variety might include some of the following arguments. Intelligent and creative professionals should be able to find better things to do with their lives than identifying and exploiting micro segments of the carbonated beverage market. As a society we should spend fewer resources on designing, producing, and marketing dozens of different variants of diet colas and more resources on educating children and improving human health. Yvon Chouinard (2005), the founder of Patagonia, writes, "When I die and go to hell, the devil is going to make me the marketing director for a cola company. I'll be in charge of trying to sell a product that no one needs, is identical to its competition, and can't be sold on its merits." Ultimately, moral judgments rest on moral principles, and a particular set of principles may give rise to a particular argument about the moral value of variety. Personally, I'm amused by variety, sometimes confused by it, but do not find variety as morally offensive as, say, the design, production, and purchase of automobiles that weigh three tons and achieve 12 miles per gallon (5 km per liter) of fuel economy.

This chapter has mostly taken the perspective of a single producer responding to various forces to increase or decrease variety. One could also analyze variety from the perspective of the entire product category. There is some empirical evidence that over the life cycle of a product category, variety increases substantially with the entry of new firms and then peaks and declines as the more economically fit firms drive out unprofitable rivals (de Figueiredo and Kyle 2006). This dynamic suggests that from a societal perspective, there may be more variety than strictly necessary to address the heterogeneous needs of consumers.

[LAB TEST]

THE MARKETING GODS MUST BE CRAZY

By Paul Lukas

For a long time, the reason to drink Diet Coke was "Just for the Taste of It." Things are a lot more complex these days as Coke marketers parse demographic segments and create drinks for each niche. There's now a new Diet Coke sweetened with Splenda and Coca-Cola Zero, which, as its name implies, has zero calories—as opposed to the regular and Splenda versions of Diet Coke, both of which have, um, zero calories. And then there's still Coke's original no-cal cola, Tab. All of which leads to some very creative marketing-speak.

		Product	Core demographic	Brand message, as found on Coke.com	Brand message from Katie Bayne, a senior VP, Coca-Cola Brands	Actual brand message, as translated by FAST COMPANY	Flavor profile, according to Scott Williamson, Coca-Cola spokesman	Flavor profile, according to admittedly unscientific FAST COMPANY taste test
DIET COKE		Launched in 1982; sweetened with aspartame **0 calories**	Very broad footprint, with marketing efforts focused on those in their late twenties to early thirties, skewing slightly female	"Diet Coke is your style, it's your sass, it's doing what makes you happy. . . . So flirt, laugh, dance, prance, giggle, wiggle—do what feels good."	"The adult cola taste that uplifts with style—it's a very stylish brand. It's upscale. It's sophistication, but an invitational sophistication."	"Tastes just as good while watching *Sex and the City* reruns as it did while watching the original episodes on HBO."	"According to lore—I've never heard this internally disputed or confirmed—it resembles what used to be New Coke."	Sweet nectar of the gods
DIET COKE w/ Splenda		Launched in May 2005; sweetened with Splenda [sucralose] and acesulfame potassium **0 calories**	30- to 40-year-olds, skewing slightly female	"For those who love the sweet and intense taste of Splenda Brand Sweetener, now there's one more way to enjoy Diet Coke!"	"An adult cola taste, it uplifts with style, and it's sweetened with Splenda, which is a sweetener people say they want. It's that simple."	"Hey, we'll sweeten it with denatured monkey sweat if that's what the carbophobe crowd wants."	"It's meant to mimic Diet Coke. But with Splenda, you will taste a difference, and the Splenda lover loves this new flavor note."	Clean and crisp but a bit short on depth. There's no there there.
COCA-COLA ZERO		Launched June 2005; sweetened with aspartame and acesulfame potassium **0 calories**	18- to 34-year-olds, skewing slightly male	"A new kind of beverage that features real Coca-Cola taste and nothing else. Nothing that could potentially get in the way of your chill."	"It's really the pause that lets them recenter in this fast-paced, time-warped world, and keep going. That's the 'just chill' part of the positioning."	"We're still trying to figure out what those crazy gen-X and gen-Y kids are into, but one thing we're sure of: They don't like the word 'diet.'"	"It's formulated to match regular Coca-Cola."	Sure enough, it really does taste remarkably like Coke.
TAB		Launched in 1963; sweetened with saccharin and aspartame **0 calories**	Urban-sophisticate baby boomers with a sense of ironic kitsch	"Tab has achieved a retro pop-culture status and has the reputation of being somewhat hard to find."	"It's continuing to meet the needs of the small but unbelievably passionate group of people who continue to love Tab, but it isn't actively marketed."	"We can't believe anyone's still buying this stuff."	"It has a strong cola flavor, with that distinctive saccharin sweetness."	Singularly metallic and synthetic in a "You can tell it's a diet drink because it totally makes you lose your appetite" sort of way.

34 FAST COMPANY September 2005

Exhibit 8-10. An analysis of four variants of diet cola offered by Coca-Cola. Source: "The Marketing Gods Must be Crazy" by Paul Lukas in the September 2005 issue of *Fast Company*, page 34.

An economic evaluation of variety could in theory address the question of whether or not variety maximizes social welfare (Lancaster 1975). While variety pursued for the economic motive of addressing heterogeneous user needs is hard to oppose, one could object to variety pursued by the producer to garner additional shelf space or to avoid direct price competition in the sales channel. As with many economic concepts, one must be careful about relying on intuition. It is possible that such actions provide incentives for producers to provide artifacts that better meet user needs. I do not know enough economics, nor have I devoted enough attention to this question, to offer a compelling argument one way or the other. Instead I leave for others the question of the extent to which variety offered by producers is a good thing for society.

Designing Variety

In this book, I address the design of many types of artifacts, including buildings, graphics, services, software, and physical goods. I consider settings ranging from an individual designing for his or her own use to an institution creating products for a large consumer market. The problem of designing variants of artifacts is most prominent in the institutional setting where a team of product designers creates a family of products for a market of many customers. In this section, I assume this context and lay out a framework for making an optimal choice of the level of variety of a product. This framework is simple and static, but is a foil against which I can articulate a set of more subtle complications and issues that face the firm.

Optimal variety

The notion of optimizing variety has its roots in economics and operations research. Ramdas (2003) provides a comprehensive review of the literature related to decisions faced by producers in managing product variety.

I can illustrate the basic idea behind the optimization of variety with additional detail on the scooter example. I provided the cost analysis in Exhibit 8-9 for two scenarios, one deck and two decks. Conceptually, I can extend this cost analysis to many decks by considering how the various costs of producing the scooter would change as variety is increased. There are two problems with this extension.

First, as variety increases, one would be less likely to use a production process like die casting, with high fixed costs per variant. Each new die for each new variant would add about $30,000 in up-front investment. If the scooter company were to offer ten different scooters using this production technology, then the required investment would be $300,000, a sum that I can assure you the company would not spend. Instead, the firm would adopt a different production process technology, in this case computer-controlled machining (*CNC machining*), which requires investment of only about $1,000 per variant, but incurs unit costs of materials, labor, and processing of about $40 per scooter deck. Process *flexibility* refers to the ability to produce additional variants of an artifact while incurring relatively lower fixed costs per variant—CNC machining is more flexible than die casting. The optimization of variety relies not only on the choice of a level of variety, but on the simultaneous choice of a production technology.

The second problem with the static cost analysis is that the production volume would not remain the same as variety is increased. Indeed, if the demand for scooters did not increase with increased variety, then there would be no motive for having more than one variant. The quantity produced is, however, a determinant of cost. This mutual dependency of variety, production process technology, costs, prices, and demand make the optimization problem tricky even when these factors can be readily modeled with mathematical expressions. One of the first such efforts was undertaken by de Groote (1994), who simultaneously considered costs, demand, production technology, and variety. Even so, he was able to do so only for a stylized model, which would be somewhat difficult to apply in practice.

Fortunately, the practical extent of variety in most settings is quite finite, and so one can consider discrete scenarios of, say 1, 2, 5, and 10 variants of the product and estimate what production process would be used, what revenues would likely be generated, and what would be the overall costs of delivering the particular level of variety. Then, one can compare total profits under the different scenarios and make an informed decision about the level of variety to offer. One such analysis is Exhibit 8-11, which for the scooter is the result of analysis, and judgment based on experience.

We should not get too carried away with our optimization, however, as the reality of design practice is that we have many more degrees of freedom in addressing this problem than simply what level of variety to offer, and the rules of the game are changing constantly. In the balance of this section, I consider several interesting complications that make designing variety an intellectual challenge.

Number of deck variations	**1**	**2**	**5**	**10**
Total quantity sold	25,000	31,000	34,000	35,000
Average price	150.00	165.00	170.00	172.00
Total revenues	3,750,000	5,115,000	5,780,000	6,020,000
Process technology	Die casting	Die casting	Die casting (2) + CNC (3)	Die casting (2) + CNC (8)
Fixed costs	52,000	86,000	89,000	94,000
Support costs	12,000	14,000	20,000	28,000
Average unit variable costs	90.00	92.00	101.00	107.00
Total variable costs	2,250,000	2,852,000	3,434,000	3,745,000
Total costs	2,314,000	2,952,000	3,543,000	3,867,000
Profit contribution	**1,436,000**	**2,163,000**	**2,237,000**	**2,153,000**

Exhibit 8-11. Revenues, costs, and profits for four different variety scenarios for the scooter example. Illustrative values in US$.

Variety is best measured in terms of attributes as well as end items

In this chapter I have mostly used variety to refer to the number of *end items* or *stock-keeping units* (*SKUs*) offered by a producer. However, this measure of variety can be deceptive. A toothpaste manufacturer offering nine different toothpaste end items consisting of the same formulation in the same tube, but placed in nine cartons printed in different languages, is behaving quite differently from a manufacturer offering nine end items comprising a gel and two paste formulations, each available in a pump and two sizes of tubes. Superficially, each offers nine variants, and yet the modes of competition, the design requirements, and the systems of production and distribution are likely to be very different for the two producers. For this reason, an analysis of variety is most useful when it considers both the number of end items and the variety offered with respect to each of the important individual attributes of the product.

The architecture of the artifact dictates what can be varied

In Chapter 6, I treat the architecture of artifacts in detail. The key idea is that a physical decomposition of an artifact into components may or may not correspond to a functional decomposition, and the nature of the mapping from structure to function is dictated by the architecture (Ulrich 1995). Variety refers by definition to differences in the attributes of the product, which can only be created by differences in structure. The architecture of the product constrains the ways in which the product can be changed, and therefore constrains the variety that can be achieved by the producer. A static optimization of variety may fail to account for dramatic changes to cost structure that could result from a fundamental change to the product architecture.

Variety is an element of competitive strategy

Taylor Randall and I (2001) studied the choices firms made in the bicycle industry with respect to product variety, production process technology, and supply chain strategy. We discovered that successful firms had made harmonious decisions across three different sets of decisions: the attributes over which variety would be offered, the production process technologies used to produce the bicycles, and the configuration of the supply chain for producing and distributing the goods. There is typically no single dominant strategy for competitive superiority. Rather, different firms may adopt different equally coherent sets of choices that provide differentiation in the market in a relatively efficient fashion.

Concluding Remarks

Variety has indeed increased in most categories in current society. This is partly the result of increasingly global markets in which firms serve highly heterogeneous consumers. It is also the result of increased production process flexibility and the associated loosening of the bonds of scale economies. In this world, design is less and less focused on the creation of a single perfect artifact and is increasingly a puzzle requiring creative problem solving and analytical judgment about product architecture, production process technology, supply chain structure, and market strategy.

Notes

[1] Why adopt arbitrary labels for concepts when more descriptive terms could be used? For most people, labels like *horizontal/vertical variety, type I/type II errors,* and *left-brain/right-brain* require rote memorization and cognitive effort every time they are used.

[2] Lancaster (1990) provides a nice discussion of variety from the perspective of economic theory.

References

Chouinard, Yvon. 2005. Let My People Go Surfing: The Education of a Reluctant Businessman. New York: Penguin.

de Figueiredo, John M., and Margaret Kyle. 2006. "Surviving the Gales of Creative Destruction: The Determinants of Product Turnover." Strategic Management Journal 27(3), 241–264.

de Groote, Xavier. 1994. "Flexibility and Marketing/Manufacturing Coordination." International Journal of Production Economics 36 (2): 153–167.

Hotelling, H. 1929. "Stability in Competition." Economic Journal 39 (153): 41–57.

Iyengar, Sheena. 2010. The Art of Choosing. New York: Twelve.

Kahn, Barbara E. 1995. "Consumer Variety-Seeking Among Goods and Services: An Integrative Review." Journal of Retailing and Consumer Services 2 (3): 139–48.

Lancaster, Kelvin. 1975. "Socially Optimal Product Differentiation." American Economic Review 65 (4): 567–585.

Lancaster, Kelvin. 1990. "The Economics of Product Variety: A Survey." Marketing Science 9 (3): 189–206.

Ramdas, Kamalini. 2003. "Managing Product Variety: An Integrative Review and Research Directions." Production and Operations Management 12 (1): 79–101.

Randall, Taylor, and K. T. Ulrich. 2001. "Product Variety, Supply Chain Structure, and Firm Performance: Analysis of the U.S. Bicycle Industry." Management Science 47 (12): 1588–1604.

Randall, Taylor, Christian Terwiesch, and K. T. Ulrich. 2007. "User Design of Customized Products." Marketing Science 26 (2): 268–280.

Schmalensee, Richard. 1978. "Entry Deterrence in the Ready-to-Eat Breakfast Cereal Industry." Bell Journal of Economics 9 (2): 305–327.

Ulrich, K. T. 1995. "The Role of Product Architecture in the Manufacturing Firm." Research Policy 24 (3): 419–440.

NINE

Conclusion

Much of this book is *descriptive*, an explanation and framework for understanding how design works in society. Yet, because design is largely a cognitive human phenomenon, and not grounded in physical reality as are the natural sciences, the practice of design varies highly. There are few if any immutable and agreed-upon laws of design. Thus, in its description, the book is also necessarily *prescriptive*, arguing that design can and should be thought of in a certain way and that the practice of design can be improved with a common understanding of principles and effective processes. A key objective of the book, toward that end, is to establish a common framework for design across all domains. I believe that such a framework can be the basis for *design literacy*, an understanding of the minimal set of principles and practices required to be effective as a designer.

For design literacy to take hold in society, I believe two basic requirements must be met. First, the elements of design must be codified. That is, a consensus must emerge about the core of what it means to be an effective designer. Second, that core body of knowledge must be disseminated through the educational and training activities of members of society.

Codification of Core Principles of Design

As a start, let me propose this list of principles and practices as the core elements of design literacy:

1. An effective design process includes these steps (Chapters 1 and 2):
    a. Sense gap
    b. Define problem
    c. Explore alternatives
    d. Select plan
2. Problem definition benefits from asking the "five whys" in order to frame the challenge at the right level of abstraction. (Chapter 3)
3. Understanding user needs is a key element of problem definition, and that understanding is usually best developed with interactive and immersive methods. (Chapter 3)

4. Exploration is a form of search (Chapter 4) whose primary goal is to expose as many diverse ways to address the design problem as possible. Each domain will have idiosyncratic heuristics and methods for exploration.
5. The design process is rarely a pure flow from first to last step, but more typically involves iteration. As a result, early and frequent prototyping and testing usually results in better outcomes. (Chapters 1–4)

These principles seem simple and may even be natural for many people, and of course much more can be said about design. However, I believe that these few basic principles are the essential distinguishing characteristics of effective design outcomes. Though these elements are simple, their application is remarkably rare in society. How many corporate teams, university committees, and government commissions stumble along in addressing challenges with no explicit problem definition, no deep understanding of stakeholder needs, and limited exploration of alternatives? How frequently do design efforts result in elaborate plans that have never been tested and will not have a pilot or prototype?

One of the reasons these defects persist in important problem-solving efforts is that the challenges are not recognized as design problems; also, facility in the basic design process is not ingrained in the approaches of many professionals. Of course, the design process becomes second nature for most who think of themselves as designers (e.g., architects, engineering designers, graphic designers, industrial designers), but most lawyers, business administrators, and politicians do not think of themselves as doing design, even though they frequently engage in problem-solving activities, beginning with a sensed gap and resulting in an artifact.

As noted in Chapter 3, these principles share quite a bit with the general problem-solving process articulated by Shewhart as the plan-do-check-act cycle (PDCA). PDCA is the foundation of most quality-improvement efforts and it is well understood by many if not most professionals even remotely engaged in quality-improvement projects. Of course, there is more to quality management than PDCA (e.g., process capability, design of experiments), yet PDCA is a central framework. Its codification was a precursor to dissemination.

Dissemination of Core Principles

I believe every member of society would benefit from being design literate. Who does not face the challenge of creating an artifact in response to a sensed gap, the essence of design? Recall that I use *artifact* in the most general sense to refer to products, services, business models, systems, and organizations that are the result of deliberate human creation.

If I'm right, then design really should be an essential component of primary and secondary education and reinforced in higher education. Design is fun for most children. My children created fascinating devices for their fourth-grade invention fairs (a marshmallow cannon and a portable popcorn popper). All eighty students participating each year in that fair were having a blast. How hard would it be to overlay a slight bit of conceptual process over that intrinsically engaging activity? Then, how about repeating an immersive design challenge at least once per semester along with increasing depth and sophistication in design tools and methods?

Somewhere between fourth grade and ninth grade (roughly ages 9–14), design projects seem to drop out of the curriculum (at least in the U.S. and in most of the other countries with which I have some familiarity). The concomitant focus on analytical rigor and the acquisition of facts may prepare students for more school, but do not necessarily enhance their abilities as problem solvers in society.

For roughly half of adults in developed economies, we get another chance at design education in college or other postsecondary educational institutions. However, few if any design courses are part of what might be considered the general education curriculum in universities. As I look at my own university's core curriculum, design does not fit neatly into one of the general education requirements. In the undergraduate college at the University of Pennsylvania, the core curriculum comprises *communication*, *analysis*, and *perspectives* plus seven "sectors" (society, history, arts and letters, humanities and social science, the living world, physical science, and natural science and mathematics). Where would you put design? Is design (or perhaps human problem solving more generally) less important than the required topics under these categories?

Well, a gap has been sensed. What's next? I recommend defining the problem, generating alternatives, selecting a plan, and iteratively refining until the gap is closed.

* * *

Acknowledgments

My Wharton colleague, Christian Terwiesch, has been a continual sounding board for the ideas in this book. Jeffrey Smith, CEO of LUNAR design, engaged in interesting discussions about the aesthetics chapter and has seriously taken on the challenge of codifying effective approaches to aesthetics in LUNAR's practice of industrial design. Steven Eppinger and I together wrote the textbook *Product Design and Development*, which is now in its fifth edition. That on-going effort to communicate ideas to students has grounded my perspective on design. My research collaborator, Laura Kornish, has been willing to read anything I send her, and she provided detailed suggestions for improving the clarity of this book.

My mother, Laurel Thatcher Ulrich, read the manuscript carefully and provided useful feedback and encouragement. My father, Gael Ulrich, is himself a lifelong designer, taught chemical process design for many years, and wrote an influential book on that subject. While I doubt I absorbed much chemical engineering as a child, I was certainly steeped in a designer's perspective on problem solving. My wife, Nancy Bentley, and I had many animated discussions about aesthetics, and the chapter on that topic is better as a result.

My friend and MIT classmate Matt Haggerty hired me in 1988 as a part-time product designer at his young company Product Genesis. I learned a great deal about the practice of design from that early professional experience. Today many of my deepest professional and personal relationships are with alumni of that pioneering firm.

Woodie Flowers, David Jansson, Warren Seering, and Patrick Winston were exceptional professors and advisors at MIT and they taught me a tremendous amount about design, teaching, approaches to research, and effective communications. I know I'm one of thousands of former students on whom they have had a huge impact.

About the Author

Karl T. Ulrich is Vice Dean of Innovation and the CIBC Professor of Entrepreneurship and e-Commerce at the Wharton School of the University of Pennsylvania. He also holds an appointment as Professor of Mechanical Engineering. He is the coauthor of *Product Design and Development* (5th Edition, McGraw-Hill, 2011), a textbook used by a quarter of a million students worldwide, and of *Innovation Tournaments* (Harvard Business Press, 2009). At Penn, he cofounded the Weiss Tech House and the Integrated Product Design Program, two institutions fostering innovation in the university community. In addition to his academic work, Professor Ulrich has led dozens of innovation efforts for medical devices, tools, computer peripherals, food products, web-based services, and sporting goods. As a result of this work, he holds more than 20 patents. Professor Ulrich is a founder of Terrapass Inc. and he is a designer of the Xootr scooter, which *BusinessWeek* recognized as one of the 50 coolest products of the twenty-first century. Professor Ulrich holds bachelor's, master's, and doctoral degrees in mechanical engineering from MIT.

Colophon

I wrote this book over a time period in which the publishing industry changed dramatically. In the middle of this period, I began doing most of my reading on a mobile computing device; first the Kindle, then the iPad, and most recently an Android-based smartphone. I'm not alone in believing that printed books for professional and academic content are becoming obsolete, and therefore the role of the traditional publisher substantially diminished. Because of these beliefs, my obsession with design details, and my inclination to learn new things, I decided to coordinate the entire process of creating this book myself. I wrote the manuscript in Microsoft Word. I created the illustrations, except where noted, in Adobe Illustrator. I inserted most images as JPEG or PNG files, with original photos captured with a Nikon D50 digital SLR camera. I hired a freelance copy editor, Sarah Weaver, who made edits directly in the Word files. Then, a member of the Wharton staff, J. P. Lacovara, created the eBook itself using the free software Sigil and Calibre. J. P. also purchased an ISBN number and handled permissions. The book, once in eBook format, is easily uploaded to Amazon and other electronic distribution channels. All of this was quite straightforward once the tools were in place and the path known. The result is a book available at very low cost in essentially every distribution channel and format to the widest possible readership.