Res Eng Design (2009) 20:41–58
DOI 10.1007/s00163-008-0056-z

ORIGINAL PAPER

The Pugh Controlled Convergence method: model-based evaluation and implications for design theory

Daniel D. Frey · Paulien M. Herder · Ype Wijnia · Eswaran Subrahmanian · Konstantinos Katsikopoulos · Don P. Clausing

Received: 18 October 2007 / Accepted: 27 July 2008 / Published online: 5 December 2008

Abstract This paper evaluates the Pugh Controlled Convergence method and its relationship to recent developments in design theory. Computer executable models are proposed simulating a team of people involved in iterated cycles of evaluation, ideation, and investigation. The models suggest that: (1) convergence of the set of design concepts is facilitated by the selection of a strong datum concept; (2) iterated use of an evaluation matrix can facilitate convergence of expert opinion, especially if used to plan investigations conducted between matrix runs; and (3) ideation stimulated by the Pugh matrices can provide large benefits both by improving the set of alternatives and by facilitating convergence. As a basis of comparison, alternatives to Pugh's methods were assessed such as using a single summary criterion or using a Borda count. These models suggest that Pugh's method, under a substantial range of assumptions, results in better design outcomes than those from these alternative procedures.

Keywords Concept selection · Multi-criteria decision-making · Decision analysis · Comparative judgment

D. D. Frey (✉) · D. P. Clausing
Massachusetts Institute of Technology, 77 Mass. Ave., Cambridge, MA 02139, USA
e-mail: danfrey@mit.edu

P. M. Herder
Delft University of Technology, Jaffalaan 5, 2628 BX Delft, The Netherlands

Y. Wijnia
Essent Netwerk B.V., Postbus 856, 5201 AW, 's-Hertogenbosch, The Netherlands

E. Subrahmanian
Carnegie Mellon University, 5000 Forbes Avenue, Hamburg Hall 1209, Pittsburgh, PA 15213, USA

K. Katsikopoulos
Max Plank Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

1 Motivation

Recent research papers in engineering design have proposed that there are some major deficiencies in core elements of engineering practice. In particular, engineering decision-making has been singled out for attention. The following quotes give a sense of the concerns being raised:

- "Multi-criteria decision problems are still left largely unaddressed in engineering design" (Franssen 2005).
- "A standard way to make decisions is to use pairwise comparisons…. Pairwise comparisons can generate misleading conclusions by introducing significant errors into the decision process… rather than rare, these problems arise with an alarmingly high likelihood" (Saari and Sieberg 2004).
- "…there exists one and only one valid measure of performance for an engineering design, that being von Neumann-Morgenstern utility… we can say that all other measures are wrong. This includes virtually all measures and selection methods in common use" (Hazelrigg 1999).

This paper seeks to challenge the idea that current engineering decision-making approaches are significantly flawed. If decision making is at the core of engineering and if we don't have or don't routinely use good decision making capabilities, then a poor track record of the

engineering profession should be observed. Yet over the past century, engineering has successfully transformed transportation, housing, communication, sanitation, food supply, health care, and almost every other aspect of human life (Constable and Somerville 2003). Studies suggest that technical innovation accounts for more than 80% of long term economic improvement (Solow 1957). How can the methods of engineering design practice be so poor and the progress resulting from engineering practice be so valuable? A principal motivation of this paper is to explore this dissonance. The paper addresses the issues more specifically by analyzing a specific design method, Pugh Controlled Convergence and its relationship to recent developments in design theory. Figure 1 illustrates how Pugh Controlled Convergence has been subject to critique either explicitly or implicitly by three recent papers. In the second layer of the diagram, we list some features of Pugh's method. Below that, we list papers that raise concerns about those features of the method. In the bottom layer, we list aspects of the model developed in this paper that are responsive to each critique.

Figure 1 guides the structure of this paper. Section 2.1 fleshes out the second layer of the diagram. In it, we describe Pugh's method in detail. Section 2.2 provides more supporting detail on the third layer of the diagram. In it, we discuss the recent research relevant to Pugh Controlled Convergence including the three papers mentioned in Fig. 1 and several others. Section 3 is related to the bottom layer of the diagram and constitutes the core of the paper. In Sect. 3, we build and explore a model of the design process. Using the framework described by Frey and Dym (2006) we construct computer executable entities meant to represent, in abstract form, the aspects we consider most essential to understand Pugh Controlled Convergence. Our model explicitly includes: (1) the role of the datum concept, (2) the convergence of expert opinion based on investigation, and (3) the generation of new alternatives. These considerations have not played a prominent role in the scholarly debate on design decision making, but it seems to us that they have a first order impact in practice. In light of these considerations, we seek to ascertain whether or not the reported undesirable behaviors of Pugh's method actually arise under realistic conditions. Section 5 comprises a discussion of these results.

2 Background

2.1 Review of Pugh Controlled Convergence

Pugh (1981, 1990) advocated that product development teams should, at an early stage in the design process (after developing specifications but before detailed design), engage in an iterative process of culling down and adding to the set of concepts under consideration. The goals of this activity are: (1) a 'controlled convergence' on a strong concept that has promise of out-competing the current market leader; and (2) a shared understanding of the reasons for the choice. We will refer to the overall process of attaining these goals as Pugh Controlled Convergence or PuCC.

A prominent aspect of PuCC is presentation and discussion of information in the form of a matrix. The columns of the Pugh matrix are labeled with a description, in drawings and text, of design concepts. The rows of the matrix are labeled with concise statements of the criteria by which the design concepts can be judged. The method requires selection of a datum, preferably a design concept that is both well understood and known to be generally

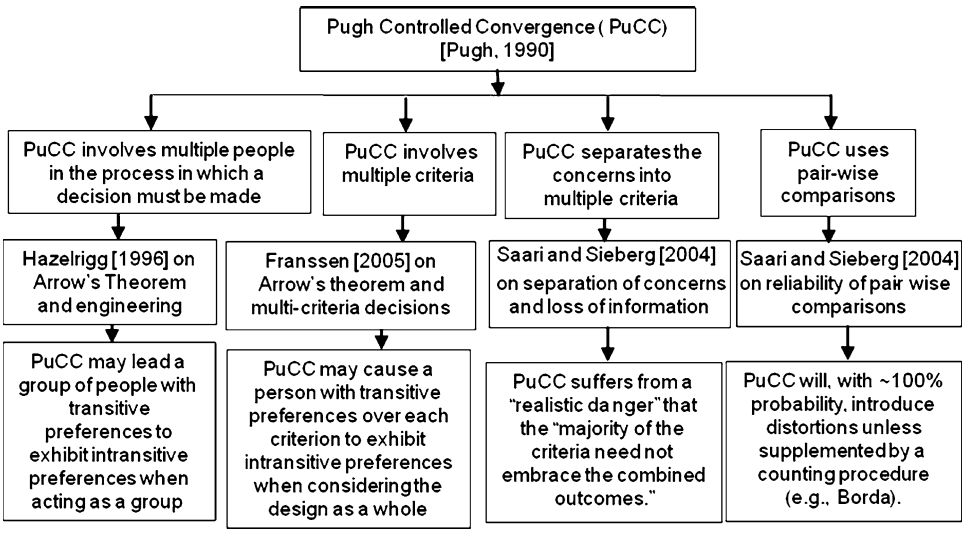

Fig. 1 Features of Pugh's method, critiques related to each feature, and our model-based approach to testing those claims

strong. Often the initial datum concept is currently the leader in the market. Evaluations are developed and entered into the matrix through a facilitated discussion among the experts. Each cell in the matrix contains symbols +, −, or **S** indicating that the design concept related to that column is clearly better than, clearly worse than, or roughly the same as the datum concept as judged according to the criterion of that row.

Academic publications on Pugh's method will often present neatly formatted tables representing a Pugh matrix. This may contribute to a misunderstanding of what is actually done. In practice, Pugh matrices are messy collages of drawings and notes. This is a reflection of the nature of early-stage design. The PuCC process is simple and coarse-grained. Observation of teams show the method is also flexible and heuristic. We assert that these are affirmative benefits, making the method fit well into its context. For example, alternatives to Pugh's method often require greater resolution of the scale (suggesting five or ten levels rather than just three) and often require numerical weighting factors. Pugh found by experience that this sort of precision is not well suited to concept design. In this paper, a model-based analysis is used to evaluate this hypothesis regarding the benefits of simplicity in the decision process and effectiveness in attaining good design outcomes.

It is important to note that there is no voting in Pugh's method. Let us consider a situation in which several experts claim that a concept is better than the datum and others disagree. In Pugh's method, a discussion proceeds in which the experts on both sides communicate their reasons for holding their views. In many cases, this resolves the issue because either: (1) facts are brought to light that some individual experts did not previously know, (2) a clarification is made about what a design concept actually entails, or (3) a clarification is made about what the criterion actually means. If that discussion leads to an agreement among the experts, then a + or − may be entered in the matrix. If the disagreement persists for any significant length of time, then an **S** is entered in the cell of the evaluation matrix. In Pugh's method, **S** can denote two different situations. It can mean that the experts agree that the concept's merit is similar to the datum or that the differences between the concept and the datum are controversial and cannot be determined yet. In this case, team members would be encouraged to find additional information necessary to resolve the difference of opinion. Pahl and Beitz (1984) have suggested an "**i**" or "**?**" should be entered to more strongly encourage investigation).

Generally, the evaluation matrix includes summary scores along the bottom. The number of +, −, or **S** scores for each concept are counted and presented as a rough measure of the characteristics of each alternative. This raises an important issue. These scores are sometimes interpreted as a means by which to choose the single winning design. This misconception is reflected in terminology—Pugh's method is most often referred to in the design literature as "Pugh Concept Selection" whereas Pugh emphasized "Controlled Convergence". The term "Concept Selection" would seem to imply that after running a matrix a single alternative will be chosen. This is not an accurate characterization of the PuCC process. The first run of the evaluation matrix can help reduce the number of design concepts under consideration, but is not meant to choose a single alternative. A matrix run can result in at least four kinds of decisions (not mutually exclusive) including decisions to: (1) eliminate certain weak concepts from consideration, (2) invest in further development of some concepts, (3) invest in information gathering, and (4) develop additional concepts based on what has been revealed through the matrix and the discussions it catalyzed. To follow up on these actions, the matrix should be run iteratively as part of a convergence process.

To illustrate how iterated runs of the evaluation matrix result in convergence, consider a real-world example. Khan and Smith (1989) describe a case in which a team designed a dynamically tuned gyroscope. The process began with 15 design concepts and 18 criteria, which we would characterize as a typical problem scale. Figure 2 depicts results from a sequence of three runs of a Pugh matrix each with a different datum concept. The figure is organized with the evaluations for all three runs of the matrix for each concept in one column with the first run on the left, the second run in the center, and the last run on the right. In the first matrix run, concepts 5 and 13 were dominated by the datum and concepts 2 and 11 were dominated by concept 12. Therefore, the set of alternatives could have been reduced by about one quarter in the first round although it appears that all these alternatives were retained for one more round of evaluation. Between the first and second matrix runs, a new alternative labeled 12a was created to improve concept 12 along one of the dimensions in which it was judged to be weak. After the second Pugh matrix was made, the team could have eliminated five more alternatives that were dominated, bringing the total of dominated designs up to nine. Figure 3 reveals that the team took advantage of the opportunity to save time and chose not to evaluate seven of the nine dominated alternatives in the third Pugh matrix. In addition, the team chose to focus on only half of the criteria. Some criteria were dropped because they did not discriminate among the alternatives and some because they were too difficult to evaluate precisely. The third matrix run did not enable any additional concepts to be identified as dominated, but did result in a final choice of concept 12a to be developed in detail. It is notable that concept 12a did not have as many positives as concept 8, but perhaps it

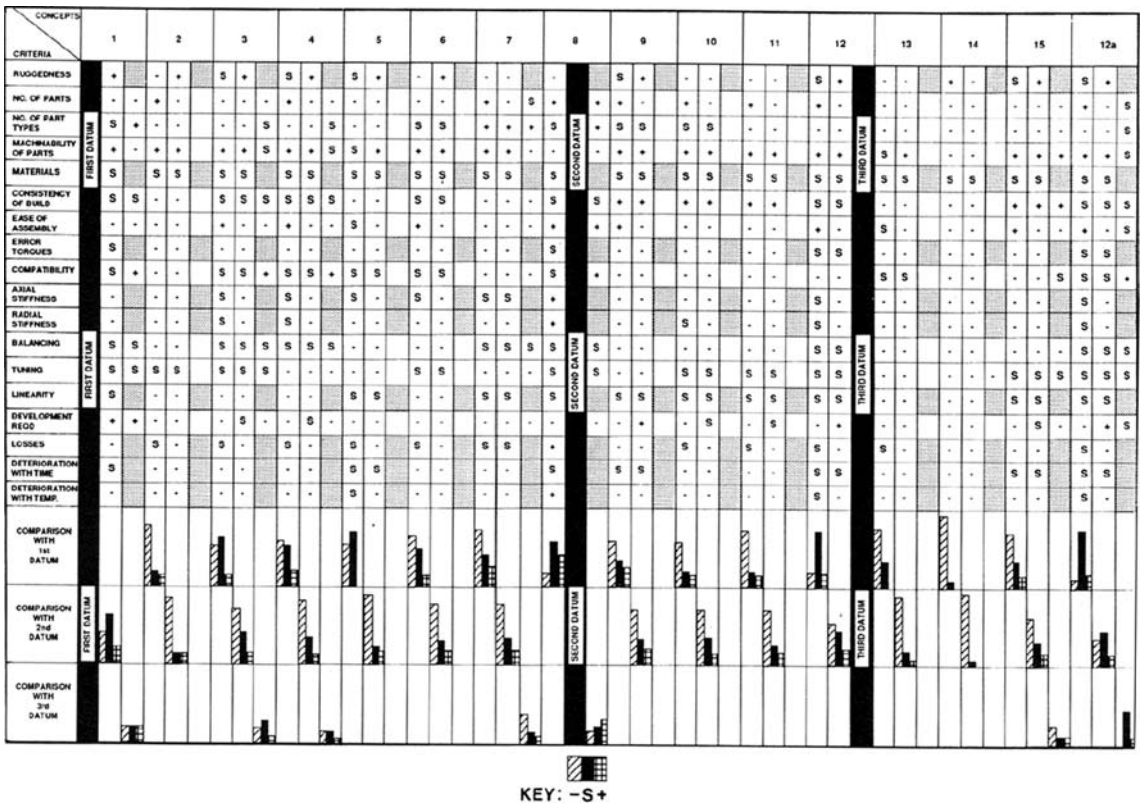

Fig. 2 Data from three runs of Pugh matrices in the design of a gyroscope [from Khan and Smith (1989)]

could be viewed as more balanced since it had no negatives in the final round. Note also that, as is common in PuCC, the concept finally chosen was not even present in the initial set of concepts considered but rather emerged through the continued creative process running in parallel and informed by the evaluation process. This sort of parallel, mutually beneficial process of evaluation and ideation was encouraged by Dym et al. (2002) and Ullman (2002) as well as by Pugh (1990).

As the case study by Khan and Smith (1989) shows, the PuCC process includes decision making, but it cannot be sufficiently modeled *only* as decision making. The process also involves learning and creative synthesis and there is no clear line when these activities stop and decision making begins. Learning, synthesis, and decision-making proceed in parallel and synergistically. The analysis and discussion of design concepts catalyzes creation of additional concepts, which in turn may simplify decision-making. This interplay among decision-making and creative work is often neglected when considering the merits of decision-making methods. Our models in Sect. 3 and 4 explicitly include these aspects of the design process.

The Pugh method is among the best known engineering design methodologies, but it seems to be used by only a modest proportion of practicing engineers. A survey of 106 experienced engineers (most of whom were working in the

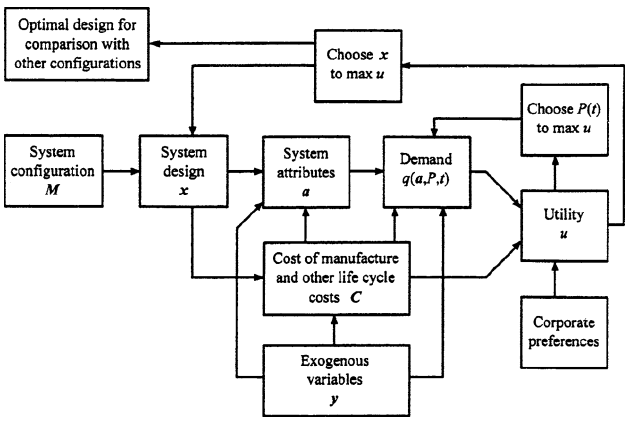

Fig. 3 A framework for decision-based engineering design (from Hazelrigg 1998)

United States) indicated that just over 15% had used Pugh Concept Selection in their work and that most of those found it useful (about 13% of the 15%) (Yang 2007). Other design methods included in the survey were FMEA, QFD, robust design, and design structure matrices which were used at work by 43, 20, 19, and 12% of respondents respectively. The survey found that a few simple techniques were used by a majority of practicing engineers including need-finding, benchmarking, storyboarding, and brainstorming. Another survey specifically focused on selection

methods (in this case, a survey of Finnish industry). This survey suggested Pugh's method is used by roughly 2% of firms (Salonen and Perttula 2005). Informal approaches labeled as "concept review meetings", "intuitive selection" or "expert assessment" were estimated to be used in about 40% of companies. These two surveys, although not conclusive, suggest that only the simplest and most flexible design techniques are used widely and that more formal design methods are generally used much less. We wish to present a case for an appropriate degree of structure. We think there is somewhat too little structure in engineering practice today and probably far too much structure is recommended in most of the design methodology literature. Later sections of this paper are intended to make this argument by comparing PuCC, a relatively simple method, with more complex alternatives. First we review some literature that presents technical objections to Pugh's method.

2.2 Pugh, utility, and Arrow's theorem

Hazelrigg (1998) has proposed a framework for decision-based design (DBD) as graphically depicted in Fig. 3. A central feature of the framework is that the choice among alternative designs is impacted by the decision maker's values, uncertainties, and economic factors such as demand at a chosen price. Hazelrigg's DBD framework requires rolling up all these diverse considerations into a single scalar value—utility as defined by von Neumann and Morgenstern (1953). Having computed this value for each alternative configuration, the choice among the design alternatives is simple—"the preferred choice is the alternative (or lottery) that has the highest expected utility" [Hazelrigg 1999].

Hazelrigg's framework for DBD is subject to much debate and continues to have significant influence in the community of researchers in engineering design. The textbook *Decision Making in Engineering Design* (Lewis et al. 2006) reflects a wide array of opinions on how decision theory can be implemented in engineering design and also demonstrates that the core ideas of the DBD framework are being developed actively.

Hazelrigg's framework explicitly excludes the use of Pugh's method of Controlled Convergence. Hazelrigg states the conclusion in broad terms explaining that the acceptance of von Neumann and Morgenstern's axioms leads to one and only one valid measure of worth for design options. Since Pugh's method does not explicitly involve computation of utility, Hazelrigg has argued that Pugh's method is invalid. Also, DBD invokes Arrow's General Possibility Theorem (Arrow 1951). Hazelrigg (1999) states "in a case with more than two decision makers or in a multi-attribute selection with more than two attributes, seeking a choice between more than two alternatives, essentially all decision-making methods are flawed".

Scott and Antonsson (1999) argue that the implications of Arrow's theorem in engineering design are not nearly so severe. A principal basis for this conclusion is that "the foundation of many engineering decision methods is the explicit comparison of degrees of preference". This line of approach to the possibility of choice is similar to Sen's who states "Do Arrow's impossibility, and related results, go away with the use of interpersonal comparisons…? The answer briefly is yes" [Sen 1998]. In combining the influence of multiple attributes, Scott and Antonsson state that "there is always a well-defined aggregated order among alternatives, which is available to anyone with the time and resources to query a decision maker about all possible combinations". The DBD framework establishes the aggregated order via expected utility, but Scott and Antonsson concluded that "the relative complexity of these methods is not justified" compared to simpler procedures such as using a weighted arithmetic mean. Pugh's method represents a further simplification and this paper seeks to determine whether this additional reduction in complexity is also justified.

Franssen (2005) attempted to counter the arguments by Scott and Antonnsen. Franssen challenges, on measure theoretic grounds, the existence of a global preference order that is determined by any aggregation of individual criterion preference values. Franssen argues that if criterion values are ordinal or interval, then the global aggregated order posited by Scott and Antonsson cannot be defined or else that it will be subject to Arrow's result. More fundamental however, is Franssen's assumption that measurable attributes of the design can never determine the designer's overall preference ordering. Franssen holds that "it is of paramount importance to realize that preference is a mental concept and is neither logically nor causally determined by the physical characteristics of a design option". Franssen concluded that "Arrow's theorem applies fully to multi-criteria decision problems as they occur in engineering design". Franssen also draws specific conclusions regarding Pugh's method:

> *…This method… can attach different global preferences, depending on what is taken as the datum…. Hence it does not meet Arrow's requirement…. It is important not to be mistaken about what Arrow's theorem tells us with respect to the problem…. What it says is that, for any procedure of a functional form that is used to arrive at a collective or global order, there are specific cases in which it will fail…. Accordingly, for any specific procedure applied, one must always be sensitive to the possibility of such failures.*

This quote by Frannsen is a major motivation for this paper. Our model-based assessment of Pugh's method of

controlled convergence will explicitly deal with the issue that the selection of the datum does make a difference in running the matrix. And, as Franssen notes, one must always be sensitive to the possibility of failures induced by one's chosen design methods. But the *possibility* of failure is not enough to justify abandoning a technique that has been useful in the past. This paper seeks to quantify the impacts of such failures and weigh them against the benefits of the PuCC process.

2.3 Pugh and pairwise comparison

Saari and Sieberg (2004) constructed an argument against all uses of pairwise comparisons in engineering design except for very restricted classes of procedures including the Borda count. Going beyond the argument based on Arrow's theorem which only claims the *possibility* of error, Saari and Sieberg make specific claims about the *likelihood* and *severity* of the errors. Saari and Sieberg propose a theorem including the statement that "it is with probability zero that a data set is free from the distorting influence of the Condorcet n-tuple data". From this mathematical statement they draw the practical conclusion that pairwise comparisons "can generate misleading conclusions by introducing significant errors into the decision process … rather than rare, these problems arise with an alarmingly high likelihood".

Saari and Sieberg claim that "even unanimity data is adversely influenced by components in the Condorcet cyclic direction". In Pugh's method, designs that are unanimously judged to be superior across all criteria will never be eliminated. Therefore the distorting effect is not always reflected in the alternative chosen, but in some other regard. Saari and Sieberg state "suppose the $A \succ B \succ C$ ranking holds over all criteria…. If we just rely on the pairwise outcomes, this tally suggests that the $A \succ B$ and $A \succ C$ rankings have the same intensity…. It is this useful intensity information that pairwise comparisons lose…". This raises an important point related to intensity of feelings. It is not enough that an engineering method should lead to selection of a good concept. It is also essential that the method should give the team members an appropriate degree of confidence in their choice. But Saari and Sieberg's proposed mathematical processing of the team members' subjective opinions may not have the desired result. We suggest that a psychological commitment to the decision may be attained more effectively by convergence of opinion rather than balancing opinions as if design were an election. As differences of opinion are revealed by the Pugh process, investigation and discussion ensue. Since we consider this an important part of engineering design, we seek to incorporate in our model the possibility that people can discover objective facts and change their minds.

A second theme in Saari and Sieberg's paper regards separation of concerns. Pugh's method explicitly asks decision makers to consider multiple criteria by which the options might be judged. Saari and Sieberg claim that such separation of the information leads to a "realistic danger" that the "majority of the criteria need not embrace the combined outcomes". Saari and Sieberg's argument for this conclusion is "Engineering decisions often are linked in the sense that the $\{A, B\}$ outcome is to be combined with the $\{C, D\}$ conclusion. For instance, a customer survey may have $\{A, B\}$ as the two alternatives for a car's body style while $\{C, D\}$ are alternative choices for engine performance". Saari and Sieberg then outline an imaginary scenario in which the survey data lead to a preference reversal due to an interaction among criteria. The survey data in the scenario suggest that although customers prefer body style A when considered separately and engine performance C when considered separately, they do not prefer the combination of those particular body styles and engine performance options. Saari and Sieberg conclude the resulting product "runs the risk of commercial failure" and that "product design decisions… could be inferior or even disastrous".

With the argument regarding separation of concerns, Saari and Sieberg may have sacrificed his claim that these events occur with high likelihood. Many inter-criterion interactions in engineering are known a priori to be too small to cause the reversals Saari and Sieberg describe. Consider a specific example in which a team designed a gyroscope and needed to consider criteria such as "machinability of parts" and "axial stiffness" (Khan and Smith 1989). The sort event that Saari and Sieberg ask us to consider is that a design concept A is better than concept B on "machinability of parts" and A is also better than B on "axial stiffness", but that the ways those two criteria combine makes B better than A overall. This sort of event seems unlikely to us. Why would hard-to-machine parts become preferable to easy-to-machine parts when the gyroscope happens to be more stiff? This example illustrates that in many pairings of technical criteria, it is safe to assume separability of concerns. A more challenging example is Saari and Sieberg's "body style" and "engine performance" pair. Clearly, a sporty body style is a better match to a more powerful engine, even if this implies more noise and lower fuel efficiency. But there is a large practical difference between interaction of components and interaction of criteria. We do not think lower fuel efficiency is actually preferred to high fuel efficiency in the presence of a sporty body style, but perhaps a louder engine sound actually is preferred. It seems to us that interactions among criteria are not large except for pairs of aesthetic criteria and that preference reversals are rare. Given the possible problems sketched here, we will evaluate (in

Sect. 4.1) how large inter-criterion interactions would have to be to lead to choice of weak concepts.

The analysis by Saari and Sieberg is not only a warning regarding potential risks, but is also presented as a guide to modifying the design process—"Once it is understood what kind of information is lost, alternative decision approaches can be designed". Unfortunately, the proposed remedies impose significant demands on information gathering and/or processing. Saari and Sieberg suggest a procedure involving "adding the scores each alternative gets over all pairwise comparisons". Let us consider what this implies for the Pugh process using the specific example in Khan and Smith (1989). The process began with 15 design concepts and 18 criteria. The first run of the matrix therefore demanded that 14 concepts be compared with the datum across 18 criteria so that 252 pairwise comparisons had to be made by the team to fill out the first evaluation matrix. If the run of the matrix was to be completed in a standard 8-h work day, then about 2 min on average could be spent by the team deliberating on what symbol should be assigned to each cell in the matrix. In reality, many of the cells might be decided upon very quickly because the difference between the concept and the datum is obvious to all concerned. However, even accounting for this, the time pressures are quite severe. Saari and Sieberg's remedy requires that every possible pairwise comparison must be made requiring 15 choose 2 pairwise combinations of concepts across 18 criteria—1890 pairwise comparisons in all. If the process is to be completed in a single work day, there would be only 15 s on average per comparison. Alternately, one might preserve the same average discussion time per cell (2 min) and allow around 63 working hours for the task rather than 8. Given this order-of-magnitude expansion of resource requirements, it is possible Saari and Sieberg's suggested remedy is more harmful than the Condorcet cycles themselves. Dym et al. (2002) prove that pairwise comparison charts provide results identical to those of the Borda count, however this approach is also time consuming. We suggest it's worth considering simpler procedures and so we make a comparative analysis of Pugh's method with the Borda count in Sect. 4.3.

2.4 Pugh and rating, weighting, and sensitivity

Takai and Ishii (2004) presented an analysis of Pugh's method including comparison with alternative approaches. The paper posits three desiderata of concept evaluation methods: (1) The capability to select the most preferred concept, (2) The capability to indicate how well the most preferred concept will eventually satisfy the target requirements, and (3) The capability to perform sensitivity analysis of the most preferred concept to further concept improvement efforts.

To evaluate the Pugh method, Takai and Ishii suggest three possible modifications of Pugh's matrix. Two of the modifications involve types of rating and weighting. One of the modifications involves computing the probability of satisfying targets. In a case study involving design of an injector for a new linear collider, they consider the merits of three alternatives over nine criteria. All four methods suggested the same design as the most preferred alternative. However, a further analysis suggested that even the most preferred concept had only an 8.9% chance of satisfying its requirements and that if availability were improved by 3% and cost reduced by 30%, then the chances of success improved to 76.8%. They conclude that the most promising approach was to quantify one's beliefs in terms of distributions, evaluate concepts by probability of satisfying targets, and perform sensitivity analysis.

The analysis by Takai and Ishii seems appropriate to situations in which the number of alternatives is small, all the alternatives are well characterized, and the possibility of generating new concepts is not available. Such a scenario is likely to arise at some stage in the convergence process, but perhaps such modifications are counterproductive in the earlier stages. If probabilistic analyses were conducted with rather coarse estimates, there may be a risk of misleading the team into false confidence. Pugh and Smith (1976) argue that numbers used in evaluation matrices are easily interpreted as similar in standing to the sorts of objective number engineers most often work with (e.g., densities, voltages, and elastic moduli). Overly precise representations create a risk of unwarranted faith in decisions based on rough estimates. It is possible that, in the early stages of design, the same time and resources needed for probabilistic analysis might be used in some more productive way. The model we propose in Sect. 3 is intended to enable exploration of such trade-offs among different emphases and different styles of work.

2.5 The psychology of pairwise comparison

To address the various critiques and the proposed improvements of PuCC, it is worthwhile to review some results from psychological research. The discipline of psychology can provide insight into what is and is not possible for humans to do or to understand. Psychology also provides information about human capacities that can be leveraged by decision making methods. This section reviews selected topics helpful to understanding later parts of this paper.

Decision field theory (DFT) is an approach to modeling human decision making. The theory acknowledges that humans make decisions by a process of deliberation which is inherently dynamic with degrees of preference

varying over time (Johnson and Busemeyer 2005). DFT models can be created that simultaneously accord with a large set of empirically demonstrated effects and have been used to analyze a variety of decision tasks including, most relevant to engineering, multi-attribute decision making under time constraints (Diederich 1997). The models described in Sect. 3 bear some resemblance to those from Decision Field Theory since they are dynamic with states varying through repeated cycles based on previous states. A difference of our approach from DFT is that we do not model decision making as emerging from weighting of valences primarily, but instead model decision making as determined by decision rules or heuristics. Psychology research has shown that such heuristics are often more robust than schemes involving weighting, especially in generalizing from experience to new tasks (Czerlinski et al. 1999).

Experimental evidence bears out the idea from Decision Field Theory that decision making emerges from adaptive sampling. Shimojo et al. (2003) showed that when presented with two faces and asked to choose the preferred one that subjects alternate between gazing at each face and begin directing more attention to the preferred one until a threshold is reached at which point a decision is made. Studies also showed that sampling and decision interact early in the process, long before actual conscious choice (Simion and Shimojo 2006) and that manipulation of sampling can influence choice (Shimojo et al. 2003). This result is in good affinity with the somatic marker hypothesis including the proposition that evaluations of complex scenarios are not explicitly represented in memory but instead correlated to bioregulatory processes (Bechara and Damasio 2000). This hypothesis poses a challenge for those who suggest decision making can always be made better through mathematical procedures since some of the information needed may not be accessible to conscious processes. This perspective from cognitive science links back to engineering design when we consider the process of rating alternatives. Saaty (2006) explains that "comparisons must precede ratings because ideals can only be created through experience" and because "comparisons are our biological inheritance". Procedures such as the Analytical Hierarchy Process are meant to take the raw data of pairwise comparison and to create interval scale measurements. In the process, inconsistencies or rank disagreements may be discovered (Buede and Maxwell 1995) and procedures have been suggested for correcting those (Limayem and Yannou 2007). Even if such inconsistencies are not present, there is still substantial uncertainty in rankings due to uncertainties in the pairwise comparisons and there exist methods for quantifying these uncertainties (Scott 2007). This paper considers the possibility that a simpler set of pairwise comparisons such as in PuCC might result in a better outcome despite uncertainties in input data and the presence of undetected inconsistencies.

Research on human perception and judgment may prove useful in evaluating results in later sections. Psychologists draw a distinction between discrimination and magnitude estimation. In a discrimination task, a human subject is asked to compare two entities and decide which has a property to a greater degree. In a magnitude estimation task, a human subject is asked to give a quantitative value for an entity along a continuous scale. Smith et al. (1984) conducted a study in which human subjects were asked to make judgments about line length under various task conditions. The study showed that human judgment is much less prone to failure (by roughly a factor of two) when two entities are compared directly rather than estimating values on a continuous scale. Katsikopoulos and Martignon (2006) studied paired comparisons for more complex criteria so that multiple cues are needed and they prove that psychologically plausible heuristics can, under some conditions, provide the same results as the optimal Bayesian computation. We suggest that these studies demonstrate an affirmative value of paired comparison and discrimination tasks. By avoiding rating and weighting, Pugh method enables engineers to consider the relative merits of alternatives in ways that are simple enough to do without external aids and also demonstrably accurate. These simplifications should make the judgments of engineers less prone to error. The implications of this hypothesis will be explored in Sect. 4.

3 A model of Pugh controlled convergence

This section presents a quantitative model of the Pugh Controlled Convergence process. The model is a highly abstract representation of the process we have observed among real teams using the method. It is important to keep in mind that "essentially, all models are wrong, but some are useful" (Box and Draper 1987). Although this model cannot hope to capture, in all its facets, how concept design actually proceeds, we envision that people can use the model to probe their beliefs about decision-making and its role in engineering design.

3.1 A model of the first round of the evaluation matrix

This section describes a basic model of the first round of an evaluation matrix. The model is stochastic, so the model is executed in many independent trials so that we can characterize the behavior statistically. In each trial, the simulation is comprised of the following four steps:

3.1.1 Create a set of design concepts to be evaluated

In the model, there are values $\mathbf{C}_{ij}$ where $i \in 1\ldots n$ and $j \in 1\ldots m$. Each value $\mathbf{C}_{ij}$ represents the objective merit of concept j on criterion i. These objective merits will influence the Pugh matrix, but the two matrices are not equivalent since $\mathbf{C}_{ij}$ is a real number and the corresponding Pugh matrix element has only three levels, **+**, **S**, and **−**. The values $\mathbf{C}_{ij}$ are sampled from random variables with distributions $\mathbf{C}_{ij} \sim N(s, 1)$ and $\mathbf{C}_i \sim N(0, 1)$, $j \neq 1$. Care is required in interpreting the use of random variables here. Random variables enable us to generate a diverse set of concepts, but the values of $\mathbf{C}_{ij}$ are fixed within each trial. The datum concept in the first run has index, $j = 1$. The intrinsic merits of the datum concept are selected from a different population than those of all the other concepts. The factor, s, represents the relative strength of the datum concept. In our model, larger values are preferred and therefore, if $s > 0$, the datum is better than the rest of the concepts on average across the many trials although it can be weak along some criteria in any particular trial. To illustrate the meaning of this parameter, consider that a value $s = 1.0$ implies the datum concept has a criterion score drawn from a population one standard deviation above the mean of criterion ratings for new concepts generated. Therefore, at a parameter setting $s = 1.0$ each new concept will improve upon the datum in about one in four opportunities.

3.1.2 Model a set of opinions held by a group of experts

In the model, there are values $\mathbf{CE}_{ijk}$ where $k \in 1\ldots o$ represent the estimated merit of design concept j on criterion i as judged by expert k. We model the expert opinion as correlated with the intrinsic merits of the design concepts, but differing from expert to expert. This is accomplished by computing the values as $\mathbf{CE}_{ijk} = \mathbf{C}_{ij}(1 + \varepsilon_{ij})$ with $\varepsilon_{ijk} \sim N(0, \sigma_{ij}^2)$. Again, these values are related to the Pugh matrix, but not equivalent to it. In particular, there are o different expert opinions of each concept's merits along each criterion, yet only one symbol will be entered in the Pugh matrix.

3.1.3 Generate the Pugh matrix

Each cell of the Pugh matrix, $\mathbf{M}_{ij}$, corresponds to a design concept j and a criterion i. The cells are determined as $\mathbf{M}_{ij} = +$ if $\mathbf{CE}_{ijk} > \mathbf{CE}_{ijk}$ for all $k \in 1\ldots o$, $\mathbf{M}_{ij} = -$ if $\mathbf{CE}_{ijk} < \mathbf{CE}_{ijk}$ for all $k \in 1\ldots o$, $\mathbf{M}_{ij} = \mathbf{S}$ otherwise. To state the same thing another way, if all experts agree that the concept is better than the datum, then a **+** is entered in that cell. If all experts agree that the concept is worse than the datum, then a **−** is entered. If there is any disagreement among the experts, then an **S** is entered.

3.1.4 Eliminate concepts based on the Pugh matrix

In actual use of the Pugh Controlled Convergence process, there is no formulaic prescription that automatically leads to the elimination of a concept. In this model, we eliminate any concept that is dominated. In other words, concept A is dominated by another concept B if, according to **M**, B is better along one or more criteria and is no worse than A along all other criteria. In PuCC, dominated concepts appear to have no advantages that could not be drawn as well or more easily from some other concept and will therefore be eliminated.

We simulate the process above to explore the influence the ability of a design team to converge by eliminating weaker concepts from consideration. In particular we wished to understand how the strength of the datum concept influences convergence. To anchor our analysis more strongly in data, we used four case studies along with the model—Khan and Smith (1989), Pugh (1990), Begley (1990), and Miller et al. (2005). Each of these publications presented a Pugh matrix from which we could infer how much convergence was possible. In each case, we considered how many concepts were dominated according to the matrix. Begley (1990) was a somewhat non-standard case study since two different datum concepts were used in forming the Pugh matrix. The case concerned steering columns, some of which enabled tilting and some of which did not. The team found it difficult to compare concepts across these two groupings. This made convergence more difficult in this case study. We did not attempt to correct for this minor discrepancy.

In Fig. 4, the convergence data from the four case studies are presented along with corresponding model-based results. Figure 4 presents results from a single application of a Pugh matrix—convergence resulting from repeated applications is addressed in Sect. 3.2. In our

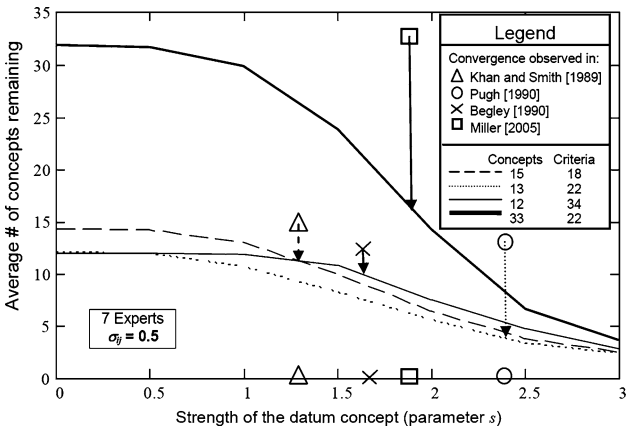

Fig. 4 The ability of the first run of the evaluation matrix to eliminate weak concepts

model, the strength of the datum was varied from zero to three in seven equal increments (s = 0, 0.5, 1.0, … 3.0). The number of initial concepts and the number of engineering criteria evaluated were set at four discrete combinations (15, 18), (13, 22), (12, 34), and (33, 22) chosen to correspond with the case studies. Each value on the curves plotted in Fig. 4 arises from 500 replications of a model with seven experts each of whom had random error in criterion judgments generated with a standard deviation of 0.5. The assumption of seven experts was based on roughly the number of disciplines reflected in the list of criteria from the case studies. The degree of error was set at 0.5 which made the number of **S** ratings in the Pugh matrices a reasonable match with those observed in the case studies. The convergence observed in each of the four case studies is graphically depicted by placing a symbol at the height corresponding to the number of initial concepts and an arrow down to the number of concepts not dominated according to the Pugh matrix. The symbols and arrows were adjusted horizontally so that the arrow heads lie upon the curve generated by the model with the number of concepts and number of criteria matching those in the case study. To emphasize the s values estimated in this way, the symbols are repeated along the x axis of Fig. 4.

A principal conclusion we draw from Fig. 4 is that datum strength (s values) above 1.0 are needed to explain the degree of convergence observed in engineering practice. All four case studies attained a fairly good degree of convergence, ranging from about 25% to about 70%. According to our model, convergence of less than 10% should be expected with datum strength at 1.0 or weaker. With a parameter value of s = 1.0, the probability is far below 0.1% of seeing four instances with convergence as large as observed in the four case studies (40% average across the sample). To make the statement somewhat more formally, a null hypothesis of $s < 1.0$ has a corresponding p-value less than 0.1%. An alternative explanation of the data is that there are actually two different populations mixed together here, some projects that clearly have very strong datum concepts and others that might have weak datum concepts. For the two cases Khan and Smith (1989) and Begley (1990), the range of simulation results are reasonably consistent ($p > 10\%$) with the null hypothesis that s = 0.

To further explore the conclusion above, we considered the consequence if a strong datum exists but cannot be identified by the team. We repeated the simulations with the datum selected at random and found that the convergence was reduced. The decrement in convergence was modest over the range of s = 1.0–2.5 where the data suggests the parameter values tend to be distributed.

To summarize, a major critique levied against the Pugh method is that the choice of the datum influences the outcome of the concept selection process. The analysis presented in this section reveals several facts relevant to this issue:

(1) In practice, the datum concept is significantly stronger than the rest of the population. Since the datum is not arbitrary in practice, it seems to us less problematic that datum selection can influence the process (for example, by slowing convergence).
(2) If there is a strong datum concept, the first round of PuCC will reduce the set of alternatives by a substantial degree, ranging from 25 to 70% in most cases.
(3) If the datum were not strong in some particular case, if Pugh's approach is followed properly, the consequence would not be a poor decision, it would be a lack of convergence in the first round. The PuCC process, as we modeled it here, will tend to retain many concepts rather than risk eliminating anything worthwhile.
(4) A single run of the Pugh matrix rarely leads to selection of a single alternative. This is to be expected as the matrix is part of an iterative process of learning and creative synthesis. The next section develops a model of the additional work required following the first run of a Pugh matrix.

3.2 A model of work between matrix runs

We saw in the previous section that the first run of the Pugh matrix eliminates only a modest number of concepts. In practice, this may be a positive feature of the method because each one of the remaining concepts exhibits potential in some dimensions. In work between the runs of the matrix, the design team may find ways to make use of all the concepts that were carried forward. Some concepts may be actively developed and others may serve as a source of ideas. The process by which the design team seeks improvements between matrix runs has been incorporated into our model and is described below.

When a large number of concepts are in play, some additional decision making is needed to set priorities for further work. This is a principal justification for summary information that is constructed at the bottom of the Pugh matrix. Concepts with a large number of + scores and relatively few − scores represent good platforms on which to build a serious contender against a strong datum. Concepts with a small number of + scores and relatively many − scores represent sources of ideas, but probably do not deserve further investment in their own right. The PuCC process does not include any formula for making these decisions. Nevertheless, we propose an algorithm so that we can implement it in our model. The team works in two ways:

Ideation—Between runs of the matrix, the team can invest time and energy in ideation—creative work focused by the information revealed in the previous matrix run. This is an important aspect of the engineering work that would normally be conducted between iterations of Pugh's evaluation matrices. Our model of this activity is based on the possibility of forming hybrids of two concepts. Sometimes one can combine different aspects of two or more concepts to form a new concept superior to any of its constituents. We assume that between runs of the matrix, one of the designs in the top 1/3 is selected at random as a basis for a hybrid. Based on the matrix **M** from the last run, a second design is selected that appears most complementary in the sense that it has strengths in just those areas where the chosen concept requires improvements. The hybrid is then formed assuming that, for each criterion i, the new value $\mathbf{C}_{ij}$ is the larger of those of two designs being merged. This is an abstract, highly simplified representation of the creative process. In reality, complex technical factors determine which combinations of concepts are feasible and which are not. We want to express in our model the possibility that such hybrids can emerge in response to the evaluation process. We seek to represent this in a reasonably realistic way so that a small number of hybrids that address some, but not all of the observed challenges we observe in experience. This model of ideation, although rough, does enable study of the interplay between creative work, evaluation, and decision making which we believe is critical to drawing an accurate picture of various concept design methods.

Investigation—Between runs of the matrix, the team can seek improved understanding of the design problem. Because resources are assumed to be constrained, we model investigation of a focused nature guided by the last Pugh matrix. Our model of this activity is that for each concept j, if it was in the top 1/3 and it earned an **S** in the previous Pugh matrix on criterion i, then for each expert k the opinion $\mathbf{CE}_{ijk}$ is refined. In addition, all the concepts receive a refined estimate in the three most influential criteria. The refined estimates are modeled by reducing the parameter σ_{ij} by a factor of two and newly sampling the expert opinion. This is meant to represent the possibility that investigation (including computation, experimentation, interaction with customers, and discussion among the experts) can lead the team to a shared understanding of the issues affecting the decision. In our model, investigation moves the criterion estimates of each expert into better alignment with the objective merits.

Figure 5 presents results from simulations conducted with ideation and/or investigation included as described above in repeated rounds of controlled convergence. The horizontal axis corresponds to the phase of the work with progression in time from left to right. We assumed that the Pugh matrix would be run three times with two periods of work between matrix runs. Each point in Fig. 4 arises from 1,000 replications of a model with 15 initial concepts, 7 experts, a moderately strong datum concept (s = 1.2), and moderately large initial variance in expert opinion (σ_{ij} = 0.5). The vertical axis represents the number of concepts under consideration. We ran two cases, one in which a single hybrid concept was formed between matrix runs, and a case with no new concepts generated. For the case including ideation between matrix runs, we plot the median and the 10th and 90th percentiles to give a sense of the variance within the population of trials. For the other case we plot only the median to avoid cluttering the Figure. The convergence observed in a real world case study in Khan and Smith (1989) is also presented for comparison.

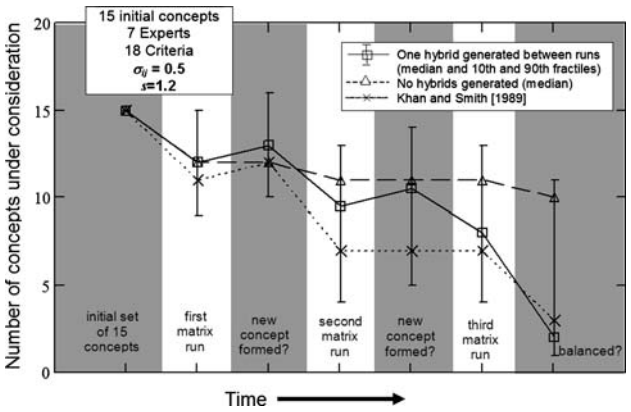

Fig. 5 The convergence of PuCC through three iterations with and without new concepts being generated

A key observation from Fig. 5 is that the model with hybrids being generated is generally consistent with the trend in Khan and Smith (1989). After the modest convergence in the first round, the degree of convergence is primarily dependant on the creation of hybrids. If hybrids are formed, subsequent work enables weaker concepts to be eliminated at a high rate so that only a few options remain after three runs of the matrix and after filtering out poorly balanced designs. Both the model and the case study are consistent with this conclusion. On the other hand, if hybrids are not formed, then convergence based on dominance will be very slow. Changing the datum does enable one or two concepts to be eliminated even if hybrids were not formed. Other approaches for trade study analysis would be needed for selection in this case. Mistree et al. (1994) presented a method for concept selection involving multiple rounds of evaluation with different datum concepts in each round. Their method involved weighting of the criteria so that summary "merit functions" could be formed. We think rating and weighting can be avoided if hybrids can be formed and that Fig. 5 supports this conclusion.

It is critical to appreciate the mechanism explaining the connection between divergence and convergence. A hybrid of two complementary designs can often dominate a substantial number of competitors. Visualizing the patterns of strengths and weaknesses in the Pugh matrix seems, based on our experience, to catalyze the creative work needed to generate new concepts that can simplify future decision making. We believe this was the reason that, in Khan and Smith (1989), so many concepts were eliminated in the second run of the Pugh matrix. Our model of hybrid generation included two such hybrid creation events, but still matches well the convergence attained by actual practitioners who reported only one hybrid being generated. Therefore, we suspect that engineers are better at creating hybrids after running Pugh matrices than we have reflected in our simulations.

Even if we acknowledge the ways that creative work can create dominant concepts, convergence by dominance alone may not suffice for convergence. According to our simulations, if there are many criteria, around half of the total concepts may remain even after three rounds of Pugh matrix runs. However, considerable additional convergence can be made once it is known that additional hybrids will not be formed. As the datum strength increases through PuCC, many designs tend to have one or two positives overwhelmed by a large number of negatives. Although not strictly dominated, poorly balanced designs can be safely eliminated after the last matrix run without sacrificing future opportunities for creative work. Our model suggests that a simple rule based on a 2:1 ratio of $-$:$+$ will eliminate a large number of the remaining concepts. At this point, either a few designs should be developed in detail, or else recourse might be made to rating and weighting or probabilistic analysis [as in Takai and Ishii (2004)] to converge to a single alternative.

4 Comparison of decision making approaches

The previous section shows that the Pugh Controlled Convergence Process, under appropriate conditions, can down-select to a small number of alternatives without resorting to voting, rating, or weighting. But we also need to explore the merits of such an approach compared to alternative procedures. The next sub-section presents an extension of the previous model to incorporate "bottom line" measures of the design outcome. Subsequently that model is used to evaluate methodological alternatives inspired by the design literature such as papers by Hazelrigg (1998), Saari and Sieberg (2004), and Takai and Ishii (2004).

4.1 A model of profitability

Let us suppose there is real scalar $\mathbf{P}_j$ which represents the overall merits of the *j*th design concept. It is convenient to think of the $\mathbf{P}$ vector as standing for profitability of the *j*th design concept if it were selected and developed.

Central to our model is a quantitative relationship between the criteria $\mathbf{C}_{ij}$ and the value of $\mathbf{P}_i$. We assume that, all other things being equal, a higher rating along one criterion should cause the overall merit to rise. However, we also want to address the issue of "separation of concerns" raised by Saari and Sieberg (2004). Our model includes the possibility that scoring best across individual criteria does not necessarily imply a design that scores best overall. It is our judgment that this does not happen often in practice, but we include it here to measure its possible impact. To include this possibility and otherwise keep the model as simple as possible, we include only two-factor interactions between pairs of criteria.

$$\mathbf{P}_j = \sum_{i=1}^{n} \beta_i \mathbf{C}_{ij} + \sum_{p=1}^{n} \sum_{\substack{q=1 \\ q>p}}^{n} \beta_{pq} \mathbf{C}_{pj} \mathbf{C}_{qj} \tag{1}$$

The sensitivity of the overall merit of any design concept to the *i*th criterion score is represented by β_i and the interactions among criteria are represented by β_{pq}. By modeling the relationship between criteria and $\mathbf{P}$ in this way, we are assuming that a full set of criteria uniquely determine the expected outcomes of the design process. In other words, we assume the expected profitability of two designs should be the same for any two concepts that score the same on all criteria.

To instantiate instances of the model in Eq. (1), we select the coefficients β from the populations $\beta_i \sim |N(0, 1)|$ and $\beta_{pq} \sim N(0, \tau^2)$. The coefficients with a single subscript are non-negative so that the criterion values more naturally correspond with the conventional symbols in the evaluation matrix (e.g., a $+$ is meant to indicate a "better" value). The parameter τ represents the relative degree of interactions between criteria. To express the notion that main effects are usually larger than interactions, we suggest $\tau \ll 1$ in a reasonable model of concept design. Increasing values of τ lead to a situation in which criterion values individually explain only a small fraction of the overall merit. Given the distributions we have chosen, interactions between criteria are equally likely to be synergistic or anti-synergistic.

At this point it is useful to discuss the concept of inter-criterion interactions which are included in our model through coefficients β_{pq}. An improvement in a criterion value such as "manufacturability" should lead to an increase in a measure of overall merit such as expected profit. An improvement in some other criterion, such as

"ease of use" should also lead to increase in a measure of overall merit. However, there may be good reasons to believe the effects of two improvements are not simply additive. In extreme cases, anti-synergistic interactions creates a risk of a ranking reversal which was emphasized by Saari and Sieberg (2004). We do not consider such inter-criteria interactions a major concern in engineering because rank reversals should be rare. However, by including coefficients β_{pq} we allow for the possibility in our model so that we can explore the influence of these effects. If there are only two criteria, then ranking reversal happens only if $\beta_{12} < -(\beta_1 + \beta_2)$. Given our model, the probability of this event is $0.5\left|1 - (2/\pi)\tan^{-1}(\sqrt{2}/\tau)\right|$. Therefore we see that the parameter τ enables the modeler to set the probability as desired, but some additional guidance is useful. According to a study by Li et al. (2006), two-factor interactions in physical experiments are typically about 20% of main effects. If τ is 0.2, the probability of a preference reversal due to an interaction is about 5% per opportunity. In other words, if we choose values of τ typical of physical experiments, then individual criteria and overall merit are consistent in roughly 95% of all instances. With good specification of the design problem, inter-criterion interactions may be smaller than interactions between physical factors, but this is a subject for further research. Those using the model can modify their assumptions in this regard or test the influence of their assumptions by changing the value of τ in the model.

4.2 Profitability of Pugh Controlled Convergence

This section is intended to represent an implementation of Pugh Controlled convergence including a model of profitability. We adapted the model described in Sect. 3.2 which included three rounds of Pugh matrices to include the profit model presented in Sect. 4.1. Also, for the purpose of simulation, the final convergence from a handful of options to a single alternative had to be forced somehow. Although Pugh emphasized that this final decision rests with the engineers and not with the matrix, we simply chose the design with the highest difference between the sum of + scores and sum of − scores in that column of the matrix, a heuristic procedure sometimes called "tallying" (Gigerenzer et al. 1999).

In Fig. 6 are plotted the results of these PuCC process simulations. The abscissa represents the average **P** of the selected concept normalized by the maximum value in the initial population of design concepts. The ordinate represents the model parameter σ_{ij} which can be interpreted as the uncertainty in the criterion scores.

A principal observation from Fig. 6 is that the possibility of continuing ideation during evaluation, as was strongly emphasized by Pugh, has a large influence on the

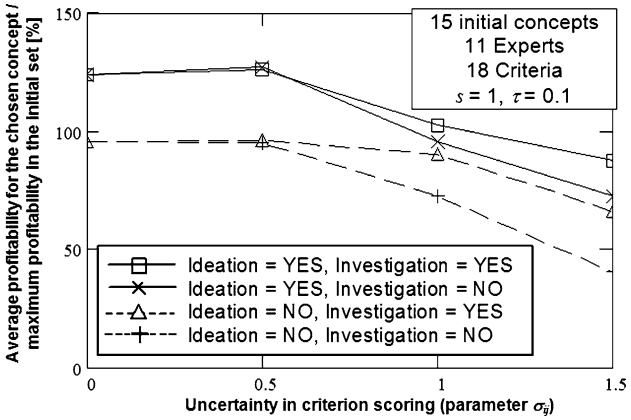

Fig. 6 The outcomes after three runs of the Pugh matrix

outcomes. Figure 5 suggests that even when σ_{ij} is unity meaning uncertainty is as large as the variations within the population, the benefits of continued ideation are larger than the decrements due to uncertainty so that one will attain average **P** values exceeding the maximum value in the initial population. The implication is that no degree of finesse applied to the decision among the fixed set of initial alternatives can compensate for failing to exploit the benefits of additional creative design work.

A second major observation from Fig. 6 is that there is very little influence of small degrees of uncertainty on the PuCC process. Figure 6 suggests that even when σ_{ij} is less than 0.5 meaning uncertainty is half as large as the variations within the population, the influence on the outcomes is very nearly zero for all four scenarios we simulated.

Our last major observation from Fig. 6 is that the benefits of focused investigation are considerable, especially when uncertainties are large. Figure 6 suggests that even when σ_{ij} is as large as unity and ideation is not used, investigation can remove the majority of the losses due to uncertainty. This is somewhat surprising since investigations in our model are conducted for only about 10% of the criterion/concept pairs. Thus, our model tends to support the notion that Pugh matrices are helpful in locating leverage points for modeling, experimentation, and information sharing among experts.

4.3 A model of the decision based design framework

This section describes an implementation of Hazelirigg's Decision Based Design framework (1998) as applied to the model presented in Sect. 4.1. The scenario simulated is similar to that in Sect. 4.2 except that there is only a single round of evaluation. Also, because the mathematics of decision making as conceptualized by Hazelrigg apply only to individual decision makers, we assume there is just one expert in this model. We formulated a single summary criterion which is that expert's estimate of expected

profitability, **PE**. We assume this estimate is related to the true profitability of the concepts but subject to uncertainty so that $\mathbf{PE}_j = \mathbf{P}_j(1 + \varepsilon_j)$ with $\varepsilon_j \sim N(0, \sigma_j^2)$. We model the decision maker as risk neutral so that he prefers the highest expected value of profit. Under this assumption, the decision is made simply by picking the largest scalar from among the 15 estimated $\mathbf{PE}_j$ values. We modified the simulation used in Sect. 4.2 to reflect these changes. We set the strength of the datum at a moderate value ($s = 1$) and we ran these simulations for a range of different degrees of uncertainty in expert judgment σ and plotted the **P** value of the selected concept normalized by the maximum value of **P** in the available set of 15 alternatives. The results are depicted in Fig. 7 with selected data from Fig. 6 also shown for comparison.

The results presented in Fig. 7 admit a simple interpretation. When the designer's uncertainty is zero, the profitability is 100% of the potential within the initial set of 15 designs. In other words, if somehow the profitability of the design concepts can be estimated accurately, then choosing the highest estimated profit will obviously maximize the profit. However, the plot shows that as the designer's uncertainty rises, profit attained drops. With a σ of 1.0, only about 75% of the potential profit will be realized on average. Note that given a σ of 1.0, the uncertainty in the evaluation of profitability is roughly as large as the variance among the profitability of the options. This is by no means an upper limit—the uncertainty involved in estimating profitability at an early stage of the design process might be substantially greater.

Figure 7 also depicts data from Fig. 6 on the performance of PuCC under the worst scenario we considered—no ideation or investigation allowed. Note that PuCC performs about 5% worse than the DBD approach at zero error. According to our model, the payoff for implementing the DBD framework is that this 5% loss might be avoided. If the resources needed and constraints imposed by DBD detract from ideation or investigation, the net effect of DBD will be negative according to our model.

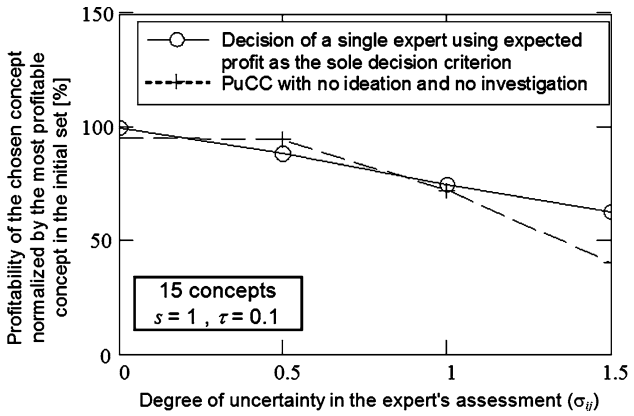

Fig. 7 The profit earned based on decisions using a single expert using only estimates of expected profit

The two decision procedures plotted in Fig. 7 offer very similar performance as a function of the parameter σ with an advantage for the DBD approach as σ rises above unity. An advantage for PuCC not shown on Fig. 7 is the consistency of results from trial to trial. For example, at $\sigma = 1.5$, PuCC has worse outcomes on average than DBD, but the variance in the outcomes is somewhat less. This is substantially due to the averaging of error that will tend to occur when groups participate in judgments. Such an effect strongly depends on our modeling assumption of independence of the error across the population of participants.

Another point is worth mentioning regarding interpretation of Fig. 7. As discussed in Sect. 2.5, the research of Smith et al. (1984) show that the reliability of human judgments is better by roughly a factor of two in discrimination tasks as compared to magnitude estimation tasks. Since PuCC is based on discrimination of each concept with a datum and the DBD framework substantially depends on magnitude estimation according to Fig. 3 (such as estimation of costs), we infer that σ should be substantially lower for PuCC than for DBD. If this phenomenon documented by of Smith et al. (1984) actually applies in engineering design as well as in the tasks they studied, PuCC might provide better results than DBD even in highly uncertain environments.

A preliminary conclusion of this comparative analysis is that internally consistent decision processes can still result in very large losses when uncertainty is high. These losses are due to lack of external correspondence of the decision maker's judgment. By contrast, PuCC may be subject to some potential for internal inconsistencies, but it enables better external correspondence in this model since it involves many experts in the decision and focuses their attention on things that are important to the decision outcome. An alternative interpretation of the simulation results is that the DBD framework was overly penalized in this model by restriction to a single decision maker. The DBD framework does not preclude the individual decision maker from gathering and incorporating the views of several experts, so perhaps the single expert should be given a lower σ value than any one of the multiple experts that are individually contributing to the PuCC process.

4.4 A Model of the Borda count

This section is intended to represent an implementation of Saari and Sieberg's suggested approach (Saari and Sieberg 2004) as applied to the model presented here. We consider the possibility of a Borda count over multiple experts using one criterion and also a Borda count over multiple criteria as judged by a single expert.

We modified the simulation from Sect. 4.3 so that 11 experts were involved in the decision, but only a single criterion, **PE**, was employed and the Borda count was used to combine the information from the experts to choose a single alternative. The results appeared virtually indistinguishable from those from one expert choosing as described in Sect. 4.3. This confirms that the Borda count generally retains the internal consistency of the DBD framework, but is also similarly sensitive to uncertainty despite involving multiple experts.

We also modified the simulation from Sect. 4.3 so that 18 criteria were used and the Borda count was used as if the criteria were voting for the winner as Saari and Sieberg (2004) described. The winner of the election was recorded and the **P** value was computed for each trial. We repeated this procedure in 1,000 probabilistically independent simulations. This process was repeated for four different values of τ.

The results are depicted graphically in Fig. 8 with the degree of interaction among criteria values plotted on the abscissa and normalized profitability on the ordinate. A conclusion based on Fig. 8 is that, if the criteria are reasonably separable ($\tau = 0.1$ or 0.2), then the Borda count over multiple criteria performs quite similarly to the Borda count over multiple experts based on profit alone. However, if criteria interact strongly ($\tau = 0.3$ or 0.4), then there appears to be a substantial advantage to using just a single summary criterion. We should warn that this assumes that somehow the experts can judge the summary measure (such as expected profitability) with only as much uncertainty as they would judge the 18 separate criteria. We would venture to say that profitability cannot be estimated as precisely as engineering criteria such as axial stiffness or ease of assembly.

Our preliminary conclusion in this sub-section is that the impact of inter-criterion interactions on PuCC is small for realistic scenarios and is outweighed by what might be lost by diverting attention away from creative work. As noted in Sects. 3, 4 a large meta-analysis of data (Li et al. 2006) showed interactions are typically 20% of single factor effects. This would correspond to $\tau = 0.2$. But this meta-study data represents interactions among physical factors. We suggest interactions among criteria will be even smaller because: (1) the team of experts are free to define criteria in such a way that they avoid large interactions, and (2) market segmentation tends to limit the degree of differences among concepts considered in PuCC which, a Taylor's series approximation suggests, will encourage a more linear mapping between criteria and overall merit.

4.5 A model of rating and weighting

This section is intended to represent an implementation of one of Takai and Ishii's (2004) proposed variants of Pugh's method as applied to the model presented here.

We modified the simulation from Sect. 4.3 so that each criterion score is estimated (as a real valued scalar) by a single expert and that linear weighting factors β_i are estimated with the same degree of uncertainty as criterion scores. The ratings and weightings are used to form a score and the concept with the highest score is selected. Note that, given this model, the scores from the rating weighting matrix differ from **P** only because of uncertainty (σ_{ij}) in criterion and weight estimates and because of non-zero criterion interactions β_{pq} (we set $\tau = 0.1$). We repeated this procedure in 1,000 probabilistically independent simulations. The results of the simulation process are depicted graphically in Fig. 9.

A preliminary conclusion based on Fig. 9 is that, if uncertainties in criterion scores are the same in both methods (PuCC and rating and weighting), then the outcomes are very similar. But again, research of Smith et al. (1984) show that the reliability of human judgments is improved by about a factor of two in pairwise comparison

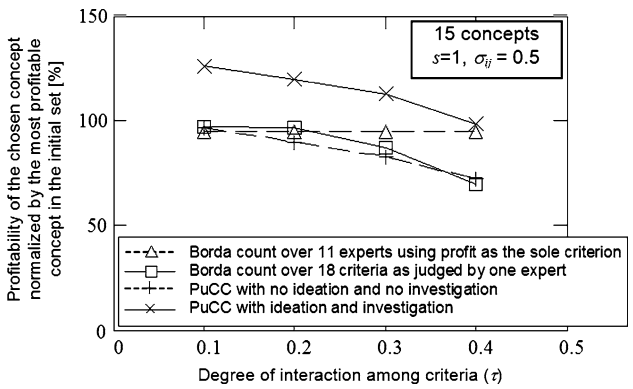

Fig. 8 The profit earned as a function of the interactions among criteria

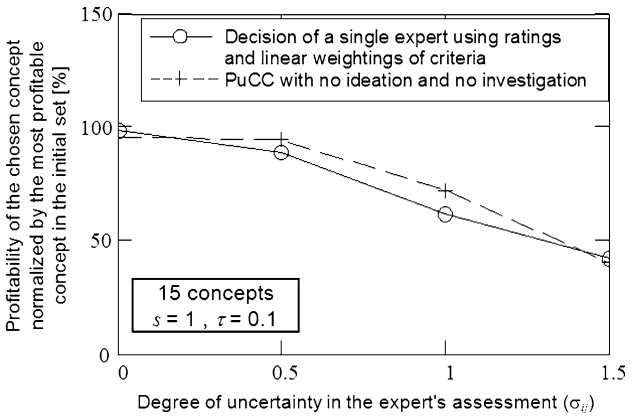

Fig. 9 The profit earned based on decisions using a single expert using ratings and linear weightings

as opposed to magnitude estimation. If that research applies here, PuCC is preferred to rating and weighting even when no new concepts are generated.

5 Conclusions

The conclusions drawn from this study must be viewed in light of the degree of model validation conducted thus far. First, as a minimum, we believe the model presented here has enough common structure with concept design scenarios to enable "surrogative reasoning"—it allows us to reason directly about the representation in order to draw conclusions about the phenomenon that it depicts (Swoyer 1991). In addition, the model is in reasonable agreement with four case studies. It is important to recognize that more validation is required before the model presented here can serve as a predictive tool. However, the model has been subject to more validation and has been evaluated with substantially more empirical data than the papers criticizing Pugh's method.

5.1 Conclusions about Pugh Controlled Convergence

The models presented here support the contention that Pugh Controlled Convergence is an effective method to apply during the concept design phase. There are risks of internal inconsistencies and distortions as emphasized by Hazelrigg, Saari and Sieberg, and Frannsen (as summarized in Fig. 1), but the model suggests that these considerations are outweighed by other issues. Engineering experience with PuCC has generally been good and the model presented here supports a positive evaluation and enables further probing into required assumptions and underlying mechanisms.

A principal conclusion of this paper is that PuCC can improve the creative process. The method encourages the team to present, in an easily interpreted visual format, patterns of information concerning the alternatives and their merits relative to criteria. This helps focus the creative work of the team on developing new concepts that can dominate other alternatives under consideration. If this can be done, then our model suggests that decision making is greatly facilitated. Because of this, we conclude that ideation and evaluation should proceed in parallel and that a major objective of PuCC is to encourage this. Our model suggests is that if just a couple new hybrid concepts emerge from insights arising from Pugh evaluation matrices, then these benefits trump the concerns about potential violations of internal consistency. On the other hand, ideation is not exclusive to PuCC. Perhaps more can be done within design frameworks such as Decision Based Design to interweave periods of concept generation among periods of concept evaluation. This would require development and testing of some practices within these frameworks that would call attention to promising opportunities for hybrid formation and reversal of negative attributes.

Another important conclusion is that uncertainty should not be taken as an immutable facet of design decision making. Engineering design, as it is normally practiced, includes a sequential, iterative process by which uncertainties are reduced through experimentation, investigation, and information sharing among experts. Methods that facilitate this learning process should be strongly encouraged. The model presented here supports the idea that PuCC can help teams target alternative/criterion pairs with high leverage in the decisions they face. In our models, reducing uncertainty in a targeted fashion improved the design outcomes. Similar observations were made by Ward et al. (1995) in studying design at Toyota and Nippondenso where multiple design options are often carried forward, concept selection is deferred, and decisions can be based on more data.

Our model supports the notion that the datum concept is important, especially in the early rounds of PuCC. The practical consequence is that datum selection should not be haphazard. An analysis of the existing competition should be undertaken to identify a concept that can serve as a yardstick for all the others (perhaps a leader in the marketplace). Our model suggests that a strong datum is likely to simplify decision making and improve the rate of convergence. This conclusion fits well into a broad historical perspective of engineering. Most every successful new design results from evolution of existing successful designs. We conclude that the central role of a strong datum concept in Pugh's method is well aligned with the evolutionary nature of most engineering.

The models presented here also support the notion that Pugh's method encourages greater objectivity in engineering decision-making. In general, people working together on an engineering project should have a substantial agreement on goals (such as whether profit is the dominant objective) and values (such as attitude toward risk). If such agreement is in place, differences of opinion on an engineering team can frequently be settled based on facts. Because of this, engineering decisions may converge as knowledge is shared and evidence is accumulated. In other words, we agree with Scott and Antonsson (1999) that there exists a well-defined aggregated order among alternatives and that the availability of this order depends on "time and resources". Pugh's method is intended to facilitate this sort of fact-based convergence. By focusing the team on criteria at an appropriate level of detail, the resulting decisions can be determined more by facts and less by emotional attachments of team members to favorite concepts. Movement in the direction of objectivity, although never realized perfectly, is to be greatly valued.

5.2 Conclusions about design theory

The analysis of Pugh Controlled Convergence enables insights into the role of economics and social choice theory as tools for understanding engineering decision making. These theories make assumptions that don't always map well into engineering. For example, Saari and Sieberg's analysis of election procedures assumes each person's stated preference ordering deserves equal consideration. This seems appropriate in a democratic election, but not so appropriate in engineering. Imagine a scenario in which an engineer believes, based on her expertise, that a particular concept is weak and a voting process results in the team selecting that concept. If the dissenting individual based her judgment on facts not known to the others, it provides little comfort that the voting process ensures that her opinion was weighted just as much as every other expert's opinion. We suggest that it would be better to spend time discussing, in concrete engineering terms, her reasons for holding her opinion rather than investing that same time in a process that prevents the distorting effects of Condorcet cycles. The results in Sect. 4.4 suggest that investigating the reasons for a difference of opinion and exploring new options in light of what is revealed is more productive than using a carefully crafted election procedure to decide the matter.

Franssen (2005) proposes that Arrow's theorem applies fully to multi-criteria decision-making because preferences are "mental concepts neither logically or causally determined by" physical parameters. The implication is that Arrow's stipulated conditions such as Unrestricted Domain and Minimal Liberty imply that preferences must be unrestricted by any demand for objectivity. In personal and political contexts, perhaps people should be unrestricted in this sense, but in an engineering context, it seems inappropriate. If an engineer is faced with a solid body of evidence showing the superiority of one alternative over another, we argue they must either conform to the evidence or else their view is irrelevant to rational engineering decision making. Our model provides a means to explore this contention. The model explicitly includes the concept of objective merits possessed by design concepts (reliability, manufacturability, performance). Such objective merits have a bearing on the bottom line outcomes. During the design process, engineers work to improve these objective merits and also the better characterize them and the way they map to summary measures (like profitability). Good correspondence of expert judgments with facts is essential to good engineering design. In our models, external correspondence breaks down in the limit that σ_{ij} becomes very large. In this case, the expert's estimates are aligned poorly with one another and with facts. Our models suggest that profit earned will drop rapidly as external correspondence breaks down. We conclude it is best to avoid a subjectivist position toward engineering decision-making.

More generally, much of research in engineering design today depicts decision making as a process that begins after the set of alternatives is closed. In this light, trade-offs take center stage. For example, See et al. (2004) state "There are always trade-offs in decision making. We have to pay more for better quality, carry around a heavier laptop if we want a larger display, or wait longer in line for increased airport security. More specifically, in engineering design, we can be certain there is no one alternative that is better in every dimension". On the other hand, on longer time scales, engineering provides better quality for less money, engineering leads to laptops that are both lighter and have larger displays, and engineering might eventually enable better airport security with shorter queuing times. We think this longer term perspective weighs strongly against use of engineering methods that place too much emphasis on trade-offs, especially in the early stages of design. Such methods may preclude opportunities to create advances beyond the current state of the art. Such advances are needed not only in breakthrough projects, but also play a part in even the most routine design work wherein performance, cost, and reliability are incrementally evolved to levels previously thought to be impossible.

5.3 Suggestion for future research

This paper has presented a model-based evaluation of Pugh Controlled Convergence. We suggest that additional work is needed in empirical validation of the results and also further exploration of the theoretical implications. Below we expand on these possibilities in turn.

The conclusions of this paper should be put to the test by means of experiments with human subjects. Two types of experimental testing seem to be possible in this context: (1) controlled experiments in the lab placing Pugh's method up against alternatives under essentially equivalent conditions of time, resources, and skill levels of the teams; and (2) full scale field tests in which real projects are conducted, some using Pugh's method and some using alternatives. The evidence from laboratory conditions can attain more precise estimates of the effects of methodological differences, but the second approach is also necessary to ensure the benefits translate to the less controlled conditions and longer time scales od authentic engineering practice.

There has been much discussion of rationality in engineering design and most of the emphasis in this discussion has been on internal consistency. Sen (1993) has argued that internal consistency demands "cannot be assessed without seeing them in the context of some external correspondence, that is, some demand originating outside the choice function itself". We suggest this represents a great

opportunity for research concerning external correspondence and its role in engineering decision-making. It seems to us that a theory of engineering design, recognizing the important role of decision making, must have something to say about not only how data are processed by an individual, but also how data are gathered via interaction with the real world. Such a theory might reveal exciting links between research in cognitive psychology and research in engineering design.

Acknowledgments This material is based upon work supported by the National Science Foundation under Grant No. 0448972.

References

Arrow KE (1951) Social choice and individual values. Wiley, New York

Bechara AH, Damasio AR (2000) Emotion, decision making, and the orbitofrontal cortex. Cereb Cortex 10(3):295–307

Begley RL Jr (1990) Steering column concept selection for low cost and weight: transactions from the 2nd symposium on quality function deployment. QFD Institute, Ann Arbor

Box GEP, Draper NR (1987) Empirical model-building and response surfaces. Wiley, Hoboken

Buede D, Maxwell DT (1995) Rank disagreement: a comparison of multi-criteria methodologies. J Multi Criteria Decis Anal 4:1–21

Czerlinski JG, Gigerenzer G, Goldstein DG (1999) How good are simple heuristics? In: Gigerenzer G, Todd PM, the ABC Research Group (eds) Simple heuristics that make us smart. Oxford University Press, New York, pp 97–118

Constable G, Somerville B (2003) A century of innovation: twenty engineering achievements that transformed our lives. National Academies Press, Washington, DC

Diederich A (1997) Dynamic stochastic models for decision making under time constraints. J Math Psychol 41:260–274

Dym CL, Wood WH, Scott MJ (2002) Rank ordering engineering designs: pairwise comparison charts and Borda counts. Res Eng Des 13(4):236–242

Franssen M (2005) Arrow's theorem, multi-criteria decision problems and multi-attribute preferences in engineering design. Res Eng Des 16(2005):42–56

Frey DD, Dym CL (2006) Validation of design methods: lessons from medicine. Res Eng Des 17(1):45–57

Gigerenzer G, Todd PM, the ABC Research Group (eds) (1999) Simple heuristics that make us smart. Oxford University Press, New York

Hazelrigg GA (1998) A framework for decision-based engineering design. ASME J Mech Des 120:653–658

Hazelrigg GA (1999) An axiomatic framework for engineering design. ASME J Mech Des 121:342–347

Johnson JG, Busemeyer JR (2005) A dynamic, stochastic, computational model of preference reversal phenomena. Psychol Rev 112(4):841–861

Katsikopoulos KV, Martignon L (2006) Naïve heuristics for paired comparisons: some results on their relative accuracy. J Math Psychol 50(3):488–494

Khan M, Smith DG (1989) Overcoming conceptual barriers—by systematic design. Proceedings of the Institute of Mechanical Engineers ICED, Harrogate

Lewis KE, Chen W, Schmidt LC (2006) Decision making in engineering design. ASME Press, New York

Li X, Sudarsanam N, Frey DD (2006) Regularities in data from factorial experiments. Complexity 11(5):32–45

Limayem F, Yannou B (2007) Selective assessment of judgmental inconsistencies in pairwise comparisons for group decision rating. Comput Oper Res 34:1824–1841

Miller K, Brand C, Heathcote N, Rutter B (2005) Quality function deployment and its application to automotive door design. Proc. IMecheE 219 part D: 1481–1493

Mistree F, Lewis K, Stonis L (1994) Selection in the conceptual design of aircraft. proc. of the 5th AIAA/USAF/NASA/ISSMO symposium on recent advances in multidisciplinary analysis and optimization, Panama City, FL AIAA-94-4382-CP

Pahl G, Beitz W (1984) Engineering design: a systematic approach. Springer-Verlag, Berlin

Pugh S (1981) Concept selection: a method that works. Proceeding of the international conference on engineering design ICED, Rome, Italy

Pugh S (1990) Total design. Addison-Wesley, Reading

Pugh S, Smith D (1976) The dangers of design methodology. First European Design Research Conference, Portsmouth

Saari DG, Sieberg KK (2004) Are partwise comparisons reliable? Res Eng Des 15:62–71

Saaty TL (2006) Rank from comparisons and from ratings in the analytical hierarchy/network processes. Eur J Oper Res 168:557–570

Salonen M, Perttula M (2005) Utilization of concept selection methods—a survey of Finnish Industry. ASME design engineering technical conferences, Long Beach

Scott MJ, Antonsson EK (1999) Arrow's theorem and engineering design decision making. Res Eng Des 11(4):218–228

Scott MJ (2007) Quantifying uncertainty in multicriteria concept selection. Res Eng Des 17:175–187

See T-K, Gurnani A, Lewis K (2004) Multi-attribute decision making using hypothetical equivalents and inequivalents. ASME J Mech Des 126:950–957

Sen A (1993) Internal consistency of choice. Econometrica 61(3):495–521

Sen A (1998) The possibility of social choice: nobel prize lecture. Trinity College, Cambridge

Shimojo S, Simion C, Shimojo E, Scheier C (2003) Gaze bias both reflects and influences preference. Nat Neurosci 6(12):1317–1322

Simion C, Shimojo S (2006) Gaze manipulation biases preference decisions. J Vis 3(9):306, 306a. http://journalofvision.org/3/9/306/; doi:10.1167/3.9.306

Smith J, Kaufman H, Baldasare J (1984) Direct estimation considered within a comparative judgment framework. Am J Psychol 97(3):343–358

Solow RM (1957) Technical change and the aggregate production function. Rev Econ Stat 39(3):312–320

Swoyer C (1991) Structural representations and surrogative reasoning. Synthese 87:393–415

Takai S, Ishii K (2004) Modifying Pugh's design concept evaluation methods. DETC2004–57512. ASME design engineering technical conferences, Salt Lake City, UT

Ullman DG (2002) Toward the ideal mechanical engineering design support system. Res Eng Des 13:55–64

von Neumann J, Morgenstern O (1953) The theory of games and economic behavior. Princeton University Press, Princeton

Ward A, Liker JK, Christiano JJ, Sobek DK (1995) The second Toyota paradox: how delaying decisions can make better cars faster. Sloan Manage Rev 36(3):43–61

Yang MC (2007) Design methods, tools, and outcome measures: a survey of practitioners DETC2007–35123. Proceedings of the ASME des eng technical conferences, Las Vegas