An Approach to Automated Scoring of Architectural Designs

Philip K. Oltman
Isaac I. Bejar
Sung Ho Kim

Educational Testing Service
Princeton, NJ 08541 USA

An automated approach to scoring architectural designs of building sites was devised. Design drawings were represented in the computer as a database consisting of a series of objects and their locations and relationships to other objects. Designs were automatically scored by: (a) determining whether or not certain basic requirements were met (such as parking, handicapped access, and building within site boundaries) and (b) determining the extent to which the drawings met criteria of efficient site design. The drawings had been scored earlier by a panel of expert registered architects. The automated technique was able to reproduce the human jurors' ratings for many drawings. In particular, drawings that the automated technique failed were quite likely to have been failed by human jurors as well.

Keywords: design evaluation, site design.

1 Introduction

The application of quantitative and artificial intelligence methods to architecture has been motivated by considerations ranging from substantive concerns, e.g., the history of architecture (Hillier etal., 1987), to the creation of productivity enhancing systems (e.g., Radford and Stevens, 1987). An unanticipated but natural application of such developments is to the assessment of architectural design expertise, which in turn can also lead to important advances in teaching, including curriculum-embedded assessment. We discuss aspects of that application in the paper and present empirical results on the feasibility of computer-based assessment of architectural design expertise.

A challenging goal in processing open-ended assessments is to score or grade the solutions automatically, i.e., without direct human assistance. It is the savings from not having to have humans grade the solutions that makes it attractive to administer tests by computer. Figure 1 provides a schematic view of the process we envision to perform automatic scoring. The candidate responds to one of several problems assigned to him or her. The resulting solution is analyzed into a low-level representation from which “features” are then computed. The resulting vector of features is mapped onto a score that classifies the answers into a “passing” or “failing” category.

U. Flemming and S. Van Wyk (eds.), ***CAAD Futures ‘93****, pp. 215-224.*

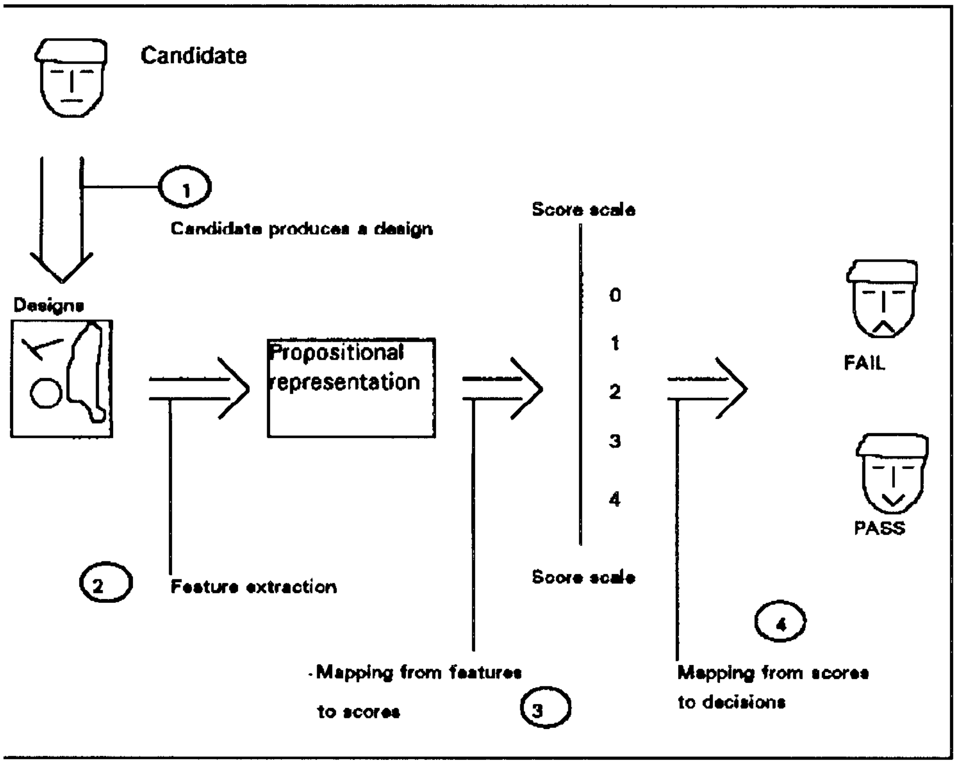
Candidate
1
Candidate produces a design
Designs
Propositional representation
2
Feature extraction
Score scale
0
1
2
3
4
Score scale
Mapping from features to scores
3
FAIL
PASS
4
Mapping from scores to decisions

Figure 1. Schematic view of the automated scoring process.

Because the notion of computer scoring is novel, and those candidates that fail might be somewhat skeptical, it is important to grade solutions in a principled fashion (Bejar, 1991), so that when necessary, the computer-generated scores can be "explained" to a candidate. Explanations at different levels are possible. In the case of architectural designs, at a modest level we can think of an explanation as a listing of the features on which the candidate's performance was less than ideal. But this raises the question of when a performance gives evidence of less than minimal competence. Clearly, judgment is required at some point and the question is how to factor that judgment into the process. The approach we will present assumes it is possible to convene a group of judges and have them rate a sample of the solutions into passing and failing categories in order to "train" the automated scoring algorithm, i.e., to capture a set of rules that will reproduce the human experts' behavior. An alternative approach which might be useful when it is not feasible to collect data from actual candidates would make use of advances in "generative design systems" with the idea of enumerating all possible solutions to a problem, e.g., Flemming (1986). In practice, we may need to generate schematic solutions, with details to be added by hand. Such generated solutions would then be graded by a panel of architects in order to capture the rationale architects use in grading solutions.

2 Method

2.1 The Design Problem

Drawings made by candidates for the site design section of an architecture licensing examination were analyzed. Candidates were asked to design a site adjacent to an existing college campus. The "bare site" before placement of the design elements is shown in Figure 2. Pedestrian traffic onto the site would flow from a pedestrian bridge leading from a large parking area and a pedestrian mall connecting to the main campus, both of which were already in place on the site. The site design was to accommodate a lecture hall, a new student center, ten service parking spaces and two handicapped spaces, dedicated open space, and a new outdoor amphitheater. Predrawn "footprints" were provided for the student center, lecture hall, and amphitheater, and candidates were asked to design the placement of these elements on the site. Trees, walkways, and open space were to be shown, and all construction was to be inside 20-foot setbacks from the property lines.

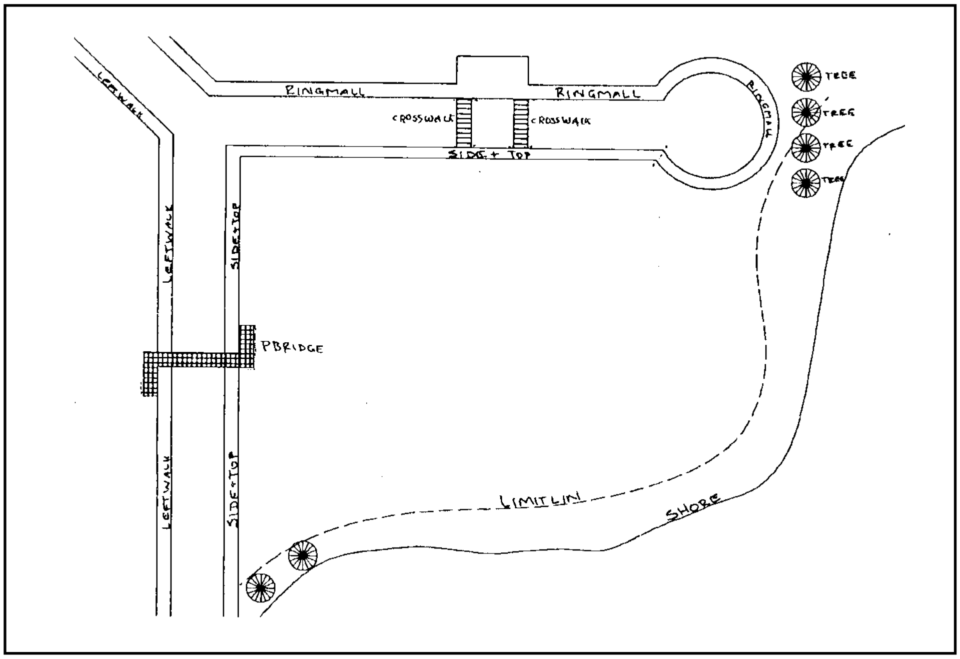

Figure 2. The "bare site" before placement of the design elements.

2.2 Drawings

An anonymous set of 180 site design drawings was made available for the project through the courtesy of the National Council of Architectural Registration Boards.

2.3 Converting the Drawings to Machine Readable Form

A typical site design drawing is shown in Figure 3. Candidates made their drawings with paper and pencil. Therefore, the first step in the study was to convert the drawings to machine-readable form.

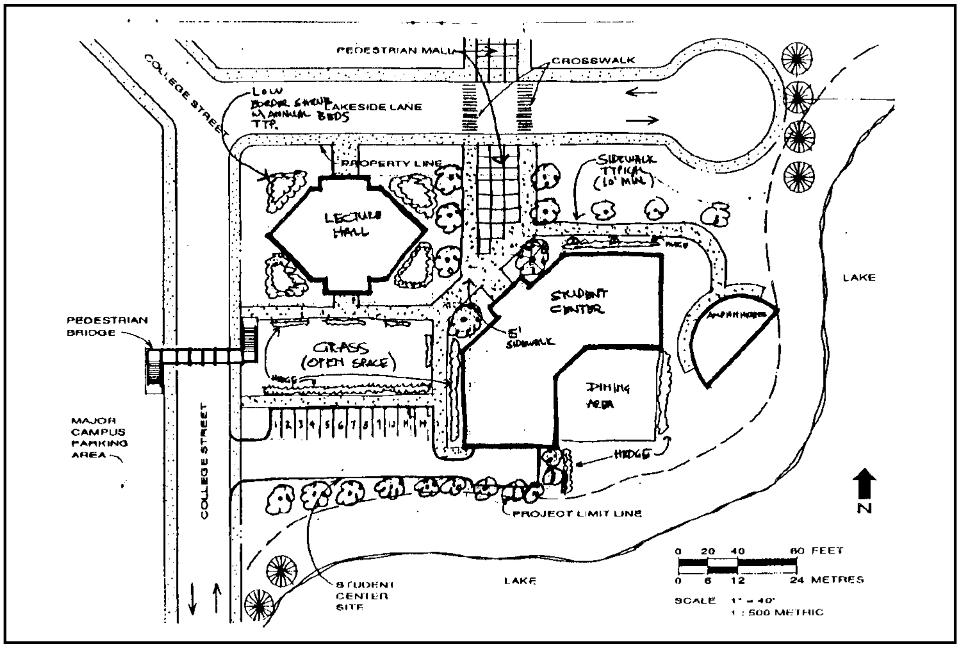

Figure 3. A typical site design drawing.

Scanning the drawings to produce bitmaps was considered unacceptable in this context. Converting the drawings to bitmaps would destroy their structure and would require “re-parsing” the bitmaps to recover the objects and their positions. Because we wanted our scoring system to work with the design logic of the drawings, we wished to preserve their structure.

Rather than scanning, we used an approach that started with the bare site as shown in Figure 2, onto which a schematic but detailed version of the candidate's design was transferred. As noted, the design exercise for the candidate consisted of placing various entities on the site, following a set of specifications. For conversion of the paper-and-pencil drawings to machine-readable form, each building or other object that the candidate was to place on the site was predrawn with computer-aided design (CAD) software, and stored in a library. The library of reusable objects consisted of a student center, a lecture hall, an amphitheater, and various vehicular and pedestrian paths and intersections of paths on the site. To convert a drawing, a coder first retrieved the bare site and displayed it. The candidate's paper drawing was placed on a digitizing tablet and calibrated by setting the coordinates of two fixed points on the site to correspond with the same points on the bare site displayed on the screen (the pedestrian bridge and the crosswalk from the mall). Next, each of the buildings, the parking lot, and the paths and their intersections and terminations were entered on the screen by using the digitizing tablet's cursor to locate a predefined reference point on each object and then retrieving the object from the library and entering it into the representation of the drawing on the screen. When this process was completed, the candidate's original drawing appeared on the computer screen in a schematic representation that preserved the basic structure of the solution (Figure 4 displays the schematic representation of the original drawing shown in Figure 3). We did not attempt to capture details such as placement of shrubbery or grass, or to represent the width of pathways from building to building. Nevertheless, the placement of each building or object, and the relationships of each to the others was preserved.

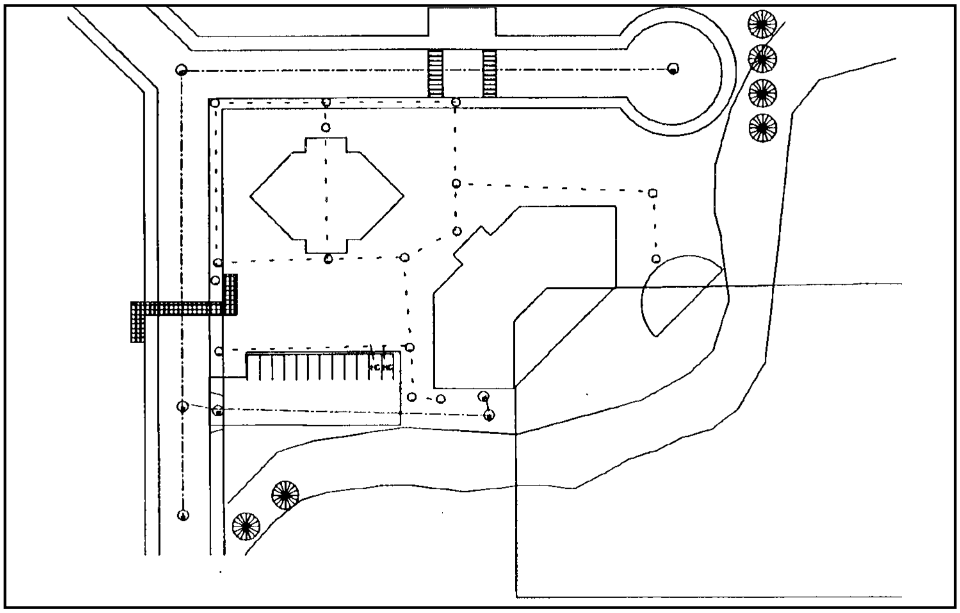

Figure 4. Schematic representation of the original drawing shown in Figure 3.

The CAD representation was stored in drawing exchange format (DXF), which is a standard protocol for drawings of this kind. This format consists essentially of a database describing each object and its position in terms of primitives such as lines, arcs, rotations, and locations.

2.4 *Drawing Processing*

Because the drawings were stored as databases containing information about the configuration of the designs, they could be queried to answer various evaluative questions. These questions were of two kinds: specification satisfactions, and relations among objects.

Specification satisfaction. A number of specifications were set forth explicitly in the instructions, while others were implied by standard practice or declared by an expert scoring committee during test development. For each drawing, each specification was either satisfied or not. The specifications used in scoring were the following:

Explicit specifications:

- Parking areas do not violate setbacks;
- Buildings do not violate setbacks;
- The correct number of parking stalls exists;
- Service drive to student union exists;

- Pedestrian access to buildings provided.

Implied specifications:

- Handicapped parking is adjacent to sidewalk;
- Handicapped pedestrians need not cross vehicular traffic lanes;
- The access to parking is located south of pedestrian bridge (as specified by expert committee).

Other specifications requiring proper use of open space and scenic views were set aside because of difficulties encountered in applying a scoring scheme for them. A requirement that the buildings relate to each other and to the pedestrian mall and the pedestrian bridge was scored using relations among buildings as described below.

Relations among objects. This category of variables consisted mainly of a series of distances between various pairs of objects on the site:

- LHSC: Distance from student center entrance to nearest lecture hall entrance;
- LHML: Distance from pedestrian mall to nearest lecture hall entrance;
- LHPB: Distance from pedestrian bridge to nearest lecture hall entrance;
- TRIAV: Average of distances among lecture hall, student center, and pedestrian mall;
- TRISD: Standard deviation of distances among lecture hall, student center, and pedestrian mall;
- BDIST: Closest point distance between student center and lecture hall;
- PKAR: Area of parking lot.

The scoring program initially queried each drawing's representation to extract a value for each of the variables, i.e., a “yes” or “no” for each of the specifications, and a numerical value for each of the relations among objects.

2.5 Scoring Rules

Exploratory analysis. Scoring rules were developed that combined the values of the variables to produce criteria for passing or failing scores. Our strategy was to compose a set of rules that would sort candidates' drawings into those that obviously failed, and those that are passed (or that must be given to human experts for further evaluation). For purposes of this investigation, we did not assume every drawing would be scorable by the computer, and therefore we characterized the scoring categories as “fail” and “possibly pass.” Criteria were selected so as to fail only those candidates who also would be failed by a human expert. Because we had data from human expert jurors, we could develop scoring rules that followed this strategy on a subset of the drawings, and then test it on the remaining drawings.

The scoring rationale was twofold. First, candidate drawings had to satisfy all, or most, of the specifications (setback rules, parking, handicapped access, etc.). Second, buildings were to relate to each other and to the access points of the site. This latter requirement was simplified and made concrete by defining it to mean that circulation of pedestrians on the site should be optimized. We selected a simple correlate of circulation to score, namely that the distances between buildings and access points would be relatively small. The underlying assumption was that circulation on the site would be more efficient when distances between facilities were not made larger than necessary.

The 180 drawings were distributed at random into three sets of 60 drawings each. Scoring criteria were developed empirically on the first set of 60 drawings, adjusting cutoffs on the specifications and relational variables so that the results of applying the cutoffs approximated the behavior of the human jurors. The goal was to detect as many of the drawings that were failed by the human jurors as possible, while minimizing the number of drawings that, while passed by the human, were failed using the cutoffs. A number of combinations of variables were explored to arrive at a satisfactory solution. The scoring rules that were adopted were as follows:

1. If the design failed two or more of the specifications, it was assigned a failed rating, and was not further considered.
2. If the distances among the major buildings on the site (the lecture hall and the student center) and the entrances to the site (the pedestrian bridge and pedestrian mall) were too large (TRIAV greater than 3.0 or LHPB greater than 2.5), the design was failed.

Confirmatory analysis. The scoring criteria were established empirically on the first random set of 60 drawings and then cross-validated on the second and third sets.

3 Results

3.1 Interjuror Agreement

All drawings had been rated previously by two expert architects; when they disagreed concerning a pass/fail decision, a third “master” juror was called in to decide the matter. For the 180 drawings in the study, the initial two human expert jurors agreed with each other 79 percent of the time on the decision to pass or fail a drawing. These results are shown in Table 1. The pattern of results in Table 1 yielded a kappa (Cohen, 1960) of 0.49 ($p < 0.01$), which has been characterized as “fair to good” agreement.

When the two initial human jurors disagreed as to whether a drawing passed or failed, a third juror was called upon to make the decision. The mean of all jurors (two, or in the event of disagreement, three) was used as the standard against which to evaluate the automated scoring algorithm.

Table 1. Agreement Between Two Human Jurors in Total Sample

		Juror 2			
		Fail	Pass	Total	kappa
Juror 1	Fail	108	17	125	0.49
	Pass	21	34	55	
	Total	129	51	180	

3.2 Agreement of Automated Scoring with the Human Jurors

Plots of ratings by human expert jurors against the variables that were generated by processing the drawings were examined using the first group of 60 randomly selected drawings. Failure of two or more of the specifications was clearly associated with failing ratings by the human expert jurors, as was the presence of relatively long distances among buildings and entrances to the site. However, the relations were not linear. That is, the plots showed that failing two or more specifications or having relatively long distances among buildings and site entrances were almost always associated with failing ratings by the experts, but passing these scoring criteria was not strongly associated with passing ratings by the experts. The data from the first set of 60 drawings suggested the following scoring algorithm:

- If the drawing fails two or more specifications OR the distances among buildings and entrances to the site are relatively large (TRIAV greater than 3.0 or LHPB greater than 2.5), then fail the drawing;
- Otherwise, pass the drawing (or send the drawing to a human expert juror for scoring).

Table 2 shows the results of setting these criteria empirically for the first 60 drawings, and the results of applying the same criteria to the second and third sets of drawings. Recall that Cohen's kappa for interjuror agreement was 0.49. The kappas for each group expressing the extent of agreement between the automated scoring algorithm and the mean of the human jurors were 0.50, 0.55, and 0.49 respectively (each with $p < 0.01$). The second and third groups represent a cross-validation of the empirical rules devised with the first group's data. Thus, it appears that the exploratory automated scoring algorithm can do as well as human jurors, in the sense that it agrees at least as well with humans as they do with each other.

4 Discussion

An efficient site design will allow convenient access to its various facilities and will not require more pedestrian and vehicular travel than is needed for such access. In addition, of course, the design must conform to the specifications stated by the client and the applicable codes and regulations. An attempt was made to encompass both circulation efficiency and specification satisfaction in our scoring algorithm.

No attempt was made to evaluate the designs for their aesthetic merit. In fact, the representation of the drawings in the computer was quite schematic. Nonetheless, sufficient information was abstracted from the drawings to permit deriving scores that related as well to human judgments as the two human jurors' ratings did to each other.

In the course of exploring the data, a number of combinations of variables and criteria produced reasonably good results. However, it became clear different strategies had different value tradeoffs. It was possible to select an algorithm with fewer "false failures," but at the expense of having to pass on a far greater number of drawings to human jurors. For the results reported above, the program had to pass on 38 percent of the drawings to human scorers, and failed six candidates who were passed by the human jurors. Using another algorithm with a different combination of variables, we has only two "false failures," but has to pass on 56 percent of the drawings. One of the advantages of automated scoring is that it can reduce the cost of scoring by eliminating the need for at least a substantial portion of the human judgment expense. However, increasing the automatically-scored portion may bring with it an increase in false failures. A machine-based scoring system can have a very large number of possible strategies available. The relative costs and benefits of adopting different algorithms must be considered in adopting such a system for operational use.

Table 2. Human Jurors Versus Scoring Algorithm in Three Samples

Group 1		**Automated Scoring Algorithm**			
		Fail	Pass	Total	kappa
Human Jurors	Fail	41	11	52	0.50
	Pass	0	8	8	
	Total	41	19	60	

Group 2		**Automated Scoring Algorithm**			
		Fail	Pass	Total	kappa
Human Jurors	Fail	32	12	44	0.55
	Pass	1	15	16	
	Total	33	27	60	

Group 3		**Automated Scoring Algorithm**			
		Fail	Pass	Total	kappa
Human Jurors	Fail	32	9	41	0.49
	Pass	5	14	19	
	Total	37	23	60	

The answer to the question of when performance on some feature is too high or too low is answered by the set of if-then rules. The question, of course, is why must no more than two criteria be failed for a solution to be classified as passing? One answer is that each rule should not be viewed independently, but rather as a set. The seemingly arbitrary cut points do not need to be interpreted independently. Rather, the scoring process is viewed as a means of capturing the extent to which a given design deviates far enough from correctness to be considered as being produced by someone who lacks minimal competence. This in some sense ducks the question, or more accurately passes it on to the psychometricians. Specifically, it is the responsibility of the psychometrician to demonstrate through whatever means that such a characterization is indeed valid. One aspect of the validation process is related to the criteria we imposed on ourselves, namely that the scoring should be principled enough to allow for an explanation of the decision. In this context, such an explanation would be a verbalization of the rules together with pointing out the flaws in the design itself. Whether such explanations would be acceptable to candidates remains to be seen. It is the case that currently candidates seldom question the decisions of human graders, but it can be reasonably expected challenge computer-based scores would more likely be challenged. We intend to investigate the acceptability of such explanations by carrying out, in effect, a Turing test. That is, critiques provided by human architects would be paraphrased into the same sublanguage into which we would paraphrase the feature representation of a solution, and given to candidates to determine if a difference exists between the computer explanation and the human explanation.

To make computer scoring of design drawings practical, an interface must be devised that will allow candidates to enter their own solutions as they produce them (Akin, 1986). Scoring rules must be made more general, so less "customizing" needs to be done with each new problem. This will likely be achieved by deriving rules for classes of design vignettes. Such rule sets will, in turn, inform attempts to introduce new problems with difficulty levels that can be determined *a priori*.

References

Akin, O., 1986. *Psychology of Architectural Design*. London: Pion.

Bejar, I.I., 1991. "A Methodology for Scoring Open-ended Architectural Design Problems," *Journal of Applied Psychology* 76, pp. 522-532.

Cohen, J., 1960. "A Coefficient of Agreement for Nominal Scales," *Educational and Psychological Measurement* 20, pp. 37-46.

Flemming, U., Coyne, R., Glavin, T., and Rychener, M., 1986. "A Generative Expert System for the Design of Building Layouts," in D. Siriam and R. Adey (eds.), *Application of Artificial Intelligence to Engineering Problems*. Berlin: Springer-Verlag.

Hillier, B. and Hanson, J., 1987. "Ideas Are in Things: An Application of the Space Syntax Method to Discovering House Genotypes," *Environment and Planning B: Planning and Design* 14, pp. 363-385.

Radford, A. D., and Stevens, G., 1987. *CAD Made Easy: A Comprehensive Guide for Architects and Designers*. New York: McGraw-Hill.