Some Tours are More Equal than Others: The Convex-Hull Model Revisited with Lessons for Testing Models of the Traveling Salesperson Problem

Susanne Tak[1], Marco Plaisier[1] and Iris van Rooij[2]

Abstract

To explain human performance on the *Traveling Salesperson problem* (TSP), MacGregor, Ormerod, and Chronicle (2000) proposed that humans construct solutions according to the steps described by their convex-hull algorithm. Focusing on tour length as the dependent variable, and using only random or semirandom point sets, the authors claimed empirical support for their model. In this paper we argue that the empirical tests performed by MacGregor et al. do not constitute support for the model, because they instantiate what Meehl (1997) coined "weak tests" (i.e., tests with a high probability of yielding confirmation even if the model is false). To perform "strong" tests of the model, we implemented the algorithm in a computer program and compared its performance to that of humans on six point sets. The comparison reveals substantial and systematic differences in the *shapes* of the tours produced by the algorithm and human participants, for five of the six point sets. The methodological lesson for testing TSP models is twofold: (1) Include qualitative measures (such as tour shape) as a dependent variable, and (2) use point sets for which the model makes "risky" predictions.

1. Introduction

The Traveling Salesperson problem (TSP) is a well-known combinatorial optimization problem. The problem consists of finding the shortest tour visiting a set of points in the plane.[1] TSP has a search space that displays an inherent combinatorial explosion: e.g., for $n = 5$ only 12 tours exist, but for $n = 14$ already more than 3 billion tours exist. This means that solving TSP by exhaustive search (i.e., by checking all tours and then selecting the shortest one) is computationally unfeasible for all but very small instances. Furthermore,

[1]Human-Technology Interaction, Eindhoven University of Technology, Eindhoven, The Netherlands
[2]Nijmegen Institute for Cognition and Information, Radboud University Nijmegen, Nijmegen, The Netherlands

Address correspondence to:
Iris van Rooij, Nijmegen Institute for Cognition and Information (NICI), Radboud University Nijmegen, Montessorilaan 3, 6525 HR Nijmegen, The Netherlands
Email: i.vanrooij@nici.ru.nl

http://dx.doi.org/10.7771/1932-6246.1028

it seems that there cannot exist any nonexhaustive algorithm solving the TSP, because the problem is known to be *NP*-hard (Garey & Johnson, 1979). Despite its computational intractability, however, humans are found to show remarkably good performance on the TSP (MacGregor & Ormerod, 1996), producing close to optimal tours with seemingly little effort and using little time even for point sets as large as 20 points or more. This observation has motivated researchers to set out to identify the human strategy for solving TSPs and implement it in a computational model. In this paper, we take a closer look at one such model, the Convex-hull (CH) algorithm model proposed by MacGregor, Ormerod, and Chronicle (2000), and its purported fit to human performance on the TSP (for altogether different modeling attempts for TSP see Graham, Joshi, & Pizlo, 2000; and Pizlo et al., 2006).

The CH model of MacGregor et al. (2000) is a formal algorithmic elaboration on the convex-hull hypothesis put forth by MacGregor and Ormerod (1996; see van Rooij, Stege, & Schactman, 2003, for an overall assessment of the empirical support for this hypothesis) that proposes that people construct solutions to the TSP by first perceiving (and mentally sketching) the boundary of the point set (called the convex hull) and then inserting one by one the interior points in the tour. To investigate the extent to which this CH algorithm captures the process underlying human performance on the TSP, MacGregor et al. (2000) compared tours produced by the algorithm (for different starting points and directions of travel) with tours produced by humans. They studied in particular random or semirandom point sets, and found that tours produced by model and participants resembled each other in terms of tour length. More precisely, it was found that the model produced a good fit for both average and minimal tour length. MacGregor et al. took these findings as providing evidence that people may be solving TSPs by a process that resembles the steps taken by their proposed CH algorithm.

Notwithstanding the intuitive appeal of the CH model and the apparent fit to human performance reported by MacGregor et al. (2000), we submit that evidence that people solve TSP by a process analogous to the CH algorithm is nonexistent at present or weak at best. The reason is that the empirical tests performed by MacGregor et al. were not strong tests of the CH model, because the tests were such that confirmation was likely to be found, even if the model were false (Meehl, 1997). First of all, it is known that for many (randomly generated) point sets human and model performance is close to optimal (MacGregor et al., 2000; van Rooij, Schactman, Kadlec, & Stege, 2006; Vickers, Butavicius, Lee, & Medvedev, 2001), making it difficult to detect differences between human and model performance in terms of tour length. Put differently, any algorithm that produces close to optimal tours for random point sets will give a reasonably good fit to human performance measured in terms of tour length. The focus on tour length as a dependent measure overlooks the fact that there may nevertheless be systematic and substantial differences in terms of the *shapes* of the tours that model and humans produce. Secondly, MacGregor et al. did not attempt to test any surprising predictions that can be derived from their model (Roberts &

Pashler, 2000). In fact, their choice for point sets was not driven by predictions that could be derived from their model at all. In sum, we feel that MacGregor et al. (2000) failed to subject their model to a strong test (Meehl, 1997) and hence strong support for their model is lacking to date.

In this paper we report on several strong tests of the CH algorithm as a model of human TSP performance, using six deliberately selected point sets and adopting tour shape as our main dependent measure. We observe that most of our tests in fact disconfirm the CH model. It seems therefore that, the CH algorithm fails to capture the human problem-solving process underlying human performance on the TSP. Despite this negative result, we believe our paper makes several positive contributions to the study of human TSP performance, not in the least because our results identify a set of characteristic properties of tours produced by humans that may guide future attempts at modeling human TSP performance. Also, we discuss several methodological lessons than can be drawn from our experiment for testing and formulating computational models of human TSP performance in general.

1.1. Overview

We start by explaining in detail the CH model of MacGregor et al. (2000) and its implementation as a well-defined algorithm (Section 2). Subsequently, we present relevant details about our participants (Section 3.1), the six point sets used in our experiment, as well as the rationale for using them (Section 3.2), and the experimental procedure (Section 3.3). In Section 4, we report on comparisons of both the lengths and the shapes of tours produced by model and participants. The nature of the observed differences is discussed for each of the point sets separately (Sections 4.1 – 4.6). In Section 5, we present an overview of how our results map to four classes of explanatory failures on the part of the model. We conclude, in Section 6, with methodological recommendations for future research on human performance on the TSP and a theoretical discussion of the research's validity and implications.

2. Implementation of the model

The MacGregor et al. (2000) model was not previously implemented in a working computer program. M. Plaisier and S. Tak have written such a program using Visual Basic 6.0. The program and its source code are available online at http://tsp.wtak.nl. In this section we describe the implementation.

We start by giving an intuitive characterization of the MacGregor et al. (2000) model: It assumes that humans generate tours starting with the *convex hull* as a base. The convex hull can be visualized as an imaginary rubber band being wrapped around the entire

point set. More formally, the convex hull is the smallest convex polygon that contains all the points (Cormen, Leiserson & Rivest, 2001, p. 947). The model consists of an algorithm that first sketches the convex hull of the presented point set. Subsequently, a random starting point and a direction of travel (clockwise or counterclockwise) are chosen, and the algorithm inserts unconnected interior points into previously formed tour segments in an iterative fashion.

More precisely, the algorithm works as per the following stepwise specification (adopted from MacGregor et al., 2000, pp. 1185-1186, but with a correction in Step 3 that is explained Appendix A):

> *Step 1*: Sketch the connections between adjacent boundary points of the convex hull.
> *Step 2*: Select a starting point and direction randomly.
> *Step 3*: If the starting point is on the boundary, the starting node is the *current node*. The arc connecting the current node to the adjacent boundary node in the direction of travel is referred to as the *current arc*. Proceed immediately to Step 4. If the starting point is not on the boundary, apply the insertion rule to find the closest arc on the boundary. Remove this arc and create two new arcs, connecting the starting point to each end of the removed arc. The end node of the removed arc becomes the current node.
> *Step 4*: Apply the insertion criterion to identify which unconnected interior point is closest to the current arc. Apply the insertion criterion to check whether the closest node is closer to any other arc. If not, proceed to Step 5. If it is, move to the end node of the current arc. This becomes the current node. Repeat Step 4.
> *Step 5*: Insert the closest node. The connection between the current node and the newly inserted node becomes the current arc. Retaining the current node, return to Step 4 and repeat Steps 4 and 5 until a complete tour is obtained.

An important design choice is which insertion criterion to use in Step 4. There are multiple criteria available to define the closest arc or point, such as the shortest Euclidean distance from the point to the arc, largest angle or cheapest insertion (MacGregor et al., 2000). Although at first Euclidean distance may seem like a sensible measure of closeness, it actually makes for a psychologically implausible insertion criterion. The reason is that this closeness measure is insensitive to the length of the current arc and the orientation of the point relevant to the arc. As a consequence, a Euclidean insertion criterion entirely ignores the increase in tour length created by inserting a point, generating many tours that are far from optimal and as such providing a poor model of human performance. The largest angle (LA) criterion overcomes this problem to some extent. According to this crite-

rion the closeness of point p to an arc (a, b) is given by the angle between edge (a, p) and edge (p, b). As a result, if arc (a, b) is longer than arc (c, d), and point p is equidistant to (a, b) and (c, d) according to a Euclidean distance measure, then according to the LA criterion p is closer to an arc (a, b) than to (c, d). The cheapest indentation (CI) criterion tackles the problem more directly, however, as it directly takes into account the added length due to inserting point p. According to the CI criterion the closeness of a point p to arc (a, b) is given the added length $D(a,p) + D(p,b) - D(a,b)$ created when inserting p in between a and b, where $D(x, y)$ is the Euclidean distance from point x to point y. The point that, when inserted into a certain arc, results in the least added length (compared to inserting other points into the same arc), is regarded *closest* to that particular arc.

MacGregor et al. (2000) considered both the LA and CI criteria, but observed that their convex-hull algorithm using the CI criterion resulted in the closest fit with human performance. For this reason, we have decided to focus our experimental investigation of the algorithm with the CI criterion. To derive predictions for the experiment, the computer program was run once for each possible starting point and each possible direction of travel (clockwise or counterclockwise). This simulates model behavior for an infinite number of runs, because—according to Step 2 of the algorithm—the starting point and the direction of travel are chosen randomly. In the cases where two or more points are at equal distance from the current arc all possibilities were explored (with each point having a probability of $1/n$ of being inserted, where n denotes the number of points at equal distance from the current arc).

3. Methods

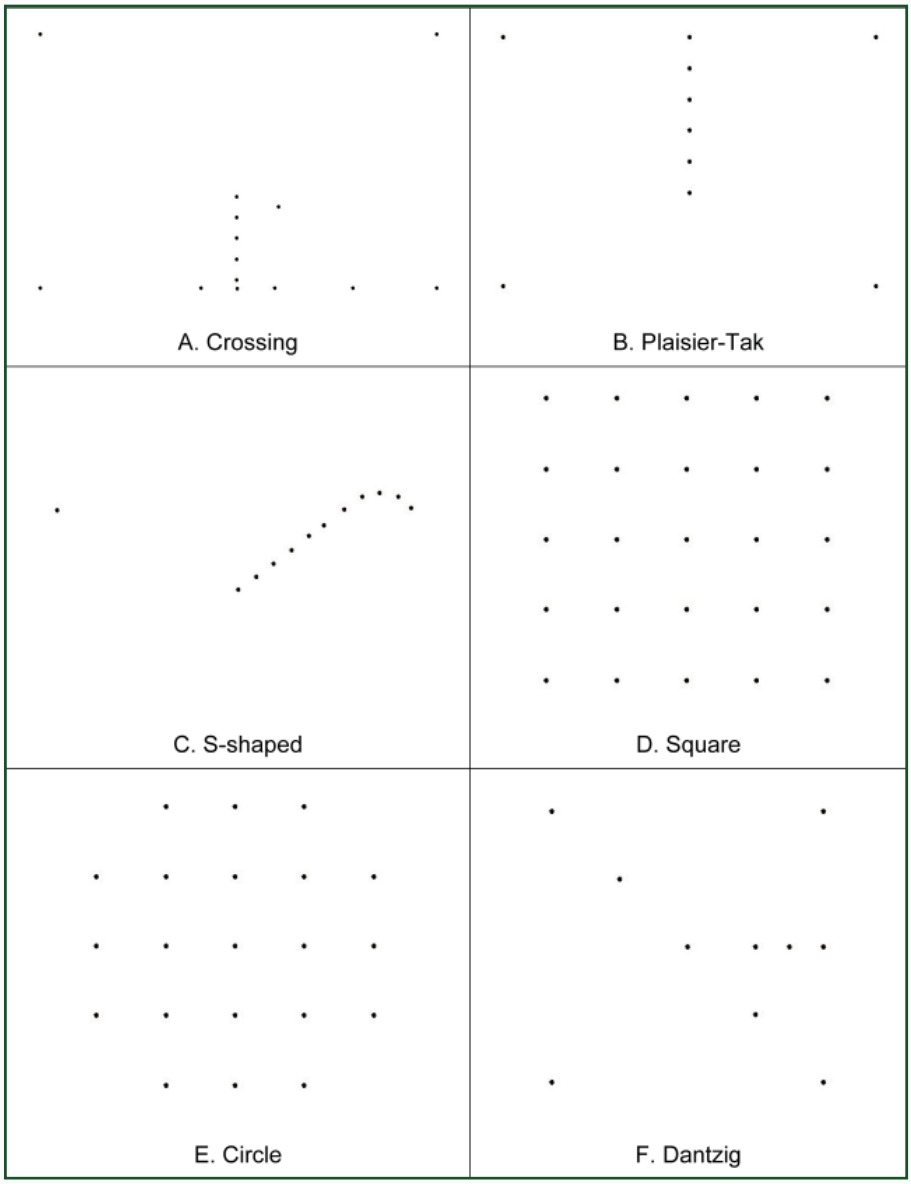
3.1. Participants

A total of 30 students of the Eindhoven University of Technology participated in the experiment. Participants included both males and females, ranging between 20-30 years of age. All participants were naïve with regard to the purpose of the study. The experiment was undertaken with the consent of each participant. Participants were paid 2 for participation.

3.2. Stimuli

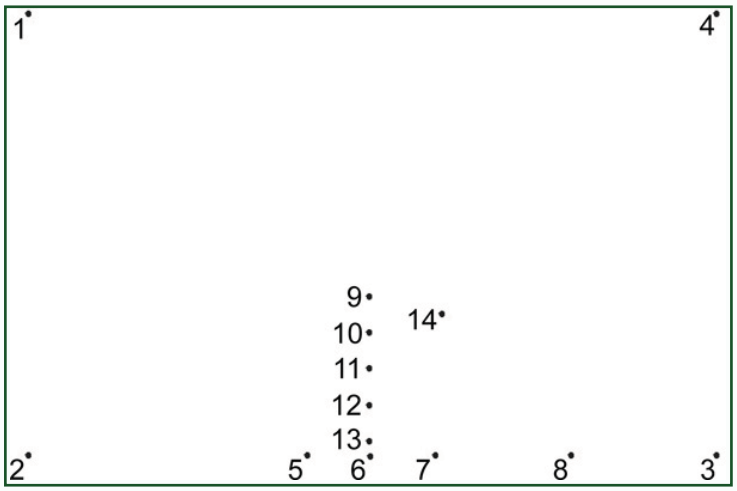
In our choice for stimuli we aimed at selecting point sets that instantiate critical tests of the CH model. In total six different point sets were used, referred to as the *Crossing*, *Plaisier-Tak*, *S-shaped*, *Square*, *Circle* and *Dantzig* point sets. The point sets are shown in Figure 1 and the coordinates for each point set are listed in Appendix B. We next explain the rationale for using each point set.

Figure 1. The six point sets used in the experiment.



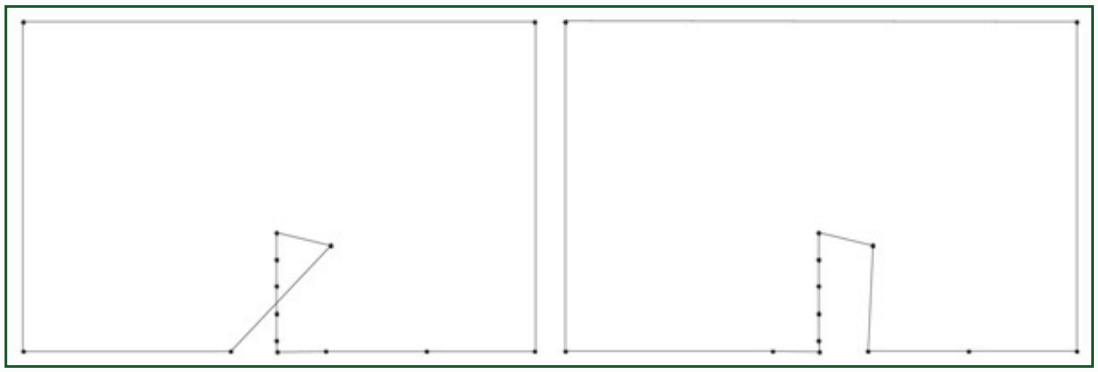
Crossing point set: A robust finding, replicated in many different studies of TSP, is that people produce relatively few tours with tour segments that cross in the plane (MacGregor & Ormerod, 1996; van Rooij et al., 2003). MacGregor and colleagues have proposed that a convex hull-based strategy can account for this finding (MacGregor & Ormerod, 1996; MacGregor et al., 2000, 2004). This raises two questions: Can the CH algorithm of MacGregor et al. (2000) produce tours with crossings? And if so, do humans and model produce tours with crossings for the same type of point sets? We discovered that the algorithm can indeed produce tours with crossings. Consider, for example, the Crossing point set, depicted in Figure 2 with the points labeled. The tour created most frequently (82%) by the model

Figure 2. The Crossing point set with labeled points for referencing (see Stimuli section for details)..



for this point set, called tour A, is shown in Figure 3 on the left. Tour A is generated, for example, if the starting point is point 2 and the direction of travel is counterclockwise. Point 13 is closer to arc 5-6 than to arc 6-3. Therefore, after having reached arc 5-6 as the current arc point 13, 12, 11, 10 and 9 are consecutively inserted. After this, 5-9 becomes the current arc. Using the cheapest insertion criterion point 14 is closest to arc 5-9 (and not arc 9-10!) and is inserted, creating a crossing. As we did not expect that participants would create this kind of tour, we selected the point set for our experiment.

Figure 3. The tour most frequently produced by the model (tour A; left) and the tour most frequently produced by participants (tour C; right) for the Crossing point set.



We remark that, although poorly visible to the naked eye, points 5, 7 and 8 in the Crossing point set are not on the convex hull. As participants may fail to notice this, we also investigated model performance when points 5, 7 and 8 are placed on the convex hull in such a way that points 2, 5, 6, 7, 8 and 3 are on a straight line. We found that the model still produces a tour with a crossing in 50% of the cases. In the Result section we test for both the original and adjusted scenario.

Figure 4. The tour shape most frequently produced by the model (tour shape A; left) and the tour shape most frequently produced by participants (tour shape E; right) for the Plaisier-Tak point set.

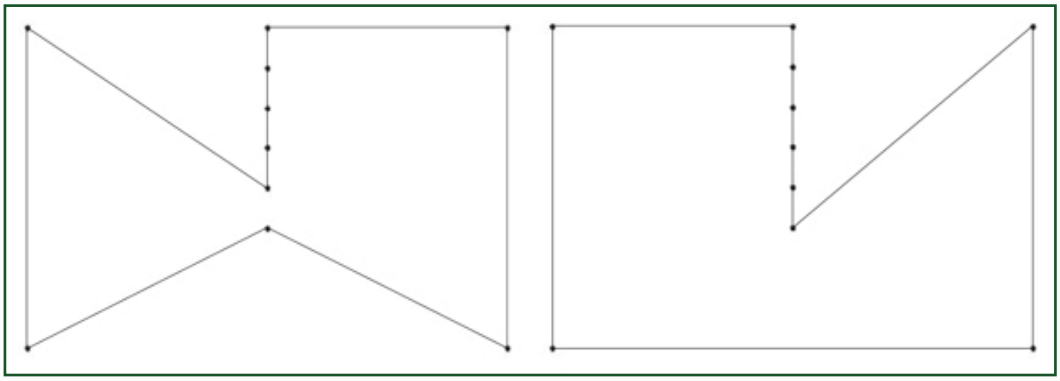
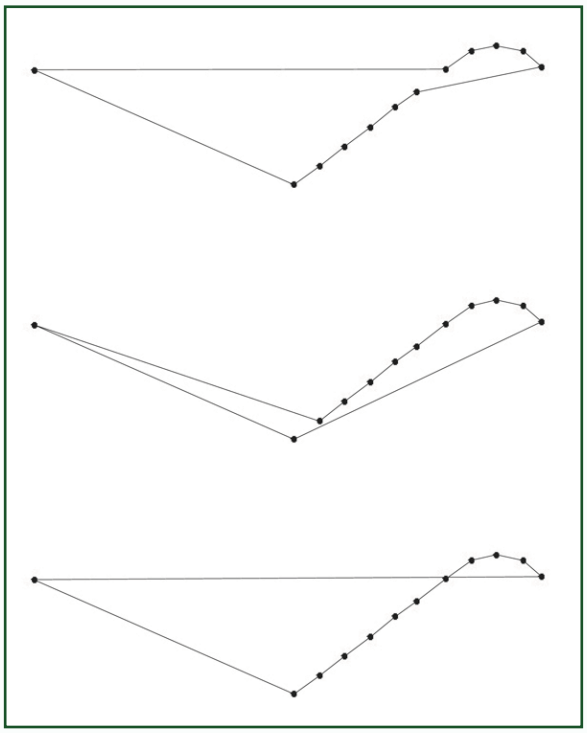


Figure 5. The tour most frequently produced by the model (tour A; top) and the two tours most frequently produced by participants (tour B, middle; tour C, bottom) for the S-shaped point set.

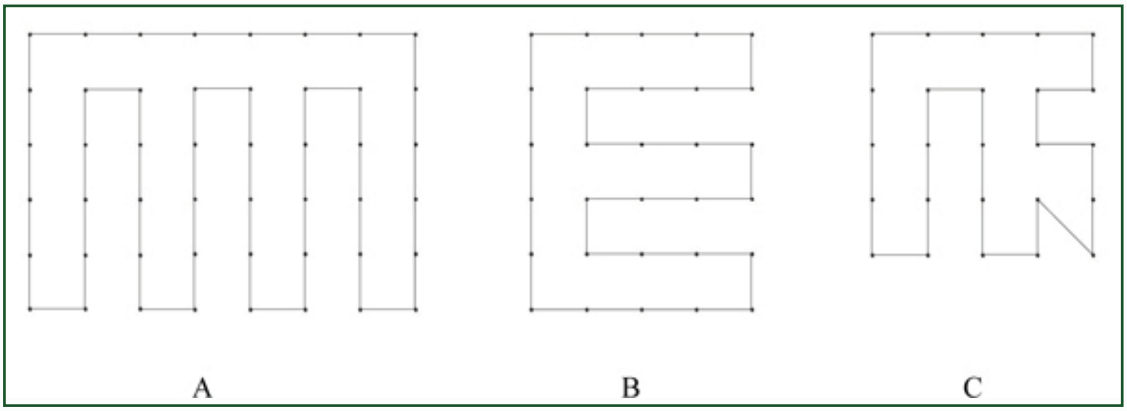


Plaisier-Tak and S-shaped point sets: MacGregor et al. (2000, p. 1189) observed that "figural factors . . . appear to influence human solutions." To investigate whether people are guided in their tour construction by the Gestalt principles of proximity and good continuation (Wertheimer, 1923), we decided to include two point sets. The first is the Plaisier-Tak point set depicted in Figure 1B. The second is the S-shaped point set adopted from MacGregor & Ormerod (1996) and depicted in Figure 1C. Based on the principles of

proximity and good continuation we expect participants to produce a suboptimal tour with a large indent for the Plaisier-Tak point set (see Figure 4 on the right) and a suboptimal tour with a crossing for the S-shaped point set (Figure 5 at the bottom). The model would produce optimal tours for both point sets (Figure 4 on the left and Figure 5 at the top).

Square and Circle point sets: MacGregor, Ormerod, and Chronicle (1999, p. 1426) proposed that "Human participants are able to capitalize upon the presence of regularity in a point array," and that "they probably use perceptual hypotheses about good figure in order to do this." This statement seems to conflate regularity of a point set with the possibility of "good figure" tours for that point set. Take, for example, the point set used by MacGregor et al. (1999; Experiment 2) shown in Figure 6A. This point set consists of a regular array of 8×6 points, and hence contains an even number of points. In general, regular arrays with an even number of points allow for symmetric, regular and relatively easy solutions (see Figure 6B for another example). On the other hand, regular arrays with an odd number of points do not. For example, a symmetric and regular "E-shaped" solution as depicted in Figures 6A and 6B is not possible for the 5×5 point set shown in Figure 6C. We decided to include the 5×5 point set in our experiment (Figure 1D). With this Square point set we aim to test the effects of point set regularity on performance when "figural goodness" of tours is not supported, and to investigate how model and humans compare in this respect. The rationale for using the Circle point set (Figure 1E) is the same as for using the Square point set; again, there is no regular and symmetric solution possible.

Figure 6. Regular point sets of 8 x 6 points (A), 5 x 6 points (B) and 5 x 5 points (C), and an optimal solution for each.



***Dantzig point set*:** Lastly, we included the Dantzig point set (adopted from Dantzig, Fulkerson, & Johnson, 1959), because according to MacGregor and Ormerod, it "represents a benchmark test, one that many relatively effective heuristics fail to solve" (1996, p. 529). The Dantzig point set is shown in Figure 1F. As far as we know the model of MacGregor et al. (2000) has not been tested on this point set before.

3.3. Procedure

The participants were tested in small groups and were seated far apart to prevent communication between participants during the experiment. The experimenter was present at all times. The point sets were presented in a paper booklet and participants were given a pencil and an eraser. The order of the point sets was randomized without repetition of orders. Participants were instructed to choose a starting point, draw a tour that passes each point at least once and return to the starting point in such a way that the overall distance traveled was minimized. An example was included in the instructions. Participants were further instructed to complete the task for each point set in the order in which they appeared in the booklet and not to take longer than a few minutes per point set. It took participants approximately 15 minutes to complete the experiment.

4. Results

Of the total 180 tours produced by the participants, 15 tours were excluded from the analyses because they did not visit all points in the point set (2 for Crossing, 1 for Square and

Table 1. Overview of descriptive statistics for model and participants per point set.

		Tour length (in cm)*					
	N	median	mean	SD	optimal	% of tours that were optimal	% of tours with crossings
Crossing							
Model	28	64.0	63.9	0.4	62.6	7.1%	82.1%
Participants	27	62.6	64.6	3.2		51.9%	7.4%
Plaisier-Tak							
Model	20	51.4	51.6	0.3	51.4	60%	0%
Participants	29	52.1	53.4	3.2		10.3%	17.2%
S-shaped							
Model	22	26.7	26.7	0.0	26.7	100%	0%
Participants	26	26.9	27.3	0.5		0%	42.3%
Square							
Model	50	57.9	57.2	1.1	56.0	42.4%	0%
Participants	27	57.9	58.4	3.0		44.4%	29.6%
Circle							
Model	42	49.0	49.7	0.8	49.0	61.9%	9.5%
Participants	28	51.8	52.2	3.9		25.0%	21.4%
Dantzig							
Model	20	41.0	41.0	0.6	40.1	10.0%	0%
Participants	28	41.8	42.5	2.3		14.3%	25.0%

* For readability, tour lengths have been converted from twips into centimeters. Here 1 cm corresponds to 567 twips.

1 for Circle) or did not form a closed path (2 for Dantzig, 1 for Plaisier-Tak, 4 for S-shaped, 2 for Square, 1 for Circle and 1 for Crossing). For the remaining 165 tours, Table 1 presents descriptive statistics per point set. Using tour length as the dependent variable, model performance was found to be significantly better than human performance for the Circle point set (Mann-Whitney U test, $p < .001$), the S-shaped point set ($p < .001$), the Plaisier-Tak point set ($p < .001$) and marginally better than human performance for the Dantzig point set ($p < .10$). There was no significant difference in tour lengths produced by model and participants for the Crossing and Square point sets.

To compare the tours produced by model and participants with respect to their shapes we classified tours as belonging to different shape classes depending on the number, size and shape of tour indentations. An *indentation* in a tour is the situation where two adjacent points on the convex hull are not directly connected by a single line segment, but instead are connected by a path of line segments that visits at least one interior point. We classified two tours as belonging to the same shape class if and only if (1) the tours were of equal length and (2) the tours had the same number of indentations and (3) there existed a one-to-one mapping of indentations in one tour to indentations in the other tour having the same surface area (see Figure 7 for an illustration). For each point set, each shape class was assigned a letter (A, B, C, etc.). For the Crossing and S-shaped point sets a tour shape class always contained exactly one unique tour shape. For the other point sets it was possible for two distinct tours to belong to the same shape class, while having a different configuration of indentations along the convex hull.

Figure 7. Illustration of two tours belonging to the same shape class for the Square point set (tour shape C). See text for details on how shape classes are defined.

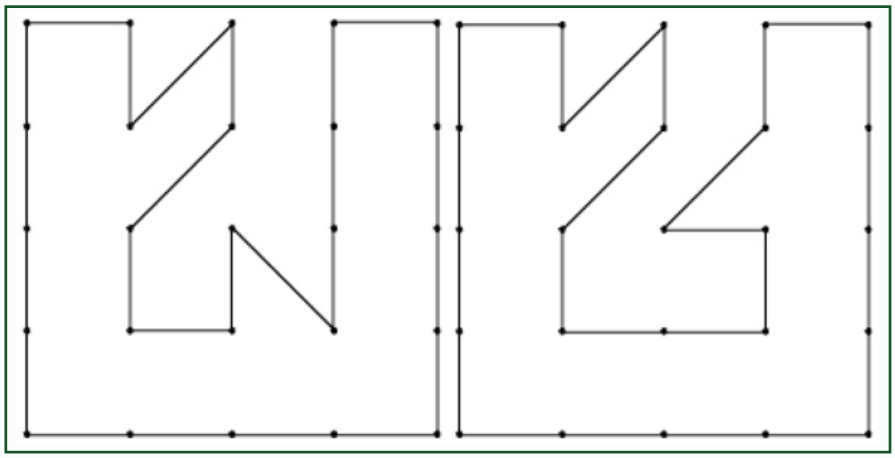


Table 2 presents an overview of the number of different tour shapes that were produced by model and participants. For readability, tour shapes that were produced no more than once in total by participants or model are collapsed into the "Other" category in Table 2. The fit between model and participants' behavior was assessed for each point

Table 2. Percentages of tours falling in different shape classes for model and participants respectively.

Crossing	A	B	C	D	E	Other
Model	82	7	7	4	0	0
Participants	0	4	52	4	7	33

Plaisier-Tak	A	B	C	D	E	F	G	Other
Model	40	20	20	20	0	0	0	0
Participants	10	0	0	0	45	10	10	25

S-shaped	A	B	C	D	Other
Model	100	0	0	0	0
Participants	0	42	38	8	12

Square	A	B	C	D	E	F	G	H	Other
Model	30	31	18	6	4	2	0	0	9
Participants	11	4	0	0	7	22	19	11	26

Circle	A	B	C	D	E	F	G	H	I	J	K	L	Other
Model	26	19	17	14	10	5	4	3	0	0	0	0	2
Participants	7	4	0	0	0	0	0	0	25	14	7	7	36

Dantzig	A	B	C	D	E	F	G	H	Other
Model	30	10	20	10	10	10	10	0	0
Participants	11	7	4	11	0	14	0	7	46

set independently. To compare human and model performance, we performed a Fisher exact test for each $2 \times n$ table of observed frequencies, where n denotes the number of different tour shapes. This test came out significant for the Crossing, Plaisier-Tak, S-shaped, Square and Circle point sets (all $ps < .001$), indicating that the distributions of tour shapes produced by model and participants are significantly different for these point sets. Only for the Dantzig problem the Fisher exact test failed to reach significance ($p > .3$).

We next discuss the nature of the observed differences for each point set in turn. For each point set, we present a figure depicting the tour shapes most frequently produced by model and participants (Figures 3–5, 8–9). A complete overview of all tour shapes that were produced by model or participants more than once is available online at http://tsp.wtak.nl.

4.1. Crossing point set

For the Crossing point set, 82% of the tours produced by the model were tour A (Figure 3 on the left), a tour with crossing line segments, while no participant produced this tour. Two participants did produce a tour that revisited a point (viz., point 6 in Figure 2; cf. van Rooij et al., 2006), hence, strictly speaking, producing a tour with a crossing, but these two tours are clearly different from the type of crossing in tour A that is produced so frequently by the model for this point set. Further, 52% of the participants produced the optimal tour (i.e., tour C in Figure 3 on the right), while the model produced the optimal tour only 7% of the time.

Recall from the Stimulus section that points 5, 7 and 8 in the Crossing point set are not on the convex hull (see Figure 2), which may be poorly visible to participants. To ensure that our results are not biased by this fact, we compared human performance on the original Crossing point set with model performance for the adjusted scenario where points 5, 7 and 8 are on the convex hull. The comparison again revealed a significant qualitative misfit between model and humans (Fisher exact test, $p < .001$), with the model still producing the tour with crossed line segments as much as 50% of the time.

4.2. Plaisier-Tak point set

For the Plaisier-Tak point set, the model produced only four different tour shapes (A, B, C and D), where tour A and C are of minimal length. Tour shape A, produced by the model 40% of the time, is shown in Figure 4 (left). Participants never produced tour shape B, C and D and they produced tour shape A only 10% of the time. Further, 45% of the participants produced the nonoptimal tour shape E (Figure 4 on the right). This suggests the Gestalt principles of proximity and good continuation at work in human tour construction, leading participants to connect all the points along the midline to each other, rather than disconnect them as is done in tour A.

4.3. S-shaped point set

As can be seen in Table 2, there is absolutely no overlap between the tours produced by model and participants for the S-shaped point set. Tours A, B and C are shown in Figure 5. The model produces the optimal tour shape A for all possible starting points and directions of travel. Participants, on the other hand, most frequently produce tour shape B (42% of the participants) and tour shape C (38% of the participants). Again, this suggests the Gestalt principles of proximity and good continuation at work in human tour construction, leading participants to create a tour with a relatively large indentation or a tour with a crossing.

Two out of the three tours in the "Other" category are similar to tour A in that the point connecting the "bump" on the right to the leftmost point is adjacent to the connecting point used in tour A. It could be argued that these tours are equivalent to tour A and that the difference in distance is too difficult to estimate for a person. However, even when these two tours are classified as 'tour shape A,' model and human performance differ significantly (Fisher exact test, $p < .001$).

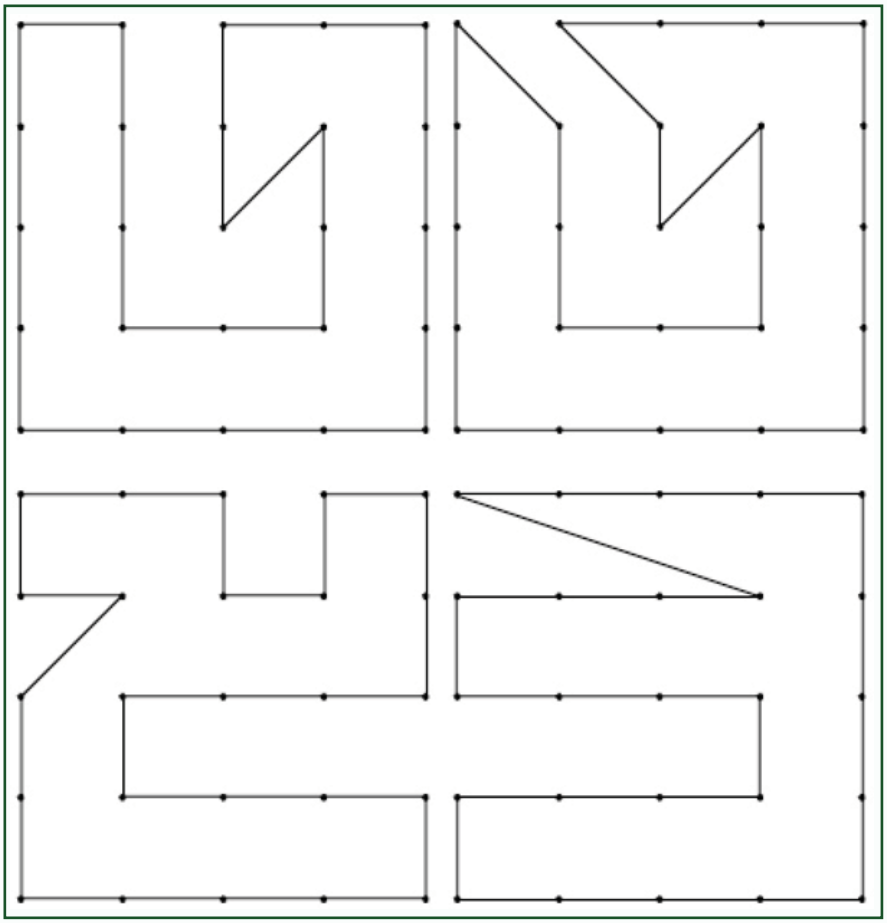
4.4. Square point set

For the Square point set, 61% of the tours produced by the model have exactly one (large) indentation. Participants produced a tour with this property in only 14.8% of the cases.[2] In general, participants tend to make tours with more and smaller indentations than tours produced by the model, producing on average 2.3 indentations (SEM = 0.23) with

a median of 2 indentations.

Four tours for the Square point set are shown in Figure 8; tour A en B (top, left and right) are the tours produced most frequently by the model and tours F and G (bottom, left and right) are tours produced most frequently by the participants. Both tours A and F are optimal. In total, 42% of the tours produced by the model are optimal and 44% of the tours produced by participants are optimal.

Figure 8. The tour shapes most frequently produced by the model (tour shapes A and B; top, left and right) and two tour shapes most frequently produced by participants (tour shape F and G; bottom, left and right) for the Square point set.



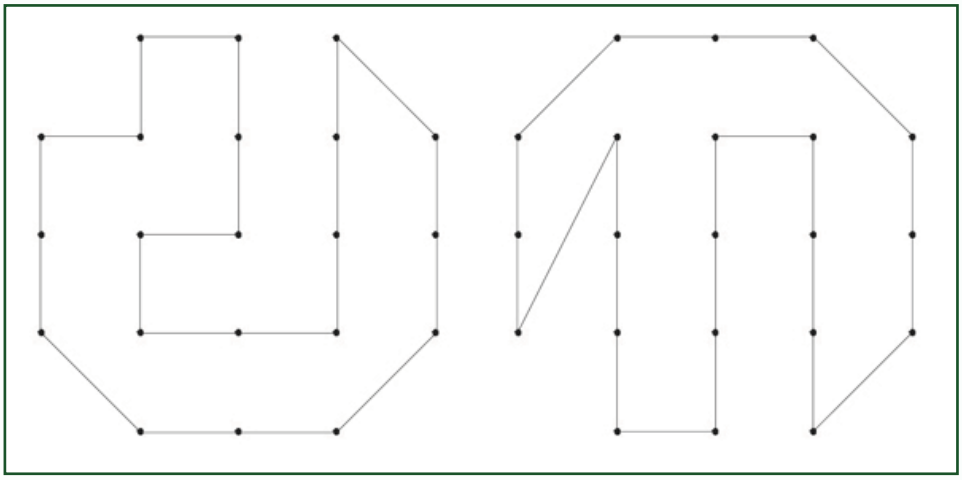
4.5. Circle point set

In Figure 9 two tours for the Circle point set are shown; tour A (left) is the tour produced most frequently by the model and tour I (right) is the tour produced most frequently by the participants. Note that even though both tours contain exactly two indentations, they differ qualitatively in terms of the shape and size of their indentations.

In total, 62% of the tours produced by the model for this point set are optimal. Participants produced an optimal tour only 25% of the time. In Section 5 we furthermore

compare these results with the model's performance on the 8 × 6 point set studied by MacGregor et al. (1999).

Figure 9. One of the tour shapes most frequently produced by the model (tour shape A; left) and the tour shape most frequently produced by participants (tour shape I; right) for the Circle point set.



4.6. The Dantzig point set

The tours produced by the model and participants for the Dantzig point set showed a large variability (see Table 2), which may be the reason Fisher's exact test did not reach significance. We did observe that model performance was marginally better than human performance for the Dantzig point set. Because the differences between model and participants were much less apparent for the Dantzig point set than for the other five point sets, we will not discuss the Dantzig point set further.

5. Summary and Interpretation

If the process underlying human performance on TSP were captured by the Convex-hull (CH) algorithm of MacGregor et al. (2000), then we would expect people to produce tours resembling tours produced by the CH algorithm. Instead, we observed large qualitative differences in the shapes of the tours produced by the CH model and human participants. From the characteristic shapes of the tours made by model and humans, and the misfit between them, we infer four explanatory failures on the part of the model: the failure to explain (1) the absence of crossings; (2) the presence of crossings and other effects of proximity; (3) effects of point set regularity on solution quality; and, (4) the approximate tour regularity of tours produced by humans for regular point sets. We discuss each of these explanatory failures below.

5.1. Absence of Crossings

It has been proposed that the CH model can explain why tours produced by humans contain relatively few lines that cross in the plane (MacGregor & Ormerod, 1996; MacGregor et al., 2000; but see also van Rooij et al., 2003). We find, however, that for a point set for which human participants do not produce any tours with crossings, namely, the Crossing point set, the model produces an overwhelming number of tours with crossings (Section 4.1). If the model cannot explain the absence of crossings for this point set, there is no reason to believe it explains the absence of crossings for other point sets either.

We acknowledge that the crossings produced by the model for the Crossing point set are a direct consequence of the particular insertion rule adopted; that is, the cheapest insertion (CI) rule. It is the particular way in which this rule defines "closest" that causes the algorithm to consider the point 14 closer to arc 5-9 than to arc 9-10, and its insertion into arc 5-9 forces a crossing (see Figure 2). Given that the CI criterion was identified by MacGregor et al. (2000) as the criterion producing the best fit, it is unclear if the model could be improved by using a different insertion criterion. We do know that the other insertion criterion considered by MacGregor et al., 2000, the largest angle (LA) criterion, cannot improve upon the cheapest insertion criterion, nor can it explain the absence of crossings in human solutions. Even though the LA criterion does not produce tours with crossings on the Crossing point set, there is no overlap whatsoever between the crossing-free tours produced by this insertion criterion and the crossing-free tours produced by the participants (this can be verified using the program available at http://tsp.wtak.nl, which also implements the option to use the LA criterion in the CH algorithm).

5.2. Crossings and other Gestalt Effects

In producing solutions to the TSP, participants seem to be guided by the Gestalt principle of good continuation (Sections 4.2 and 4.3). For the Plaisier-Tak point set this led many participants to produce a tour that connects all the points on the midline to each other, resulting in tours with a single indentation (Figure 4; right). In contrast, the CH algorithm tends to produce tours that "break up" the points along the midline so as to create tours with two indentations (Figure 4; left). For the S-shaped point set it led many participants to connect the points along the "s-shaped path" to each other, either forcing them to make a crossing (Figure 5, bottom; cf. MacGregor et al. 2004, Experiment 3) or forcing them to create a large indent in the tour, presumably so as to avoid making a crossing (Figure 5; middle). The CH algorithm, on the other hand, connects the points for the S-shaped point set exactly in a way required for tour optimality, completely ignoring the principles of good continuation and proximity.

Parenthetically, we note that our findings on the S-shaped point set provide empirical support for the *crossing-avoidance hypothesis* put forth by van Rooij et al. (2003), which states that participants aim at avoiding crossing lines when trying to solve the TSP. We

see no plausible reason for why participants would construct the middle tour depicted in Figure 5 other than that participants are making some form of trade-off between "the goal to connect proximate points" and "the goal of avoiding crossing lines" in favor of the latter over the former. Admittedly, 42.3% of the participants did produce a tour with a crossing for the S-shaped point set, which is much more than the 6% typically found for (semi)random point sets (van Rooij et al. 2003). Moreover, also for the Plaisier-Tak, Square, Circle and Dantzig point sets we observed more tours with crossings than was the case in other experiments (compare our Table 1 to Table 2 in van Rooij et al., 2003). Does this mean that our data refute the crossing-avoidance hypothesis? We do not think so, for two reasons. First, crossing-avoidance need not be the only goal that people have; presumably "goodness of figure," "connecting proximate points" and "optimality" are other plausible goals that people may pursue. Second, aiming for a goal does not always guarantee you reach it. As also noted by MacGregor et al. (2004) the "look ahead" required for crossing avoidance can be computationally too taxing at times, so that even a strategy that has as its primary goal to avoid crossings may yet produces tours with crossings. Be that as it may, we did not explicitly set out to test the crossing-avoidance hypothesis, but we thought it interesting to note that even a point set specifically constructed to "trick" people into making a crossing (MacGregor & Ormerod, 1996) often fails to do so.

5.3. Effects of Point Set Regularity

Our results for the Square and Circle point sets showed that although regularity of point sets may indeed facilitate human performance as predicted by MacGregor et al. (1999), the degree of facilitation seems to depend in part on the regularity of the optimal tour (Sections 4.4 and 4.5). The Square and Circle point set were as regular as, and even smaller than, the 8×6 point set used by MacGregor et al. (1999). Yet fewer participants were able to find an optimal tour for the Square and Circle point sets (44.4% and 25%, respectively) than for the 8×6 point set (85%) (MacGregor et al., 1999, p. 1422). Moreover, the quality of human performance on regular point sets is not explained by the CH algorithm; the model produces 42%, 62% and 13% optimal tours for the Square, Circle and 8×6 point sets, respectively.[3] If anything, there seems to be a *negative* relationship between model and human performance on such regular point sets.

5.4. (Approximate) Tour Regularity

We observed systematic differences in the number and size of the tour indentations produced by the CH algorithm and participants for the regular point sets (see also Sections 4.4 and 4.5), with the CH algorithm producing many tours with one large indentation (possibly accompanied by one or more small indentations). Participants instead tended to produce tours with indentations of intermediate sizes. This difference in qualitative tour shape produce by model and humans was most pronounced for the Square point

set. For this point set we found that the CH algorithm produced a tour with a single large indentation in 61% of the cases, while participants seldom made such a tour. To see if this finding was specific to rectangular point sets with an odd number of points, we inspected the model's behavior on the 8×6 point set of MacGregor et al. (1999). We found that also for this regular point set the model mostly produces tours with one large indentation, accompanied by one or more small indentations. This is in strong contrast to human performance, which is characterized by the E-shaped form depicted in Figure 6A (see MacGregor et al., 1999, p. 1422). Notably, the tours most frequently produced by participants on both the Square point set (Tours F and G; Figure 8, bottom left and right) and the Circle point set (Tour I; Figure 9, right) seem to be approximations to this E-shaped form under the constraints imposed by the odd number of points (cf. Figure 6C). In other words, it looks as if several participants are trying to produce a tour shape that resembles the best tour for an even-numbered regular point set. The reason for this tendency may be that such tours are considered more regular or otherwise "good form" (cf. Koffka, 1935) and that participants prefer to construct tours with that property (cf. Ormerod & Chronicle, 1999). Whatever the reason, it is clear that the CH algorithm cannot explain the phenomenon, as it tends to produce tours that in no way approximate a regular E-shaped form, but instead contain large irregular indentations.

6. Methodological Afterthought

As we argued in the Introduction, the tests performed by MacGregor et al. (2000) were weak tests of the CH model (i.e., tests that have high probability of yielding confirmation, even if the model is false; see Meehl, 1997), and such tests provide only weak or negligible support.[4] The strong tests (i.e., tests that have low probability of yielding confirmation, unless the model is true) that we performed in fact disconfirmed the model. We end this paper by discussing some methodological lessons that can be derived from this for testing models of TSP in general. We will also consider potential criticisms that our strong tests of the CH model are too strict and that they do not do justice to the model as it was originally intended by MacGregor et al. (2000).

It is common practice in psychological research to submit one's model only to one or more weak tests. As this methodology makes it difficult, if not impossible, to build a genuine theoretical understanding of mental processes, we submit that TSP modeling research would be better off if it did not adopt this standard psychological methodology. In particular, we draw two important lessons from our own study that may help contribute to a more rigorous testing practice for models of TSP in general:

1. TSP modelers and researchers may do well not to focus exclusively on quantitative measures, such as tour length, as the dependent measure. Instead it may be better to also take into account the characteristic shapes of the tours produced

by humans (cf. MacGregor et al., 1999, 2004).

2. TSP modelers and researchers may do well not to focus exclusively on replicating the high quality of human performance and consider successful replications as support for their models (see also Graham et al., 2000; Pizlo et al., 2006). Instead it may be better to evaluate TSP models on the basis of risky predictions; after all, if a model passes a test based on a risky prediction, then this counts as genuine support for the model (see also Meehl, 1997; Roberts & Pashler, 2000).

Few researchers would object to the claim (made in point 2 above) that confirmed risky predictions constitute strong support for a model. However, if risky predictions are *dis*-confirmed opinions may vary about what to make of these disconfirmations. We consider here two possible objections to our claim that the strong tests that we have reported in this article actually disconfirm the CH model.[5] We present brief replies for each possible objection.

Objection 1. *The CH-algorithm may be a good model of human performance on random point sets, albeit it less descriptive of human performance on regular or otherwise nonrandom point sets.*

MacGregor, Ormerod and Chronicle also made a suggestion to this effect, when they wrote that "the model is unlikely to produce a particularly good fit to human solutions to highly patterned TSPs," but "for random or relatively irregular TSPs . . . the model provides a very reasonable approximation" (MacGregor et al., 2000, p. 1189).

Granted that models typically have restricted domains of application (e.g., models of TSP are not necessary models of Tower of Hanoi), we see no plausible justification for why a model of human TSP performance would have random point sets—or any other restricted set of point sets—as its domain of application.[6] Humans understand the TSP instruction for any given point set (as far as we know no participant ever complained not to understand the task for a particular point set), and the speed and effortlessness with which humans construct solutions for TSP seems to apply for random and regular point sets alike. An explanatory process model of human TSP performance should therefore be able to provide a "very reasonable approximation" of human performance for *all* point sets, as well as make insightful why tours produced by humans tend to have the particular shapes that they do.

Besides being unmotivated, domain restriction as suggested by MacGregor et al. (2000) seems to make for a poor research strategy as it risks that we unduly protect non-veridical process models of TSP from disconfirming evidence, simply by excluding (potentially) disconfirming point sets from the domain of application. Admittedly, it may be the case that humans use qualitatively different heuristics for solving different subclasses of TSP instances, and the CH algorithm may be one such heuristic used by humans for a

subclass of randomly generated instances (albeit it not used for other instances). Though we cannot rule out this possibility, it does not mean that building partial TSP models for separate classes of sets of point sets is necessarily a good methodology for understanding human TSP performance in general. First of all, such a methodology risks a proliferation of ad hoc and unrelated partial TSP models for different, unmotivated subclasses of point sets. Each such a partial model can be made close to immune for disconfirmation by simply further restricting its domain of application. Secondly, we are in the end still left with the question of how humans decide *which* heuristic to use *when* (see also Cooper, 2000; Newell, 2005). To prevent "a homunculus problem of needing a meta-heuristic to select the appropriate 'tool for the job'" (Newell, 2005, p. 12) we think it best to try to build unified models of *general* human TSP performance from the start and to hold current models of TSP against this general standard—as we did with the CH model in our experiment.

> **Objection 2.** *Models should not be tested against the "ideal" of perfect fit, because no model is perfect. Model and reality will always disagree to some extent, so given enough power, a difference between model and reality can always be found.*

Although we agree with the statement as such, we think it is nevertheless sensible to claim that some models are further from reality than others. Moreover, some models can even be relatively close to reality (i.e., the models may be approximations of reality in some relevant sense, which is after all what MacGregor et al. [2000] claimed for their model). In this case a model is said to have high verisimilitude. Our point is *not* that the CH model gives an imperfect fit to human data, but rather that for certain instances of the TSP the CH model fits human data so poorly that we are led to believe the model has low verisimilitude.

Note that the observed differences between model and human performance that we reported in Section 4 and reviewed in Section 5 are not demonstrations of mere statistically significant misfit (we agree such misfits are probably detectable for any model, given sufficient statistical power), but rather large qualitative and systematic differences in the tour shapes produced by humans and the CH model. That is, we find that the model consistently makes tours that people do not, and vice versa. As shown in Table 2, for five of the six point sets there is minimal to no overlap between the tour classes characterizing human tours and those of the model, with relatively large agreement *within* the two groups (keep in mind that there exists an explosive number of possible tours for each point set). When model and human performance is almost maximally dissimilar, as is the case here, it seems justified to infer the model has low verisimilitude.

In closing, we emphasize that our negative conclusion about the validity of the CH model is not the only—and perhaps not even the most important—message to take home from our article. We have illustrated how a process model of human problem-solving that has high intuitive appeal and an apparently strong empirical base, can be shown to be

nonexplanatory and of low verisimilitude if (and perhaps only if) we take the model at face value, derive critical predictions for particular point sets and submit them to strong tests. The importance of such critical and rigorous methodology has already been noted by others for psychology in general (e.g., Meehl 1997; Roberts and Pashler, 2000). With our illustration we hope to help motivate the same rigorous methodology for model testing in the psychology of problem solving, and TSP-solving in particular. In addition, we hope that the shape characteristics that we have identified in tours produced by humans for the Plaisier-Tak, S-shaped, Square and Circle point sets may serve as explanatory targets for future TSP modeling attempts.

Author note

We thank Michael Lee and an anonymous reviewer for helpful comments on an earlier version of this article. Parts of this research were presented at the 46th Annual Meeting of the Psychonomic Society in Toronto, November 2005 and at the Annual Meeting of the Society for Mathematical Psychology in Vancouver, August 2006. The reported experiment was performed by M. Plaisier and S. Tak as part of a student research project in Human-Technology Interaction at the Eindhoven University of Technology, and the article was written while I. van Rooij was at Eindhoven University of Technology. The authors gratefully acknowledge the financial support of the Department of Human-Technology Interaction and the J. F. Schouten School at Eindhoven University of Technology.

References

Cooper, R. (2000). Simple heuristics could make us smart but which heuristics do we apply when? *Behavioral and Brain Sciences, 23*, 746.

Cormen, T. H, Leiserson, C. E., & Rivest, R. L. (2001). *Introduction to algorithms* (2nd ed.). Cambridge: MIT Press.

Dantzig, G. B., Fulkerson, D. R., & Johnson, S. M. (1959). On a linear-programming, combinatorial approach to the traveling-salesman problem. *Operations Research, 7*, 58-66.

Flood, M. M. (1956). The traveling-salesman problem. *Operations Research, 4*, 61-75.

Garey, M. R., & Johnson, D. S. (1979). *Computers and intractability: A guide to the theory of NP-Completeness*. New York: W. H. Freeman.

Graham, S. M., Joshi, A., & Pizlo, Z. (2000). The traveling salesman problem: A hierarchical model. *Memory & Cognition, 28*, 1191-1204.

Koffka, K. (1935). *Principles of Gestalt psychology.* New York: Harcourt Brace.

Lee, M. D., & Vickers, D. (2000). The importance of the convex hull for human performance on the traveling salesman problem: A comment on MacGregor and Ormerod (1996). *Perception & Psychophysics, 62*, 226-228.

MacGregor, J. N., Chronicle, E. P., & Ormerod, T. C. (2004). Convex hull or crossing avoidance solution heuristics in the traveling salesperson problem. *Memory & Cognition, 32*, 260-270.

MacGregor, J. N., & Ormerod, T. (1996). Human performance on the traveling salesman problem. *Perception & Psychophysics, 58*, 527-539.

MacGregor, J. N., Ormerod, T. C., & Chronicle, E. P. (1999). Spatial and contextual factors in human performance on the traveling salesperson problem. *Perception, 28*, 1417-1427.

MacGregor, J. N., Ormerod, T. C., & Chronicle, E. P. (2000). A model of human performance on the traveling salesperson problem. *Memory & Cognition, 28*, 1183-1190.

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. In L. L. Harlow, S. A. Mulaik, and J. H. Steiger (Eds.), *What if there were no significance tests?* (pp. 393-425). Mahwah, New Jersey: Lawrence Erlbaum Associates.

Newell, B. R. (2005). Re-visions of rationality? *Trends in Cognitive Sciences, 9*, 11-15.

Ormerod, T. C., & Chronicle, E. P. (1999). Global perceptual processes in problem-solving: The case of the traveling salesperson. *Perception and Psychophysics, 61*, 1227-1238.

Pizlo, Z., Stefanov, E., Saalweachter, J., Li, Z., Haxhimusa, Y., & Kropatsch, W. G. (2006). Traveling salesman problem: A foveating pyramid model. *Journal of Problem Solving, 1*, 83-101.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review, 107*, 358-367.

van Rooij, I., Schactman, A., Kadlec, H., & Stege, U. (2006). Perceptual or analytical processing? Evidence from children's and adult's performance on the Euclidean traveling salesperson problem. *Journal of Problem Solving, 1*(1), 44-73.

van Rooij, I., Stege, U., & Schactman, A. (2003). Convex hull and tour crossings in the Euclidean traveling salesperson problem: Implications for human performance studies. *Memory & Cognition, 31*, 215-220.

Vickers, D., Butavicius, M. A., Lee, M. D., & Medvedev, A. (2001). Human performance on visually presented traveling salesman problems. *Psychological Research, 65*, 34-45.

Wertheimer, M. (1923). Laws of organization in perceptual forms. In Ellis, W. (Trans.), *A source book of Gestalt psychology* (pp. 71-88). London: Routledge & Kegan Paul. (Original work published 1938).

Appendix A

Correction of Step 3 in the algorithm description by MacGregor et al. (2000).

The original formulation of Step 3 in the algorithm specification was as follows:

> Step 3*: If the starting point is on the boundary, the starting node is the *current node*. The arc connecting the current node to the adjacent boundary node in the

direction of travel is referred to as the *current arc*. Proceed immediately to Step 4. If the starting point is not on the boundary, apply the insertion rule to find the closest arc on the boundary. Connect the starting point to the end node of the closest arc which is in the direction of travel. This node becomes the current node.

When we tried to implement the algorithm using this form of Step 3 in a computer program we encountered a problem: When the starting point is an interior point a complete tour is never obtained. Consider the situation shown in Figure A1 where the starting point is the point denoted by "s" and the direction of travel is clockwise. According to Step 3*, the starting point is connected to the top left node, which is the end node of the closest arc in the direction of travel. The current arc then becomes the arc denoted by C (left), as the current arc is "the arc connecting the current node to the adjacent boundary node in the direction of travel" (Step 3*). The algorithm then proceeds on the convex hull. When arc C is the current arc (middle) the algorithm checks whether the unconnected interior point is closer to any other arc. The point is closer to another arc, namely arc 1, and the algorithm moves on to the next arc on the boundary. The algorithm will be stuck going around and around on the boundary as it never again reaches arc 1 (right).

As this cannot have been the intention of the authors of the model, we corrected Step 3 as done in the description of the algorithm in Section 2. The new situation is shown in Figure A2. The starting point is the same point as in the previous example. In this case, the arc closest to the starting (interior) point is removed and two new arcs are added connecting the starting point to the end points of the removed arc (left). The current node again becomes the top left node and the current arc is the arc connecting the current node to the adjacent boundary node in the direction of travel. After arc 5 is connected, the current arc does not become the arc connecting the current node to the adjacent boundary node in the direction of travel, as this arc does not exist anymore, but instead

Figure A1. Solution according to the literal interpretation of MacGregor et al.'s (2000) description of the algorithm. The algorithm connects the starting point to the end node of the closest arc in the direction of travel (left), continues on the convex hull (middle), but is stuck going around on the convex hull forever (right).

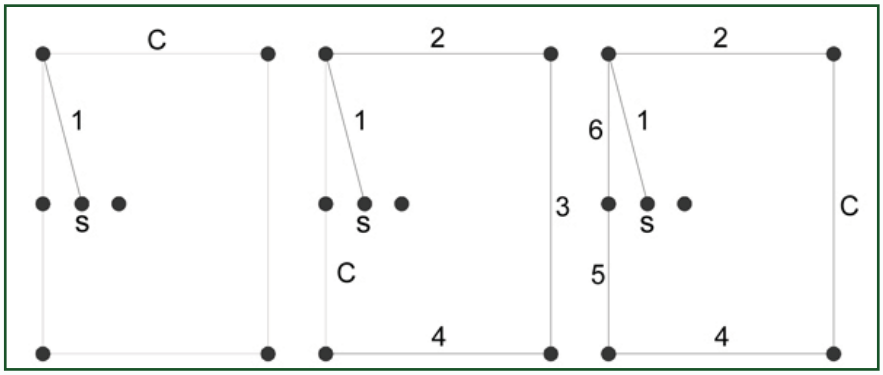
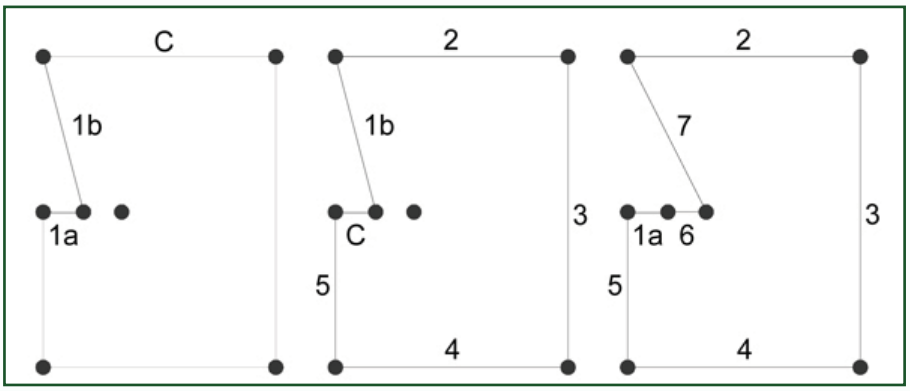


Figure A2. Solution according to the corrected description of the algorithm. The starting point is inserted in the closest arc in the direction of travel, creating two new arcs. After reaching arc 5 as the current arc 1a becomes the current arc and eventually the tour is completed.



becomes arc 1a (middle). The algorithm moves on to arc 1b and, as the last unconnected interior point is closest to 1b the point is inserted here, creating arc 6 and 7 and completing the tour (right).

Appendix B

Coordinates (in twips) for the point sets used in the experiment

Crossing: 500,100; 500,6200; 10000,6200; 10000,100; 4350,6201; 5220,6203; 6120,6202; 8000,6201; 5200,4000; 5200,4500; 5200,5000; 5200,5500; 5200,6000; 6210,4245 || **Plaisier-Tak:** 750,500; 4500,500; 8250,500; 4500,1125; 4500,1750; 4500,2375; 4500,3000; 4500,3625; 750,5500; 8250,5500 || **S-shaped:** 893,2114; 4513,3711; 4878,3453; 5213,3179; 5578,2905; 5928,2631; 6232,2418; 6628,2099; 6993,1840; 7342,1764; 7707,1840; 7966,2069 || **Square:** 2000,500; 3250,500; 4500,500; 5750,500; 7000,500; 2000,1750; 3250,1750; 4500,1750; 5750,1750; 7000,1750; 2000,3000; 3250,3000; 4500,3000; 5750,3000; 7000,3000; 2000,4250; 3250,4250; 4500,4250; 5750,4250; 7000,4250; 2000,5500; 3250,5500; 4500,5500; 5750,5500; 7000,5500 || **Circle:** 3250,500; 4500,500; 5750,500; 2000,1750; 3250,1750; 4500,1750; 5750,1750; 7000,1750; 2000,3000; 3250,3000; 4500,3000; 5750,3000; 7000,3000; 2000,4250; 3250,4250; 4500,4250; 5750,4250; 7000,4250; 3250,5500; 4500,5500; 5750,5500 || **Dantzig:** 2100,600; 6900,5400; 5700,4200; 4500,3000; 5700,3000; 6300,3000; 6900,3000; 3300,1800; 2100,5400; 6900,600.

Notes

1. In this paper we consider only the Euclidean version of the TSP.

2. We have excluded here four tours made by the participants with one indentation but also a crossing, because the model does not make any crossings for this point set.

3. To obtain an estimate for the number of optimal tours for the 8×6 point set the program was run ten times. The percentage of optimal tours ranged from 10% to 16%.

4. Recall that we argued that the fit in observed tour lengths of model and human performance may have been an artifact of the fact that both model and humans on average perform close to optimal on randomly generated point sets. Using our current data we can test this hypothesis. We find that the lengths of the tours produced by the participants for our six point sets are on average about 4% above optimal, which reveals worse performance on these point sets than on randomly generated instances of TSP of comparable size (e.g., van Rooij et al., 2006, found that tour lengths produced by adult human participants is on average about 2% above the optimal tour length for randomly generated point sets of 15 points). Notably, the high quality performance of the CH algorithm is not similarly attenuated by nonrandomness, as the algorithm's performance is on average about 2% of optimal even for our six point sets. For the CH algorithm to be a viable model of the human problem solving process its performance should have been as much affected by nonrandomness, but evidently it is not.

5. We are grateful to Michael Lee for bringing the second of these possible objections to our attention.

6. Lee and Vickers (2000) made a similar point. They argued that in building confirming evidence for their convex hull hypothesis MacGregor and Ormerod (1996) artificially restricted their point sets to include only point sets with a relatively large number of points on the convex hull, making the convex-hull perceptually more salient than can be expected to hold for point sets in general. To explain human performance on the *Traveling Salesperson problem* (TSP), MacGregor, Ormerod, and Chronicle (2000) propose that humans construct solutions according to the steps described by their convex-hull algorithm. Focusing on tour length as the dependent variable, and using only random or semirandom point sets, the authors claimed empirical support for their model. In this paper we argue that the empirical tests performed by MacGregor et al. do not constitute support for the model, because they instantiate what Meehl (1997) coined "weak tests" (i.e., tests with a high probability of yielding confirmation even if the model is false). To perform "strong" tests of the model, we implemented the algorithm in a computer program and compared its performance to that of humans on six point sets. The comparison reveals substantial and systematic differences in the *shapes* of the tours produced by the algorithm and human participants, for five of the six point sets. The methodological lesson for testing TSP models is twofold: (1) Include qualitative measures (such as tour shape) as a dependent variable, and (2) use point sets for which the model makes "risky" predictions.

Paper submitted on June 6, 2006.
Final version accepted on Nov 9, 2007.