By Albert Madansky

Evaluating paired comparisons, maximum difference and traditional ranking

Editor's note: Albert Madansky is vice president of the Analytical Group Inc., a Scottsdale, Ariz., research firm. He is also H.G.B. professor emeritus of business administration at the Booth Graduate School of Business, University of Chicago. He can be reached at albert.madansky@analyticalgroup.com. To view this article online, enter article ID 20101002 at quirks.com/articles.

snapshot

The author examines a host of data interpretation issues related to the use of paired comparisons and maximum difference surveys.

Market researchers have always been troubled by the difficulty of the task of obtaining reliable responses when the respondent is asked to rank a large number of items.[1] In the early days of market research, when personal interviewing was the primary mode, various devices, such as the sort board, were recommended as aids in eliciting reliable responses. As market research moved toward telephone interviewing and Web-based surveys the difficulties associated with the ranking task were exacerbated, and two alternative elicitation techniques have been espoused: paired comparison and maximum difference. Though these techniques have simplified the response task, the concomitant move, from face-to-face interviewing to responding to a live interviewer by phone and finally to responding to a computer screen, has successively removed the effect that the interviewer has on the responses. For, as will be seen later, both the paired comparisons and the maximum difference set of queries can lead to inconsistent responses. And with the successively greater distancing of the interviewer from the respondent, there is greater likelihood that these inconsistencies will be unchecked. (Of course, the computer-assisted interview can, if programmed properly, check for such inconsistencies and proceed with an interrogation designed to rectify the inconsistent responses. But this requires special programming attention in the design of the interviewing script.)

The purpose of this note is, by means of an example, to illustrate the data interpretation issues associated with each of these methods. What we will see is that, unless they are carried out in full, in both the paired comparisons and maximum difference surveys, the preference proportions observed or inferred from the interview about some of the pairings will be based on sample sizes that are far short of the full sample size of the study. Moreover, these sample sizes are randomly determined, and so preference proportions based on these observations will not have the statistical properties of ordinary proportions. Consequently, standard statistical

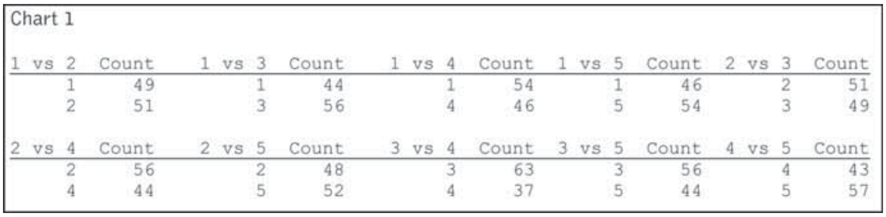


Chart 1

1 vs 2	Count	1 vs 3	Count	1 vs 4	Count	1 vs 5	Count	2 vs 3	Count
1	49	1	44	1	54	1	46	2	51
2	51	3	56	4	46	5	54	3	49

2 vs 4	Count	2 vs 5	Count	3 vs 4	Count	3 vs 5	Count	4 vs 5	Count
2	56	2	48	3	63	3	56	4	43
4	44	5	52	4	37	5	44	5	57

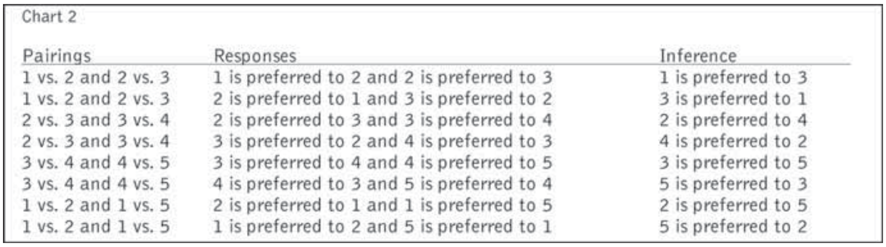


Chart 2

Pairings	Responses	Inference
1 vs. 2 and 2 vs. 3	1 is preferred to 2 and 2 is preferred to 3	1 is preferred to 3
1 vs. 2 and 2 vs. 3	2 is preferred to 1 and 3 is preferred to 2	3 is preferred to 1
2 vs. 3 and 3 vs. 4	2 is preferred to 3 and 3 is preferred to 4	2 is preferred to 4
2 vs. 3 and 3 vs. 4	3 is preferred to 2 and 4 is preferred to 3	4 is preferred to 2
3 vs. 4 and 4 vs. 5	3 is preferred to 4 and 4 is preferred to 5	3 is preferred to 5
3 vs. 4 and 4 vs. 5	4 is preferred to 3 and 5 is preferred to 4	5 is preferred to 3
1 vs. 2 and 1 vs. 5	2 is preferred to 1 and 1 is preferred to 5	2 is preferred to 5
1 vs. 2 and 1 vs. 5	1 is preferred to 2 and 5 is preferred to 1	5 is preferred to 2

inferences made from these proportions (based on assumptions of binomial distributions with fixed sample size) may not be correct.[2] Indeed, it's not clear whether these proportions are unbiased estimates of the population preference proportions.

My example is based on a sample of 100 rankings of five items. There are 5!=120 different possible rankings, and Appendix I (see article ID 20101002 at quirks.com) contains the frequency of occurrence of each of these rankings, so that the reader can use the data and try other combinations of pairings or best/worst elicitations than those illustrated in this article to see what results one would get. The rankings are the order of preference of the five items, with the convention that the items are ranked from most preferred to least preferred. Thus 13542 means that Product 1 is most preferred (it got listed first), followed by Product 3 (it got listed next), then by Product 5 (it got listed third), then by Product 4 (it got listed fourth), and that Product 2 is the least preferred (it got listed last).

Easy to answer

The idea behind using paired comparisons instead of ranking is that the task of answering the question, "Which do you prefer, Item A or Item B?" is easy to answer. When there are five items there are 10 different pairs about which one can ask this question. As a baseline, Chart 1 is a table of what the responses of our sample of 100 would have been if they were asked about all 10 paired comparisons. (Being that the sample size is 100, you can also read these counts as percentages.)

But being interrogated about all 10 paired comparisons is also tedious. And so market researchers may ask the respondent to do only a subset of the 10 paired comparisons, and logically infer what the respondent would have said if he/she were presented with the remainder of the paired comparisons. One popular subset is what I call the daisy chain subset. An example in this case would be these five sets: 1 vs. 2, 2 vs. 3, 3 vs. 4, 4 vs. 5, and 1 vs. 5.

From this one might, for example, attempt to infer what the respondent would have done on the 1 vs. 3 comparison by looking at the responses to 1 vs. 2 and 2 vs. 3. If Item 1 is preferred to Item 2 in the first pairing and Item 2 is preferred to Item 3 in the second pairing, then logically the respondent would prefer Item 1 to Item 3. Chart 2 is a listing of all the inferences that could be made based on the responses to two of the paired comparisons.

One might also attempt to infer what the respondent would have done on the 1 vs. 3 comparison by looking at the responses to 1 vs. 5, 4 vs. 5, and 3 vs. 4. If Item 1 is preferred to Item 5 in the first pairing and Item 5 is preferred to Item 4 in the second pairing, and Item 4 is preferred to Item 3 in the third pairing, then logically the respondent would prefer Item 1 to Item 3. Chart 3 shows a listing of all the inferences that could be made based on the responses to three of the paired comparisons.

Of course one can only make indirect inferences about what the respondent would have done on the 1 vs. 3 pairing from only a subset of responses to the five pairings in the daisy chain (e.g., from the responses to the 1 vs. 2 and 2 vs. 3 pairings or from the responses to the 1 vs. 4, 4 vs. 5, and 3 vs. 4 pairings, if they are as given above). In some cases one cannot make any logical inference about the results of 1 vs. 3 (e.g., from a respondent who says that 2 is preferred to 1 and 2 is preferred to 3).

Chart 4 shows the results of applying the kind of logic described above to infer what the results would have been in the five pairs that were not part of the daisy chain.

Note that the number of inferences on the pairs not directly compared is less than 50 percent of the respondents. Note also that the inferred percentages are contrary to those based on the full set of rankings. For example, in the case of 2 vs. 5, Item 2 was ahead of Item 5 in 61.29 percent of the inferences, whereas in reality only 48 percent of the sample prefer Item 2 to Item 5.

Which is most and least preferred

The underlying idea in this mode of questioning is that one presents a subset of the items to the respondent and, instead of asking the respondent to rank the items, the respondent is asked to tell the interviewer which is the most preferred and least preferred of these items. In our example of five items, it is a trivial feat to infer the relationship between all the items if the subset is of size three. So let's consider the case where we present the respondent a subset of four items. There are five possible subsets that may be presented: 1234, 1235, 1245, 1345 and 2345. One can infer from the responses that each of the items not designated as either best or worst is ranked lower than the best and higher than the worst. For example,

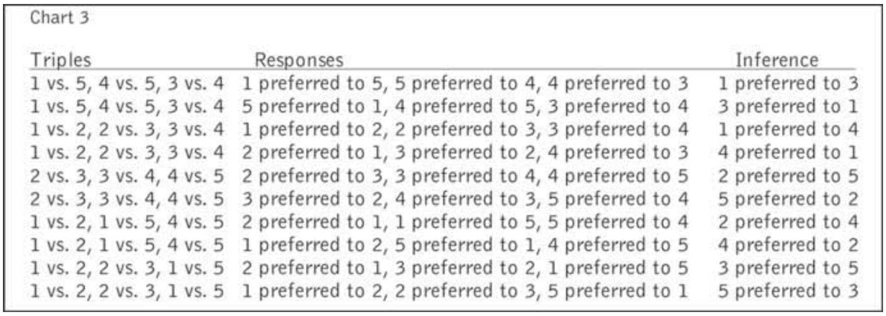


Chart 3

Triples	Responses	Inference
1 vs. 5, 4 vs. 5, 3 vs. 4	1 preferred to 5, 5 preferred to 4, 4 preferred to 3	1 preferred to 3
1 vs. 5, 4 vs. 5, 3 vs. 4	5 preferred to 1, 4 preferred to 5, 3 preferred to 4	3 preferred to 1
1 vs. 2, 2 vs. 3, 3 vs. 4	1 preferred to 2, 2 preferred to 3, 3 preferred to 4	1 preferred to 4
1 vs. 2, 2 vs. 3, 3 vs. 4	2 preferred to 1, 3 preferred to 2, 4 preferred to 3	4 preferred to 1
2 vs. 3, 3 vs. 4, 4 vs. 5	2 preferred to 3, 3 preferred to 4, 4 preferred to 5	2 preferred to 5
2 vs. 3, 3 vs. 4, 4 vs. 5	3 preferred to 2, 4 preferred to 3, 5 preferred to 4	5 preferred to 2
1 vs. 2, 1 vs. 5, 4 vs. 5	2 preferred to 1, 1 preferred to 5, 5 preferred to 4	2 preferred to 4
1 vs. 2, 1 vs. 5, 4 vs. 5	1 preferred to 2, 5 preferred to 1, 4 preferred to 5	4 preferred to 2
1 vs. 2, 2 vs. 3, 1 vs. 5	2 preferred to 1, 3 preferred to 2, 1 preferred to 5	3 preferred to 5
1 vs. 2, 2 vs. 3, 1 vs. 5	1 preferred to 2, 2 preferred to 3, 5 preferred to 1	5 preferred to 3

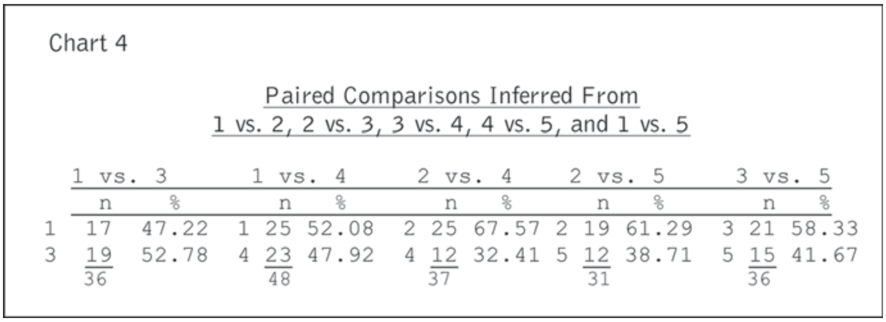


Chart 4

Paired Comparisons Inferred From
1 vs. 2, 2 vs. 3, 3 vs. 4, 4 vs. 5, and 1 vs. 5

1 vs. 3			1 vs. 4			2 vs. 4			2 vs. 5			3 vs. 5		
	n	%		n	%		n	%		n	%		n	%
1	17	47.22	1	25	52.08	2	25	67.57	2	19	61.29	3	21	58.33
3	19	52.78	4	23	47.92	4	12	32.41	5	12	38.71	5	15	41.67
	36			48			37			31			36	

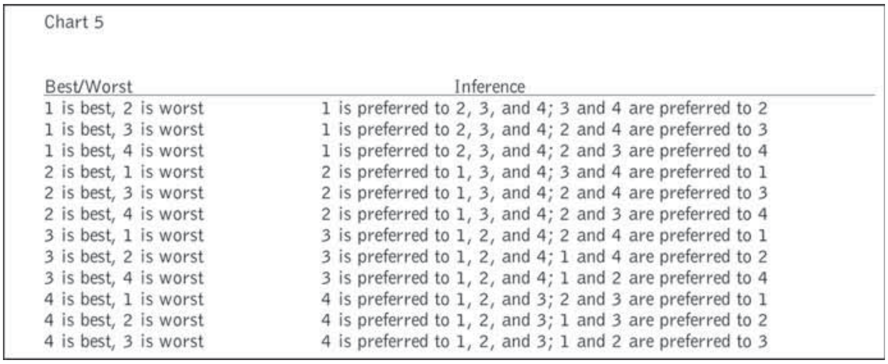


Chart 5

Best/Worst	Inference
1 is best, 2 is worst	1 is preferred to 2, 3, and 4; 3 and 4 are preferred to 2
1 is best, 3 is worst	1 is preferred to 2, 3, and 4; 2 and 4 are preferred to 3
1 is best, 4 is worst	1 is preferred to 2, 3, and 4; 2 and 3 are preferred to 4
2 is best, 1 is worst	2 is preferred to 1, 3, and 4; 3 and 4 are preferred to 1
2 is best, 3 is worst	2 is preferred to 1, 3, and 4; 2 and 4 are preferred to 3
2 is best, 4 is worst	2 is preferred to 1, 3, and 4; 2 and 3 are preferred to 4
3 is best, 1 is worst	3 is preferred to 1, 2, and 4; 2 and 4 are preferred to 1
3 is best, 2 is worst	3 is preferred to 1, 2, and 4; 1 and 4 are preferred to 2
3 is best, 4 is worst	3 is preferred to 1, 2, and 4; 1 and 2 are preferred to 4
4 is best, 1 is worst	4 is preferred to 1, 2, and 3; 2 and 3 are preferred to 1
4 is best, 2 is worst	4 is preferred to 1, 2, and 3; 1 and 3 are preferred to 2
4 is best, 3 is worst	4 is preferred to 1, 2, and 3; 1 and 2 are preferred to 3

from the best/worst responses to the 1234 subset we can make the inferences shown in Chart 5.

And one can make up similar tables of inferences from the best/worst designations in each of the other subsets 1235, 1245, 1345 and 2345. Appendix II (see article ID 20101002 at quirks.com) lists for our example the inferences about pairs of items that can be made from the best/worst designation from each of these five subsets.

Appendix II illustrates that somewhere between 75 to 90 of the 100 respondents' data are used in estimating the proportion of respondents who preferred Item A to Item B. Moreover, there is no fixed relationship between the estimates made using this method and the estimates made from the rankings themselves. For example, though Item 3 is preferred to Item 5 in 56 percent of the responses, the inference from the 1235 subset is 55.52 percent, and it is 58.33 percent from the 1345 subset and 59.30 percent from the 2345 subset.

No inferences can be made about the item missing from the subset. If one presented the respondent with all five subsets, one could net out the inferences from each of the five and produce the estimates made from the rankings themselves. But what if one only presents a few of the subsets? We consider one reasonable example, namely the presentation of the 1234, 1235 and 2345 subsets. Since all the subsets involving Items 2 and 3 are represented, the net inference from presentation of these three subsets will produce the fraction corresponding to the result of the 100 paired comparisons of Items 2 and 3. But what about the other nine pairs? Chart 6 shows the results.

Again, there is no fixed relationship between the estimates made using this method and the estimates made from the rankings themselves. For example, though Item 2 is preferred to Item 5 in 48 percent of the responses, the inference from the net of these three subset is 50.54 percent.

Chart 7 is a recap of the results of our example, where the columns labeled “pct” give the percent of the sample who preferred the first of the pair of items. So, for example, in the 25 line we see that 48 percent of the sample of 100 preferred Item 2 to Item 5, whereas in the paired comparisons daisy chain sample we could only infer preference between these items from 31 respondents and of those 61.29 percent preferred Item 2 to Item 5 and in the four maximum difference sets-of-four we could infer preference between these items from 93 of the respondents and of those 50.54 percent preferred Item 2 to Item 5.

The paired comparisons daisy chain design leads to small n on which to base inferences about the unpaired sets of items. The maximum difference can, with only three iterations, produce a larger n, but not necessarily pick the winner (as illustrated by the 2 vs. 5 inference). Moreover, in both designs one does not know in advance what the sample size will be for any of the comparisons that are not explicitly part of the design. And, as stated earlier, neither of these sets of percentages are unbiased estimates of the proportion in the population that prefer the first of the paired items.

All of this is of course based on the assumption that the respondent is logical, in that if he/she says that Item 1 is preferred to Item 2 and Item 2 is preferred to Item 3 then the respondent prefers Item 1 to Item 3. But, as is often discussed in the psychological literature, this may not be the case.

A simple example is the following. Suppose a respondent is presented with and asked in a paired comparison study to compare three pizzas:

Pizza 1: salami and onions
Pizza 2: pepperoni and garlic
Pizza 3: anchovies and mushrooms

His response to the 1 vs. 2 comparison is, “I prefer Pizza 1 to Pizza

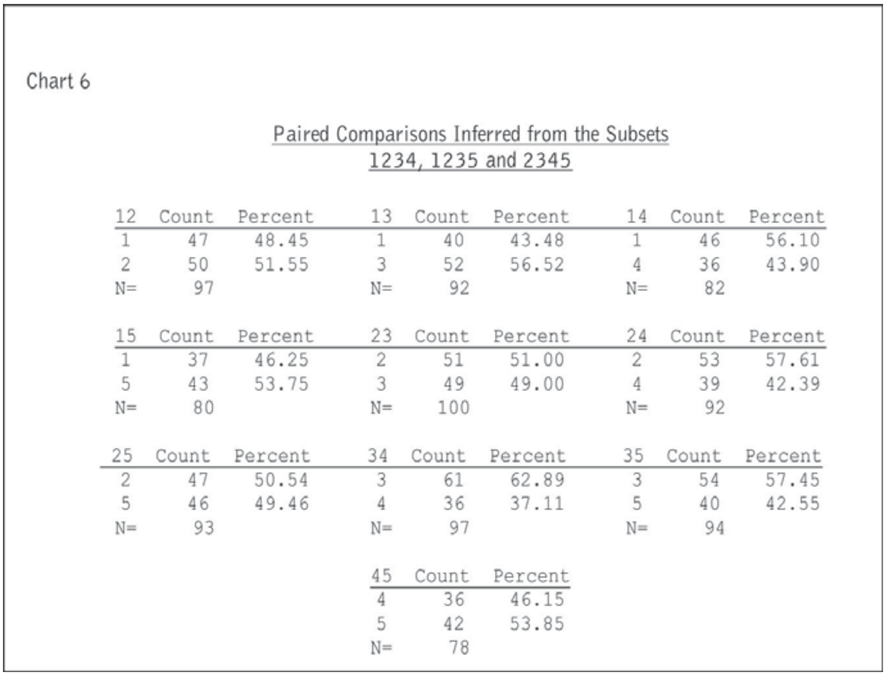


Chart 6

Paired Comparisons Inferred from the Subsets
1234, 1235 and 2345

```
 12  Count  Percent     13  Count  Percent     14  Count  Percent
  1     47    48.45      1     40    43.48      1     46    56.10
  2     50    51.55      3     52    56.52      4     36    43.90
 N=     97              N=     92              N=     82

 15  Count  Percent     23  Count  Percent     24  Count  Percent
  1     37    46.25      2     51    51.00      2     53    57.61
  5     43    53.75      3     49    49.00      4     39    42.39
 N=     80              N=    100              N=     92

 25  Count  Percent     34  Count  Percent     35  Count  Percent
  2     47    50.54      3     61    62.89      3     54    57.45
  5     46    49.46      4     36    37.11      5     40    42.55
 N=     93              N=     97              N=     94

                        45  Count  Percent
                         4     36    46.15
                         5     42    53.85
                        N=     78
```

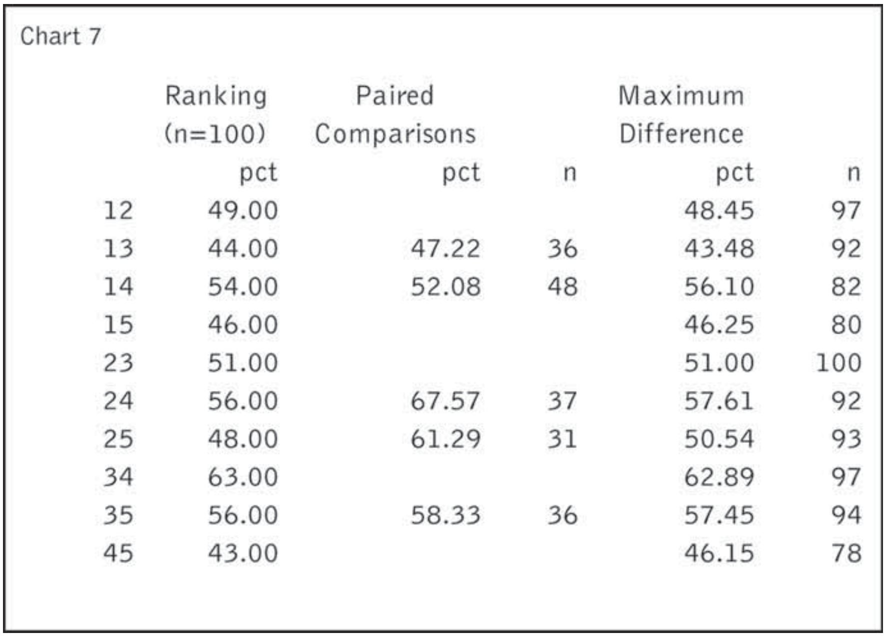


Chart 7

	Ranking (n=100)	Paired Comparisons		Maximum Difference	
	pct	pct	n	pct	n
12	49.00			48.45	97
13	44.00	47.22	36	43.48	92
14	54.00	52.08	48	56.10	82
15	46.00			46.25	80
23	51.00			51.00	100
24	56.00	67.57	37	57.61	92
25	48.00	61.29	31	50.54	93
34	63.00			62.89	97
35	56.00	58.33	36	57.45	94
45	43.00			46.15	78

2." His response to the 2 vs. 3 comparison is, "I prefer Pizza 2 to Pizza 3." His response to the 1 vs. 3 comparison is "I prefer Pizza 3 to Pizza 1."

Here's the rationalization for this intransitive set of responses. The respondent likes onions much more than garlic and garlic much more than mushrooms. Also, the respondent likes anchovies a little better than pepperoni and pepperoni a little better than salami and likes anchovies a lot better than salami. Suppose finally that the respondent makes his choice by first comparing the primary toppings, and, if there is little difference between the primary toppings, uses his preference on the secondary topping as the decider.

When Pizza 1 is compared to Pizza 2 the respondent says, "I like pepperoni a little better than salami but not enough to decide on Pizza 2 on that basis. Since I like onions a lot better than I like garlic, I'll choose Pizza 1."

When Pizza 2 is compared to Pizza 3, the respondent says, "I like anchovies a little better than pepperoni but not enough to decide on Pizza 3 on that basis. Since I like garlic a lot better than I like mushrooms, I'll choose Pizza 2."

Finally, when Pizza 1 is compared to Pizza 3, the respondent says, "I like anchovies way better than salami, so regardless of the secondary topping, I'll choose Pizza 3."

Also, it is well-known that the addition of an element in the set of alternatives may change the rankings of the prior elements. For example, if asked, "Which would you order in this restaurant, chicken or steak?" one might reply "Steak." But if asked "Which would you order in this restaurant, chicken, steak or fish?" one might reply "Chicken."

Here's the rationalization for such a response. When faced with the choice of only chicken or fish in a restaurant one might reason, "I prefer chicken but chicken is more difficult to prepare than steak, so to be on the safe side I'll choose steak."

Now, when faced with the choice of chicken, steak or fish, one might reason, "Fish is even more difficult to prepare than chicken. Since fish is on the menu, this is a signal that there's a good chef here, so I'll now order the chicken." So even explicit responses in a maximum difference mode will not be consistent with the responses in a paired comparison.

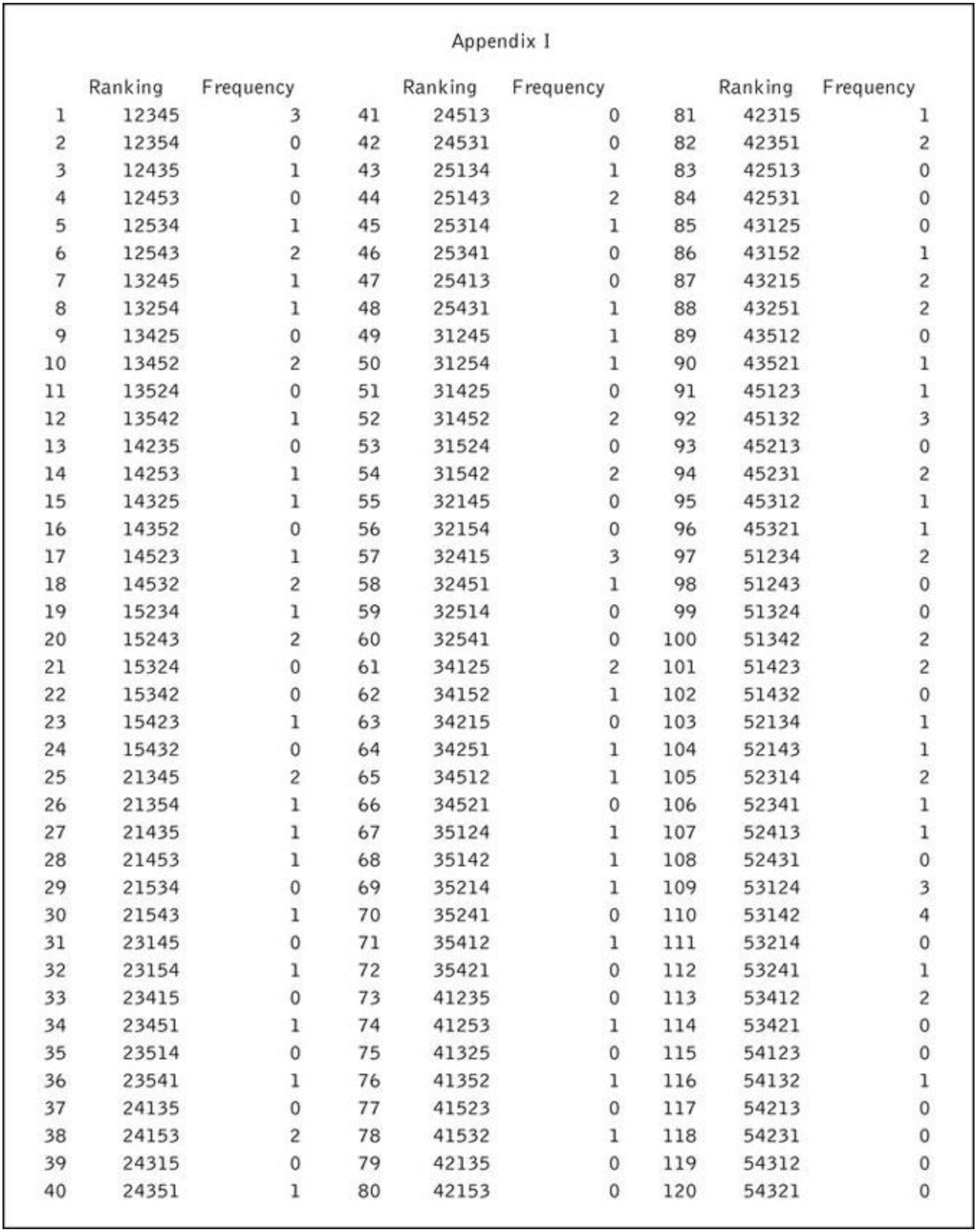
Not an easy substitute

The bottom line is that, though it may be easier to implement multiple paired comparisons and/or maximum difference subsets, one should recognize that they are not an easy substitute for the traditional rankings. | Q

Footnotes

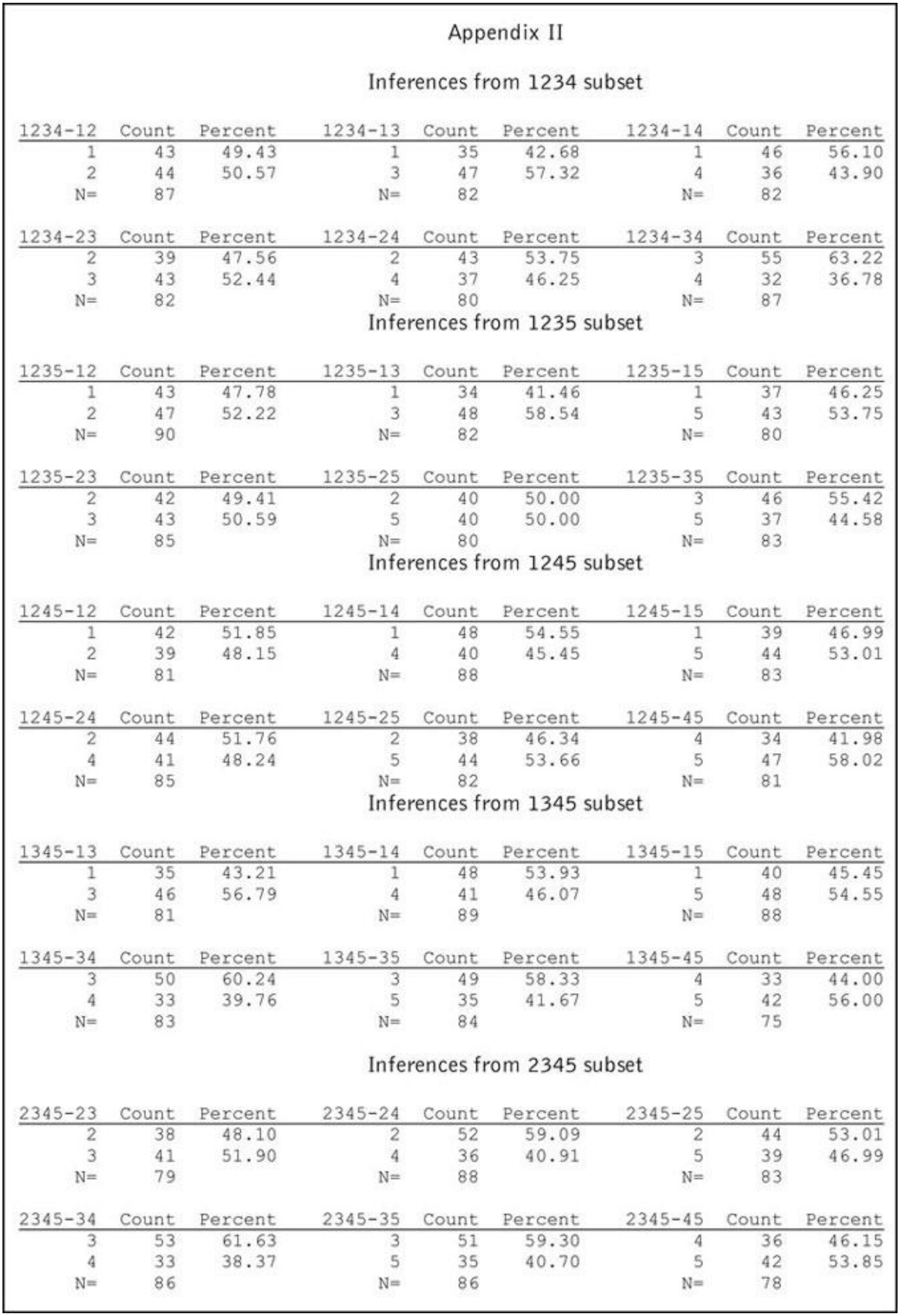
[1] Sometimes the ranking task is bypassed by asking the respondent to rate a large number of items, with the rankings inferred from the ratings. There are many additional issues that can be raised when this manner of obtaining rankings is used, due to problems associated with administering ranking questions in various interviewing modes, problems associated with the validity of the rating scale that is being used and to the problems associated with converting from ratings to rankings. We do not discuss these issues in this article.

[2] Only the use of subsets of sets of items in a maximum difference interrogation or of pairs of items in a paired comparisons setting to estimate preference proportions is considered in this article. We do not consider herein any other uses of these survey frameworks, such as for estimation of utilities in a conjoint setting.



Appendix I

	Ranking	Frequency		Ranking	Frequency		Ranking	Frequency
1	12345	3	41	24513	0	81	42315	1
2	12354	0	42	24531	0	82	42351	2
3	12435	1	43	25134	1	83	42513	0
4	12453	0	44	25143	2	84	42531	0
5	12534	1	45	25314	1	85	43125	0
6	12543	2	46	25341	0	86	43152	1
7	13245	1	47	25413	0	87	43215	2
8	13254	1	48	25431	1	88	43251	2
9	13425	0	49	31245	1	89	43512	0
10	13452	2	50	31254	1	90	43521	1
11	13524	0	51	31425	0	91	45123	1
12	13542	1	52	31452	2	92	45132	3
13	14235	0	53	31524	0	93	45213	0
14	14253	1	54	31542	2	94	45231	2
15	14325	1	55	32145	0	95	45312	1
16	14352	0	56	32154	0	96	45321	1
17	14523	1	57	32415	3	97	51234	2
18	14532	2	58	32451	1	98	51243	0
19	15234	1	59	32514	0	99	51324	0
20	15243	2	60	32541	0	100	51342	2
21	15324	0	61	34125	2	101	51423	2
22	15342	0	62	34152	1	102	51432	0
23	15423	1	63	34215	0	103	52134	1
24	15432	0	64	34251	1	104	52143	1
25	21345	2	65	34512	1	105	52314	2
26	21354	1	66	34521	0	106	52341	1
27	21435	1	67	35124	1	107	52413	1
28	21453	1	68	35142	1	108	52431	0
29	21534	0	69	35214	1	109	53124	3
30	21543	1	70	35241	0	110	53142	4
31	23145	0	71	35412	1	111	53214	0
32	23154	1	72	35421	0	112	53241	1
33	23415	0	73	41235	0	113	53412	2
34	23451	1	74	41253	1	114	53421	0
35	23514	0	75	41325	0	115	54123	0
36	23541	1	76	41352	1	116	54132	1
37	24135	0	77	41523	0	117	54213	0
38	24153	2	78	41532	1	118	54231	0
39	24315	0	79	42135	0	119	54312	0
40	24351	1	80	42153	0	120	54321	0



Appendix II

Inferences from 1234 subset

1234-12	Count	Percent	1234-13	Count	Percent	1234-14	Count	Percent
1	43	49.43	1	35	42.68	1	46	56.10
2	44	50.57	3	47	57.32	4	36	43.90
N=	87		N=	82		N=	82	

1234-23	Count	Percent	1234-24	Count	Percent	1234-34	Count	Percent
2	39	47.56	2	43	53.75	3	55	63.22
3	43	52.44	4	37	46.25	4	32	36.78
N=	82		N=	80		N=	87	

Inferences from 1235 subset

1235-12	Count	Percent	1235-13	Count	Percent	1235-15	Count	Percent
1	43	47.78	1	34	41.46	1	37	46.25
2	47	52.22	3	48	58.54	5	43	53.75
N=	90		N=	82		N=	80	

1235-23	Count	Percent	1235-25	Count	Percent	1235-35	Count	Percent
2	42	49.41	2	40	50.00	3	46	55.42
3	43	50.59	5	40	50.00	5	37	44.58
N=	85		N=	80		N=	83	

Inferences from 1245 subset

1245-12	Count	Percent	1245-14	Count	Percent	1245-15	Count	Percent
1	42	51.85	1	48	54.55	1	39	46.99
2	39	48.15	4	40	45.45	5	44	53.01
N=	81		N=	88		N=	83	

1245-24	Count	Percent	1245-25	Count	Percent	1245-45	Count	Percent
2	44	51.76	2	38	46.34	4	34	41.98
4	41	48.24	5	44	53.66	5	47	58.02
N=	85		N=	82		N=	81	

Inferences from 1345 subset

1345-13	Count	Percent	1345-14	Count	Percent	1345-15	Count	Percent
1	35	43.21	1	48	53.93	1	40	45.45
3	46	56.79	4	41	46.07	5	48	54.55
N=	81		N=	89		N=	88	

1345-34	Count	Percent	1345-35	Count	Percent	1345-45	Count	Percent
3	50	60.24	3	49	58.33	4	33	44.00
4	33	39.76	5	35	41.67	5	42	56.00
N=	83		N=	84		N=	75	

Inferences from 2345 subset

2345-23	Count	Percent	2345-24	Count	Percent	2345-25	Count	Percent
2	38	48.10	2	52	59.09	2	44	53.01
3	41	51.90	4	36	40.91	5	39	46.99
N=	79		N=	88		N=	83	

2345-34	Count	Percent	2345-35	Count	Percent	2345-45	Count	Percent
3	53	61.63	3	51	59.30	4	36	46.15
4	33	38.37	5	35	40.70	5	42	53.85
N=	86		N=	86		N=	78	