21. SCORING MODELS

Social science data collection uses questionnaires composed of items constructed to elicit informative responses. The most common format asks respondents to select from among a set of alternative categories. The construction of useful alternative categories is essential to the success of the items. Once categories have been analyzed in order of increasing "strength" of response, a scoring model must be used to bring responses into a measurement construction process. The analysis of alternative scoring models is the focus of this primer.

Item formats for such instruments are of two general types:

1. Dichotomous responses such as

Agree	Disagree
1	0

and its variants.

2. Polytomous responses such as

Strongly Agree	Somewhat Agree	Somewhat Disagree	Strongly Disagree
4	3	2	1

and its variants.

These choices begin as discrete categories at a nominal level of measurement. Dichotomous choice of categories such as agree/disagree requires that such alternatives be mutually exclusive with no possibility that selecting one category overlap with another category.

The categories, when polytomous, usually have an intended order that ascends or descends across the alternatives. The Likert scale (Likert, 1932) is the most familiar form of instrument design.

Although polytomous responses begin as categories, they are usually intended as "ordinal" by the direction implied in the labeling of the alternatives.

The most informative ordering of categories, however, is not always as originally intended. Sometimes there is ambiguity about the order and position of the categories. Sometimes the categories are carelessly worded or implausible. This produces an experienced order which is different from what was intended. Consider this arrangement taken from U.S. News & World Report dated November 7, 1994:

How do you feel about Bill Clinton?

		Code A	Code B
Hopeful	30%	5	2
Disappointed	19%	2	0
Disgusted	16%	1	0
Uncertain	14%	3	1
Neutral	9%	4	1
Enthusiastic	7%	6	2
Angry	5%	0	0

There is no single, obvious order to this sequence of seven categories. From "neutral", the sequence moves downwards to "enthusiastic" and finally to "angry" or upwards to "disgusted", "disappointed" and then "hopeful"! If this is the sequence that occurred in the respondents' questionnaire, it is quite possible that some respondents became confused. If they did not look carefully at all categories, they might not consider all the alternatives available.

Codes A and B suggest two possible "scorings" for these categories. The use of either of them could give a sequential order better than the arrangement presented.

It is always necessary to determine the extent to which the categories were used by respondents in the way they were intended. Respondents often interpret questions and categories differently from the way in which they are intended. Confusion can also result from ambiguous directions.

Preliminary analysis must compare the intended order of alternatives and respondent behavior. The data cannot be accurately interpreted until a useful scoring model has been determined. This is the reason pilot studies are recommended. The best pilot study investigates the category ordering in the scale before it is used to gather the main data. In any case, a scoring model must be found before the main analysis can be undertaken. Some researchers immediately proceed to code the categories and analyze the responses before ascertaining the most useful scoring model. This leads to misinterpretation of data.

A full scale analysis consists of examining every reasonable scoring model to determine which one produces the most useful results.

Consider the Fear Survey Schedule (FSS) by Wolpe & Lang (1969). The FSS has 108 items. Each item is to be answered on a five-point scale:

0	1	2	3	4
Not At All	A Little	A Fair Amount	Much	Very Much

There are fourteen additional scoring models to be considered.

	FSS CATEGORIES				
Original	0	1	2	3	4
Alternative 1	0	0	0	0	1
Alternative 2	0	0	0	1	1
Alternative 3	0	0	0	1	2
Alternative 4	0	0	1	1	1
Alternative 5	0	0	1	1	2
Alternative 6	0	0	1	2	2
Alternative 7	0	0	1	2	3
Alternative 8	0	1	1	1	1
Alternative 9	0	1	1	1	2
Alternative 10	0	1	1	2	2
Alternative 11	0	1	1	2	3
Alternative 12	0	1	2	2	2
Alternative 13	0	1	2	2	3
Alternative 14	0	1	2	3	3

With 15 different scoring models, the question is, "Which one works best?"

We must study how the people used these categories. This can be done by examining the results for each of the alternative scoring models. Statistics from Rasch analysis provide explicit information about which scoring model is most useful with these data.

The Rasch statistics for the analysis of scoring models are:

1. The item and person separation statistics which are ratio scale equivalents to person and test reliability.

2. The item and person unbiased or "adjusted for error" standard deviations.

3. The fit statistics computed for item, person, and scoring category.

4. Finally, if we have supplemental information, we can go beyond the data and evaluate which scoring model best separates known groups of persons.

We evaluate the differences according to their standard errors and determine whether these differences are significant. But we need to use the model standard error as modified by misfit to accomplish this task.

When a scoring model shows considerable person misfit, this must be taken into account so as to produce a measure error that is increased by the amount of misfit. Any scoring model that increases the person separation statistic or the adjusted standard deviation is more efficient. Increase in either of these statistics gives us an index by which to judge the efficacy of scoring alternatives.

These statistical tools provide explicit ways to evaluate which of the alternate scoring models is most useful.

Category reduction, from more numerous categories to less, frequently provides a more efficient scoring model. Many researchers plan more categories than are used by respondents or useful to define the variable. Too many options are the result of idealized expectations rather than real experience. Fewer categories are often more efficient. Multiple categories are often "nonexistent" and can be modeled more effectively by dichotomous scoring.

We analyze all 15 scoring models of the FSS, using output from *BIGSTEPS* (Wright & Linacre, 1992) to determine which model is most useful.

Table 21.1 gives the data for the 15 scoring models. Column 1 gives the model code. Column 2 indicates the steps in the model. The alternative models are arranged by step. Column 3 gives the number of iterations (UCON) to a converged solution. Columns 4 and 5 give the standard errors for items and persons. Columns 6 and 7 give the person and items infit statistics. Column 8 shows the number of items identified beyond a standardized misfit statistic of 2.0. Columns 9 and 10 give the person and item separation statistics. Columns 11 and 12 give the item and person quotients resulting from the ratios of item and person errors to their standard deviations (Column 4/Column 6 and Column 5/Column 7).

Scoring Model #8 has been highlighted to identify it as the most efficient one. Columns 11 and 12 show this model to have produced the highest values for the adjusted ratio previously described. The person and item reliabilities for this model are 0.97 and 0.98 which are as good as for any other model. The number of misfitting items, while not the lowest, is less than for 11 of the other models.

This model contains only one step, thus indicating that the FSS can be scored efficiently as dichotomous. The more detailed ratings supplied by the authors do not correspond with how the respondents view the scale. Simple identification of fear is sufficient and attempts to discriminate further are unproductive. The FSS functions efficiently in this mode and provides a model that is consistent with the data.

Our example illustrates the need to investigate the possible scoring models in any scale before analyzing the data further. Failure to take into account the influence of scoring model choice confuses subsequent data analysis.

Table 21.1

Wolpe Fear Scale

Scoring Models Analysis

By Mark Stone

1	2	3	4	5	6	7	8	9	10	11	12
SCORING	STEPS IN	UCON	ISEP	PSEP	IINSD	PINSD	# Items	ISEPR	PSEPR	ISEP/	PSEP/
MODEL	MODEL	# ITERATIONS					Out			IINSD	PINSD
		[T.0.2]	[T.3.1]	[T.3.1]	[T.3.1]	[T.3.1]	[T.3.1]	[T.3.1]	[T.3.1]	[C4/C6]	[C5/C7]
1 SM 00001	1	4	1.9	1.6	0.13	0.07	4	0.78	0.71	14.62	22.86
2 SM 00011	1	4	3.3	2.6	0.12	0.12	8	0.92	0.87	27.50	21.67
4 SM 00111	1	3	5.4	4.0	0.11	0.16	11	0.97	0.94	49.09	25.00
8 SM 01111	1	4	7.4	5.3	0.09	0.17	9	0.98	0.97	82.22	31.18
3 SM 00012	2	8	3.2	2.3	0.17	0.27	8	0.91	0.84	18.82	8.52
5 SM 00112	2	3	5.2	3.8	0.16	0.30	11	0.96	0.94	32.50	12.67
6 SM 00122	2	8	5.2	3.8	0.13	0.25	10	0.96	0.93	40.00	15.20
9 SM 01112	2	20	7.4	5.5	0.18	0.40	15	0.98	0.97	41.11	13.75
10 SM 01122	2	9	7.5	5.8	0.18	0.35	20	0.98	0.97	41.67	16.57
12 SM 01222	2	7	7.9	6.0	0.14	0.30	21	0.98	0.97	56.43	20.00
7 SM 00123	3	14	5.0	3.6	0.19	0.36	14	0.96	0.93	26.32	10.00
11 SM 01123	3	10	7.2	5.4	0.25	0.51	26	0.98	0.97	28.80	10.59
13 SM 01223	3	4	7.7	5.7	0.19	0.44	21	0.98	0.97	40.53	12.95
14 SM 01233	3	10	7.6	5.7	0.19	0.40	22	0.98	0.97	40.00	14.25
15 SM 01234	4	15	7.4	5.5	0.25	0.51	22	0.98	0.97	29.60	10.78

MEASUREMENT ESSENTIALS

2nd Edition

BENJAMIN WRIGHT

MARK STONE

WIDE RANGE, INC.
Wilmington, Delaware