Rating Scale Categories:
Dichotomy, Double Dichotomy, and the Number Two

Mark H. Stone, Ph.D.
Adler School of Professional Psychology
Chicago, IL

My conjecture is that dichotomies in rating scales are more useful than multiple ratings. This conjecture implies that most multiple ratings can be reduced to a useful natural dichotomy making construction of multiple ratings futile. Why do I maintain such a conjecture when most rating scale practice uses multiple categories?

Personality Inventories

First, I illustrate my point by reminding the reader that the most utilized of all standard personality inventories is the Minnesota Multiphasic Personality Inventory, the famous MMPI. It uses a dichotomy, true/false, for response alternatives. MMPI protocol allows a "?" or "cannot say" response as an alternative. But the directions ask the test administrator to encourage the respondent to return again to such responses and to decide in which direction to mark the answer. The goal is to eliminate middle responses.

I don't argue that just because the test authors recommend this, it is correct. I only remind the reader that if multiple ratings had been found to be more advantageous, you can bet they would appear on the test protocol. I suggest this has not occurred because after decades of use the dichotomy still works.

Indeed, multiple ratings have not been found to add information, but rather provoke noise. When the number of "?" responses is high it is a sign that the validity of the entire test is in question. Graham (1987, p. 19) says, "... the validity of a resulting protocol with many omitted items should be questioned..." and "... encourage individuals to try to answer previously omitted items, most people will complete all or most of the items." Graham says the same in his text on the revised edition MMPI2 (Graham, 1993). The MMPI Manuals for both editions recommend the same procedures.

We see that "forcing" a dichotomy is standard administrative practice for the two editions of the MMPI and the same can be said for the competing personality inventories, the Millon Clinical Multiphasic Inventory, California Personality Inventory and 16PF.

The earliest edition of the MMPI produced each item separately, printed on a card, and the patient placed cards sorted "true" in one box and those sorted "false" in the other. I have always considered this process an intelligent procedure for patients inasmuch as most of the people taking the MMPI are less then optimally functional. Any strategy that assists them ought to be promoted. Sorting is a tactile activity as well as a cognitive one that is advantageous to the subject. It gives the respondent the opportunity to "handle" the question and physically sort as opposed to marking responses with a pencil on an answer sheet. The size of the response window has progressively decreased over the years. I doubt this has brought much advantage to respondents. The main impetus for answer sheets is that the original sorting routine was troublesome to score for psychologists. Today's streamlined answer sheet can be quickly scanned. Good for psychologists. Bad for subjects.

From appraising well-known personality inventories, we observe that patients are asked to make dichotomous decisions to each item. To inquire into motivation and other confounding variables behind their responses takes us away from the problem at hand, but requiring a true/false or yes/no response clearly seems the most useful way to collect responses from patients. For people under stress, this is the most reasonable expectation and solution. Of course, personality inventories are not rating scales, but the problem of determining a valid response alternative is common to Likert scales and personality inventories, and the latter have promoted the dichotomy for more than 50 years with little motivation to change.

I think this example adds support to my conjecture, but taken alone it is not an overwhelming argument for advocating a dichotomy. What adds more evidence to my conjecture comes from the reasoning of individuals about the status of a dichotomy in general. There are several quotes worth thinking about.

Karl Menninger in his book on Number Words and Number Symbols says

> *"Two has a special status and is not just a number like any other in the number sequence, but instead is that extra ordinary number ... ."*

He then goes on to say that the number two has more significance then we might assume today in the era of big numbers. It occupies a unique place after "one." But it is not only the second numeral in our counting system. Two suggests something beyond "one more" because at this juncture we enter upon the idea of contrasts, comparisons, and opposites.

The proverbial essay question that teachers frequently give to students often requires "contrast and compare" in some form or another. We pursue many tasks efficiently and effectively by dichotomous grouping, particularly when they are vo-

luminous and tend to overwhelm us. Consider the following categories and just one of the dichotomous groupings that can result.

Spelling: words spelled phonetically vs. words that are not.
Grammar: regular verbs vs. irregular ones.
Math: plane geometry vs. spherical.

Alfred Adler and other psychologists have suggested that a dichotomy is generally the haven of the perplexed, the neurotic, and the primitive mind. The dichotomy comes forth whenever we feel pressured or at risk. At such times we formulate response alternatives by a dichotomy, not by imagining an array of alternatives. So whenever respondents do not know how to answer an item they respond by falling back on a dichotomy.

Jung also thought two had a special value.

"Two is the first number because, with it, separation and multiplication begin, which alone make counting possible."

What the number two brings us is a phenomenon that is omnipresent:

- from the body: two eyes, two ears, two hands, two feet, two kidneys, two lungs.
- from nature: male/female, night/day, sun/moon.
- from contrasts: old/young, right/left, up/down, plus/minus, hit/miss.
- from mythology: god/goddess, two in one — twins, the Egyptian double lion, named Routi.

Given this ubiquity for opposites, are we not more attuned to a dichotomy than to any other system?

Edward Edinger (1995) expands Jung's point in discussing Moby Dick.

A major theme of Moby-Dick is the problem of opposites. As we proceed we shall encounter numerous antitheses: alienation and inflation, courage and cowardice, strength and weakness, black and white, good and evil, the bounded land and the boundless sea, height and depth, the universal and the particular, Christian and pagan, primitive and civilized, the outer word and the inner soul, spirit and matter, destiny and free will, love and hate, calm and turbulence, delight and woe, orthodox and heretic, reason and madness, God and man. (p. 30)

Paul Tillich, the philosopher/theologian adds this point,

"Philosophical ideas necessarily appear in pairs of contrasting concepts, like subject and object, ideal and real, rational and irrational."

Tillich reminds us that ideas are "paired," that for every point we conjecture an opposite.

Lastly, C.S. Peirce, the American philosopher/logician expresses in a more comprehensive view the totality of what is found in the first three numbers.

"First is the conception of being or existing independent of anything else. Second is the conception of being relative to, the conception of reaction with something else. Third is the conception of mediation. ... The origin of things, considered not as leading to anything, but in itself, contains the idea of First, the end of things that of Second, the process mediating between them that of Third."

What these thinkers have to say about "two-ness" and the dichotomy is more than idle speculation. They are speaking about a phenomenon that permeates our thinking about the number two and a dichotomy. We see most concepts in terms of dichotomies — pairs, opposites, and contrasts.

George Miller (1956, p. 82) offers commentary that is relevant in his paper entitled, "The magical number seven, plus or minus two: Some limits on our capacity for processing information." Miller defines "amount of information" as variance which is a dimensionless quantity. He goes on to say,

"When we have a large variance, we are very ignorant about what is going to happen. If we are very ignorant, then when we make the observation we get a lot of information. On the other hand, if the variance is very small, we know in advance how our observation must come out, so we get little information from making the observation." (p. 82)

The key point from Miller which applies to rating scales is whether or not we "get a lot of information." This can only occur with multiple ratings when a two-step model is shown empirically to be more informative than a one-step model, and threesteps is shown to be more informative than two steps. Instead, the construction style for most Likert scales seems to be slapping as wide a range of response alternatives as possible to a varied collection of poorly worded items. Such a process cannot produce information.

From this state of ignorance it is possible to "collect data," but the quality of such responses is unknown and suspect. Not knowing how a person will answer an item is an entirely different problem from not knowing what the possible response alternatives might mean to a range of respondents. In the former situation we have the state of ignorance prior to knowing the outcome. In the second situation we are simply ignorant of how to build a response alternative that is meaningful. We might want to read the thermometer with scientific dispassion, but we do not construct a thermometer dispassionately! We give its construction our best attention. There is a big difference between these two states of ignorance, and there appears to be misplaced credence in believing that "ignorance" expresses the desired state of neutrality in scientific work. If we propound ignorance do we produce knowledge or only become more confused?

There is one response scheme that is popular on rating scales. It builds on a double dichotomy of four alternatives. A common example is "Strongly Agree, Agree, Disagree or Strongly Disagree." Miller (1956) informs us that "Two bits enables us to decide among four equally likely alternatives" (p. 83). As the number of alternatives increases by a factor of

two, one more bit of information is added. Consequently, eight alternatives equals three bits, which is about as many response alternatives as are ever found on a rating scale.

Miller says,

> *"It is interesting to consider that psychologists have been using seven-point rating scales for a long time, on the intuitive basis that trying to rate into finer categories does not really add much to the usefulness of the ratings" p. 84.*

He goes on to cite four experiments in which a good observer can identify about four intensities, about five durations, and about seven locations. Miller argues that our nervous system gives us a finite limit to our capacity for making judgments. This limitation does not vary much from one sensory attribute to another.

His article concludes by saying (1) we have definite limitations of absolute judgment (2) chunking helps and is the only way we can address this limitation. In his summary, Miller suggests,

> *"the recoding that people do seems to me to be the very lifeblood of the thought processes" p. 95.*

With four alternatives we must solve two dichotomies. The first one is 1-2 vs. 3-4 followed by deciding between 1-2 or 3-4, or else the item is resolved as a single dichotomy 1 vs. 2-4 or 1-3 vs. 4. We solve a double dichotomy of four responses by chunking the problem into two groupings of two each — two successive dichotomies — or else form it into a single dichotomy.

Lastly, Miller proffers his theory as

> *"a yardstick for calibrating our stimulus material and for measuring the performance of our subjects" p. 96.*

His conclusion of a natural limit of three bits makes eight alternatives the maximum according to his evaluation of four physiological and memory studies. He concludes that the practical span of alternatives is, in fact, much smaller than eight. On the basis of his studies we are advised to reduce rather than expand the number of ratings. Miller infers that through chunking and recoding we resolve a large number of alternatives into a smaller number. The process may occur so quickly with some items as to make us think it is a single solution, but whenever we have to pause and deliberate over multiple ratings, it is clear that chunking and regrouping are operating.

We need to be aware of the limitations of our nervous system and not offer the possibility of multiple ratings when, in fact, they are not easy to resolve. Multiple ratings have to be demonstrated as empirically operating, not imagined to do so. It is doubtful that we can actually cope systematically with many alternatives. What we learn from Miller's investigations is that the dichotomy is not easily transcended.

Support for my conjecture of the dichotomy also comes from considering the practice of rescoring response alternatives. I present two examples, showing in both of them that the rescoring of four alternatives is efficiently reduced to two.

The first example concerns the Beck Depression Inventory. It was administered in the Adler clinic to 266 non-clinical subjects and 153 clinically depressed persons. This scale of 21 items has four responses to each item indicated by 0, 1, 2, or 3. James Natter and I recoded these responses to a dichotomy of 0 = 0 and 1 = 1, 2, 3 which produced a dichotomous scoring model that differentiated between clinical depressed and non-clinical subjects better than the original category scale. A second rescoring dichotomy 0 = 0, 1 and 1 = 2, 3 was not as discriminating as the first, but still better than the original scale. The first dichotomy also produced better differentiation between persons attempting suicide or not in the depressed sample than did the original scale. Natter (1994) concluded that the original BDI scale is less effective than the dichotomy for differentiating pathology.

The second sample includes responses of 233 outpatient subjects in the Adler clinic taking the Wolpe-Lange Fear Survey Schedule II (1969). This scale is a self-report list of 108 items to which respondents endorse the amount of unpleasant feelings associated with each. **Table 1** gives the complete rescoring analysis for each of the 15 models.

Column 1 gives the scoring code.
Column 2 gives the steps in the model.
Column 3 PSEPR is the person separation reliability.
Column 4 PSEP is the person separation index.
Column 5 ISEP is the item separation index.
Column 6 UCON is the number of iterations for convergence.
Column 7 PINSD is the person infit standard deviation.
Column 8 IINSD is the item infit standard deviation.
Column 9 is the number of items identified beyond a standardized misfit of 2.0.
Column 10 ISEPR is the item separation reliability.
Column 11 PSEP/PINSD is the ratio of person separation to the person infit standard deviation.
Column 12 ISEP/IINSD is the ratio of item separation to the item infit standard deviation.

Examination of the results shows that model 01111, a one-step model, and model 01122, a two-step model, were better than the original model 01234, a four-step model. Model 01222 does better than any other two, three, or four-step models in ISEP and PSEP, but does produce misfit in 21 of the 108 items. Model 01111, however, while losing some ISEP and PSEP saves 12 of these items. This model is efficient. The ISEP and PSEP indices are among the highest values for several models. The number of fit items, although not the lowest, is less than eleven other models. Model 01111 contains only one step and indicates that the FSS can be efficiently scored as dichotomous. Columns 11 and 12 produce their highest values for the dichotomous model.

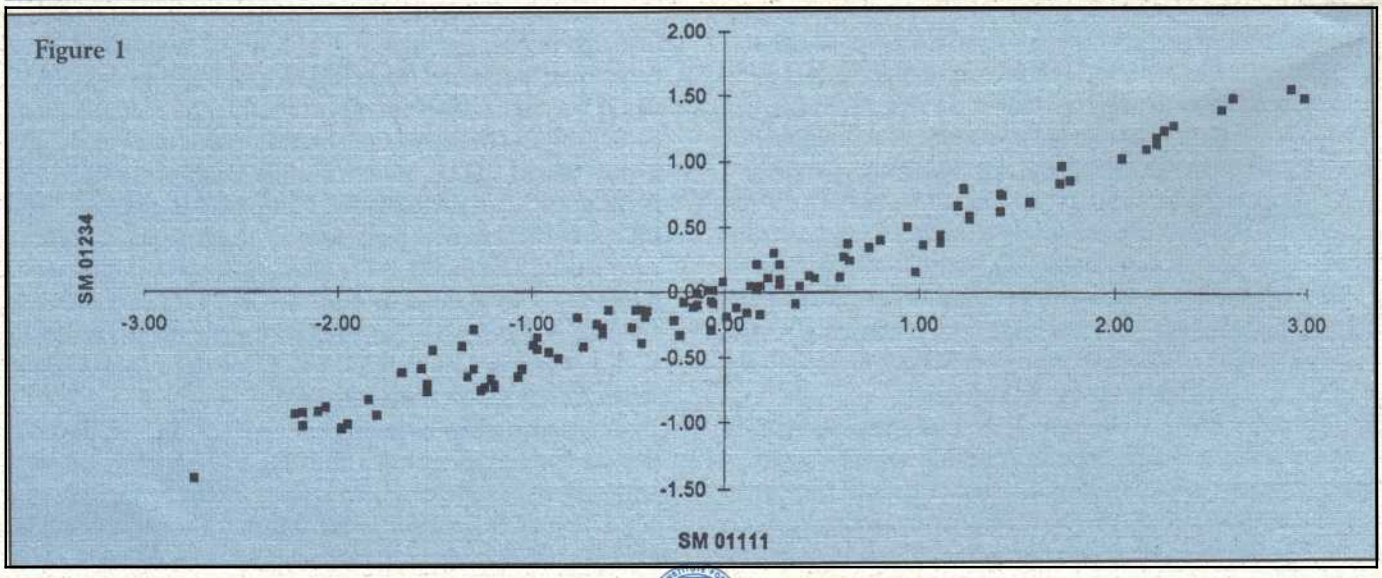
Comparing the dichotomous model of 01111 to the two-step model 01222 produces a PSEP ratio of 5.3/6.0 = .88 indicating the dichotomous model is 100(5.3/6.0) = 88% efficient of the best scoring model of the fifteen. The original model 01234 is 100(5.5/6.0) = 91% of model 01111, but at the cost

Table 1

Wolpe Fear Scale
Scoring Models Analysis
Mark Stone

	1	2	3	4	5	6	7	8	9	10	11	12
	Scoring Model	Steps in Model	PSEPR	PSEP	ISEP	UCON # Its	PINSD	IINSD	# Items Out	ISEPR	PSEP/ PINSD	ISEP/ IINSD
			[T. 3.1]	[T. 3.1]	[T. 3.1]	[T. 0.2]	[T. 3.1]	[T. 3.1]	[T. 3.1]	[T. 3.1]	C4/C7	C5/C8
1	SM 00001	1	0.71	1.6	1.9	4	0.07	0.13	4	0.78	22.86	14.62
2	SM 00011	1	0.87	2.6	3.3	4	0.12	0.12	8	0.92	21.67	27.50
3	SM 00012	2	0.84	2.3	3.2	8	0.27	0.17	8	0.91	8.52	18.82
4	SM 00111	1	0.94	4.0	5.4	3	0.16	0.11	11	0.97	25.00	49.09
5	SM 00112	2	0.94	3.8	5.2	3	0.30	0.16	11	0.96	12.67	32.50
6	SM 00122	2	0.93	3.8	5.2	8	0.25	0.13	10	0.96	15.20	40.00
7	SM 00123	3	0.93	3.6	5.0	14	0.36	0.19	14	0.96	10.00	26.32
8	SM 01111	1	0.97	5.3	7.4	4	0.17	0.09	9	0.98	31.18	82.22
9	SM 01112	2	0.97	5.5	7.4	20	0.40	0.18	15	0.98	13.75	41.11
10	SM 01122	2	0.97	5.8	7.5	9	0.35	0.18	20	0.98	16.57	41.67
11	SM 01123	3	0.97	5.4	7.2	10	0.51	0.25	26	0.98	10.59	28.80
12	SM 01222	2	0.97	6.0	7.9	7	0.30	0.14	21	0.98	20.00	56.43
13	SM 01223	3	0.97	5.7	7.7	4	0.44	0.19	21	0.98	12.95	40.53
14	SM 01233	3	0.97	5.7	7.6	10	0.40	0.19	22	0.98	14.25	40.00
15	SM 01234	4	0.97	5.5	7.4	15	0.51	0.25	22	0.98	10.78	29.60

Figure 1



of 15 items! A plot of the dichotomy vs. the original model in **Figure 1** shows that the person measures for the two models are consistent (r = .98). Simple identification of fear is sufficient. Attempts to discriminate further are not useful. The FSS functions very well when scored as a dichotomy.

In both examples, reduction to a dichotomy was a reasonable alternative. This is useful to know before beginning further study of the data. It might prove useful to retain the original format for administration, but it is clear that the original model is only a conjecture of what the authors imagined, not what occurred.

Conclusions

1. Personality measurement has employed the dichotomy for more than 50 years as a response alternative. The dichotomy has worked in this field.

2. The dichotomy is a fundamental phenomenon of mind according to those who have given it thought — Menninger, Adler, Jung, Edinger, Tillich, and Peirce. It operates most noticeably when we are overwhelmed by experience that needs reduction.

3. Miller explains that multiple ratings are limited by our span of comprehension and that we reduce multiple indices by chunking and regrouping, especially when overwhelmed.

4. Scoring model analysis indicates that dichotomous models are as good or better than some original scales. These examples show that devising multiple ratings requires more than attaching a rating scheme to an item.

Based upon my conjecture, here are some suggestions for scale construction.

1. Put yourself in the respondents' role and carefully determine what their responses might be. Utilize the psychology of human behavior to determine how respondents might behave. Don't just slap a rating scale to an item.

2. Write a strategic number of carefully crafted items that contribute to the construction of a unidimensional variable. Don't ask every question you can imagine.

3. Begin with a dichotomy and forget about having multiple ratings until a well-defined variable has been constructed. If you think it might be useful, expand the rating alternatives and evaluate the results.

4. Analyze all the scoring models to see how each is working. Don't begin with a rating scale and pretend it works.

To construct a good scale we first need to address the intent of the scale regarding person response behaviors, not write items. We need to identify the characteristics of the intended respondents. This will guide how items should be written and response model alternatives.

The major question is, "Do author and respondent models coincide?" If we do not make a careful analysis of responses we will never know the answer to this question. Many researchers accept the responses according to their intention without bothering to make an analysis of respondent behaviors. Scoring models need to be evaluated to determine how respondents view the scale. My conjecture suggests we should accept the preeminence of the dichotomy as the operating model until other alternatives can be demonstrated.

Ben Wright has suggested that the scale is a "conversation" between the author(s) of the scale and the respondent(s). This is a useful model for scale construction and it reiterates the idea that the first task in scale construction is not to write items, but to address the possible range of relevant person behaviors that could occur. I have suggested a number of steps to follow in item construction, but want to emphasize that planning for respondent behavior should always precede item writing.

The next step is creating a response format. I argue that rather than create the typical Likert response format, use a dichotomy to investigate whether a variable has been achieved. When a variable has been successfully constructed, investigate whether or not the measures are enhanced by a more complex scoring format. Proceeding in a step-by-step approach is more sensible than beginning with a more complex response scheme that may not work. When in doubt, keep it simple. Use a dichotomy.

References

Edinger, E. (1995). Melville's Moby-Dick: A Jungian commentary. Toronto: Inner City Books.

Graham, J. (1987). The MMPI: A practical guide. (2nd Ed.), New York: Oxford.

Jung, C. (1938). Psychology and religion. Volume 11. Collected Works of C.G. Jung. Princeton, NJ: Princeton University Press.

Linacre, J. & Wright, B. (1997). BIGSTEPS. Chicago: MESA Press.

Menninger, K. (1969). Number words and number symbols. New York: Dover.

Natter, J. (1996). Psychometric properties of the Beck Depression Inventory. Unpublished dissertation.

Peirce, C. (1940). in Philosophical writings of Peirce. J. Buchler (Ed.), New York: Dover.

Tillich, P. (1962). Existentialism and psychotherapy in psychoanalysis. in Existential Philosophy H. M. Ruitenbeak (Ed.), New York: Dutton.

There are three stages to the life of revolutionary scientific ideas. They are initially rejected as outrageous heresies, then they are recognized as brilliant discoveries, and finally they are assumed to be the way things have always been.

William James (paraphrased)