Defining Professions with Rasch Paired-Comparisons: New Instruments for Better, Faster and Easier Task Analyses

Introduction

Task analyses are vital components to ensuring the validity of testing programs. Nowhere is this more evident than in the world of high-stakes testing. Licensure examinations are entryways into practice for many professionals while at the same time serving as a protection for the public from practitioners who are unqualified.

Licensure and certification examinations typically portray themselves as "criterion-referenced" such that the content presented on the examination is a reasonable reflection of the content knowledge required of practicing professionals. The development of a practically based examination typically begins with a gathering of professionals who talk and reflect about what they consider to be the content of the profession. As important as this first step in establishing content-validity may be, it does not provide the critical, objective evidence to support their claim. The task analysis provides critical evidence to support validity.

A task or "job" analysis attempts to dissect a profession into its component parts. For instance, what exactly does a practicing physician do during their day? Do they take blood pressure readings? Do they diagnose strep throat? Do they mend broken bones? The task analysis evaluates the profession through a series of representative, single "tasks" performed by working professionals on a routine basis.

The typical task analysis instrument asks the participant to review a series of presented tasks and answer two questions. The primary question included in these surveys addresses frequency. Professionals are asked to rate "how often" a specific task is performed. The rating scale used depends largely upon the profession but for the purposes of this explanation often runs from "more than once a day" to "once a year" or "never". An additional question regarding *importance* is often added to the survey. Because the number of times a task is performed does not necessarily relate directly to its criticality, participants are often asked to evaluate the task on the basis of how important it is to their job and their other tasks in general.

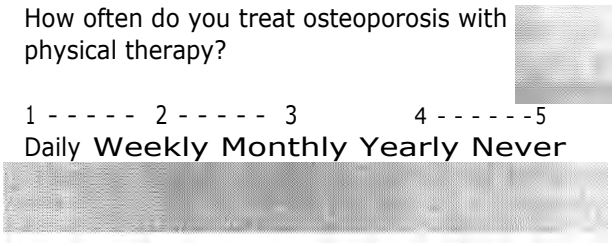
About the Author

Gregory E. Stone

Paper presented at IOMW-XI, New Orleans, 2002

Data for both importance and frequency are most often gathered using traditional Likert scales. The following example illustrates the traditional means for measuring frequency:

How often do you treat osteoporosis with physical therapy?



Similar scales are used for importance.

While Likert scales work well for this purpose, they may also increase the practical difficulties associated with the task analysis efforts. Likert decisions take a long time to make, particularly when considering there may be more than one hundred tasks to examine. Each task requires the respondent to evaluate the content and make a targeted speculation regarding how often (or how important) a task is considered in relation to their overall job. Careful answers require careful contemplation and time may become a limiting factor.

Time is of particular importance when working with volunteers. Most often these studies are conducted using volunteers in the field. While volunteers are usually anxious to help, their enthusiasm fades when presented with a project that takes a considerable amount of time to complete. With a dearth of enthusiasm comes the problem of low return rates. Rates of below 30% are not uncommon when completing such analyses. Additionally, those that do return the surveys may become so tired (or bored) with responding that they leave many answers blank or worse yet, fill in answers without careful consideration. Bad data is worse than no data, but either condition signals trouble for the researcher!

Faced with such practical dilemmas we began to look for ways to make the process of completing questionnaires simpler and more convenient, and at the same time ensure that the collecting of data is adequate for our purpose. A lesser-used statistical model offered a possible solution. The hypothesis in our initial and subsequent evaluations was that paired comparison developed instrumentation might improve the collection of data in two ways. First, we hypothesized that the use of Likert scales might increase the level of relative, unstable speculation. We wondered how easy it really was for a physician to think back and remember accurately whether a specific procedure was performed once a week or once a month. For instance, if a measles

epidemic occurred directly before data collection but had not occurred earlier in the year, how would the respondent make their choice? Second, we hypothesized that the decision made concerning a pair might be easier to make and may therefore result in the collection of data more precise that that collected through traditional instruments.

Both hypotheses would be tried and tested in this and earlier reports with great success. Our report should not be considered as a groundbreaking methodology for performing task analyses. The thought process and methodology behind the creation of specific questions, and the way the data are used to develop a content outline or table of specifications are different and tangential considerations. This investigation primarily assists in the development of remarkably useful instrumentation that is helpful in gathering data within differing theoretical frameworks. It speaks to the powerful simplicity of the Rasch-derived models and illuminates how elegantly data can be collected and analyzed outside of the tradition associated with such enterprises. It beautifully illustrates how thinking "out of the box" and using tools of remarkable power and flexibility can lead us to discovery. We owe much to our communities for inspiring such development and I wish to mention that the seed for this development was planted at a 1998 meeting of the Chicago Objective Measurement Table (COMET), affectionately referred in its early renditions as the "eating" table for reasons known well to those of us who maintain more logits on our waistlines than is the norm.

This investigation assists in the developmeni of remarkably useful instrumentation that may be helpful in gathering data within any number of differing theoretical frameworks.

Research Preparations

As suggested, we began with a traditionally developed, task analysis that included items addressing both *frequency* and *criticality (importance).* The instrument would be created for a high-stakes osteopathic board's review of their profession.* The board consisted of a 7-member panel of expert, working physicians. Eligibility to participate in this project was governed by certification. All 430 board-certified participants were eligible. From this group a self-selecting sample would be taken. Self-selection can be a serious problem, particularly when small sample sizes are encountered. It was decided that should a non-stratified sample be obtained, additional over sampling would be necessary. We shall see later that such was not necessary.

During a daylong meeting of the board, and through the use of pre-meeting "homework" a total of 76 common tasks were selected to represent the profession. Each task was considered to be part of one of five specific, related "subcategories". The methods behind the development of these 76 tasks and the manner in which the data would be used after basic analysis to develop the table of specifications and content outline will not be discussed as they are not germane to the central purpose of this paper. Instead, our focus is the construction of instrumentation and the evaluation of the subsequent data matrix.

Instrumentation

Development of the instrumentation involved the structuring of each of the 76 tasks within a systematic framework of pairs. As described earlier, the 76 tasks were representative of 5 distinct subcategories. It was important in creating the task pairs to elicit as much information as possible.

Tasks were arranged into the 5 subcategories by virtue of related content. Within each subcategory tasks were ordered according to *hypothesized* frequency. Pairs were assembled to allow for comparisons both across and within subcategories.

Sample Construction of Task Pairs

***A** selection of tasks in topics X and V:*

	Content Category X	Content Category Y
Hypothesized Most Frequent	Item 1	Item A
	Item 2	Item B
Hypothesized	Item 3	Item C
Least Frequent	Item 4	Item D

May produce task pairs arranged as:

Item 1 vs. Item A	(Across groups)
Item 4 vs. Item D	
Item 1 vs. Item 2	(Within group)
Item 3 vs. Item 4	
Item A vs. Item B	
Item C vs. Item D	

The pairs were arranged randomly on the developed survey instruments. Two response items adjoined each pair of tasks. The first asked respondents to consider the task pair and indicate which task was performed more frequently within their medical practice. Unlike the more specific judgment needed to answer the Likert-scale item, respondents need only determine which task was "more" and which "less".

For reasons of confidentiality and security, we cannot reveal all specific information discovered or employed in the project, but we will include as much non-sensitive material as is practical and available.

The second question asked respondents to consider the two tasks in relation to their practice as a whole, and then make a determination as to which was more and which was less important. Unlike the Likert-scale item, the decision is again simplified to a dichotomous comparison.

Task A	vs.	Task B	Which task is performed more frequently in your practice?	Considering your overall job, which task is most important?
4 Culture for definitive diaanosis of tonsillitis	vs.	Treat corneal abrasions wit antibiotics	o A o B	o A o B
Treat corneal abrasions with antibiotics	vs.	Biopsy a palpable mass in the breast	o A o B	o A o B
Manage acute otitis media	vs.	Use X-Ray to assess foreign body in nose	o A o B	o A o B
Use X-Ray to assess foreign body in nose	vs.	Treat a simple fracture	o A o B	o A o B
Examine (dgitally) for pain in lower right quadrant of abdomen	vs.	Employ OMT for treatment of back pain associated with pregnancy	o A o B	o A o B
Perform flexible sigmoidoscopy	vs.	Manage cervicitis with vaginal cream	o A o B	o A o B
Biopsy a palpable mass in the breast	vs.	Debride ulcer (of the foot) ,	o A o B	o A o B

The theory regarding the supposed simplicity of responding to a paired comparison is more explicit when examining a less speculative construct like weight. The comparison of weight is straightforward and easily identifiable. Suppose we pick up two stones, one in our right hand and the other in our left. Which of the following questions will be easiest for us to answer?

Question one asks us to guess the weight of each stone, in pounds and ounces.

Question two asks us to compare the weights of each stone and determine which stone is heavier.

Even in the case of stones weighing nearly alike, the latter is clearly an easier question because the measurement scale is less detailed. Rather than guessing a specific weight, we can simply respond as "more" or "less". Broader measurement tools with smaller degrees of speculation about precision cannot help but produce a lower measurement error.

The problem with "more" and "less" ratings is that they may be too broad and too general. Indeed, a single rating or more or less does not go far in responding to our need. Yet if we consider each decision as a single observation, and combine that observation with many others, both overlapping and independent, we will discover a richly defined pattern observed by many other researchers using similar paired comparison designs to explore issues from geographic distance to patient well being.

The Data Matrix

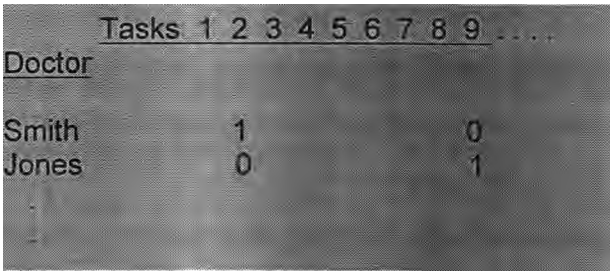
Each task pair within the instrument is considered to be unique and independent. This consideration is different from traditional Likert arrangements. When preparing Likert obtained data for analysis, the data are typically arranged such that each person is presented with a single, sequence of responses. On the other hand, paired-comparison models view each task pair as a single and unique observation. Unique observations are arranged to express each person's response to each task pair as a separate parcel of information. The following example illustrates the process for constructing a functional matrix.

Consider the first task pair presented to our sample group of physicians. The participants were asked to determine which of the following two tasks were performed most frequently in their practice:

Task 2: Culture for definitive diagnosis of tonsillitis
Task 9: Treat corneal abrasions with antibiotics.

The first respondent, Dr. Smith, reported that task 2 was the task performed most frequently. We transform Dr. Smith's response dichotomously for *each* response. In coding the response we would offer a "1" for task 2 (most frequent) and a "0" for task 9 (least frequent).

On the other hand, the second respondent, Dr. Jones, reported that task 9 as the task performed most frequently. We would transform Dr. Jones' response in the same dichotomously fashion: a "0" for task 2 (least frequent) and a "1" for task 9 (most frequent).



To create the matrix, we first arrange the tasks in a standard, defined order (see example matrix above). Once arranged, we enter the recoded data generated by each respondent. In this instance, we have placed a 1 in the column for task 2 and a 0 in the column for task 9. Similar recoding and data arranging for each task pair presented is completed to develop the final matrix.

It is evident from this arrangement that the matrix can and does grow very large, very rapidly. Computerized programs such as Microsoft Access© and Excel() can and should be employed to automate data arrangement as it is a considerably time consuming task.

The Limitation of Paired-Comparisons

Use of task pairs provides an extensive amount of information from which precise task rankings (orderings) can be developed. Unfortunately, rankings have a basic limitation in that they lack a zero point. Without a zero, or similar reference, it is impossible to assign a more specific value to more and less. For instance what does it mean to be "most" frequent? Once a day? Once a month? Once every minute? The lack of specificity may become troublesome when trying to define a test specification table for an exam using data from the job analysis.

Our solution to this vexing problem was to create a hybrid instrument; by adding six tasks presented in the traditional Likert item format. The six tasks were chosen to represent hypothesized "extremes" of most and least, as well as tasks hypothesized to fall within the mid-range of frequency. Data collected from this section of the instrument would allow the assignment of distinct measurements to rankings generated by the paired-comparison items. This hybrid construction was well suited for its purposes.

Analysis of the Data

The matrices were analyzed using the Winsteps© Rasch Analysis software, considering the data as a series of simple dichotomies.

Person measures obtained through analysis of unique pairs are not useful because each person response to each task pair is considered an independent observation and thus, within the matrix, it appears as though each person answered a one-item examination.

Because it is the job and not the individual that we are most interested in understanding the task statistics were our focus. The statistics produced using our model proved to be extremely useful, demonstrating exceptionally high reliability thanks in part to the large number of cases involved when data is formatted for the pairs wise analysis. Model separation was over 8 suggesting that the tasks were spread distinctly across the spectrum of our model ruler.

The two questions presented in the instrument (frequency and importance) were analyzed separately. Both sets of global performance statistics are presented in the figure below.

A clear difference is observable from even a quick glance at frequency and importance. Importance paints a much finer picture, with clearer detail than does frequency.

FREQUENCY PERFORMANCE STATISTICS

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	36.7	73.4	.00	.38	1.00	.0	.97	-.1
S.D.	27.6	36.8	3.65	.19	.02	.2	.09	.3
MAX.	126.0	170.0	6.17	1.01	1.05	.4	1.10	.4
MIN.	1.0	24.0	-8.58	.18	.88	-.8	.65	-1.3

REAL RMSE .43 ADJ.SD 3.62 SEPARATION 8.47 Task RELIABILITY .99
(MODEL RMSE .43 ADJ.SD 3.62 SEPARATION 8.48 Task RELIABILITY .99

IMPORTANCE PERFORMANCE STATISTICS

	RAW SCORE	COUNT	MEASURE	MODEL ERROR	INFIT MNSQ	INFIT ZSTD	OUTFIT MNSQ	OUTFIT ZSTD
MEAN	42.7	79.6	.00	.24	1.00	.0	.97	.0
S.D.	24.4	31.3	3.96	.10	.01	.2	.06	.2
MAX.	134.0	188.0	5.54	.98	1.05	.4	1.10	.3
MIN.	9.0	20.0	-7.21	.22	.90	-.5	.65	-1.0

I REAL RMSE .35 ADJ.SD 3.94 SEPARATION 14.55 Task RELIABILITY .99
(MODEL RMSE .35 ADJ.SD 3.94 SEPARATION 14.88 Task RELIABILITY .99

Separation is greater for importance and error is lessened. We theorize that such a difference is easier to observe in paired comparison analysis than in traditional Likert analysis because of the extremely large sample sizes created through the model. We speculate that this difference has gone largely unnoticed and would have continued to hide from view were it not for the methodological eccentricity of paired comparisons.

It is our hypothesis that the nature of the questions creates the difference. Frequency asks for what "is" - - what exists, what can be observed and counted. Importance asks for what "may be" - that which is created by the construct each respondent has developed relative to his or her professional experience. Importance may lend itself to easier gradation. For instance, frequency labeled as daily-weekly-monthly is fixed and easier to divide into meaningful chunks. It may also be easier for the respondent to answer in extremes (e.g. always or never). Importance on the other hand is less known. It exists only in the creator's mind and philosophy. Participants are likely to be reluctant to label anything as "not at all important" because many (most) may feel that *everything is important* for a profession. An indication of "less frequent" does not preclude content

from an exam, but an indication of lesser importance may seem pejorative to respondents. Additionally, a rating of "less important" may appear to lessen the omniscience of the professional and create an unconscious ego-conflict.

The map of persons and items presented on this page is a yardstick allowing us to understand how physicians in practice view and rate the importance of tasks (content) within their job. The map is by far the easiest way to initially review the content. It demonstrates in a simple way how each of the tasks, within each of the content subcategories works together to complete a snapshot of a career.

By itself, the paired-comparison generated map tells us a detailed story and helps to explain what it means to be a physician. However grand the tale, it lacks a foundation. Clearly task 2U is the least important while 6A is as the most important. But how can we understand what 6A represents on its own, without the aid of its counterpart? How can we quantifiably define 6A? Likert tradition offered its own simple answer.

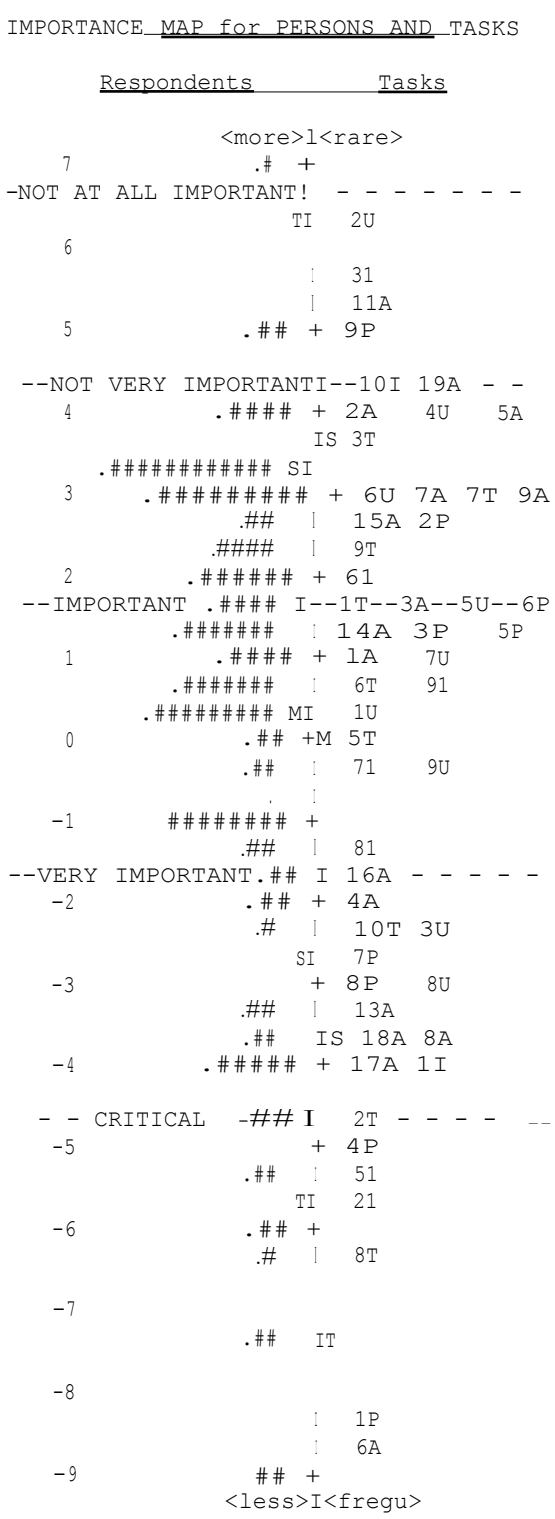
To the original, Winsteps© map (right) we added the key elements defined by the data gathered from the 6 Likert-style items employed. Using the traditional data and the robust Winsteps© program, we were able to place specific, quantitative labels onto our ruler. With labels in place, the picture became much more complete. Our understanding grew immensely simply by the inclusion of these 6 simply questions.

From this point, one of several methodological framework could be employed to finish the task analysis, develop content blueprints, tables of specification, and ultimately link content to practice. There are many well outlined schemes to conduct such an investigation but they are not in the purview of this paper. Our focus has been instead to explore the efficacy of using the paired-comparison item type in the collection of task data. The instrumentation appears to work well and provides ample data from which decisions can be made.

Are There Any Benefits?

Analysis of paired comparison data is far more complicated than using traditional Likert models. Furthermore, in order to allow for concrete, quantification of more and less, standard Likert items must be included on the instrument. What then would be the benefit of such a design?

Task analyses for this same profession had been completed a number of years earlier. At that time, traditional instrumentation was employed. The questionnaires were 15 pages long and took approximately 5-6 hours to complete. The response rate after two mailings was slightly more than 13%. Low response rates pose problems for any survey, and a task analysis in particular.



Understanding the Tasks: Each number represents a specifically defined task, while each letter (A, P, T, I and U) represents one of the 5 subcontent areas.

Our instrument design included 6 (very full!) pages containing both paired and Likert questions. On our demographic survey, we asked respondents to indicate the amount of time they spent in answering the questions. The average time spent was 1 hour and 50 minutes. Time is of the essence in questionnaire design, and the paired/Likert hybrid managed to reduce the average completion time by at least 270%.

That reduction in completion time itself was fairly impressive, but our next observation was even more important. As with many small organizations, volunteers were used as respondents. Without an incentive, volunteer participants easily baulk at taking half of their day to complete such a project. Our survey, requiring a much shorter time to complete, was returned by 82% of our sample. We believe the unexpectedly high response rate was directly connected to the simplicity of design and shortness of required time commitment. It is our contention that a response rate from volunteers worthy of paid participants is certainly worth the added effort to analyze the end product.

We have also demonstrated the effectiveness and statistical relevance of the methodology. The benchmarks of performance, including reliability, error and separation lend credence to this practice long established in the world of statistics but so often overlooked in measurement circles.

Want high reliability, clear data spread and an 82% response rate? Next time, consider the use of a paired-comparison design to complete your task analysis. The paired instrument is efficient and worthy of consideration regardless of the models employed to link content with the examination.

Acknowledgements

This paper was made possible by a generous grant from:

MetriKs Consulting Limited, LLC.
Making Measurement Meaningful tm

and with research support from the:

FLORIDA DEPARTMENT OF HEALTH