The Origin and Development of Rating Scales

Michael C. Rodriguez

Educational Psychology
University of Minnesota

January 2014

Revised August 2016

The Origin and Development of Rating Scales

Rating scales are everywhere. They appear on surveys, applications for athletic club memberships, student evaluation of teaching forms at colleges and universities, product surveys following online purchases, the automated telephone satisfaction question following a conversation with an airline customer service representative, and even in the emergency room: "On a scale of one to ten, where one is no pain…." We use rating scales informally on a regular basis. Consider the options to the question: "How would you like your steak?" Under most circumstances, options range from Rare to Well Done. Consider the simple act of judging how you look before facing the world in the morning: "Not Presentable–Acceptable–Hot!"

When entered into Google, "rating scale" yielded over 1.6 million results. When entered into Google Scholar, 1.3 million results were found. Searching the University of Minnesota online library of articles, "rating scale" yielded nearly 200,000 results. In *PsychInfo*, the keyword "rating scale" yielded 17,725 results and as part of the title, 2,227 results. It is a ubiquitous topic and method.

Although they are commonplace in every field and virtually every aspect of daily life, from education to employment settings, to retail and health care, there are a great deal of misconceptions about the origin and use of rating scales. Consider for example the definition provided by Wikipedia:

> *A rating scale is a set of categories designed to elicit information about a quantitative or a qualitative attribute. In the social sciences, common examples are the Likert scale and 1-10 rating scales in which a person selects the number which is considered to reflect the perceived quality of a product.* (http://en.wikipedia.org/wiki/Rating_scale)

Notice this suggests that the target of a rating scale is an attribute, limiting its function. It also suggests that attributes can be qualitative or quantitative, characteristics typically applied to the measure of the attribute rather than the attribute itself. Also, perhaps the most common misnomer involved in the rating scale dialogue, there is the reference to the so-called *Likert scale*.

The voluminous literature on the origin and early uses of rating scales is wide and deep. This essay provides a glimpse at the earliest contributions to that body of work. The history of rating scales is traced, starting with the work of scientists of the mid-1800s, the formal introduction of the rating scale in 1910 and the graphic rating method in the 1920s, a sample of early research on improving rating scale methods, the application of rating scales to the measurement of attitudes, and an in-depth look at the contributions of Rensis Likert. Although the goal was to be comprehensive, this review of the origin and development of rating scales is certainly not exhaustive.

Predecessors of the Rating Scale

Francis Galton has been attributed with the honor of being the first psychometrician (Ludlow, 1998). Galton (1879) conducted "psychometric experiments" where he employed the emerging methods of psychometry, "the art of imposing measurement and numbers upon operations of the mind" (p. 149). Galton likely was reflecting on the work of German psychophysicists working on the science of psychometry in the mid-to-late 1800s (Ludlow). Ludlow reported that the term psychometrics first appeared in the work of J.R. Buchanan, who was investigating psychological properties of persons. This was in opposition to craniology (or craniometry, associated with anthropometry), the approach of measuring cranium features to make inferences about intelligence, temperament, and other human characteristics.

J.R. Buchanan (1854) lectured on psychometry in the 1840s, as published in his lecture outlines (not the contemporary psychometrics of psychology and education). He lectured on the role of psychometry in the investigation of the neurological system, the placement of professionals, arranging marriages, the functioning of the brain, and in selecting and forming friendships.

> *The Psychometry, or mind-measuring of the Gallian system was merely a rude system of craniology, sketching boldly and roughly the profile of a character appropriate to the skull, which the individual often failed to realize practically from the want of full and systematic mental cultivation. The Psychometry of the Neurological system determines the actual power of the organs by the impression which they give of their vital energy to an impressible and intuitive person, Hence the new Psychometry differs from the old Cranoscopic sketching as much as a cast … of the face differs from a penciled profile. Our Psychometry has also the advantage that it is entirely independent of the cranium, and applies with as much facility to the absent, the dead, or the ancient, as to the present.* (pp. 86-87)
>
> *The influence of Psychometry will be highly valuable, also, in the selection from candidates for appointments to important offices, and in the judicious arrangement of matrimonial unions.* (p. 125)
>
> *The analytical view of the brain derived from experimental psychometric investigation is no doubt the true scientific view of man. But those who look at this analysis need to have sufficient synthetic power in their own minds to conceive the separate organs as parts of a harmonious whole.* (p. 274)
>
> *In the formation of friendships, our natural Psychometric capacity is generally sufficient to enable us to choose a suitable friend. Still there are many errors in the selection of friends, and many vague painful doubts of their character, from which Neurology might relieve us.* (Appendix, p. 10)

In his 1879 work, Galton conducted free-association experiments, investigating the rate at which ideas formed, the frequency of repeated associations, and other characteristics of associations and persons. In his writings during this time, he contemplated ideas that resemble our notions of classical test theory, reliability, and even validity (Ludlow, 1998).

As an example of his contributions to measurement, in 1883, Galton (2004, reprint) asked his participants to describe their mental representations of objects using a 5-point scale of very faint, faint, fair, good, or vivid. Galton also published a 9-point scale for rating the clearness of one's mental imagery, for example, participant recall imagery of their breakfast table. The nine levels included (Galton, pp. 64-65; actual labels with adapted descriptions):

1. Highest. Brilliant, distinct, never blotchy.
2. First suboctile. The image once seen is perfectly clear and bright.
3. First octile. I can see the object as well in all particulars as if it was before me.
4. First quartile. Fairly clear and fairly represented. Well defined.
5. Middlemost. Fairly clear. Brightness at least from one-half to two-thirds of the original.
6. Last quartile. Dim, certainly not comparable to the actual scene.
7. Last octile. Dim and not comparable in brightness to the real scene.
8. Last suboctile. Rarely able to recall the object with distinctness.
9. Lowest. Almost no association of memory with objective visual impressions.

In a similar approach, Karl Pearson (1906), Galton's protégé, published a 6-point rating scale to rate mental ability in his investigations of correlates of intelligence, including the classification categories of (1) very dull, (2) slow-dull, (3) slow, (4) slow-intelligent, (5) intelligent, and (6) quick intelligent. He also used a 7-point scale dividing "quick intelligent" into two categories, including the following definitions (p. 107; actual labels with adapted descriptions):

1. Very dull. A mind capable of holding only the simplest facts, incapable of reasoning about relationships between facts.
2. Slow dull. A mind capable of perceiving relationships between facts with continuous effort.
3. Slow. A mind advancing in general, but very slowly, with time and effort.
4. Slow intelligent. A mind slow generally, possibly more rapid in some fields.
5. Intelligent. A mind ready to grasp and capable of perceiving facts in most fields.
6. Capable. A mind less likely than the specially able to originate inquiry, quick in perception.
7. Specially able. A mind especially bright and quick both in perception and reasoning, including regarding novel facts.

As a side note, from these studies, Pearson (1906) argued that the results of his work provided clear and consistent evidence of no association between intelligence and external physical measurements, particularly with regard to the size and shape of one's head.

By 1910, E.L. Thorndike was developing his scale for measuring the quality of handwriting. He claimed that "any measurement of the quality of handwriting may be made more accurately and conveniently with the scale, either actually present or held in memory, than without it" (Thorndike, 1910, p. 128). The unit of measurement for the scale equaled approximately one-tenth of the difference between the best and worst of a sample of formal writing products from 1000 children in grades five to eight. This scale was purported to possess a ratio-level of measurement, as a score of zero was deemed to represent hand-writing of absolutely no merit, and where the difference between a score of 5 and 6 was equal to the difference between a score of 12 and 13, and a score of 16 was twice the score of 8, and so on.

Thorndike also provided some criterion-referenced interpretations such that the worst quality of writing observable from fourth-grade children was at a score of 5, whereas the best writing quality of eighth-grade children was a score of 17. Quality 7 was "nearly the worst writing of fifth-grade children" (p. 89). The scale provided a book of model samples of each level of quality rated by competent judges, so that it "extends from a quality, better than which no pupil is expected to produce [score of 18], down to a quality so bad [score of 4] as to be intolerable, and probably almost never found, in school practice in the grammar grades" (p. 89).

Thorndike's (1910) writing quality scale provided a guide for rating the quality of handwriting (cursive writing), where the rater compared a sample of writing to the examples in the scale, each with an assigned level of quality from 4 to 18. And, in completing the ratings of writing quality, Thorndike recommended the use of several samples of writing, each rated independently by multiple judges. In classic Thorndike style, he concluded his presentation of the handwriting scale by arguing:

> *The entire history of the judgments of the merit of handwriting supports that claim that if a number of facts are known to vary in the amount of any thing which can be thought of, they can be measured in respect to it. Otherwise, I may add, we would not know that they varied in it. Wherever we now properly use any comparative, we can by ingenuity learn to use defined points on a scale.* (p. 69)

Among the many scales developed by Thorndike (1913) was a measure of the merit of drawings by children 8 to 15 years old. The scale included sample drawings exemplifying 14 levels of merit from zero to 17 degrees. As in the scale of measurement for handwriting, zero merit indicated the absence of merit or failure to represent the intended object: failure to inform, portray, or please.

This scale was also further developed to achieve more uniformity in drawing compositions represented by different units of the scale (Childs, 1915). This effort was to support the use of measures of achievement in drawing in the Indiana city schools, but also to respond to a limitation described by Thorndike in that the scale he developed did not provide for accurate comparisons of all types of drawing, for example those including human figures, animals, landscapes, designs, and other characteristics. It is also interesting to note the many purposes put forth for the measurement of drawing achievement in schools, including (a) to establish norms of ability for each grade and (b) to determine growth in ability from grade to grade.

The Formal Introduction of the Rating Scale

E.C. Elliott (University of Wisconsin) is attributed with the first formal suggestion of a scheme for rating individuals in 1910. His *Scorecard for Measuring the Merit of Teachers* employed numerous traits, each of which was assigned value or credit, so that their sum reached 100 points (Rugg, 1921a). Although several researchers have written about the Elliott scales, they were not recoverable in an original form published by Elliott.

Crawshaw (1916) published a version of the Elliott rating scale, in explaining his modification of the scale so that it more closely applied to teachers of industrial arts. The Elliott scale he published was the *Provisional Plan for the Measure of Merit of Teachers* (see Figure 1). The ratings were determined by assigning and subtracting levels of deficiencies from suggested values (weights) for each characteristic, resulting in determined values that were summed across items.

General Instructions.

Deduct from possible 10; very slight, 2; slight, 4; marked, 6; very marked, 7; extreme, 8. (Possible 20, 40, 60, 80, or 100, in same proportion.)

Total efficiency = Total Individual Efficiency *plus* Total Directed Efficiency.

Minimum standard for approval; according to the standards and exigencies of the school or school system.

Individual Efficiency—800 units

	Suggested values	Deficiencies	Determined values
I. PHYSICAL EFFICIENCY—80 units	(80)		
1. Impressions—general	10	……..	……..
2. Health—general	20	……..	……..
3. Voice	20	……..	……..
4. Habits—personal	10	……..	……..
4. Energy and endurance; power of relaxation	20	……..	……..
II. MORAL—NATIVE EFFICIENCY—100 units	(100)		
1. Self-control	20	……..	……..
2. Optimism—enthusiasm	20	……..	……..
…			

Figure 1. A portion of the Elliott rating scale of teacher merit, as formatted in the source. (Source: Crawshaw, 1916)

Monroe and Clark (1924) reviewed the state-of-the-art in measuring teaching quality (efficiency), reporting that prior to 1910, teaching quality was measured based on a "general impression method" (p. 3). Prior to the 1920s, at least three methods were proposed, including a score-card method, the man-to-man comparison scale (described below), and the use of standardized test scores of the teacher's students (still debated today). They referenced over a dozen researchers that worked to identify the "essential traits or characteristics of successful teachers" (p. 3) beginning in 1905. They identified the work of Elliott and his 1910 scorecard based on 42 essential traits for successful teaching, as the first attempt to measure teacher quality.

Elliot assigned a maximum number of points for perfection in each trait. The rater was to deduct points for deficiencies, based on the extent of deficiency. For example, points deducted for specific levels of deficiencies could be 2, very slight; 4, slight; 6, marked; 7, very marked; 8, extreme (as seen in Figure 1).

As a side note, on the issue of measuring teacher quality, Monroe and Clark (1924) argued for the inclusion of student achievement in such measures:

> *A teacher's academic and professional training, experience, intelligence, personal or social qualities, interest in teaching, and other traits are merely means to an end, namely, the engendering of achievements in school children. Thus the measure of a teacher's efficiency should be based upon the achievements which he engenders.* (p. 14)

In 1911, Rugg (1921a) participated in the training of school administrators to rate teachers using the Elliott scale. After several observers rated ten teachers, correlations among raters rarely exceeded .20. Rugg suggested that the arbitrary weights on each trait and the lack of an external standard resulted in little to no validity to support score interpretation or use.

Researchers appeared to be very interested in the proposed rating scales of Elliott, but were also quick to offer modifications and improvements.

Early Advances in Rating Scales

By 1914, Witham was among a growing group of researchers arguing for the use of rating methods to support "the reduction of guess work in the important function of rating teacher and schools" (p. 267). Witham then presented a 46 item rating scale, each with three levels of teacher knowledge, skills, or abilities. The three level descriptors changed across the items as each was tailored for the specific item. Several examples are provided in Figure 2. Notice items #26 and #27, two among 16 items covering ability to teach school subject areas (including two blanks where the supervisor could write in unique or more specific subjects). These included reading, writing, spelling, geography, drawing, nature, history, arithmetic, and others.

Witham argued that teacher scores could be summed within a teacher and averaged across teachers to estimate the efficiency of the school. Weights were assigned to each of six broad areas, summed, and divided by three to estimate overall efficiency. Weights were a function of importance assigned to each of the six areas determined by a survey of superintendents. Depending on the area, the + scores were assigned full weight (15, 30, or 180 points, depending on the importance of the area), the *a* values were assigned ⅔ of the + weights, and the – values were assigned ⅓ of the + weights.

1. Morals	+ Uplifting influence on others. *a* Upright but not influential. – Questionable Character.
2. Leadership	+ Among students and in community. *a* Among students only. – Lacking.
3. Personality	+ Magnetic. *a* Not magnetic but able to command respect and attention. – Too quiet or too talkative.
26. Reading	+ Knows and applies the best special methods. *a* Shows good natural ability; is weak on special methods. – Shows no methodology and not much natural ability.
27. Writing	+ Knows and applies the best special methods. *a* Shows good natural ability; is weak on special methods. – Shows no methodology and not much natural ability.

Figure 2. Example items from Witham's *Measuring Scale for Teacher Measurement*. (Source: Witham, 1914)

Johnston (1917) introduced a teaching evaluation scale including ten qualities of teaching efficiency. Johnston referenced the work of Elliott and others in creating this tool (Figure 3).

A number of other uses of rating scales were described by researchers at this time. Terman (1915) investigated the mental hygiene of exceptional children (both gifted and cognitively disabled). He sought to compare the Stanford-Binet scores with three other indicators of school success, including teacher ratings of the quality of a student's school work using a 5-point scale (very inferior, inferior, average, superior, very superior), teacher estimates of student intelligence also rated on a 5-point scale, and grade progress.

Instructions:
Score each of your elementary teachers in each of the items listed below. Items A and B are to be scored first for each teacher, on the basis of the knowledge that you already have of the teacher and her work.
…
The ratings are to be made on a scale of 1 to 10. "1" will be understood as exceptionally good, "10" as very poor, and the other numbers as intermediate degrees. Please write the figures representing your ratings plainly in the rectangles opposite each item for each teacher.

	Teacher 1	Teacher 2	Teacher 3	Teacher 4	Teacher 5
A. Estimate of the *total* efficiency of the teacher (social, moral, educational, etc.) in her relations to the school, the community, etc.	….	….	….	….	….
B. General estimate of the teacher's actual *teaching ability.*	….	….	….	….	….
C. Specific items rated on basis of class (recitation):					
1. Speech. (Modulation and quality of voice and rate and enunciation of speech).	….	….	….	….	….
2. Governing skill. (Are the pupils serious or flippant, natural or constrained?)	….	….	….	….	….
3. Use of English (by teacher and pupils).	….	….	….	….	….
4. Teacher's skill in the organization of material of the recitation.	….	….	….	….	….
…					

Figure 3. A portion of Johnston's *Test of Teaching Efficiency*. (Source: Johnston, 1917)

At this point in time, W.D. Scott, a professor of psychology at Northwestern University, was working on methods to select employees. He criticized existing methods of employee selection and promotion as absurd, "based upon an inadequate estimation of technical ability of the applicant" (Scott, 1916, p. 182). He argued that the methods of the first approach to making selection more scientific was a form letter requesting the previous employer to rate the candidate regarding characteristics of services provided (see Figure 4).

Please place a check mark in the space below that indicates the character of his service:

	Good	Fair	Unsatisfactory
Work	( )	( )	( )
Conduct	( )	( )	( )
Ability	( )	( )	( )
Character	( )	( )	( )

Figure 4. Character rating items from the Scott form letter of recommendation. (Source: Scott, 1916).

Scott (1915) argued that the method of measuring mental age introduced by Binet and Simon "rendered a great service to mankind" (p. 94). Scott (1916) developed a series of tests to inform selection and promotion, including Physical Condition (based on physician report), Native Intellectual Ability (based on a series of mental tests), and Technical Ability (another series of tests depending on the technical requirement of the job, which was sales for the purpose of this article).

Miner (1917) needed to support the employment office at the Carnegie Institute of Technology in Pittsburg, Pennsylvania, in their system of recommending students and graduates for employment in a variety of positions. He found that he could not use the Thorndike approach with rankings because it required an individual to be ranked by the same several judges as the individuals adjacent to them in the final order (Thorndike, 1916, offered a solution to this problem, although it was complicated). He needed a method to estimate relevant traits of students completing their studies at the Carnegie Institute.

Miner presented a method of estimating abilities for practical decisions, including employment, admission, and promotion, in classrooms, retail settings, factories, or business offices. His particular purpose was to meet the need for recommending graduates of the Carnegie Institute of Technology for employment in a wide range of occupations, as a way to supplement transcripts. The method he employed consisted of rating each person by placing a dot on a line which was divided into five categories. He attributed this approach to work being done at Teachers College, Columbia University. The Teachers College approach was to use five levels of ratings (superlative, excellent, satisfactory, fair, poor). He argued for the use of a dot on a line to avoid the qualitative significance attributed to the labels, a method which would come to be known as graphic rating (see Figure 5).

Instructions:

Will you please rate the student named below for the traits indicated. Place a dot along the line after each trait, grading the student as finely as you care to. Please give the rating independently without consulting others.

Among members of *the average senior class in this student's course and school* the student would rank in the

	Lowest 5th	Fourth 5th	Middle 5th, Average		Second 5th	Highest 5th
Common sense						
Energy						
Initiative						
…						

Figure 5. A sample student rating sheet. (Source: Miner, 1917)

Miner (1917) argued that this approach accomplished the following advancements: (a) a normative group presents a clear standard for rating relative skill level of each member, (b) ambiguous qualitative labels are avoided, (c) although 5 categories are generally employed, judges can make discriminations as fine as they wish given the underlying continuum, and (d) units of measurement can be measured as finely as desired and easily transformed into standard deviations.

It is important to note that Miner (1917) found that judgments of the order of students regarding merit remained the same whether the scores were based on fifths as compared to measurements made in tenths or millimeters. He argued that if an order of merit is all that is required, scores based on simple fifths were adequate. He also argued that at least two judgments should be included to secure reasonable reliability of results.

By this time, measurement of person characteristics advanced to serve large-scale purposes. Achilles and Achilles (1917) accepted the challenge to identify military personnel for specific roles needed to support the war efforts. They took advantage of a list of "Qualities for Rating of Executive Ability" developed by Gowin of New York University in the context of successful business executives. They employed the list of qualities and asked army officers to rank order each quality from 1 to 14: "Consider that every man probably possesses a certain degree of each of the qualities, and rank them according to the desirability of their predominance in any given man for his success in military life" (Achilles & Achilles, p. 306). They found a high degree of agreement in rankings from officer candidates (with little military experience) and officers.

However, more importantly, Achilles and Achilles (1917) found that at least seven qualities were ranked similarly, not allowing for the desired level of discrimination. They argued that the task was too abstract, that the qualities being ranked were not clearly defined. They suggested that similar qualities are needed in many professions, and the target of the qualities, *success in military life*, could lead to many interpretations. Some officers also suggested that the most important qualities for military success were not on the list (e.g., leadership, ability to judge men).

What Achilles and Achilles did next approached the design of a rating scale. They reported to use a method introduced by Miner (1917) to estimate abilities of personnel. They asked members of a platoon to rate each other and themselves, on the qualities studied above, where each member of the platoon was rated in terms of being in the lowest 5th, the fourth 5th, middle 5th, second 5th, or highest 5th of the company. Each platoon member rated each of the other members on each quality with respect to their normative standing in the company. Scores were assigned to each normative category with the lowest 5th being scored 5 and the highest 5th being scored 1 – the smaller the score, the more favorable the rating (see Figure 6).

Instructions:

Please rate the candidate named above for the traits indicated, keeping in mind employment in military service. Give the rating independently without consulting others.

Among the members of the company the candidate would rank in which fifth? Indicate the position in each trait by placing a dot along the line grading the candidate as finely as you can.

	Lowest 5th	Fourth 5th	Middle 5th, Av		Second 5th	Highest 5th
Judgment						
Initiative						
Aggressiveness						
…						

Figure 6. Military platoon member rating form. (Source: Achilles & Achilles, 1917)

"True efficiency in war, as in industry, consists largely in getting men into the right places—in assigning them to those positions where each can serve with greatest effectiveness" (Strong, 1918, p. 130). The Committee on Classification of Personnel in the Army, established in 1917, was charged with the task of efficient and effective placement of recruits, directed by Walter Dill Scott,

Director of the Bureau of Salesmanship Research at Carnegie Institute of Technology (Strong). Among the other members of the committee were E.L. Thorndike and members of the Army Alpha team, including Bingham, Yerkes, and Terman. The committee developed a method for rating officers and candidates for commission.

Kelly (1919) summarized the characteristics of ratings to maximize reliability of these efforts, including:

1. An average of several ratings is more reliable than a single rating and less susceptible to personal bias;
2. An average of independent ratings is better than a consensus of opinions;
3. Raters must understand the daily lives of those being rated; and
4. Ratings from different judges must be on the same scale.

He argued that the test of classification accuracy would come from a high correlation between the rating scores and a measure of performance (i.e., criterion-related validity evidence).

The So-Called Man-to-Man Rating Method

Thorndike (1920) referenced a 1915 study of employees of the General Electric Company and Westinghouse Electric Company – a rating study of relevant traits of success. He noted the possibility of a halo effect creating a high degree of correlation among the independent ratings of multiple traits. He provided a method of creating a rating scale where the anchor points along the rating scale included the names of men (officers). This method became known as the man-to-man comparison scale, which Rugg (1921a) claimed moved rating methods toward an objective science.

Creating the scale required the identification of the officer with the highest level of a given quality (such as leadership), the lowest level of the quality, and one about half way between the highest and lowest – to mark the low, middle, and high points of the scale; then to identify the officer who is half way between the middle and the highest and between the middle and the lowest, so that there are five names demarcating the lowest, low, middle, high, and highest levels of the quality. To use the scale, a given officer is rated based on a comparison with the anchor-named men, and given a score based on the name of the closest officer; if the rated officer is equally between two points, he is given half a point.

This method was used to create the *Army Rating Scale* (see Figure 7; Rugg, 1921a), which through extensive use in 1917-1918, engendered serious questions about score reliability and validity of use. Rugg reported the findings from research conducted during wartime on the quality of the rating scale. Early in this process, evidence suggested that the scale was not being used properly. Subsequent studies under more controlled conditions also found

difficulty in securing reliable estimates of character from rating scales (Rugg, 1921b). Rugg (1921b) suggested that among the reasons why independent ratings varied so much were (a) lack of acquaintance with the individual being rated, (b) tendencies to rate high or low, (c) and the complexities of the characteristics being rated and their importance and relevance in the eyes of the rater.

Additional evidence of rating scale qualities, limitations, and potential was presented by Rugg (1922a, 1922b) in a series of experimental studies. Rugg (1922b) offered several recommendations to enhance the quality of rating scale scores: (a) use the average judgments of several competent judges; (b) use a scale that is as practical as the intended use – in educational settings, this is often a simple diagnostic rating of deficient, absence of trait, or mediocre levels (not a 5 or 7 category scale that is too refined); (c) secure objective ratings of important relevant characteristics (social and dynamic traits).

I. PHYSICAL QUALITIES Physique, bearing, neatness, voice, energy, endurance. Consider how he impresses his command in these respects.	Highest ..... 15 High ..... 12 Middle ..... 9 Low ..... 6 Lowest 3
I. INTELLIGENCE Accuracy, ease in learning; ability to grasp quickly the point of view of commanding officer, to issue clear and intelligent orders, to estimate a new situation, and to arrive at a sensible decision in a crisis.	Highest ..... 15 High ..... 12 Middle ..... 9 Low ..... 6 Lowest 3
II. LEADERSHIP Initiative, force, self reliance, decisiveness, tact, ability to inspire men and to command their obedience, loyalty and cooperation.	Highest ..... 15 High ..... 12 Middle ..... 9 Low ..... 6 Lowest 3
...	

Figure 7. Portions of *The Army Rating Scale* (Source: Rugg, 1921a). Names of men were placed in the blanks associated with the Highest to Lowest categories.

Characteristics of the rating scale method were described by Paterson and Ruml (1920), researchers at the Scott Company, to counter critics of the summated rating scale method as though it were attempting to add apples and oranges, that personal characteristics are sufficiently different such that ratings of them should not be summed together. They argued that the rating scale

secures a numerical measure of disparate qualities of which is correlated with general value in a particular line of work. The measures then are measures of varying reliability of the individual's general value. These several measures, each inferentially diagnostic of general value, can logically be summated. (p. 80)

Graphic Rating Methods

By 1920, the graphic rating method was introduced by the Scott Company (Freyd, 1923). Two contributions were promoted in support of this method (Hayes & Patterson, 1921), including freedom from quantitative limits in rating individuals and the allowance of any level of discrimination in those ratings. Hayes and Patterson reported on the experimental development of the graphic rating method with professionals in several fields, finding the ratings to yield high correlations between multiple judges and high inter-rater reliabilities over time and intra-rater reliabilities over time, where all correlations tended to be greater than .65.

In exploring the state-of-the-art of graphic rating scales, Freyd (1923) reviewed existing rating scales. He described the rating scale used by Downey, which scored the reaction in Resistance to Opposition Test, which required test takers to write their names with eyes shut, while the test administrator placed an obstruction under the pen requiring the test taker to exert pressure to continue writing. Downey used an 11-point scale, which he called the decile scale, to describe the reaction in resistance with the following labels (Freyd, p. 85):

10. Strong pressure against obstacle…
9. Very strong counter-pressure on level…
8. Very rapid and energetic dodging…
7. Very deliberate but gentle counter-pressure…
6. Evasive reaction…
5. Very mild counter-pressure….
4. Strong pressure AFTER URGING and READJUSTMENT…
3. Moderate pressure after urging…
2. Moderate counter-pressure after urging…
1. Feeble pressure after urging…
0. Absolute passivity in spite of urging…

Another scale in use at the time was Plant's scale for rating attention. In this example, it is a 10-point scale for use by nurses in psychiatric hospitals to rate the attention of their patients (Freyd, 1923, p. 86):

1. Stuporous.
2. Can't hold attention long enough to do even commonest things such as completely dressing self or eating a meal.

3. Dresses self but can't hold attention long enough to do any particular work.
4. Can do only childish pieces of work. Cannot fit a picture puzzle of more than 15 or 20 pieces.
5. Can do only childish pieces of work if they are new. Will do very long and complicated pieces of work along lines he has been working on—as picture puzzles.
6. Can sew for half an hour or so. With the men—those who can play a game of checkers or billiards but does nothing requiring a longer time. Leaves task half finished—to take up some other task.
7. Remains interested in a piece of work until the end of the day, but next morning has forgotten it or has no interest in it.
8. Will work for a day, or day and a half, on a piece of work, and finish it.
9. Often stops, even for days, in a task requiring a long time but goes back to it over and over again until it is finished.
10. Plans and carries out a piece of work requiring a long period of time, as weaving a rug or making a piece of pottery.

Freyd (1923) presented several examples to argue that to obtain the most accurate and reliable ratings, the methods used to obtain the ratings must be refined. He did this to evaluate the quality of graphic rating scale methods, stating that "there are innumerable possibilities in the way of methods of rating" (p. 88). He described 11 methods in use to secure ratings of individuals. The graphic rating method was one where "the rating is indicated by a check along a straight line, under which are printed descriptive phrases indicative of varying degrees of the trait, from one extreme to the other" (p. 88). The two features of this method, the use of a line on which a rating is drawn and the use of descriptive terms, had both been in use prior to 1920. An example of this is a scale developed by Freyd (Figure 8).

Instructions for using the rating scale

1. Let these ratings represent your own judgments. Please do not consult anyone in making them.
2. In rating this person on a particular trait, disregard every other trait but that one. Many ratings are rendered valueless because the rater allow* himself to be influenced by a general favorable or unfavorable impression which he has formed of the person.
3. When you have satisfied yourself on the standing of this person in the trait on which you are rating him, place a cheek at the appropriate point on the horizontal line. You do not have to place your check directly above a descriptive phrase. You may place your check at any point on the line.

…

3. Does he appear neat or slovenly in his dress?

……

Extremely neat and clean. Almost a dude	Appropriately and neatly dressed.	Inconspicuous in dress.	Somewhat careless in his dress	Very slovenly and unkempt

9. How does he impress people by his physique and bearing?

……

Looked down on	Unimpressive physique and bearing	Noticeable for good physique and bearing	Excites admiration. Very impressive

Figure 8. Example graphic rating scales developed by Freyd (1923).

The Freyd graphic rating scales were scored using a stencil that was placed beneath the line. The stencil was divided into 20 equal intervals, numbered from 1 to 20. The score given to the rating was based on the number of the space in which the check was made on the line. It was interesting to note that if two or more checks were made on the line, perhaps indicating uncertainty, the average of the ratings (midpoint) was used as the rating score.

Freyd (1923) argued that there were several advantages to the graphic rating method, including:

1. Simple to understand,
2. Requires little motivation to complete,
3. Can be completed quickly,
4. Can be easily scored,
5. Eliminates interpretation of direct quantitative terms, and
6. Discriminations can be made as finely as desired.

He noted that the number of rating scale points could vary given the particular use and need for discrimination by the users, including scores for example

from 1 to 5 or from 1 to 100. More importantly, the method provides for comparable ratings without the requirement that every rater knows every member of the group being rated (as in the Man-to-Man Rating method).

From a series of experiments employing graphic rating scales, Freyd (1923) offered guidelines for the design of the scales:

1. Define the trait to be rated, recognizing that what we often wish to rate is composed of several separate traits, and use behavioral features to make the definitions concrete;
2. Determine the extremes of the trait;
3. Pose a question to introduce the rating task;
4. Use a line long enough to support the use of a stencil for scoring, but not longer than five inches;
5. Use a single continuous line with no breaks or divisions;
6. Use three to five descriptive labels along the continuum of the line;
7. The extreme descriptive labels should not be so extreme as to make them implausible;
8. The descriptive label associated with the neutral or average position should be located in the center of the scale;
9. If there are five descriptive labels, the intermediate ones (values 2 and 4) should be closer in meaning to the central label than to the extremes;
10. Descriptive labels should be universally understood, avoiding slang;
11. Terms such as average, very, extremely, excellent, good, fair, or poor, should be avoided – with a preference for terms that express varying degrees of a trait (e.g., use fastidious instead of extremely neat, or use slovenly instead of very careless in dress);
12. Descriptive labels should be short and to the point;
13. The labels should be in small print to allow sufficient white space for separation; and
14. Alternate the location of the favorable extreme, or essentially alternate positive/negative orientation of the scales to avoid response sets.

It is very interesting to note that Freyd (1923) suggested that to create spread in the distribution, guideline #9 above should be used – or that another way to accomplish the same result is to make the intervals on the scoring stencil smaller in the center of the scale and wider at the extremes.

Note that this stretching of the scale metric near the extreme values is similar to what happens to number-correct scales when they are scaled through the Rasch model – the nonlinear transformation of number-correct scores stretches the raw-score distribution at the lower and upper extreme values, illustrating the ordinal nature of raw-scores and the idea that it takes more of the trait to move a single unit on the raw-score scale near the extremes.

Perhaps the most salient aspect of these guidelines is that they are, for the most part, represented in the most current guidelines for developing rating scales or more generally, survey design guidelines, with the exceptions of 8, 9, and 14, for which contemporary researchers suggest otherwise (see for example, Haladyna & Rodriguez, 2013).

Evaluating the Rating Scale

In 1921-22, Rugg published a series of articles seriously questioning the meaningfulness, appropriateness, and usefulness of ratings of human characteristics, essentially critiquing the validity of resulting inferences. He asked: "Can human character be 'rated' on point scales accurately enough for practical uses in education?" (1921a, p. 8). He answered affirmatively, if the ratings are done under rigorous conditions. To be rigorous, the rating given to the person should meet these conditions:

1. be the average of three independent ratings on an objectified scale (as in the man-to-man comparison scale),
2. the scales are equivalent and ratings are made by trained raters, and
3. the three raters are well acquainted with the person being rated.

He then argued that these conditions generally are not attainable in public schools. Rugg (1921a, 1921b, 1922a, 1922b) completed a deep analysis of rating scale scores, in terms of reliability and validity, as well as the utility, practicality, and functionality in educational settings. In doing so, he presented several example scales being used during these times (Figures 9 and 10).

	1	2	3	4	5	6	7
	Very inferior	Inferior	Below Average	Average	Above average	Superior	Very superior
I. Physical qualities:							
Physique							
Neatness							
Energy							
…							

Figure 9. A check sheet for rating elements to be summed into a composite score (source: Rugg, 1922b).

I. Skill in Teaching	Low	Aver	High
To what extent:			
Does he know the subject matter of his own and related fields:			
1. In subjects like history, geography, etc., does he make effective use of material outside the text book			
2. Does he relate lessons to material in other fields and use illustrations outside his own subject (e.g., mathematics and science)			
Does he select subject matter effectively for class reading and discussion			
…			

Figure 10. A self-diagnosis and improvement chart (Source: Rugg, 1922b).

Although Rugg (1921a) argued that the *Army Rating Scale* approach would not function well in educational settings, researchers from Teachers College, Columbia University (Chassell, 1924) attempted to employ the method in kindergarten classrooms in 1921, a study that began just before Rugg published his critique. The task was to rate kindergarten children regarding their readiness to be promoted to first grade. Prior to this, the practice was to rank students within their group – which was deemed insufficient for such a decision. The traits to be rated from the *Army Rating Scale* were modified by the school principal to fit the kindergarten context and further modified by teachers (content experts), including the four trait areas of habits of work, participation, cooperation, and responsibility. The scale was fixed so that 25 was assigned to the highest position, 15 to the middle, and 5 to the lowest, so that across four traits, the scale scores ranged from 20 to 100 points.

Overall, the use of the *Army Rating Scale* method in kindergarten settings over the course of three years was successful. Among the more mature kindergarten children, teacher ratings during the kindergarten year were successful in predicting desirable habits and attitudes among the children a year later in first grade (Chassell, 1924).

Around this time, Symonds (1924) challenged the claim that the graphic rating scale allowed for any level of discrimination desired by the rater (as suggested by Freyd, 1923). Symonds investigated the extent to which reliability was related to the coarseness of the rating scale – the number of rating scale points. He suggested that just like a physical measurement scale is limited by the natural limits of eyesight and ability to observe fine distinctions, the rating scale is potentially limited to the extent that judges are able to discriminate among finer levels of a given trait. He argued that the relevant index of ability to discriminate among rating scale points was the coefficient of reliability of

scores. Symonds found that although the rating scale had several advantages, it did not necessarily facilitate finer distinctions in rating a given trait – as a method in itself, it did not contribute to increased reliability. He demonstrated how a test with 100 or more points could result in a reliability of .91, whereas a test of 14 points had a reliability of .90 – suggesting that scales with many more points may not offer more accuracy.

Symonds (1924) argued that the optimal number of intervals for rating scales was seven. He argued that more scale points were not supported given the small increases in reliability that might be achieved. Similarly, rating scales with fewer points suffered from a noticeable loss of reliability.

Improvements of the Rating Scale Method

Rating scale methods were soon compared to ranking of individuals within a known group, which was the more common approach. Some argued that rakings provided for more definitive discrimination among a group of individuals since each was compared to all others, relative discriminations (Symonds, 1925). Symonds found that both rating scale and ranking methods yielded scores of similar reliability. He then presented a typical rating scale (Figure 11) and offered a revised rating scale (Figure 12), to capture the benefits of both rating scale methods and ranking methods.

Instructions:

…

4. Place a cross somewhere on the line running from "very high" to "very low" to indicate this child's standing in each quality. *You may place your cross at any point on the line.* It is not necessary to locate it at any of the division points or above any descriptive phrase.

…

HEALTH—Is he generally healthy and vigorous?

Bad	Poor	Average	Good	Excellent

Leadership—Does he take the lead in school affairs or does he follow others?

Always Follows others	Rather tends to follow	Average	Rather tends to be a leader	Masterly, not easily influenced

Figure 11. Example of a standard rating scale. (Source: Symonds, 1925).

Instructions:

…

4. Place a cross in one of the compartments running from “very high” to “very low” to indicate each child’s standing in the quality.

…

8. Try to let the percentages guide you as to the number of crosses to fill in each compartment.

Trait – Health

Is he generally healthy or vigorous?

Pupil	4% Very Bad	11% Bad	21% Poor	28% Average	21% Good	11% Very Good	4% Excellent
Charles…							
William…							
George…							

Figure 12. Example of a revised rating scale, combining elements of rating and ranking methods. (Source: Symonds, 1925).

In addition, Symonds (1925) investigated the magnitude of the halo effect on ratings – the extent to which ratings on specific traits were influenced by a general impression of the individual being rated. In a series of studies, he found noticeable effects on inter-rater partial-correlations of trait ratings (controlling for overall composite scores) due to the halo effects. For example, he found increases in correlations as much as .25 or more because of the systematic presence of the halo effect across ratings. He offered several reasons for large halo effects, including where the trait was (a) not easily observed; (b) not salient or commonly observed; (c) not clearly defined; (d) based on interaction with others, not simply a personal behavior; and (e) of high moral importance.

A series of studies ensued investigating the optimal number of rating scale points, and methods for estimating rating scale score reliability using rater consistency across items within a scale (internal consistency) versus inter-rater reliability of total scale scores (Furfey, 1926), as well as more general investigations regarding the estimation of rating scale score reliability (Remmers, Shock, & Kelly, 1927). Additional studies followed attempting to improve the rating scale tasks (Stoddard & Ruch, 1926).

A Sample of Early Published Rating Scales

Numerous rating scales began to appear in the mid-1920s and early 1930s. Most of these scales employed various numbers of rating scale points and were scored by summing the points and computing the average rating.

In 1928, Clara Brown, at the University of Minnesota in Minneapolis, published the *Rating Scale for Teachers of Home Economics*. The measure was intended to be used as a tool to measure the teaching ability of home economics teachers. The initial form contained 58 items. In an early version, the category labels of poor, fair, average, good, and superior were used. The headings were later removed to increase objectivity, settling on a 10-point scale with three anchors; the anchors were specific to the item, as seen in Figure 13. The scale was scored by summing the points across the items (1 to 10) and taking the average for each of the three sections, then summing the three sections and dividing by 3 for an overall average. The three sections included (a) Organization of Work, (b) Technique of Teaching, and (c) Personal Qualities and Abilities.

Instructions:

…

Each section includes a list of items which are described on three levels. However, it is possible to make finer distinctions by checking anywhere along the line, thus giving values from 1 to 10.

I. ORGANIZATION OF WORK

1	2	3	4	5	6	7	8	9	10	Score

1. OBJECTIVES – DEFINITIVENESS

Vague	Fairly definite	Very definite

…

5. SEQUENCE OF WORK

Units of work disconnected	Fairly good continuity	Work starts with wholes; New problems grow out of what precedes

II. TECHNIQUE OF TEACHING

1	2	3	4	5	6	7	8	9	10	Score

1. ORDERLINESS OF ROOM

Very disorderly	Fairly orderly	Very orderly

…

8. TEACHER PREPARATION

Poor	Fairly adequate	Thorough

Figure 13. Example items from the *Rating Scale for Teachers*. (Source: Brown, 1928).

Brown suggested that her work was informed by that of Leo Brueckner, who published *Scales for the Rating of Teaching Skill* in 1927. Brueckner (1929) criticized teacher rating scales at the time for being too general and not specific enough to the subject matter in which the teacher worked – he argued that teaching techniques were specific to the subject. He argued that in terms of teacher improvement, curriculum-specific information was more helpful. He also addressed issues related to prejudice, tradition, and attitudes of raters – of particular interest to Brueckner were personal attitudes of the rater regarding the instructional methods used by the teacher being rated (a form of bias). To accomplish the elimination of such attitudes, he introduced the method of Courtis (and his unpublished *Standards of Methods*), who recommended first

classifying the teacher based on the method used in teaching, then to rate the teacher according to the skill in using that method (regardless of the attitude of the rater toward that method). Once classified in terms of teaching method, the teacher could be rated on a scale from 1 (excellent) to 9 (failure).

The *Purdue Rating Scale for Instructors* was published by Brandenburg and Remmers (1928; see also Remmers, 1927). This measure used a 100-point rating scale with three anchors on 10 traits of teaching. A stencil was used for scoring. Scores were computed by summing the total number of points across all items and taking the average. Sample items are found in Figure 14.

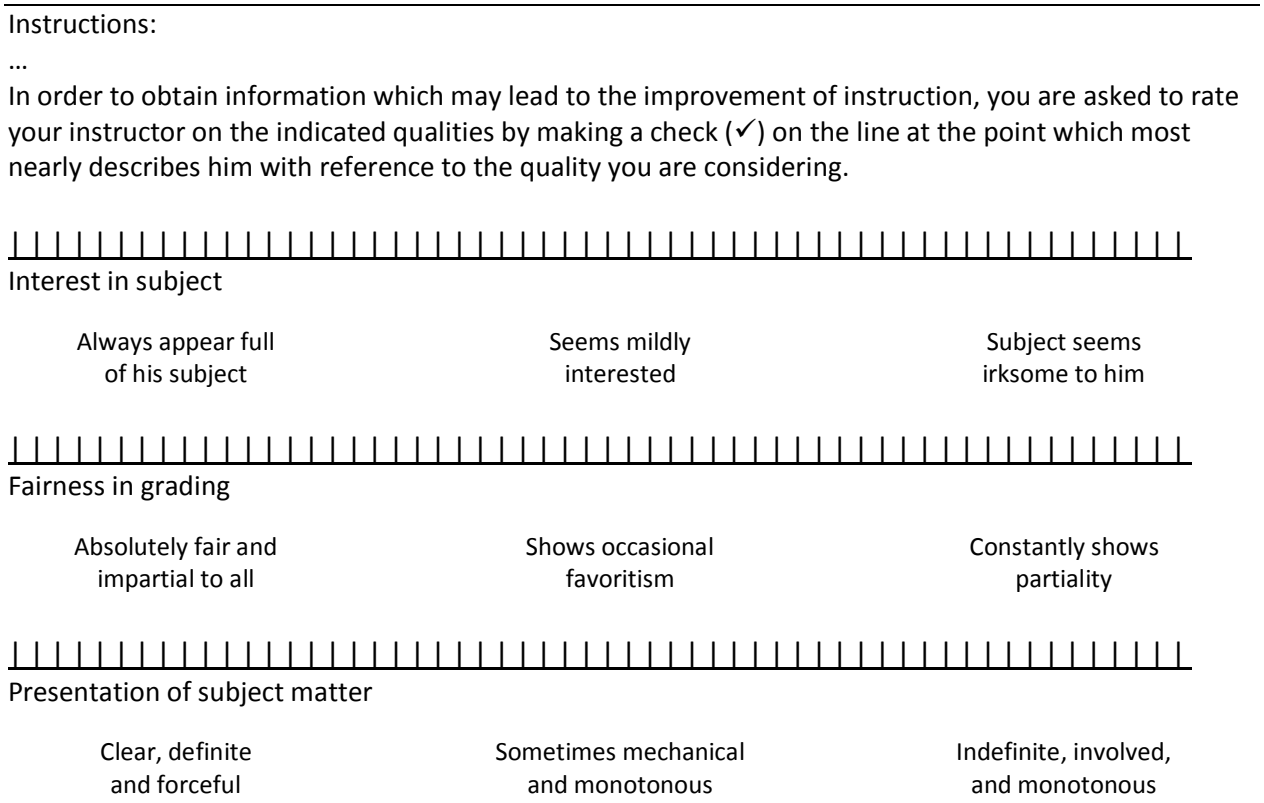

Figure 14. Example items from the *Purdue Rating Scale for Instructors.* (Source: Remmers, 1927)

Baker (1930) published the *Detroit Adjustment Inventory*. This measure asked students to report on their behavior, or in terms of the title of the inventory: *Telling What I Do*. This used a 3-point rating scale. The descriptors associated with each point were tailored to the specific item and were randomly ordered across the items (points from low to high were not in the same order across the items to avoid inspiring response tendencies). The "ideal" response was worth 3 points; a "doubtful" response was worth 2 points; an "undesirable" response was worth 1 point. No answer or selection of all three responses was awarded 0

points. There were two forms appropriate for grades 4 to 6 and 7 to 9. The scale scores were based on the total summed score. A scoring key was provided that assigned scores of 1, 2, or 3 to the responses a, b, or c. Each item was followed by parentheses where the administrator entered a score based on the response choice, as seen in Figure 15.

1. At night			
a. I go to bed late	b. I go to bed early	c. I don't go at all	( )
2. In school			
a. I don't look out of the window	b. I sometimes look out	c. I look out all the time	( )
52. At home			
a. I make much noise	b. I am noisy sometimes	c. I keep very quiet	( )
80. When my folks go away			
a. I am good sometimes	b. I always try to be good	c. I am often bad	( )

Figure 15. Example items from the *Telling What I Do* scale. (Source: Baker, 1930)

By the early 1930s, at least 50 rating scales had been published on teaching and teacher quality alone (Hildreth, 1933). Hildreth listed 50 teachers' rating scales in her bibliography, published between 1914 and 1932. In total, over 3,500 measures were catalogued in her bibliography, among which at least 40 had the term "rating" in the title (not including the teacher rating scales) including psychological and educational measures.

The Measurement of Attitude

In the late 1920s, L.L. Thurstone, through the Behavior Research Fund, Illinois Institute for Juvenile Research in Chicago, accepted the challenge to measure attitude on a linear continuum. He argued that only those attitudes for which individuals could be compared in terms of more or less could be so measured. To do so, individuals could be asked to endorse (or reject) opinions, opinions located at different positions in accordance with the attitude the individual would express. The challenge in measuring attitude in this way was in the definition of the unit of measurement. Thurstone (1928) conceded that an attitude is a complex human characteristic which cannot be entirely described by a single numerical index. He defined attitude as "the sum total of a man's inclinations and feelings, prejudice or bias, preconceived notions, ideas, fears, threats, and convictions about any specified topic… admittedly a subjective and personal affair" (p. 531). He also defined opinion as "a verbal expression of

attitude" (p. 531), stressing the idea that opinion is restricted to verbal expression to the extent that the opinion indicates an attitude. We are not necessarily interested in the specific opinions, but in the extent to which opinions indicate attitude. Opinions are the means for measuring attitude.

Thurstone (1928) discussed many of the challenges in using opinions as indicators of attitude. He noted that any given opinion may not be consistent with an attitude, that opinions and behaviors may not always be consistent, that attitude may change, and that external pressures may affect the expression of attitude (honesty, social pressures). To some extent, these challenges can be addressed in the design of the measure of attitude. First and foremost, the attitude of interest must be clearly defined; it must be restricted sufficiently to be amenable to measurement along a continuum. It must be stated so that we can describe individuals as having more or less of the attitude (e.g., being more strongly in favor of capital punishment, or being more religious).

Thurstone described the goal of attitude measurement resulting in a linear unidimensional scale. Based on the location of several endorsed (or rejected) opinions, the scale location of an individual should convey at least three characteristics: (a) the mean position one occupies on the scale (attitude continuum), (b) the range of opinions the individual is willing to endorse, and (c) the one opinion that most closely represents the individual's attitude. From such a design, four types of inferences are appropriate from an attitude scale, including (a) the average attitude of an individual, (b) the range of opinions one is willing to endorse, (c) the relative popularity of each attitude in the scale for a given group, and (d) the degree of variability in attitudes of the group.

To address the unit of measurement challenge, Thurstone introduced a method of locating statements of opinion on a continuum, where their relative position can be determined, thus providing a metric for the scale (continuum). This is what has come to be known as Thurstone scaling. Briefly the method involves developing a list of statements, perhaps 100 or more, and asking a large group of judges, perhaps several hundred, to arrange them in rank order based on the specific attitude variable – typically in terms of severity or extreme levels of attitude. The proportion of judges who consider a statement to be more or less representative of the attitude, relative to other statements, can be estimated. The psychological scale separation between two statements is measured in terms of these comparative judgments. Because of the practical demands of such an approach, Thurstone offered a compromise procedure – identify 10 to 20 opinion statements, gather a few judges to rank order them, and count the number of endorsements for each statement.

The first measures of attitude were intended to measure attitudes regarding militarism, prohibition, and the church. The method used to create these scales involved the following:

1. Define the attitude variable to be measured.
2. Collect a wide variety of opinion statements related to the attitude variable. Opinion statements can be obtained by having several groups of people write out their opinions on an issue. A literature search can be used to identify statements of opinion. Give special attention to the development of neutral statements.
   a. Statements should be brief.
   b. Statements should be amenable to endorsement or rejection.
   c. Statements should lend themselves to a location on the continuum.
   d. Statements should not be double-barreled.
3. The resulting statements are then written on small cards, one statement per card.
4. Two to three hundred judges arrange the statements in 11 piles, ranging from opinions most strongly affirming an attitude to most strongly negative. There is a middle pile for neutral statements.
5. Calculate the scale value for each statement. The pile location of each statement is then plotted against the proportion of judges that assigned it to each pile. Thorndike used the diagram in Figure 16 to illustrate the principle, but it is simply heuristic. Notice the curve for statement A suggests that no one classified the statement below pile 3, half the judges classified it below pile 6, and none of them classified it above pile 9. Statement A was classified by all judges within piles 3 to 9.
6. Eliminate ambiguous, irrelevant statements. Select about 20 statements for the final scale.
7. The scale value for a given statement is the point on the scale continuum where half of the judges consider it to be located. From Figure 16, we see that statement C has a scale value of 1, B is at 4, A at 6, and E at 10.

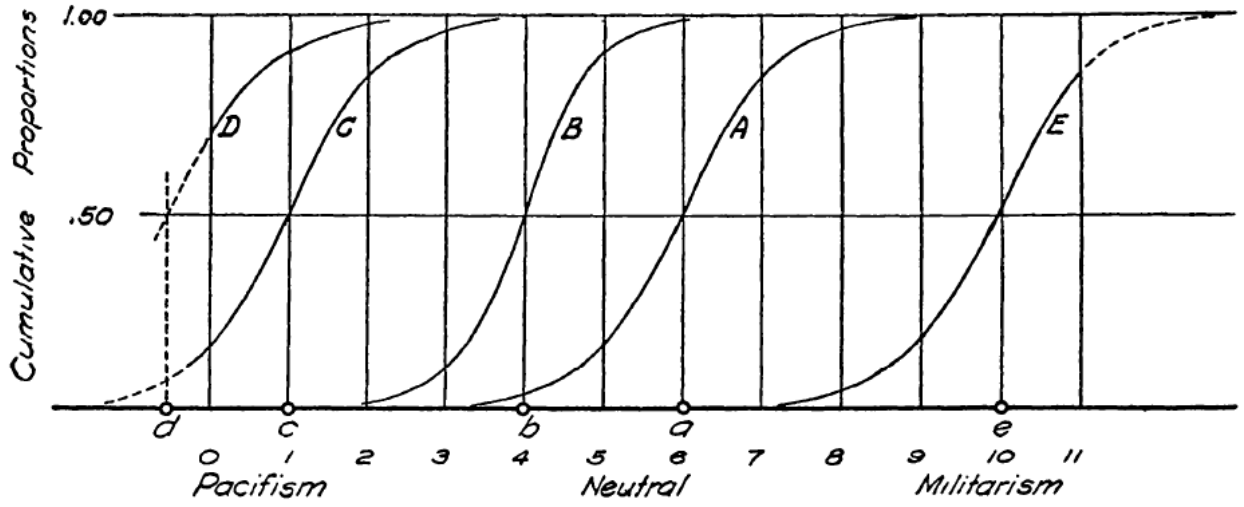

Figure 16. An illustration of Thurstone's principle of scaling statements on a continuum. (Source: Thurstone, 1928)

Normally, scale values would not align perfectly with an integer value as in this heuristic example, but could occur at any value between 1 and 11.

It is interesting to note the similarity of this graph and the contemporary graphs of item characteristic curves in item response theory – the probability of endorsement of an item increases as one's attitude (trait) increases, monotonically. The resulting scale can be designed to maximize the intended qualities – interval levels of measurement, by doing the following (Thurstone, 1928):

1. Select the final set of statements so that they approximate evenly spaced scale values on the continuum.
2. Eliminate statements that create too much dispersion on the continuum (essentially the flat line or nondiscriminating statements).
3. Eliminate statements that can be endorsed or rejected because of factors irrelevant to the attitude variable (e.g., statements that may be offensive to some groups).

The design of measures of attitude using Thurstone scaling relies on a functional rating scale employing 11 points or categories where the extreme categories are defined (extreme values of attitude, positive and negative). This allows the developer to scale each statement, resulting in a scale value for each. The measurement of attitude is then completed by asking respondents to endorse or reject each statement in the measure. The respondent's score is the average scale value of the statements that are endorsed. This essentially places persons on the same scale as the statements of opinion (again, note the resemblance to IRT). Thurstone also noted that the tolerance a person conveys on a particular issue can be estimated by the standard deviation of the scale values of the endorsed statements – an index of indifference perhaps, where the extreme occurs when individuals endorse all statements, locating themselves across the entire continuum of attitude.

In the early 1930s, at least three methods of measurement were prominent in the literature and practice, including (a) the rating scale, (b) the questionnaire, and (c) the objective test. Garrett and Schnick (1933) summarized these three methods. The rating scale methodologies have been introduced above. The questionnaire was defined as a "systematic report of an individual's thoughts, attitudes, or experiences" (p. 122). The two techniques are the same when the rating scale consists of questions of attitudes or facts. At the time, questionnaires had been used by psychologists to assess a person's level of adjustment; in studies of personality, attitudes, and beliefs; to measure interests in such things as books, sports, vocations, social activities; and by sociologists to measure home conditions, occupational status, cultural level, and social environments. Finally, objective tests differed from rating scales and questionnaires in the scoring—test scores were a function of amount completed or time taken to complete. Garrett and Schnick recognized that there were

many cases where such distinctions were not important, that the method of scoring was often a matter of convenience.

The Contributions of Rensis Likert

Rensis Likert received a BA in sociology in 1926 from the University of Michigan, although he began his studies there as a civil engineering student. He then went to Columbia University to obtain a PhD in social psychology, which he received in 1932. He subsequently published his dissertation, *A Technique for the Measurement of Attitudes*, as a monograph in *Archives of Psychology*. His work on the measurement of attitudes began in 1929 with his advisor, Gardner Murphy.

In this work, Murphy and Likert set out to measure common attitudes of interest at that time, including international relations, race relations, economic conflict, political conflict, and religion. In his dissertation work, Likert employed the attitude areas of race relations, international relations, and economic conflict. He compared three scoring methods, the Sigma method (which assumed attitudes were normally distributed and assigned weights based on the observed distribution across the options), Thurstone scoring, and the simpler summed rating-scale method employing the numeric labels assigned to each category. The simpler method achieved reliability comparable to that reported by Thurstone with half the number of items.

The work of Murphy and Likert originally began as an investigation into the extent to which character traits, or attitudes, were a function of specific independent characteristics or related components of a general unified characteristic. Murphy and Likert hypothesized that the attitudes measured in their work, race relations, international relations, and economic conflict, would yield highly specific factors. Their *Survey of Opinions* was administered to over 2000 undergraduate students in nine universities – although their analyses typically involved 650 of these cases. The questionnaire employed four formats, including a Yes-No response (examples 1-2 below), a multiple-choice format (example 3 below), propositions with a rating scale response from strongly approve to strongly disapprove (examples 4-5 below), and a series of newspaper-based scenarios describing conflicts with an outcome to which the students selected a degree of approval as in the previous format (example 6 below). Examples of each are provided in Figure 17.

Regarding Figure 17, the numbers in parentheses associated with each response option were not seen on the original Survey, but are included here to indicate the score value used when employing the simple rating-scale scoring method. Likert did not use numeric labels on the questionnaires administered to college students.

1. Do you favor the early entrance of the United States into the League of Nations?

YES	?	NO
(4)	(3)	(2)

2. Ought the United States to consult other nations in making her immigration laws?

YES	?	NO
(4)	(3)	(2)

3. How much military training should we have?

   (a) We need universal compulsory military training. (1)
   (b) We need Citizens Military Training Camps and Reserve Officers Training Corps, but not universal military training. (2)
   (c) We need some facilities for training reserve officers but not as much as at present. (3)
   (d) We need only such military training as is required to maintain our regular army. (4)
   (e) All military training should be abolished. (5)

4. All men who have the opportunity should enlist in the Citizens Military Training Camps.

Strongly Approve	Approve	Undecided	Disapprove	Strongly Disapprove
(1)	(2)	(3)	(4)	(5)

5. The United States, whether a member or not, should co-operate fully in the humanitarian and economic programs of the League of Nations.

Strongly Approve	Approve	Undecided	Disapprove	Strongly Disapprove
(5)	(4)	(3)	(2)	(1)

6. As a result of inflammatory press dispatches, mobs in a small Latin-American country have repeatedly attacked United States flags and torn them to shreds. The United States citizens feel that their lives are in danger. MARINES ARE SENT TO PROTECT THE LIVES AND PROPERTY OF THESE CITIZENS.

Strongly Approve	Approve	Undecided	Disapprove	Strongly Disapprove
(1)	(2)	(3)	(4)	(5)

Figure 17. Example items from the Murphy and Likert *Survey of Opinions*. (Source: Likert, 1932)

Murphy and Likert proposed not only to compare these formats, but methods of scaling and scoring. In Likert's dissertation work, he further analyzed these data. The methods of scaling and scoring measures of attitudes included sigma scoring (essentially standardized scores bounded between -3 and +3, facilitated by tables created by Thorndike), whereby each statement in the attitude measure received a sigma score; the total attitude scores were based on the mean or median of statement scores. For example, the data in Figure 18 were presented by Likert (1932) regarding one item in the Internationalism Scale:

Alternative	Strongly Approve	Approve	Undecided	Disapprove	Strongly Disapprove
Percent responding	13%	43%	21%	13%	10%
Sigma value	-1.63	-0.43	0.43	0.99	1.76
Numerical label	1	2	3	4	5

Figure 18. Sample response pattern for an example item with associated Sigma values and corresponding category numerical labels. (Source: Likert, 1932)

This sigma method was compared to the Thurstone scaling method, described above, which required the use of hundreds of judges to scale items. The sigma method proved to be easier logistically and resulted in comparable reliabilities of scores. Likert then argued that another appropriate comparison method involved the use of category numeric labels. In Likert's work, a value of 1 was always assigned to the negative end of the sigma scale and a value of 5 was always assigned to the positive end of the scale (see Figures 17 and 18). The score assigned to the individual's level of attitude was based on the average of the numerical category values based on the responses to all of the items – although in Likert's work, he noted that the number of statements to which each individual responded was equal, so he used the sum of the numerical label scores instead of the mean (which are mathematically equivalent).

The reliability of the simpler scoring method based on the sum of the 1 to 5 numeric labels (values) was the same as the sigma method, which was an improvement over the state-of-the-art Thurstone scaling method. Likert also found that the summed scores correlated nearly perfectly with the sigma method scores. In addition, he found that the summed scoring method yielded similar levels of score reliability as Thurstone scaling, but with fewer items – or higher reliability when using the same number of items. Likert's findings were a significant contribution to the methods of scoring and scaling, particularly at a time when computations of reliability, correlations, and factor analyses were laboriously completed by hand.

An interesting point was presented by Likert (1932) when he took the then famous Thurstone-Droba War scale, for which all the items were previously Thurstone scaled, and administered it to compare the Thurstone scores with the summed scores. In assigning the numeric labels 1 to 5 to each item, he found four statements for which it was not possible to use the Strongly Agree to Strongly Disagree alternatives. One such item was: "Compulsory military training in all countries should be reduced but not eliminated" (p. 34). He noted that this item was double-barreled; a person who opposes compulsory military training would *disagree* with "not eliminated" but a person who favors compulsory military training would *disagree* with the "reduction" part. So both individuals who oppose or favor compulsory training would disagree. At any rate, using the Thurstone-Droba War scale, he found the summed scores yielded equivalent reliabilities with nearly half the number of items.

Likert also argued that this simple method of scoring allowed for the inclusion of statements using different response formats, including the [Yes] [?] [No] response options, scored 4, 3, 2. To be complete in this review, Likert also tested alternative scoring methods for the [Yes] [?] [No] response options, including score values of 1, 3, 5, and a modified sigma method. He found that all three methods yielded the same results, so argued for the simpler 2-3-4 scoring method for 3-point rating items and 1 to 5 for the 5-point rating items. Although the summated rating scale methods of scoring had been used for decades prior to Likert's comparative study, they had not been used in the measurement of attitudes.

Considering the original purpose of the study, Likert (1932) reported to find high generality and less specificity in social attitudes regarding internationalism, imperialism, and race relations. Although there were specific attitudinal differences across items with each attitude measure, particularly between respondents from different colleges in different regions of the country (e.g., northern colleges versus southern colleges), the overall general component found in responses across items was significant.

A few comments need to be made regarding Likert's work in the context of previous work on the development of rating scales. As we know, survey developers and researchers, as well as most individuals in academia who do survey-related work, commonly refer to rating-scale items as *Likert-type* items.

In my own experience with graduate students (and their advisors) completing theses and dissertations that involve survey work, students (and their advisors) commonly refer to their items as Likert-type items. During student defense meetings with the committee members, I invariably ask: "What about these items makes them Likert-type items?" The student and other members of the committee then engage in a discussion of what features of the items make them *Likert-like*. They refer to the fact that there are 5-point rating values, or that there is a middle neutral position, or that the first and last categories are

labeled, or that it's because the labels range from strongly agree to strongly disagree. A general puzzled look comes over everyone and they turn back to me. I then ask: "Has anyone read Rensis Likert's 1932 monograph?" In my 15 years in the academy, I have yet to meet someone who has read the paper (although I haven't asked everyone I've met).

Likert did not introduce the rating scale, nor did he claim to do so in his writing. He did not suggest that rating scales need to be 3-point or 4-point or 5-point scales. He even suggested in his summary of the methods that multiple-choice response options could be used. He selected response scales that fit the question. He did not suggest that the category labels need to be agree to disagree. He did not recommend that only certain categories (e.g., the extremes) should be labeled; in the attitude measures he used, all categories were labeled, whether 3-point, 5-point, or multiple-choice response options. He did not use numeric labels. What he did was suggest that simple numeric labels could be assigned, giving each category a consecutively increasing value for use in scoring, which resulted in reliability similar to sigma scoring, and employing the same number of items, a reliability higher than that obtained through Thurstone scaling and scoring. He did not require the scores be based on average ratings, but used summed scores, much like the many rating-scale developers before him.

Likert was interested in the use of rating scales as a way to measure attitudes simply, avoiding the laborious methods in place at that time. The simple use of the numeric labels for scoring provided adequate reliability with fewer items than Thurstone scaling (Likert, 1932; Likert, Roslow, & Murphy, 1934). He identified, although not explicitly, limitations in this approach. He noted that Thurstone scaling and the sigma method of scoring were likely to yield interval measures, and made no comment of the interval or linear nature of the scores based on numeric-label scoring. He noted some item specificity that appeared to be a function of region of the country (an issue of measurement invariance). However, in his summary of the methods, he noted that cultural background may result in different clusters or hierarchies of items – emphasizing the point that each item should contribute to the total score (Likert used the item-total correlation to evaluate internal consistency). He argued

> *it is certainly reasonable to suppose that just as an intelligence test which has been standardized upon one cultural group is not applicable to another so an attitude scale which has been constructed for one cultural group will hardly be applicable in its existing form to other cultural groups.* (p. 52)

Regrettably he did not explore this in his analyses, but presented some discussion based on psychological orientations and their effect on item response patterns given region of the country, particularly regarding dramatic differences in the northern versus southern participant opinions regarding race relations.

Unfortunately, strong methods of investigating measurement invariance or differential item functioning were not available in the 1930s. We would have to wait until the availability of latent trait theory methods, such as Item Response Theory and Structural Equation Modeling, to begin rigorous investigation of measurement invariance, differential item functioning, and related techniques.

Likert's contributions to the methods of attitude measurement were significant, providing a leap forward. However, to refer to rating scale items as Likert-type items or Likert-type scales ignores the decades of development by researchers and practitioners that preceded him.

Epilogue

A great deal of work has occurred since the groundbreaking efforts of these early 20th Century researchers. The trained survey item writer will recognize a great deal of good advice from the early rating scale developers. Much of the early advice has been further studied.

These findings have been used to develop evidence-based guidance for survey item writers, coupled with strong measurement advice. Some of the strongest evidence-based guidance can be found in comprehensive resources. Two important summaries of experimental research on survey item format effects include Shuman and Presser (1981) and Tourangeau, Rips, and Rasinski (2000). Presser et al. (2004) presented a comprehensive treatment of evaluating questionnaire quality. Three texts in particular provide comprehensive guidance on item development (Dillman, et al., 2009; Krosnick & Presser, 2010; Haladyna & Rodriguez, 2013).

And, of course, a great deal of additional work is currently underway.

References

Achilles, P.S., & Achilles, E.M. (1917). Estimates of the military value of certain character qualities. *The Journal of Applied Psychology, 1*(4), 305-316.

Baker, H.J. (1930). *Detroit adjustment inventory*. Bloomington, IL: Public School Publishing.

Brandenburg, G.C., & Remmers, H.H. (1928). *Purdue rating scale for instructors*. Lafayette, IN: Lafayette Printing.

Brown, C. (1928). *Rating scale for teachers of home economics*. Minneapolis, MN: University of Minnesota Press.

Brueckner, L. (1929). *Scales for the rating of teaching skill* (2nd ed.). Minneapolis, MN: University of Minnesota Press.

Buchanan, J.R. (1854). *Outlines of lectures on the neurological system of anthropology, as discovered, demonstrated and taught in 1841 and 1842.* Cincinnati, OH: The Office of Buchanan's Journal of Man (no copyright). Retrieved at https://openlibrary.org/books/OL7035578M/Outlines_of_lectures_on_the_neurological_system_of_anthropology

Chassell, C.F. (1924). The army rating scale method in the kindergarten. *Journal of Educational Psychology, 15*(1), 43-52.

Childs, H.G. (1915). Measurement of the drawing ability of two thousand one hundred and seventy-seven children in Indiana city school systems by a supplemented Thorndike scale. *The Journal of Educational Psychology, 6*(7), 391-408.

Crawshaw, F.D. (1916). Organization in the teaching of manual and industrial arts. *Industrial Arts Magazine, 5*(2), 47-52. Retrieved at http://books.google.com/books?id=qtlLAAAAYAAJ

Dillman, D. A., Smyth, J. D., & Christian, L. M. (2009). *Internet, mail, and mixed-mode surveys: The tailored design method* (3rd ed.). Hoboken, NJ: Wiley.

Freyd, M. (1923). The graphic rating scale. *Journal of Educational Psychology, 14*(2), 83-102.

Furfey, P.H. (1926). An improved rating scale technique. *Journal of Educational Psychology, 17*(1), 45-48.

Galton, F. (1879). Psychometric experiments. *Brain: A Journal of Neurology, 11*, 149-162.

Galton, F. (2004). *Inquiries into human faculty and its development*. Corrected proof, G. Tredoux (Ed.). Published at the online Galton archives at http://galton.org/. Retrieved at http://www.galton.org/books/human-faculty/index.html

Garrett, H.E., & Schneck, M.R. (1933). *Psychological tests, methods, and results*. New York, NY: Harper & Brothers, Publishers.

Haladyna, T.M., & Rodriguez, M.C. (2013). *Developing and validating test items*. New York, NY: Routledge.

Hayes, M.H.S., & Patterson, D.G. (1921). Experimental development of the graphic rating method. *Psychological Bulletin, 18*(2), 98-99.

Hildreth, G.H. (1933). *A bibliography of mental tests and rating scales*. New York, NY: The Psychological Corporation.

Johnston, H.J. (1917). Scientific supervision of teaching. *School and Society, 5*(112), 181-188. Retrieved at http://books.google.com/books?id=LV4VAAAAIAAJ

Kelley, T.L. (1919). Principles underlying the classification of men. *Journal of Applied Psychology, 3*(1), 50-67.

Krosnick, J. A., & Presser, S. (2010). Question and Questionnaire design. In P.V. Marsden & J.D. Wright (Eds.), *Handbook of survey research* (2nd ed., pp. 263-313). United Kingdom: Emerald.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of Psychology, 140*, 44-53.

Likert, R., Roslow, S., & Murphy, G. (1934). A simple and reliable method of scoring the Thurstone attitude scales. *Journal of Social Psychology, 5*, 228-238.

Ludlow, L.H. (1998). Galton: The first psychometrician? *Popular Measurement*, 13-14.

Miner, J.B. (1917). The evaluation of a method for finely graduated estimates of abilities. *Journal of Applied Psychology, 1*, 123-133.

Monroe, W.S., & Clark, J.A. (1924). Measuring teaching efficiency. *University of Illinois Bulletin, 21*(22), 3-26. Retrieved at https://www.ideals.illinois.edu/bitstream/handle/2142/32447/measuringteachin25monr.pdf

Paterson, D.G., & Ruml, B. (1920). The extension of rating scale theory and technique. *Psychological Bulletin, 17*(2), 80.

Pearson, K. (1906). On the relationship of intelligence to size and shape of head, and to other physical and mental characters. *Biometrika, 5*(1/2), 105-146.

Presser, S., Rothgeb, J. M., Couper, M. P., Lessler, J. T., Martin, E., Martin, J., & Singer, E. (2004). *Methods for testing and evaluating survey questionnaires*. New York, NY: Wiley.

Remmers, H.H. (1927). The Purdue rating scale for instructors. *Educational Administration and Supervision, 6*, 399-406.

Remmers, H.H., Shock, N.W., & Kelly, E.L. (1927). An empirical study of the validity of the Spearman-Brown formula as applied to the Purdue rating scale. *Journal of Educational Psychology, 18*(3), 187-195.

Rugg, H. (1921a). Is the rating of human character practicable? *Educational Psychology, 12*(8), 425-438.

Rugg, H. (1921b). Is the rating of human character practicable? *Educational Psychology, 12*(9), 485-501.

Rugg, H. (1922a). Is the rating of human character practicable? *Educational Psychology, 13*(1), 30-42.

Rugg, H. (1922b). Is the rating of human character practicable? *Educational Psychology, 13*(2), 81-93.

Schuman, H., & Presser, S. (1981). *Items and answers in attitude surveys*. New York, NY: Academic Press.

Scott, W.D. (1915). The scientific selection of salesmen. *Advertising and Selling, 25*(5), 5-6, 94-96.

Scott, W.D. (1916). Selection of employees by means of quantitative determinations. *Annals of the American Academy of Political and Social Science, 65*, 182-193.

Stoddard, G.D., & Ruch, G.M. (1926). Ratings of Downey Will-Temperament traits. *Journal of Applied Psychology, 10*(4), 421-426.

Strong, E.K. (1918). Work of the Committee on Classification of Personnel in the Army. *Journal of Applied Psychology, 2*(2), 130-139.

Symonds, P.M. (1924). On the loss of reliability due to coarseness of the scale. *Journal of Experimental Psychology, 7*(6), 456-461.

Symonds, P.M. (1925). Notes on rating. *Journal of Applied Psychology, 9*(2), 188-195.

Terman, L.M. (1915). The mental hygiene of exceptional children. *Journal of the Proceedings and Addresses of the National Education Association, 53*, 945-951. Retrieved at https://archive.org/details/addressesproce1915natiuoft

Thorndike, E.L. (1910). The measurement of the quality of handwriting. *Teachers College Record, 11*(2), 86-151.

Thorndike, E.L. (1913). The measurement of achievement in drawing. *Teachers College Record, 14*(5), 345-383.

Thorndike, E.L. (1916). The technique of combining incomplete judgments of the relative positions of N facts made by N judges. *The Journal of Philosophy, Psychology and Scientific Methods, 13*(8), 197-204.

Thorndike, E.L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology, 4*(1), 25-29.

Thurstone, L.L. (1928). Attitudes can be measured. *American Journal of Sociology, 33*(4), 529-554.

Tourangeau, R., Rips, L., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge, MA: Cambridge University Press.

Witham, E.C. (1914). School and teacher measurement. *Journal of Educational Psychology, 5*(5), 267-278.

About the Author

Michael C. Rodriguez received his BA in Psychology from the University of Minnesota, Morris; MA in Public Affairs from the Hubert H. Humphrey Institute of Public Affairs at the University of Minnesota, Twin Cities; and PhD in Measurement and Quantitative Methods from Michigan State University, East Lansing. He is a professor of Quantitative Methods in Education in the College of Education and Human Development (CEHD) at the University of Minnesota and has been a member of the faculty since 1999. He holds the Campbell Leadership Chair in Education and Human Development and is a member of the University's Academy of Distinguished Teachers.

Dr. Rodriguez is the recipient of the Award for Outstanding Contributions to Postbaccalaureate, Graduate, and Professional Education (2009), the U of M Morris Alumni Association's Distinguished Alumni Award (2008), the CEHD Alumni Society's Robert H. Beck Faculty Teaching Award (2008), the International Reading Association's Albert J. Harris Research Award (2005), and the CEHD Community Service Award (2004). He is also a recipient of the 2010 President's Volunteer Service Award from the United States President's Council on Service and Civic Participation.

He is an active member of the National Council on Measurement in Education (Board of Directors, 2009-2012) and the American Educational Research Association. His areas of research interests include item writing, item response models, and applications of multi-level modeling and meta-analysis to practical issues in educational measurement. He also has substantive interests in early literacy and youth development.

Visit his course website at http://www.edmeasurement.net