Measuring and Scoring Latent States:

History, Practice, and Implications for Medical Research

White Paper

David A. Andrae

Brandon Foster

ENDPOINT OUTCOMES

Abstract

Medical researchers measure a multitude of aspects of people's health to diagnose conditions, assess treatments, and evaluate benefits, costs, and risks. Many measures involve quantitating unobservable, i.e., latent, states. These sorts of measurement problems pose unique challenges and have specialized techniques for developing scores to represent latent states being evaluated.

Techniques for developing scores to represent latent states have a long history. The current paper explores historical foundations of generating such scores through highlighting achievements in the fields of psychophysics, intelligence testing, and academic achievement testing.

The historical underpinnings of modern scoring are used to guide the reader through some of the finer points of different classical and modern psychometric scoring techniques and how similarities and differences in approaches manifest, including implications for medical research.

Keywords

psychometrics, scoring, Classical Test Theory, Item Response Theory, patient-reported outcomes, latent variables

Introduction

Measurement abounds in modern medical research. Often, measurement in medicine is straightforward, e.g., children's heights and weights as surrogates for developmental milestones. Other data are more difficult to measure such as one's level of depression. A convergence on health care quality improvement, drug development that includes the patient's perspective of efficacy, and the delivery of personalized medicine has increased focus on measuring patient's inner thoughts, feelings, and other internal states. These data fall under the umbrella of patient-reported outcome (PRO) data. A central challenge in measuring PRO data is that data are often non-observable. Yet, the process for measuring PRO data seems deceptively easy to most: simply ask patients questions; patients provide answers; data are collated, and sum scores are generated. It is likely because of this simplification of the process of measurement that the use of PRO measures has become ubiquitous in medical research and continues to proliferate. What is often lost is that the techniques used to measure patients' inner states are recent additions to scientific inquiry. Most of the theoretical predicates and technical advancements used to measure inner states were developed outside of medical research. As a result, the field needs clarity with respect to historical roots for these methods, assumptions of modeling and scoring techniques, and implications for research and practice.

The current paper provides a relatively recent historical foundation for the philosophy, theory, and practices that have influenced current applied psychometrics used in contemporary medical research. Every attempt has been made to be concise and the treatment is by no means exhaustive. Instead, the topics covered have been judged by the authors to be some of those most-relevant to the practice of measuring patient-reported data within medicine. Admittedly, there will be topics that, although important, are not discussed here for the sake of brevity.

The paper is organized into two main sections. First, the history of measuring latent variables is briefly surveyed, by focusing on six individuals who contributed to the practices in use today for medical research. Secondly, a discussion in which we outline the strengths and weaknesses of each methodological approach is presented. Our goal is to shed light on a topic that is often seen as a "black box" by many researchers and practitioners not familiar with psychometrics.

A note on terminology: in this paper we will use the term *latent* to refer to any non-observable variable. For example, one's subjective experience of pain is a latent variable. We contrast this idea with that of a *manifest* variable, i.e., one that is observable or corporeal in some sense. A manifest variable that functions as a corollary to subjective pain experience is the pain rating on a numeric rating scale from 0 to 10. We will also refer to *scores* to indicate quantitative values that represent latent variables so that we can treat them as manifest.

Historical Perspective

Western development of probability theory and the applications of such theory relevant to the measurement of latent variables can be traced back to the efforts of Quetelet—who applied the Gaussian law of errors to different human data—in the early 19th Century.[1] The uptake of measuring latent states progressed, albeit slowly, through the 19th Century through diverse efforts, the first of which we will turn to involve the interaction between mind and matter.

Psychophysics and the quantitation of sensory stimuli

The field of psychophysics grew out of early 19th Century advances in physiology combined with the rapidly expanding body of research in electromagnetics from contemporary physicists such as Georg Ohm. Psychophysical experiments represented some of the first direct attempts to measure the relationships between physical stimuli and perceptual experiences to those stimuli. Gustav Fechner, the father of psychophysics, was trained in medicine, but his academic career focused on bringing the mathematical rigor of the physics of Ohm's laws to understand the interaction of mind and matter.[1,2] From 1830 onward, his attempts to formulate the mathematics of sensation were marked with experimental methods involving the recording of data from hundreds of trials of stimulus-response pairs. Fechner's seminal work, *Elemente der Psychophysik* (1860), outlined the first psychological scoring scheme. Fechner developed a mathematical model for sensory perceptions in which the unit of "just noticeable differences" (JND) was used to gauge intensity of two stimuli presented in succession, as a unit of measure in determining relative perceptual responses. In Fechner's experiments, the JND represented the smallest threshold within a certain task for which a participant in the experiment could distinguish one stimulus level from another.

Through exhaustive experimentation, Fechner derived mathematical relationships for different sensory modalities, e.g., judgments on how heavy two sets of weights are. The relationship between physical stimulus and psychological response in generalized form is now known as the Weber-Fechner Law, which states that stimulus intensity and the perception of that intensity are nonlinearly related. Specifically, the perceived intensity of a stimulus is an exponential function of the actual stimulus intensity. While the details of this relationship vary by sensory modality and were extensively studied for more than a century after Fechner's death, the concept of relating tangible measurements to subjective experiences—latent states—has influenced much of psychological and psychometric research and formed a basis for much of what we think of today as measurement. Over 150 years later, two direct ways we see the impact of Fechner's work today are built into the vision and hearing tests we use to check the health of these senses.

To summarize, Fechner was able to assess a latent state with a known stimulus level. Later we will see that the field of psychometrics expanded to the accommodate the use of unknown stimuli. Further, Fechner's work became a benchmark for both experimentation and mathematical analysis of stimulus response data. This became a foundation for building a science of measuring latent variables.

Spearman and the beginnings of Classical Test Theory

The measurement of latent variables made huge strides due to the publication of two papers in consecutive years just after the turn of the 20th Century. Both papers focused on an attempt to quantify intelligence. Charles Spearman—author of the first—is arguably better known for his long-standing argument with Karl Pearson about how to best calculate the correlation coefficient, but he was also keenly interested in measuring intelligence.

Spearman developed both his theory and technique for measuring intelligence over decades of research.[3] In 1904 he published his theory of intelligence, which provided a foundation for True Score Theory, i.e., the so-called classical test theory.[4][5] Spearman's work posited that intelligence was a general psychological construct, which was reflected in the approach he used to measure it. As he analyzed more data with his correlational techniques, he substantiated the idea of a general factor, g, as the basis for intelligence. In his treatment, differences between different intelligence tests were simply due to the error with which each test measured "true" intellectual ability, g.

In honing his theory of intelligence, one of Spearman's major methodical insights was that observed correlations were attenuated from true correlations if the data contained error. Equation 1 represents Spearman's conceptualization of intelligence measurement.

$$X = \mathrm{T} + \mathrm{E} \qquad (1)$$

In the equation above, Spearman decomposed the observed score, X, on a given measure into a "true" score, T—one measured perfectly—and a random component of error E. Using the terminology from above, X is the manifest variable representing the latent variable under study, T, and its latent error, E. In the modern sense, T and X can be either individual items or composites of items. In Spearman's theory of intelligence, differences between different intelligence tests were simply due to the error, E, with which each test measured "true" intellectual ability, g. This algebraic representation of a score and error would eventually become the basis for the unidimensional factor analysis model.[6]

Spearman and the eventual CTT theories did not state how a score is comprised. Scores, according to this philosophy, were unified quantities that were immutable in the sense that any constituent parts, measured or not, only matter in their contribution to T. Following this philosophy, using multiple items to measure a latent variable is only useful to the extent that such items contribute to X.

This sort of approach where measurement is unified into a single score works well for instances where a strong underlying construct explains the latent variable; however, this framework may have limitations when the latent construct under study is complex or has diffuse components. This is because, by the simple decomposition in Equation 1, any variability not due to T is pooled with the E term. So, if we were to measure two individuals' levels of fatigue by administering a paper-based questionnaire, the respondent with dyslexia may score lower than their non-dyslexic counterpart even if both have the same true fatigue score, i.e., the dyslexia contributes to the error term for one respondent, but not the other. This is because in the example given, the individual differences between two students are marginalized by the model itself. This amounts to averaging across different latent states. Spearman was aware of this criticism of his methodology, but for intelligence he maintained that g was strong enough a construct that the errors not accounted for did not matter.[7] This may be the case for certain constructs, but others may warrant a different approach.

Spearman gave the field of psychological measurement a simple mathematical framework that distilled the essential components of measurement—the "real" value, X, and deviations from that value due to the inherent error, E, in the measurement. As we will see, his unidimensional view of intelligence, although important and impactful, was not the only means by which to approach measurement of latent states.

Binet and multiple indicators

The second paper on intelligence testing relevant to psychometrics in medical research was published in 1905 and took a different approach to theorizing and measuring intelligence than Spearman. Alfred Binet, a polymath, is best known for developing measures of intelligence. In 1904, Binet was asked to help develop a means of identifying children in Paris schools who needed remedial education. Binet synthesized the previous half-century of psycho-physiological research findings of those such as Paul Broca in physiology, Fechner, and Galton in biology and statistics, among others, to a practical application of psychological theory and measurement.

As Binet's remit in helping measure and identify remedial students had practical implications for individuals, he developed his approaches and conceptualized the problem of intelligence measurement differently from his contemporaries. Binet's earliest work attempted to measure intelligence through physiological correlates—specifically, craniometry measurements espoused by Broca—but these efforts were unsuccessful.[7] While his contemporaries in other research areas pursued efforts to measure psychological phenomena similarly to physical measurement, Binet, frustrated by the lack of progress in his craniometry work, took a different approach. Binet came to see the complexity of intelligence measurement not as a simple linear decomposition, as Spearman had, but as a series of measurements that would be used to hone in on an unseen–latent—state.[5] His idea that a construct like intelligence was best measured with multiple "tests," or what we would now call items, was innovative because he recognized the nuance of a complex latent variable and sought to combine and standardize the measures into a coherent and interpretable scale. He and his colleague, Theodore Simon, developed a scale of 30 items that spanned a variety of mental functions and they published this scale in the 1905 paper. The scale would later be called the Binet-Simon scale, versions of which are still in use today.[7,8] The scale employed a variety of items that related to different aspects of intelligence together to form an index—

what would become the intelligence quotient (IQ). By analogy, if we were to measure the size of a table, Spearman's method might be likened to measuring the length of the table whereas Binet's method would be to measure length, width, height, load capacity, etc. of that same table.

As Binet and Simon point out, their scale was not a measure of intelligence in the manner of taking, say, a ruler to record length, as they note the multidimensionality of a concept like intelligence. Their position was that the scale could be used for the purpose of classifying students, as equivalent to a measure. The resultant scale formed a basis for the uses—and abuses—of intelligence testing, as well as all psychological measurement to this day.[7] While intelligence testing has had a varied past, we can appreciate the researchers' attempts to measure latent variables and how these early efforts shaped psychometric theory and practice today.

Binet's work laid a foundation for psychological measurement to come by embracing a means for using multiple indicators—i.e., pieces of a complex latent state—could be measured and then combined to form an index for that latent state. We see his influence today not only in intelligence testing, but in any questionnaire that employs multiple items to measure complex constructs such as quality of life, satisfaction, or academic achievement. Binet's use of multiple tasks to assess complex constructs may seem intuitive, today, but they were innovative for the time and their use opened a myriad of possibilities regarding the measurement of complex latent variables.

Thurstone and scaling

Leon Thurstone was at the forefront of psychological research from the 1920s until his death in 1955. Although his work influenced all areas of psychological measurement, arguably his two most-lasting achievements were in the areas of scaling and the statistical estimation of multidimensional psychological constructs.

Thurstone, like Binet, was well-aware of the developments in psychophysics that Fechner had made more than a half-century previously. Thurstone used the template of the psychophysicists for his method of scaling other psychological constructs than sensory stimuli and perceptual responses.[5,7] Thurstone developed rigorous, and at times, tedious, methods for relating responses to posed questions about behaviors and feelings, to known statistical distributions.[9] For example, he applied his scaling to the

comparison of two statements, each of which had "yes" versus "no" response options. Over a large series of items presented this way, Thurstone was able to relate the respondents' data to normal i.e., Gaussian, distributions.

Thurstone's innovative methodology allowed complex qualitative data to be represented with mathematical eloquence—much like Fechner's JND. The method was also theoretically grounded and was based on a literature that pre-dated almost all other psychological research—specifically the psychophysical experiments of Fechner that resulted in the Weber-Fechner Law. His work has since laid the foundations for much of psychological measurement, both theoretical and applied. By relating item responses to a probability distribution rather than a coded set of "yes" versus "no" responses, Thurstone's work would later catalyze Item Response Theory (IRT) methods and other so-called modern test theory methodologies and models.

Another of his innovations was to extend factor analysis into a multidimensional latent model. Spearman is credited with the seminal ideas of factor analysis, but he maintained a unidimensional mindset.[5] Thurstone's multidimensional factor models were extensions of Binet's multi-item measurement that actually led an entire field of factor analysis research by inventing factor rotation, among other aspects of modern factor models. He was also interested in the measurement of intelligence; however, he rejected Spearman's idea of a general intelligence. Thurstone maintained that a single general intelligence factor was insufficient to account for the intercorrelations between different tests. Instead, Thurstone posited a multidimensional mind, and used different factor analytic techniques to show what he believed to be were seven distinct primary mental abilities. With his acute mathematical abilities, he formulated a means by which multiple factors could be extracted and placed on equal footing. His factor analytic work used geometric analogies to show that the location of axes in a factor model could be rotated to define primary mental abilities. Previous methodological advancements that used the factor-analytic framework to understand intelligence, like those of Spearman, relied on the methods of principal components. These methods allocated the majority of variance to a general factor and anything left over was remaindered into a residual grouping, i.e., specific factors that were considered nuisances with respect to the general factor of interest. Thurstone developed factor analytic methods that didn't assume a hierarchy from general to specific factors.[7] Thurstone's research in this area led to much of what is practiced today in the rotations employed within exploratory factor analysis.

Thurstone's legacy to measurement is vast; however, two aspects of his research impact the practice of measuring latent states in medical research. First, he expanded the experimental procedures of psychophysics beyond measuring the response to physical stimuli to purely psychological phenomena, e.g., attitudes. This represented a huge step forward in terms of being able to quantify latent states. Secondly, he expanded the unidimensional factor analysis model to allow multiple factors to co-exist within a given theoretical framework. Although seemingly at odds with his contemporaries who advocated for a unidimensional conceptualization of intelligence, Thurstone's methods and models were actually able to reconcile a theory of general intelligence with that of multiple intelligences through a hierarchical framework. In medical research, the latter point laid the foundations for subscales within patient-reported questionnaires.

Likert and measuring attitudes

Although most of the previous work in measurement of latent variables focused on whether tests or items were answered correctly—as would be customary in academic testing situations—Rensis Likert was focused on the measurement of personality traits, publishing a treatise on measurement of attitudes in 1932.[10] Whereas Spearman and Binet were concerned with measuring intelligence as an individual trait—something that is relatively stable over time—Likert was interested in a latent state—a construct that is transient depending on the situation. He described an attitude as a tendency of response dependent upon the stimuli presented. One could expand on this idea by stating the stimuli realized by the individual at the time assessed. This sort of framework implies that the measurement of latent states is less direct than in a psychophysical experiment where the level of stimulus intensity is known or in an academic setting where the satisfactory answer to an intelligence test is also known. Consequently, the difficulty for Likert was two-fold: he needed to devise a way to measure latent traits that could vary and score those variations by means other than "right" and "wrong" responses.

Because of his research needs, Likert was not only concerned with the measurement of attitudes; he became involved in the debate regarding how to scale such measurements. Although he understood the procedures and results of Thurstone's work, he was not convinced that all the assumptions and methodology were needed to achieve adequate scaling of latent constructs such as attitudes. He saw Thurstone's methods as overly intensive and sought a simpler means of scaling.

As a result, Likert developed a coding system that resulted in items being combined into a score for attitudes on a certain topic. Likert focused his 1932 paper on the scoring of three scales: attitudes toward internationalism; attitudes of whites toward blacks; and attitudes toward imperialism. Item responses were coded in one of two general forms: "Strongly Disapprove" (1), "Disapprove" (2), "Undecided" (3), "Approve" (4), and "Strongly Approve" (5); versus "Yes" (2), "?" (3), "No" (4). Of course, the directionality of the coding was dependent on the item stem and its content, but the general idea was very similar to the graded items that are almost universal in patient-reported data collected in medical research today, with the exception of the coding of "Yes," "No," or "?" to items.

In devising a scoring system, Likert noted that for his scales, most items' frequency distributions resembled a normal distribution, so it followed that he, as did Thurstone, assumed normality of items in his scoring. Despite this assumption, Likert noted that certain items were non-normally distributed, but still justified the use of the normal distribution as a model. He proposed two scoring systems. The first was more statistical and involved, the Sigma method; the latter, the Simple method, is likely more familiar to those who use and score patient-reported data.

The Sigma method involved placing all items on a comparable scale prior to some sort of summarization. Likert calculated what amounted to a Z-score for each item and then either averaged or summed the items. The responses, then, were scaled to be $\pm 3\sigma$, i.e, within 3 standard deviations of the mean [$Z \sim Gaussian(0,1)$]. Higher values were taken to mean "more" of an attitude toward a given topic, e.g., internationalism.

Alternatively, Likert endeavored to simplify even his own Sigma method to, as he phrased it, save considerable work. In his Simple method, he assigned 1 to the "negative" end of the response options and 5 to the "positive." Then the average or sum was taken as a score. This is well-known today as a typical raw score for patient-reported data. Likert did point out that the sum was taken for those data in which all individuals had the same number of responses, the implication being that there were no missing data. In his comparisons of the Sigma and Simple methods, Likert showed data for small samples that correlated very highly (all $\rho \geq 0.990$). Therefore, he pragmatically concluded that the Simple method was useful in most situations.

Likert's method has proven popular and is the one of choice in current patient-reported data for medical research items—at least in terms of generalized response options and their coding. The appeal of Likert's scoring method also seems to have a foothold as averaging or summing items is understandable to most who would field items with loose ordering of low to high. What has been generalized, however, from Likert's original work is his method across many contexts of use with little regard for the assumptions Likert made and how his method of combining items may be affected by divergences from these assumptions.

Likert's work in scaling attitudes was motivated by his need to simplify the process that Thurstone had established. His focus on the measurement scales of single items and how they would combine yielded a framework that is ubiquitous in latent state measurement in medical research today. Whereas much of the previous work on scaling and measurement of psychological constructs such as attitudes focused on "yes" versus "no" responses, Likert's use of ordered categorical responses that assumed an underlying normal distribution, as we do with the polychoric correlation, paved the way for a flexible and pragmatic means of constructing item responses.[11] His scoring schemes, for his areas of application, showed that simpler methods may be equivalent to more mathematically involved procedures under certain assumptions—an observation that is often invoked in scoring algorithms for psychometrically-based medical research instruments today. Unfortunately, although the form of Likert's attitude measures is widely applied in modern medical research, the assumptions Likert outlined for his Sigma and Simple scoring methods are often ignored.

Lord and Novick and the theory of scores

Frederic M. Lord and Melvin R. Novick, with contributions from Alan Birnbaum, compiled a text in 1968 that included a meticulous detail of CTT and its assumptions as well as the developments in what would become IRT—it was one of the first complete treatments of test theory.[12] Further, at a time when most CTT advocates remained somewhat ambiguous regarding the use of multiple indicators or items versus total scores, Lord and Novick extended the mathematical development of true score theory to include composites. From this treatment, Lord and Novick made two salient points that are germane to the application of CTT to measuring patient-reported outcomes in medical research.

First, CTT scoring is valid provided the items are parallel, i.e., that the amount of variation in an item score due to the true score is the same for all items and that the expected value for items is the same. This means that the error terms across items are independent, i.e., that the variance of the true score, T, is proportional to the square of the length of the test, $\sigma_{\mathrm{T}}^2 \sim n^2$. At the same time the variance of the errors is proportional to the length of the test, $\sigma_{\mathrm{E}}^2 \sim n$. This means that, from a true score theory approach, as the length of an instrument increases, the associated error estimates become a smaller relative amount to the total. It is also why reliability, $\rho_{XX'}$, estimates increase if more items are part of the total score. For the parallel-item case:

$$\rho_{XX'} = \frac{\sigma_{\mathrm{T}}^2}{\sigma_{\mathrm{T}}^2 + \sigma_{\mathrm{E}}^2} \quad (2)$$

So, as the variance of T, σ_{T}^2, becomes a larger proportion of the total variance with increased test length, $\rho_{XX'}$ becomes larger with an upper limit of 1 when $\sigma_{\mathrm{E}}^2 = 0$. In practical terms, this means that CTT scores from larger item sets demonstrate better reliability because more items make up the composite. Of note here is that adding items to a measure is only a benefit to increasing the reliability estimate of a test if the test is parallel. This is because the estimates of variances can be affected by the choice of item score coding, as we will see below. If the items are not parallel, the inferred correlation between errors will, in turn, increase the error variance and thereby attenuate any benefit of a longer instrument.

Secondly, Lord and Novick describe the following relationship for estimating the true score, T:

$$\hat{\mathrm{T}}_j = \bar{X} + \rho_{XX'}\left(X_j - \bar{X}\right) \quad (3)$$

That is, $\hat{\mathrm{T}}_j$, our estimate of the true score for respondent j, is composed of the mean of X, the observed score, and the reliability of X, $\rho_{XX'}$, times the deviation of the observed score from its mean. Stated differently, $\hat{\mathrm{T}}_j$ is a regression estimate based on the distribution of X where the regression coefficient is the reliability of X. An important implication of Equation 3 is that there will be regression to the mean for the $\hat{\mathrm{T}}_j$ within the classical framework—specifically, this means that observed test scores for individual values, X_j, that are below the mean will be overestimated and values above the mean will be underestimated. Therefore, the smaller the value of $\rho_{XX'}$, the more an estimated value will shrink toward $\bar{X}$.

Both points above are a consequence of the marginalization of information contained in the items—i.e., because X is a summary of T and it is essentially a regression estimate, the summarization throws away information. The items, themselves, although important as the building blocks of a score, are simply elements for the estimation of T, the true score, through the observed score X—usually the sum of item responses or their mean.

Subsequent research by Lord and Novick, and the well-known psychometrician Georg Rasch began to take a different approach. The new wave of psychometric methods developed decomposed model parameters uniquely for individuals and items based on patterns of responses to items.[5] To do this decomposition, item response models were formulated as the probability of a response value given item characteristics. All items were placed on a common probability scale. These modern psychometric models were extremely flexible. In addition to accommodating the combining of items for total scores, more realistic and complex scenarios in the item response data could also be adopted. These scenarios included differential weighting of items and, in the case of academic tests, parameters for guessing correct answers.

Take for example a frequently encountered model in modern medical research using PROs, the logistic form for discrete responses, as seen in Equation 4.[12]

$$\Pr(x_i|\theta) = \frac{1}{1+\exp\ [-(a_i\theta+d_i)]} \qquad (4)$$

The model in Equation 4 allows for the scaling of individuals on a latent continuum, θ, through a, a slope parameter that weights the item according to its contribution to estimating θ, and d an intercept parameter that indicates how difficult a given response is.

Equation 4 looks quite different from Spearman's model presented in Equation 1 and rightly so. Each uses a different approach to understand latent states are considerably different: one reifies the observed, coded responses into a total score and the other uses item-level information to estimate latent states. In using the item-level information, Equation 4 allows for more precise measurement of the latent construct under study by preserving more of the item-level information marginalized in the CTT approach.[5] In practice, the information lost in the CTT approach could result in multiple different scores estimated for the IRT models for the same single total score generated from the CTT approach.

Through intensive mathematical treatment of how item responses could be modeled—employing both classical and modern approaches--Lord and Novick laid the groundwork for an explosion of research in the half-century since the publication of their seminal text. They introduced cogency to the somewhat fragmented field of psychometrics at the time, as well as bringing to bear the innovations seen in previous generations of research: understanding the items by parameterizing them, much as Fechner understood the physical stimuli in his experiments; using multiple sources of information within each item to explain complex phenomena, much like Binet; placing mathematical treatments of measurement theory at the forefront, similar to Thurstone; and building new frameworks from the grounded principles of true score theory that were pioneered by Spearman and others. Lord and Novick and the early researchers of IRT took grounded theory and research and extended it to give a flexible framework to carry on measurement of latent states.

Measurement in Medical Research

But how do the achievements of those who developed psychometrics apply to measurement in modern medical research? Measurement, as has been defined in psychometrics and psychology, is quantitation based on a system or procedure.[13][14] While it may seem simple to attach a number to a latent attribute, quantitating a latent variable is a process that can involve several steps. While highly dependent on a given situation, we attempt here to generalize the process by definitions. Specifically, we will focus on the process by which the relationship between the latent variable and the eventual quantity is determined by the scale and is manifested as the score.

Much has been written regarding the development of scales for medical research—a recent PubMed search resulted in almost 2500 articles having the search terms "psychometric" and "patient centered" or "patient reported." Less has been written about the generation of scores and their different assumptions and properties. So: from this rich research history what can be learned about the current practices of measurement of patients' latent states? Two basic forms of scores and their implications for measurement of patient responses inform the understanding of scores.

Spearman's and Binet's influences are clear. We tend to measure patient-reported data with multi-item measures and one or more scores. Thurstone's influence often comes through in patient reported data

in the form of sub-scales. For example, the EORTC's QLQ-C30. Version 3.0 of this scale, for example, has one Global Health Status score, five functional scores, and nine scores for symptoms.[15]

Most scoring for medical measurement tends to fall into the CTT set of approaches. Whether raw item responses are employed or they are standardized to a common value set, either averaging or summation is common in scoring schemes for patient-reported data. These summation methods for scoring are in the CTT family of approaches. Consequently, CTT scoring regimes treat all items equally within a summary score, i.e., give equal weight to all items. For example, two patients can be assigned the same sum score irrespective of the specific item response patterns they endorsed. The unweighted sum score is then treated simply as a unit to be decomposed by both its value and the pooled error of the items used in the measure of the latent quantity under study. This simplified view of how items work together to measure a latent variable is plausible when items measure that latent quantity in an approximately equal manner. However, if individual items show differential properties then the assumptions of the CTT framework break down. As an example, if we compare a hypothetical Item 1, $x_1 = \tau_1 + \epsilon_1$, with a hypothetical Item 2, $x_2 = \tau_2 + \epsilon_2$, then under CTT $\epsilon_1 = \epsilon_2$, but if $\epsilon_1 > \epsilon_2$, or vice versa, then the items are showing differences with regard to their sensitivity to measure $\mathrm{T} = \tau_1 + \tau_2$.

There are fewer instances in which Lord's influence can be seen in current medical measurement. Although IRT and Rasch methods are gaining popularity, they are still not the norm. One exception is with the PROMIS project in which item banks are employed so that patient scores can be generated based on IRT modeling frameworks.[16]

Simple scores

To understand the differences between CTT and IRT scores we need to delve more deeply into how different scores are constructed. We define simple scores as those in the True Score/Classical Test Theory traditions. Specifically, a sum score, defined in Equation 5, is simply the sum of coded responses, x, for each item, i.

$$X_{sum} = \sum_{i=1}^{n} x_i \qquad (5)$$

Equation 5 represents a simple mathematical operation used on coded data to generate an unweighted sum score. Although there exist virtually limitless different ways to combine n items into a score, we have

focused on the sum here because it is still commonly used. The inherent assumptions of this method, however, are not as straightforward as the score generation. If the items summed are binary, $x \epsilon \{0,1\}$, then the sum represents a total number "correct" or, more generally, the higher the number of items endorsed the more of the construct is measured. Alternatively, if items summed are polytomous, $x \epsilon \{0,1, \ldots, k\}$, then more assumptions must be made to create valid scores, as is the case with Likert's simple scoring. The basic assumptions that need to be understood to generate scores with response scale data are:

1. The k response categories for each item should have a natural ordering that corresponds to the latent construct being measured. To illustrate, take a rating of pain intensity with a minimum of 0 and maximum of 10, and integer steps in between. This 11-point numeric rating scale, under the X_{sum} approach, treats a value of 0 as indicating less pain than a rating of 5.
2. The X_{sum} approach treats item responses as interval-level values, i.e., there is a natural zero value and the numerical intervals correspond to the same increase on the latent variable being measured.[14] To continue the pain rating example, the interval from ratings 0 to 1, 1 to 2, etc. mean *exactly the same thing* in terms of their increase in pain intensity. This also implies that values on different items that are to be summed have unit increases that are equivalent. This is partially because each item carries equal weight in determining the sum score, but because of the reliance on the numeric coding, no distinction is made between integer value differences. This latter concept is what Lord and Novick term τ-equivalent.[12] Such an assumption means for a multi-item pain instrument, a pain intensity item and a discomfort item would carry equal weight in the sum score and that the change interval of 1-point would be the same for both pain and discomfort.
3. Errors between any two sets of items should be independent (i.e., uncorrelated) strong or essentially τ-equivalent. The items are parallel in measuring the true score, T. If errors are not independent, then adding coded ratings together to create CTT sum scores would introduce bias into the scores because the common error between two (or more) ratings would be counted twice in the sum. The implication, then is that as the more a set of items has overlapping, i.e., correlated errors, then the bias in measurement of T increases.

Statistical researchers in education and psychological measurement have been well aware of the assumptions and limitations of both the number correct and Likert simple scoring means of combining

item responses and the associated limitations with doing so since the foundational work of Spearman and Binet.[5] These limitations motivated Thurstone and eventually Lord and, simultaneously, Rasch to determine model-based means of representing latent variables.[17]

Model-based scores

Scores based on models within the IRT traditions are founded on the idea that an item response is an indicator of respondents' position within a probability distribution of the construct of interest, θ. The models estimate the value of θ in what we would call a model-based score. The score is considered model-based because information from the items (i.e., item parameters) is used in conjunction with each respondent's pattern of responses to the items to estimate each respondent's θ. For the current treatment of model-based scoring, we will assume a graded response model as in Equation 6.[14]

$$\Pr(x_i = k|\theta) = p_i = \frac{1}{1+\exp[-(a_i\theta+d_{i,k})]} - \frac{1}{1+\exp[-(a_i\theta+d_{i,k+1})]} \qquad (6)$$

Where $x_i \in \{0,1,2,\ldots,k\}$ is an observed response on a rating scale to item i, a and d are slope and location parameters, respectively, and k indexes the code for actual responses. The probability of a given item response, then, is defined by that specific response and the next-highest coded category. This formulation is essentially a mixed-effects version of the adjacent-categories logistic regression model with inclusion of $\theta \sim Gaussian(0,1)$ as a random effect. It is also an extension of Equation 4 beyond binary responses.

The individual item response probabilities then can be combined to form a likelihood of response patterns by the usual means of combining individual probabilities. Equation 7 defines the likelihood of a response $\boldsymbol{u}$.[18]

$$\mathrm{L}(\boldsymbol{u}|\theta) = \prod_{i=1}^{n} p_i \qquad (7)$$

The likelihood equation is simply the product of the individual probabilities of the items, p_i, given by Equation 6. The equation assumes that the probability of response to the items, p_i, are conditionally independent given θ, meaning that the relationship between items are due to θ, and errors in estimating θ are independent across items. This assumption is called local independence and is an important contrast to the τ-equivalence assumption in the CTT approach. Local independence implies that a

statistical dependence exists between the item response data and estimated parameters only; whereas τ-equivalence is an assumption about the dependency of item responses and fixed model parameters that are not estimated (i.e., the factors loadings are assumed equal instead of variable).

Although many estimators exist for estimating scores using IRT models, two common methods are listed here. First, the *expected a posteriori* (EAP) displayed Equation 8.[20]

$$\hat{\theta}_{EAP} = \frac{\int_{-\infty}^{\infty} \theta L(\boldsymbol{u}|\theta)\phi(\theta)d\theta}{\int_{-\infty}^{\infty} L(\boldsymbol{u}|\theta)\phi(\theta)d\theta} \qquad (8)$$

Distilling Equation 8 to its essence, $\hat{\theta}_{EAP}$ is simply the *mean* of the posterior distribution of the likelihood of responses for each individual. Because the models report a probability of response, each individual's score has its own distribution. The prior distribution, $\phi(\theta)$, is usually the unit normal, but the scoring equation does not restrict this choice. The convenience of $\phi(\theta) \sim Gaussian(0,1)$ is that the posterior distribution of scores will take on the properties of the normal distribution, making score interpretation somewhat straightforward.

Also, of note is the similar *modal a posteriori* (MAP) estimate of the score. This estimator is calculated by solving the partial differential equation in Equation 9 for θ.[18]

$$\hat{\theta}_{MAP} = \frac{\partial}{\partial\theta} \frac{L(\boldsymbol{u}|\theta)\phi(\theta)}{\int_{-\infty}^{\infty} L(\boldsymbol{u}|\theta)\phi(\theta)d\theta} = 0 \qquad (9)$$

The $\hat{\theta}_{MAP}$is similar to $\hat{\theta}_{EAP}$ except that while $\hat{\theta}_{EAP}$ represents the mean of the posterior distribution, the $\hat{\theta}_{MAP}$ is the posterior distribution's mode. For the $\hat{\theta}_{MAP}$, we find the mode of the likelihood by differentiating with respect to θ, i.e., finding the tangent to the likelihood and where this tangent is equal to zero—indicating the maximum value of the distribution, which is the mode of that distribution. In practice, when a normally distributed prior distribution, $\phi(\theta)$, is assumed for θ, $\hat{\theta}_{EAP}$ and $\hat{\theta}_{MAP}$ will be very close in their values, as the Gaussian distribution's mean is equal to its mode. Both Equations 8 and 9 are Bayesian estimates in that they take prior information about the latent distribution, θ, and the estimated probability of response, i.e., the likelihood, $L(u|\theta)$, to determine an estimate.

Score meaning and consequences

Scoring algorithms all make assumptions and marginalize data to allow interpretation of individual and group summaries. Although the CTT and IRT scoring outlined above accomplish this goal by different paths, they arrive at similar places in terms of estimates. CTT scores make more assumptions and marginalize data more than IRT methods. The marginalization occurs with CTT scores because responses are treated as numbers, but the responses may not have the properties of interval or ratio data. Given that the calculation of a CTT score is agnostic to the patterns of responses to particular items, the composite of these values may behave in unexpected ways, leading to less precision in measurement of T. Once the sum or average of items is computed, the items' individual contributions are lost. For example, just three items with five response levels each has $5^3 = 125$ possible response patterns, but only 13 possible sums. To illustrate another way that the items' contributions to measurement are marginalized, we can look at the instrument's level of information. Modern psychometrics often measures a model's precision with the concept of information, I, we define it here in Equation 10 for CTT scores.[6]

$$\mathrm{I}_X = \frac{1}{\sigma_{\mathrm{E}}^2} \quad (10)$$

For CTT scores, then, the information is inversely proportional to the error variance across all items. Consistent with the overall CTT approach, this is the sum of individual item errors. As a consequence of the CTT framework it is assumed that both the items contribute equally to the measurement of a latent construct and errors within an instrument are treated as equal. From this, all items contribute equal information to the measure.

IRT formulations of information, I, allow for each item to contribute more or less information to the measure as a function of the item probabilities of responses. More intensive mathematics are required to define and calculate information in this context. The generalized IRT information framework is defined in Equation 11.[21]

$$\mathrm{I}(\theta) = -\sum \left\{ \frac{\partial^2}{\partial \theta^2} \, log[L(\boldsymbol{u}|\theta)] \right\} L(\boldsymbol{u}|\theta) \quad (11)$$

The first difference we can see in Equation 11 as compared to Equation 10, similar to the scoring formulae, is that we have the information based on probability rather than the composite of coded responses. Accordingly, the complexity of the computation is much greater and require more mathematically

intensive fitting algorithms. Secondly, information in an IRT framework is a function of θ—meaning that the information can vary depending on the latent variable value—rather than a sum of coded item responses as in CTT. This usually means that the information, i.e., precision, of the IRT scores is better in the middle of the distribution. Third, as was the case with the EAP and MAP scoring algorithms, the item probabilities are weighted as in Equations 4 and 6 and thus, can contribute differentially toward the precision of measurement of θ.

Discussion

The historical figures mentioned in this paper are but some of the scientists who contributed to the development of latent variable measurement. The field, as noted above, began slowly, but has burgeoned into many different parallel fields such as education, psychology, and medicine. The impacts of the early work in psychophysics, intelligence testing, and scaling of attitudes can be seen in our measures today in terms of the forms of items and their responses, the construction of instruments, and how we quantitate those responses into scores that function as measures of latent traits. Ultimately, determining whether a patient's condition improves upon receiving a pain treatment or that she does not suffer a loss in health-related quality of life while undergoing chemotherapy are important concepts for medical researchers and clinicians to understand and having scores to measure these and other latent traits provides tools to gain such understanding. The current paper was written to give some historical context to some of the main approaches that are employed in medical research studies, especially pertaining to the measurement of patient-reported data. As the goal of generating scales or scores is to quantify some "thing," it is important to note that measures, whether the width in inches of one's computer screen or the quality of life index of a cancer patient, are not the same despite the operationalization we give them. Reification of scores into the constructs they represent is an error both well-known and pervasive in measurement. No matter the complexity of their construction, scores and scales are mathematical abstractions that represent and marginalize the actual "thing" they represent. In the case of measuring latent traits, these mathematical abstractions are, of course, necessary to fulfill the remit of analysis that is required for quantitative answers to research questions.

While there are seemingly wide differences between the CTT and IRT approaches to scoring, it is the experience of these authors that often scores from both camps agree quite well when the number of

items contributing to scores is large and the reliabilities within those items sets is high. The correspondence between CTT and IRT approaches is also enhanced when the items are generated with good qualitative research principles and techniques. However, it is notable that it is also our experience that comparisons between CTT and IRT scores do not take measurement error into account. We recommend the literature on score attenuation and change scores.[12,13,22]

We, therefore, cannot underscore enough one advantage that IRT approaches have over CTT scores. Importantly, in practice a researcher probably will not know when departures from the CTT assumptions are impacting a score. IRT models give researchers a plethora of information about items as they individually and collectively measure latent traits. Additionally, IRT modeling, by placing items on a probability scale, allows flexibility of item responses and types to be scored together. These qualities are not inherent to CTT due to its reliance on treating scores as unitary, equally-contributing values. IRT approaches, therefore, can be simplified in terms of generating scores that are more like their simple CTT counterparts. Taking an IRT framework and summarizing it into a simpler form may be useful, for example, in generating a specific patient score in clinical practice, but in terms of validation, the consistency of methods used to derive scores must be maintained—including understanding the assumptions and their consequences of any given model. Methods do exist for marginalizing EAP scores and relating them to sum scores and even hybridizing the sum of items with different response options and formats.[23] However, such methods should be used with the knowledge of their consequences, e.g., the potential for loss of precision in the middle of the score distribution would still occur for EAP sum scores.

Conclusions

CTT approaches have traditionally served as the sole basis for development of PRO measures in medical research for many reasons including the ease of generating a summary score. We, however, fully advocate IRT approaches as alternatives and compliments to CTT. Thus, we recommend that if a CTT score is desired for a measure, that IRT be employed as part of the validation of the measure to best understand how the items function together, provided that the CTT modeling framework also be evaluated in the validation process. We acknowledge that there are instances when modern psychometric methods are not plausible, e.g., in rare diseases where sample sizes are, by definition,

small. Researchers in these scenarios may be limited to CTT methodologies and should still abide by grounded, empirical research methods to bolster their measurement precision. While the methods for validation depend on the research area and available patients, some guidance has been offered here and elsewhere as to how one can make use of various methods.[24] However, by expanding researchers' toolkits with IRT and other modern psychometric and statistical methods, the hope is to further measurement and understanding of patient experiences by more-fully using the information collected by the items within our PRO measures.

Abbreviations

CTT: Classical Test Theory; EAP: expected a posteriori; IQ: Intelligence quotient; IRT: Item response theory; JND: just noticeable differences; MAP: modal a posteriori; PRO: Patient-reported outcome

References

1. Stigler S. The History of Statistics: The Measurement of Uncertainty before 1900. Cambridge, MA: Belknap Press of Harvard University; 1986.
2. Hothersall D. History of Psychology. New York: McGraw-Hill, Inc.; 1995.
3. Spearman C. "General Intelligence," Objectively Determined and Measured. Am J Psychol 1904;1(2):201–92.
4. Spearman C. The proof and measurement of association between two things. Am J Psychol 1904;15(1):72–101.
5. van der Linden WJ. Introduction. In: van der Linden WJ, editor. Handbook of Item Response Theory: Volume 1, Models. Boca Raton, FL: CRC Press; 2016. p. 1–12.
6. McDonald RP. Test theory: a unified treatment. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.; 1999.
7. Gould SJ. The Mismeasure of Man. Revised an. New York: W. W. Norton & Company; 1996.
8. Binet A, Simon T, (tr. Kite, ES. Méthodes nouvelles pour le diagnostic du niveau intellectuel de anormaux (New methods for the diagnosis of the intellectual level of subnormals). L'Année Psychol 1905;11:191–244.
9. Jones L V., Thissen D. A History and Overview of Psychometrics. In: Handbook of Statistics, Vol. 26. Elsevier; 2007. p. 1–28.
10. Likert R. A technique for the measurement of attitudes. Arch Psychol 1932;22(140):5–55.
11. Olsson U. Maximum likelihood estimation of the polychoric correlation coefficient.

Psychometrika 1979;44(4):443–60.

12. Lord FM, Novick MR. Statistical Theories of Mental Test Scores. Reading, MA: Addison-Wesley Pub. Co; 1968.
13. McDonald R. Test theory: A unified approach. Mahwah, NJ: 1999.
14. Stevens SS. On the theory of scales of measurement. Science (80-) 1946;103(2684):677–80.
15. Fayers PPM, Aaronson NKN, Bjordal K, Groenvald M, Curran D, Bottomley A. EORTC QLQ-C30 Scoring Manual (3rd Edition) [Internet]. EORTC; 2001. Available from: https://www.eortc.be/qol/files/SCManualQLQ-C30.pdf
16. Cella D, Riley W, Stone A, et al. The patient-reported outcomes measurement information system (PROMIS) developed and tested its first wave of adult self-reported health outcome item banks: 2005-2008. J Clin Epidemiol 2010;63(11):1179–94.
17. von Davier M. Rasch model. In: Handbook of Item Response Theory. Boca Raton, FL: CRC Press; 2016. p. 31–48.
18. Samejima F. Estimation of latent ability using a response pattern of graded scores. Richmond, VA: The William Byrd Press; 1969.
19. Samejima F. Graded Response Models. In: van der Linden WJ, editor. Handbook of Item Response Theory Modeling: Volume 1, Models. Boca Raton, FL, FL: 2016. p. 95–107.
20. Bock RD, Aitkin M. Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika 1981;46(4):443–59.
21. Samejima F. Graded response model based on the logistic positive exponent family of models for dichotomous responses. Psychometrika 2008;73(4):561–78.
22. Cronbach LJ, Furby L. How we should measure change--or should we? Psychol Bull 1970;74(1):68–80.
23. Thissen D, Pommerich M, Billeaud K, Williams V. Item Response Theory for Scores on Tests Including Polytomous Items with Ordered Responses. Appl Psychol Meas 1995;19(1):39–49.
24. Cappelleri JC, Lundy JJ, Hays RD. Overview of classical test theory and item response theory for the quantiatative assessment of items in developing patient-reported outcomes measures. Clin Ther 2014;36(5).