Paper Presented at the AARE Annual Conference, Fremantle 2007

Australian Association for Research in Education

Cav07156

Measurement issues in the use of rating scale instruments in learning environment research

Associate Professor Robert Cavanagh (PhD)
Curtin University of Technology
Perth, Western Australia

Address correspondence to:

Associate Professor Rob Cavanagh
Department of Education
Curtin University of Technology
GPO Box U1987
Western Australia 6845
Email: R.Cavanagh@curtin.edu.au
Phone: +61 8 9266 2162
Fax: +61 8 9266 2547

This report is an abridged version of a previously published paper - Cavanagh, R.F., & Romanoski, J.T. (2006). Rating scale instruments and measurement. *Learning Environments Research*, 9(3), 273-289.

Abstract

The history of learning environment research is characterised by the creation and application of rating scale instruments to elicit attitudinal data from students and teachers about the learning environment. For several decades, the data from these instruments have been subject to various types of statistical analysis. Typically, such analyses are applied as part of the instrument development process as well as to inform answering of questions about the attributes of learning environments, the influences on learning environments, and the temporal stability of learning environments. Notwithstanding the widespread use of statistical techniques in quantitative learning environment research, these techniques do not necessarily reflect the knowledge and methods developed in the fields of objective measurement and rating scale analysis. It is therefore timely and appropriate to examine the implications of objective measurement theory and practice for the use of rating scales in learning environment research. The objective of this paper is to apply a measurement perspective to examine the issues involved in the development of ratings scales and in the analysis of the data they elicit. An argument is made that using raw scores from rating scale instruments for subsequent arithmetic operations and applying linear statistics is less preferable than using measures.

Rating scale instruments and measurement

Introduction

The paper is organised into three major sections. The first section is an examination of measurement in the human sciences. The second section considers the issues incumbent in ensuring data elicited by rating scale instruments are measures or correspond to measures. The third section provides evidence for the assertion that using raw scores from rating scale instruments is less preferable than using measures for subsequent arithmetic operations and applying linear statistics.

Measurement in the human sciences

Bond and Fox (2001) drew attention to the different approaches typically applied to measurement in psychology and the behavioural sciences in contrast to the physical sciences. They asserted that "… psychometricians, behavioural statisticians, and their like conduct research as if the mere assignment of numerical values to objects suffices as scientific measurement" (p. 2). Alternatively, they viewed scientific measurement as the objective abstraction of equal units in order for measures to be reproducible and additive. In advancing an argument for constructing objective measures in the human sciences, Bond and Fox (2001) asserted: "Abstractions of equal units must be created and calibrated over sufficiently large samples so we are confident in their utility. Then these abstractions can be used to measure attributes of our human subjects" (p. 3).

The notion of scientific measurement was qualified by Wright (1984), "… what physical scientists mean by measurement requires an ordering system and the kind of additivity illustrated by physical concatenation" (p.1). Physical concatenation is simplistically demonstrable by joining the ends of sticks to concatenate length, or by piling bricks to concatenate weight (see Campbell, 1920). Wright (1997) elaborated on his view of scientific measurement by considering the history of social science measurement and noted the theoretical requirements which govern the practical success of measurement. These were: measures are always inferences obtained by stochastic approximations of one dimensional quantities; and, measures are counted in abstract units, the fixed sizes of which are unaffected by extraneous factors. Measures are inferences because actual data is never complete: "… any attempt at data collection and methods which require complete data in order to proceed cannot by that very requirement be methods of inference" (Wright, 1997, p. 2). In order to create abstract units, a mathematical function is required, a function "…which governs the inferential stochastic process so that its parameters are either *infinitely divisible* or *conjointly additive* i.e. *separable*" (Wright, 1997, p. 2). Stochastic processes refers to the use of measurement models in which both the difficulty of the tasks assigned to subjects and the ability of subjects to perform the tasks are conjointly considered but are separable. These processes can be contrasted with deterministic processes in which the performance of a subject is assumed to be a measure of the subject's ability. That is, the difficulties of the task(s) used to test the subject's ability are not quantified (see Bond and Fox, 2001). This is exemplified by classical test theory which "… models the statistical nature of the scores and focuses attention on the consistency of results from the instrument (i.e., its reliability)" (Wilson, 2005, p. 88). The notion of a one-dimensional quantity concerns an instrument or a set of tasks/items eliciting data on only one trait of the subjects.

In order to create a scale that measures a variable in accord with the theoretical requirements for measurement, Wright and Masters (1981) identified seven measurement criteria:

First, each item should be evaluated to see whether it functions as intended;
Second, the relative position (difficulty) of each valid item along the scale that is the same for all persons should be estimated;
Third, each person's responses should be evaluated to check that they form a valid response pattern;
Fourth, each person's relative score (attitude or achievement) on the scale should be estimated;
Fifth, the person scores and the item scores must fit together on a common scale defined by the items and they must share a constant interval from one end of the scale to the other so that their numerical values mark off the scale in a linear way;
Sixth, the numerical values should be accompanied by standard errors which indicate the precision of the measurements on the scale; and
Seventh, the items should remain similar in their function and meaning from person to person and group to group so that they are seen as stable and useful measures.

The development of measures in the human and social sciences was strongly influenced by the Danish mathematician Georg Rasch (Wright, 1997). In 1953, Rasch investigated ways to compare past performances on different tests of oral reading. He developed a measurement model for dichotomous observations utilising a ratio measure of person ability and a ratio calibration of item difficulty (Rasch, 1960). Wright (1997, p. 14) recounted that:

> "By 1960 Rasch had proven that formulations in the compound Poisson family, such as Bernoulli's binomial, were both sufficient and, more surprising, necessary for the construction of stable measurement. Rasch had found that the 'multiplicative Poisson' was the only mathematical solution to the second step in inference, the formulation of an objective, sample [free] and test free measurement model".

When data produced from application of an instrument conforms to the Rasch Model, the above requirements for measurement and indeed objective measurement are satisfied: "The ability of the Rasch Model to compare persons and items directly means that we have created person-free measures and item-free calibrations, as we have come to expect in the physical sciences …" (Bond and Fox, 2001, p. 203). Hence the Rasch Model can be applied to ascertain the measurement capacity of an instrument. Importantly, in testing how well the data fit the model, it is possible to identify particular items that contribute to misfit and this enables the instrument to be refined so that it measures the latent trait of the persons better.

Measurement using rating scale instruments

Polytomous rating scale instruments present respondents with a series of alternative response categories for an item so that the respondent can indicate a degree of affirmation with a stem statement (the item). For example, a Likert scale ranging from 'all' (scored 4), 'nearly all' (scored 3), 'plenty' (scored 2), to 'none' (scored 1). Polytomous response categories are in contrast to dichotomous response categories in which only agreement or disagreement are possible. In the above example of a four-point rating scale, Wright and Linacre (1989) noted: "… the inarguable order of these labels from less to more can be used to represent them as a series of steps" (p. 1); and that "the observation of 'none' can be counted as zero steps up this rating scale, 'plenty' as one step up, 'nearly all' as two steps up and 'all' as three (p. 1)". In making this observation, the distinction between categorical and ordinal data becomes obvious and the data obtained from using these four categories should be ordinal. However, ordinality is not sufficient in itself for measurement: "This counting has nothing to do with any numbers or weights with which the categories might have been tagged in addition to or instead of their labels" (Wright and Linacre (1989, p. 1). The steps between the categories do not necessarily represent equal increments in what is being measured and the data cannot be assumed interval since the counting of steps says nothing about distances between categories (see Merbitz, Morris and Grip, 1989). For example, in the case of the above four-point scale, a score of '4' does not necessarily indicate double the amount implied in a score of '2'. Wright and Linacre (1989) commented on the connection between observations such as scores and the notion of a measure:

> "The original observations in any science are not yet measures in this sense. They cannot be measures because a measure implies the previous construction and maintenance of a calibrated measuring system with a well-defined origin and unit which has been shown to work well enough to be useful" (p. 2).

Further, they defined a 'measure' as "… a number with which arithmetic (and linear statistics) can be done, a number which can be added and subtracted, even multiplied and divided, and yet with results that maintain their numerical meaning" (p. 2).

This examination of rating scales focussed on response scale data, but as was noted in the previous section on measurement in the human sciences, the notion of measurement is also applicable to scores obtained from the multiple items comprising an instrument. As a result, Wright and Masters (1982) incorporated the theoretical requirements for measurement into a set of four measurement criteria for rating scale instruments: Uni-dimensionality; qualification; quantification; and linearity.

Uni-dimensionality: Data measures a single or dominant trait. The measurement of any object or entity describes one and only one attribute of the object measured (see Thurstone 1931).

Qualification: Data can be compared. Guttman (1950) noted with regard to instrument scales:

> "If a person endorses a more extreme statement, he should endorse **all** less extreme statements if the statements are to considered a scale" ... "We shall call a set of items of common content a scale if [and only if] a person with a higher rank than another person is just as high or higher on **every** item than the other person" (p. 62).

Compliance with Guttman's requirement can be tested by the Rasch Model which requires rank ordering of both person ability and item difficulty. When the data fit the Model, the items are ordered in relation to person ability. The more difficult items are affirmed by only the persons with higher ability whereas the easier items are affirmed by persons with lower ability as well as by those with higher ability.

Quantification: Variables are measured in common units. However, a unit of measurement is not necessarily a thing such as a piece of yardstick. Alternatively, a unit of measurement is always a process of some kind which can be repeated without modification in the different parts of the measurement continuum (see Thurstone 1931, p. 257). In a Rasch Model analysis of rating scale data, an attitudinal trait of persons is estimated in logits (logarithmic units). A logit is the logarithmic odds of a person affirming that statements of attitudes apply to himself/herself, and for all persons, the logit is the common unit of quantification.

Linearity: Data is positioned on a line or scale. Measurement implies a linear continuum of some sort such as length, price, volume, weight, or age (see Wright, 1997). When the idea of measurement is applied to scholastic achievement, for example, it is necessary to force the qualitative variations into a scholastic linear scale of some kind (see Thurstone and Chave 1929, p. 11). The process of forcing qualitative variations such as responses to Likert scale categories into a linear scale requires application of a stochastic measurement model to plot measures of person ability and item difficulty on one scale.

The Rasch Model described in the previous section was the first in what Bond and Fox (2001) viewed as the Rasch family of models. It is known as the simple logistic model or as the dichotomous model. The dichotomous Rasch Model mathematically expresses the probability of obtaining any given score as an exponential function of the difference between two parameters; person ability and item difficulty. The Rasch Rating Scale Model is an extension of the dichotomous model and includes a third parameter concerning the probability of a person preferentially choosing a particular response scale category.

The ordinality of response scale data for an individual item can be tested in relation to the person ability and item difficulty parameters. This requires estimation of the 'thresholds' between adjacent response categories. A threshold is the person ability location level (logit) at which the probabilities of persons choosing two adjacent response categories are equal. For a four-category response scale there are three thresholds: One for the first and second categories (e.g. 'none' and 'plenty'); another for the second and third categories (e.g. 'plenty' and 'nearly all'); and, a third threshold for the third and fourth categories (e.g. 'nearly all' and 'all'). The Rasch Rating Scale Model tests that the ordering of thresholds is in accord with respondent ability to affirm the items. A respondent with high ability should have a greater likelihood of selecting the more affirmative categories (e.g. 'nearly all' and 'all'). From an instrument refinement perspective, examination of thresholds shows how well the response categories provided respondents with appropriate choices and also whether the categories for particular items confounded the respondents (see Andrich, 1996). Such information can inform modification of the response scale categories (see Cavanagh, Waldrip, Romanoski, Fisher and Dorman, 2005), or rewriting of items (stem statements) in order to avoid confusing respondents in their selection of categories (Bond and Fox, 2001).

The case for constructing and using measures

When Wright and Linacre (1989) defined a 'measure', they drew attention to measures being required for conducting arithmetic operations and applying linear statistics. Similarly, Fraenkel and Wallen (2004) stated with regard to parametric tests:

> "It turns out that in most cases parametric techniques are most appropriate for interval data, while non-parametric techniques are most appropriate for ordinal and nominal data. Researchers rarely know for certain whether their data justify the assumption that interval scales have actually been used" (p. 241).

The salient point for quantitative learning environment research is that over many decades, the data from rating scale instruments has been subject to a wide range of linear statistical analyses. However, the data from the instruments used in this research have rarely been scrutinised in terms of the criteria that characterise measurement (see Waugh and Cavanagh, 2002). While the reasons for this are beyond the scope of this paper, the non-use of measures and interval data in general for statistical analyses is likely due to: First, confusion between counts and measures; and, second, that treating raw scores as measures sometimes seems to work (Wright and Linacre, 1989).

First, regarding counts and measures, the meaning assigned to the size of the steps between rating scale category data enables differentiation between counts and measures. Wright and Linacre (1989) concluded:

> "… our raw counts, as they stand, are insensitive to any differing implications of the steps taken. To get at these implied step sizes we must construct a measuring system based on a coordinated set of observed counts. This requires a measurement analysis of the inevitably ordinal observations which always comprise the initial data in any science" (p. 3).

Second, treating raw score counts as measures and using these for conducting regressions and other interval-level statistical analyses to test hypotheses can produce 'meaningful' results. For example, when raw scores are used in regression analysis, the variance in a dependent variable can often be accounted for by the effect of independent variables. Wright and Linacre (1989) explained this as a consequence of the monotonic relationship between scores and measures when data is complete and unedited, and also because correlation analyses of scores and the measures they may imply will be quite similar. However the monotonicity holds only when data are complete, that is, when every subject encounters every item, and no unacceptably flawed responses have been deleted. They concluded:

> "This kind of completeness is inconvenient and virtually impossible to maintain, since it permits no missing data and prevents tailoring item difficulties to person abilities. It is also no more necessary for measurement than it would be to require that all children be measured with exactly the same particular yardstick before we could analyze their growth. Further the approximate linearity between central scores and their corresponding measures breaks down as scores approach their extremes and is strongly influenced by the step structure of the rating scale" (p. 5).

Consequently, raw scores are not necessarily measures and it is desirable for any set of raw scores to be tested to ensure correspondence with linear measures prior to conducting statistical analyses (see Merbitz, Morris and Grip, 1989). Further: "Whatever the outcome of such a verification it is clearly preferable to convert necessarily nonlinear raw scores to necessarily linear measures and then to perform the statistical analyses on these measures" (Wright and Linacre, 1989, p. 6). To test the veracity of these assertions, Romanoski and Douglas (2002) used Monte Carlo simulations to determine the psychometric conditions under which differences between raw scores and Rasch transformations of those raw scores were detectable through two-way analysis of variance. They concluded: "The findings demonstrated the inherent inadequacy of untransformed raw scores for two-way analysis of variance" (p. 429).

Concluding comments

One purpose of this paper was to examine the implications of applying objective measurement theory to the development and use of rating scale instruments. This theory is based upon the assumptions implicit in item response theory and use of stochastic measurement models in contrast to classical test theory, true-score theory and use of deterministic models. While the thesis of the paper concerned the benefits of using stochastic analytic techniques, the authors were well aware of the strong advocacy amongst psychometricians for use of only deterministic techniques (see Waugh and Chapman, 2005). Accordingly, the paper has **not** presented an argument against using methods such as exploratory factor analysis, confirmatory factor analysis, structural equation modelling or hierarchical linear modelling. Instead, it has argued that these linear statistical analyses can work better when the data analysed are interval. Consequently, the issue arising for quantitative learning environment researchers could be seen as not so much a debate about the merits of different approaches towards instrument construction and data analysis, rather as concerning how to maximise the veracity of the inferences derived from empirical observations.

Finally, the paper has argued for the construction of measures in quantitative learning environment research and shown how this is attainable by application of objective measurement theory and stochastic measurement models such as the Rasch Rating Scale Model. While the argument was grounded on well-established principles of scientific measurement and on demonstrating the merits of using a measurement approach in rating scale construction, application of a measurement approach in scale construction should not be simplistically construed in terms of statistical analyses necessarily ensuring observations are transformable into measures or correspond to measures. Indeed, the requirements for objective measurement in the human sciences and the characteristics of human science measures are predicated on application of a process that identifies the qualitative differences between individuals and then makes judgements about how these differences can best be quantified.

References

Andrich, D. (1996). *Category ordering and their utility. Rasch Measurement Transactions,* 9(4), 464-465.

Bond, T.G. and Fox, C.M. (2001). *Applying the Rasch model: Fundamental measurement in the human sciences.* Mahwah, N. J.: Lawrence Erlbaum Associates, Publishers.

Campbell, N.R. (1920). *Physics: The elements.* London: Cambridge University Press.

Cavanagh, R.F., Waldrip, B.G., Romanoski, J.T., Fisher, D.L., and Dorman, J.P. (2005). *Measuring student perceptions of classroom assessment.* Paper presented at the 2005 Annual Meeting of the Australian Association for Research in Education: Parramatta.

Fraenkel, J.R. and Wallen, N.E. (2004). *How to design and evaluate research in education.* New York, N.Y.: McGraw Hill.

Guttman, L. (1950). The basis for scalogram analysis. In Stouffer, S.A. (Ed.). *Measurement and Prediction, Volume 4*. (pp. 60-90). Princeton, N.J.: Princeton University Press.

Merbitz, C., Morris, J., and Grip, J.C. (1989). *Ordinal scales and foundations of misinference.* Arch Phys Med Rehabil, 70, 308-332.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Chicago, IL: MESA Press.

Romanoski, J. and Douglas, G. (2002). Rasch-transformed raw scores and two-way ANOVA: A simulation analysis. *Journal of Applied Measurement,* 3(4), 421-430.

Thurstone, L.L. (1931). Measurement of social attitudes. *Journal of Abnormal and Social Psychology,* (26), 249-269.

Thurstone, L.L. and Chave, E.J. (1929). *The measurement of attitude.* Chicago, IL: University of Chicago Press.

Waugh, R.F. and Cavanagh, R.F. (2002a). Measuring parent receptivity towards the classroom environment using a Rasch measurement model. *Journal of Learning Environments Research,* 5(3), 329-352.

Waugh, R.F. and Chapman, E.S. (2005). An analysis of dimensionality using factor analysis (true-score theory) and Rasch measurement: What is the difference? Which method is better? *Journal of Applied Measurement,* 6(1), 80-99.

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers.

Wright, B.D. (1984). Despair and Hope for Educational Measurement, *Contemporary Education Review,* 3(1), 281-288.

Wright, B.D. (1997). *Measurement for social science and education: A history of social science measurement.* Chicago, IL: MESA Psychometric Laboratory.

Wright, B.D. (1997). Fundamental measurement for outcome evaluation. *Physical medicine and rehabilitation: State of the Art Reviews,* 11(2), 261-288.

Wright, B. and Masters, G. (1981). *The measurement of knowledge and attitude*. Chicago, IL: University of Chicago, Department of Education.

Wright, B.D. and Linacre, J.M. (1989). *Observations are always ordinal; measurements, however, must be interval.* Chicago, IL: MESA Psychometric Laboratory.

Wright, B.D. and Masters, G.N. (1982). *Rating scale analysis.* Chicago, Il: Mesa Press.