Validating Quantitative Measurement Using Qualitative Data: Combining Rasch Scaling and Latent Semantic Analysis in Psychiatry

Rense Lange\textsuperscript{1,2}
\textsuperscript{1}Integrated Knowledge Systems, Inc., 107 Carefree Drive, Chatham, IL 62629, USA
\textsuperscript{2}ISLA, Rua Cabo Borges, 55, 4430-646 Vila Nova de Gaia, Portugal

E-mail: renselange@earthlink.net

Abstract. An extension of concurrent validity is proposed that uses qualitative data for the purpose of validating quantitative measures. The approach relies on Latent Semantic Analysis (LSA) which places verbal (written) statements in a high dimensional semantic space. Using data from a medical / psychiatric domain as a case study - Near Death Experiences, or NDE - we established concurrent validity by connecting NDErs qualitative (written) experiential accounts with their locations on a Rasch scalable measure of NDE intensity. Concurrent validity received strong empirical support since the variance in the Rasch measures could be predicted reliably from the coordinates of their accounts in the LSA derived semantic space ($R^2 = 0.33$). These coordinates also predicted NDErs age with considerable precision ($R^2 = 0.25$). Both estimates are probably artificially low due to the small available data samples (n = 588). It appears that Rasch scalability of NDE intensity is a prerequisite for these findings, as each intensity level is associated (at least probabilistically) with a well-defined pattern of item endorsements.

1. Overview

Issues of validity have a long-standing history in measurement within the social sciences, as it is may not be clear that questionnaires and tests indeed capture what they are intended to measure, a requirement often called construct validity (Messick, 1989). A test is said to show face validity if test-takers’ (respondents’) subjective hypotheses concerning a tests’ purpose correspond to the designers’ intentions. While important in tests’ public acceptance (Holden, 2010), face-validity is not considered to be authoritative. Instead, most serious validation research relies on outside statistical criteria: predictive validity reflects the possibility of predicting particular behavioral outcomes, and concurrent validity revolves around a measure’s correlation with existing measures of the same construct. Borsboom and Markus (2013) have criticized these approaches for confusing consistency and truth, and others have argued that validity should revolve around the suitability of an instrument for application in actual practice (Fisher, 2003). Rasch scaling aids in this respect as it provides a variety of fit statistics to identify items that fail to show construct invariance and persons for whom the items do not combine to form linear measures (Michell, 2003).

The preceding notions regarding validity stand in stark contrast to the aims of qualitative psychology which focuses on understanding behavior from informants’ perspective within a dynamic
and negotiated reality (see e.g., Minchiello, 1990). According to Miles and Huberman (1984, p. 225) “In qualitative research, numbers tend to get ignored. After all, the hallmark of qualitative research is that it goes beyond how much there is of something to tell us about is essential qualities.” Exceptional in this respect are qualitative studies (Lange, et al., 2014) that encode their data in a format suitable for multi-faceted Rasch scaling (Linacre, 2013). However, most qualitative research efforts report their findings in the informants’ subjective language thereby making it difficult to judge their validity, as it is difficult to connect the findings to those obtained by other researchers.

Using data from a medical domain (psychiatry) - specifically, the reporting of near-death experiences (NDE) - this paper aims to show that purely qualitative data can be transformed into a form that is suitable for the validation of quantitative measures using unsupervised learning. The approach is based on algorithms that were recently developed in Artificial Intelligence and Natural Language Processing to perform Latent Semantic Analysis (LSA) using large data sets (see below). LSA takes respondents’ qualitative (written) accounts of their own subjective opinions, insights, and interpretations of events, and places these inside a high-dimensional semantic space. Also available is a Rasch scaled questionnaire that assesses important aspects of this domain. Standard statistical techniques will then be used to predict respondents’ locations on this dimension. Thus, the preceding can be seen as a special case of establishing concurrent validity by connecting qualitative and quantitative data. The approach involves considerable simplification, as LSA reduces the richness and complexity of data valued by qualitative researchers. It is not clear if this simplification actually enhances or detracts from the overall effectiveness of the approach.

2. Latent Semantic Analysis
Assuming that we have $d$ documents and $t$ relevant terms one can construct a matrix $X$ whose rows $l, \ldots, i, \ldots, t$ represent a dictionary of selected terms (or “tokens”), and whose columns $l, \ldots, j, \ldots, d$ represent the documents containing these terms. Each entry $X_{ij}$ reflects the weighted frequency with which dictionary term $i$ occurs in document $j$. The weighting does not affect the following overview, and the topic is further addressed in Section 3.

It has been known since last century that any matrix $X$ can be written as the product of three matrices:

$$X = TSD',$$  

(1)

where $T$ has size $t \times d$ and $D$ has size $d \times d$, both with orthonormal columns (i.e., $T' T = I$ and $D' D = I$), while $S$ is a $d \times d$ diagonal matrix. The matrices $T$, $D$, and $S$ are referred to as the left and right singular vectors and the diagonal matrix of singular values, respectively, and together these define $X$’s Singular Value Decomposition (SVD). By convention, the diagonal elements of $S$ are positive and ordered in decreasing magnitude.

The utility of SVD lies in the fact that keeping just the $k$ largest singular values (i.e., setting $S_{k+1} = S_{k+2} = \ldots = S_d = 0$) yields a matrix $X^*$ which increasingly approximates $X$ for greater $k$. Doing so produces two useful results: First, $X^*$ is a matrix of rank $k$ which is closest to $X$ in a least-square sense. Second, since the corresponding columns in $T$ and $D$ will be multiplied by 0, they can simply be deleted, yielding $T^*$ and $D^*$ with sizes $t \times k$ and $d \times k$, respectively.

LSA (see, e.g., Landauer, Folz, and Laham, 1998) uses SVD to reduce the dimensionality of the semantic space needed to represent a set documents from $d$ to $k$. The choice of $k$ is important as it should be chosen such that $X^*$ captures mainly the “real” data in $X$, while omitting the sampling error. There are no obviously proper ways of achieving this goal, and researchers therefore experiment with different choices of $k$ to achieve best results. In practice, the initial dimensionality $t$ is in the order of several thousands, whereas the reduced space has a dimensionality of $k = 300-500$.

Of course, the approach would be of little use if Eq. 1 has to be recomputed each time a new document $y$ is encountered. However, assuming that the model is correct (i.e., $X = X^*$), a document's
reduced coordinates $D_y$ (with $k$ dimensions) is a simple function of its original coordinates $X_y$ (in $d$ dimensions, $d > k$) as follows (Deerwester et al., 1990, p. 399):

$$D_y = X_y^T S^{-1}$$

While the mathematics behind SVD has long been known, the computation of the $T$, $S$, and $D$ of sufficient sizes needed for actual text analysis applications has become feasible only with the advent of modern computers and the development of efficient algorithms. Fortunately, recent years have seen a rapid development of extremely powerful methods based on the use of sparse matrices, and capable of distributed and/or parallel processing. For instance, Řehůřek and Sojka (2010) describe the Python based Gensim software capable of analyzing enormous data sets on a standard personal computer.

Table 1: Rasch Locations (logits) of NDE items for “True NDErs”

| Question                                                                 | Location |
|-------------------------------------------------------------------------|----------|
| Did scenes from the future come to you?                                | 1.69     |
| Did scenes from your past come back to you?                            | 1.30     |
| Did you seem to be aware of things going on elsewhere?                  | 1.17     |
| Did you see deceased or religious spirits?                              | 0.68     |
| Were your thoughts speeded up?                                         | 0.64     |
| Did you suddenly seem to understand everything?                         | 0.32     |
| Were your senses more vivid than usual?                                 | -0.06    |
| Did you come to a border or point of no return?                        | -0.11    |
| Did you seem to encounter a mystical being or presence?                 | -0.20    |
| Did you see, or feel surrounded by, a brilliant light?                  | -0.30    |
| Did you feel a sense of harmony or unity with the universe?            | -0.55    |
| Did you seem to enter some other unearthly world?                       | -0.61    |
| Did you have feelings of joy?                                          | -0.70    |
| Did you feel separated from your body?                                  | -0.74    |
| Did time seem to speed up or slow down?                                 | -1.07    |
| Did you have a feeling of pleasantness?                                | -1.47    |

3. Term Weighting

It is possible to base LSA directly on the frequency by which term $i$ occurs in document $j$. However, in practice better results are obtained by transforming the between term-document co-occurrence matrix into a locally/globally weighted term frequency - inverse document frequency (tf-idf) matrix with real-valued entries. Such transformed values increase proportionally with terms’ occurrence within a document, but - in order to control for the fact that some words are more common than others - these values decrease proportional to the logarithm of the frequency of the term across the entire corpus. For instance, assume that there are just three documents $d_1$, $d_2$, and $d_3$ and that some term $z$ occurs 3 times in $d_1$ and once in $d_2$. Then the term frequency $tf = 3+1+0=4$ and the logarithm of the inverse document frequency $idf = \ln 3/2 = 0.41$ (using natural logarithms), and hence $tf\cdot idf_z = tf * idf = 4 * 0.41 = 1.64$.

4. Case Study: Near-Death Experiences

It is well-known to physicians and mental health professionals that people (adults and children alike) when suddenly faced with their own death experience, a distinctive state of consciousness in which their existence is seemingly unbounded by a physical body or earthly environs. Such near-death
experiences (NDE) are among the most potent of psychological episodes (for an overview, see e.g., Holden, Greyson & James, 2009). In fact, the Diagnostic Manual of Statistical and Mental Disorders contains the V-Code category "Religious or Spiritual Problem" in part to acknowledge and guide clinicians in addressing the impact and aftereffects of NDEs and related experiences (Lukoff, 1998).

Greyson (1983) quantified NDE intensity by means of a series of questions derived from statements of individuals who shared their accounts following such experiences. Specifically, NDErs (i.e., those reporting an NDE) were asked to rate the occurrence of 16 different experiences in terms of three ordered categories which generically represent ‘not present,’ ‘mildly or ambiguously present’ or ‘definitively present.’ Indicative of its measurement validity, the NDE proved to be Rasch scaleable (Lange, Greyson, Houran, 2004).

The questions constituting Greyson’s NDE measure are listed in Table 1, together with their item difficulty in logits. Note that items at the lower end of this scale (i.e., with the highest endorsement) refer to experiences of peace, joy and harmony, followed by finding insight and mystical or religious experiences at intermediate levels, while items with the highest values (least endorsed) refer to an awareness of events occurring in a different place or time. By the very nature of Rasch scaling, this sequence defines a true hierarchy such that events mentioned higher in this sequence become salient only after of those lower in the sequence have already been reported - at least probabilistically. In other words, there exists a well-defined relation between NDE intensity and the verbal, i.e., qualitative, contents of the NDE.

Naturally, those who have not had near death experiences cannot be expected to provide meaningful information about NDE. Therefore, Table 1 shows the item locations as computed for those that were classified psychiatrically as having had “True NDE” only (cf, Greyson, 1983).

Table 2: Summary of multiple linear regression runs

| Number of dimensions | 1-150 | 1-200 | 1-250 | 1-300 | 1-350 |
|----------------------|-------|-------|-------|-------|-------|
| \( R^2 \)            | 0.49  | 0.56  | 0.64  | 0.70  | 0.75  |
| \( \text{Adjusted } R^2 \) | 0.29  | 0.29  | 0.33  | 0.31  | 0.28  |

An independent sample of 833 NDE accounts was available based on which a semantic space was constructed using Řehůřek and Sojka’s (2010) Gensim software. This space used a dictionary with 1500 tokens (terms) that were deemed relevant, and the resulting 1500 x 833 token x document matrix was decomposed according to Eq. 1. A total of 588 of the 833 people also had completed Greyson’s NDE scale, and their Rasch NDE intensity estimates were combined with the their accounts’ coordinates in the semantic space as computed using Eq. 2. Using the \textit{glm} procedure provided by the R language, the first 50, 100, …, 400 coordinates (ordered by importance) were used to predict respondents’ estimated Rasch NDE intensity using standard multiple regression.

Not surprisingly, Table 2 shows that using increasingly complex semantic spaces adds predictive information, as \( R^2 \) increases continuously when more predictors are added. Consistent with other research, the corresponding \( R^2 \) values when adjusted for sample size and number of predictor variables reach a maximum (i.e., 0.33) for about 250 predictor variables and this value is perhaps a more appropriate estimate of the quality of prediction. Nevertheless, the finding that at least 1/3 of the variation in respondents’ quantitative Rasch NDE measures can be predicted from their qualitative NDE accounts strongly supports the notion that qualitative data can be used to validate quantitative variables. The multiple correlation between variance explained by the first 250 factors and their weights in the multiple regression equation is essentially zero \( (r = -0.06) \), i.e., the semantic factors do not contribute to prediction in proportion to the variance they explain in the SVD context.

It is noted that respondents’ age could reliably be predicted from the semantics from their
NDE accounts ($R^2 = 0.25$). However, the multiple correlation with NDErs gender proved negligible.

6. Conclusion
A standard approach to establishing content validity is to correlate the responses to one set of questions to sets of questions purporting to measure the same construct. This approach need not be limited to comparisons between quantitative measures since that Latent Semantic Analysis can be used to capture the meaning of free style written descriptions in order to establish a correlation with other a Rasch scaled variable. It is my conjecture that one variable’s Rasch scalability may well prove crucial in this context, as Rasch scalability defines an “item hierarchy” which essentially imposes an orderly progression on the semantic labels corresponding to variables’ quantitative properties.

It should be noted that analysis of textual data is not limited by the size of the text corpus, economic factors, or hardware constraints. For instance, Řehůřek and Sojka’s (2010) Gensim used here is freely available under the OSI-approved GNU LGPL license. This software is capable of analyzing the complete English Wikipedia by decomposing a sparse matrix with 3.9 million rows x 100,000 dictionary terms with about 760 million non-zero entries in about 13 hours on a standard PC.

References
Borsboom, D. and Markus, K.A. (2013). Truth and evidence in Validity Theory. Journal of Educational Measurement, Vol. 50, No. 1, pp. 110–114.
Deerwester, S., Dumais, S.T., Furnas, G.W., Landauer, T.K. and Harshman, R. (1990). Indexing by Latent Semantic Analysis. Journal of the American Society for Information Science. 41, 391-407.
Fisher, W. P., Jr. (2003) Mathematics, measurement, metaphor, metaphysics: Part I. Implications for method in postmodern science. Theory & Psychology, 13(6), 753-90.
Miles M.B. and Huberman A.M. (1984). Qualitative Data Analysis: A Sourcebook of New Methods. Newbury Park, CA: Sage.
Greyson, B. (1983). The Near-death Experience Scale: Construction, reliability, and validity. Journal of Nervous and Mental Disease, 171, 369-375.
Holden, J. M., Greyson, B. and James, D. (eds.). (2009). Handbook of Near-Death Experiences: Thirty Years of Investigation. Santa Barbara, CA: Praeger/ABC-CLIO.
Holden, Ronald B. (2010). "Face validity". In Weiner, Irving B.; Craighead, W. Edward. The Corsini Encyclopedia of Psychology (4th ed.). Hoboken, NJ: Wiley. pp. 637–638.
Landauer, T. K., Foltz, P. W. and Laham, D. (1998). Introduction to Latent Semantic Analysis. Discourse Processes, 25, 259-284.
Lange, R., Dorsey, K., Wells, N., Roberts, N. and Verhulst, M. (2014). Rasch analysis of students’ qualitative evaluations of faculty professionalism. Manuscript submitted for publication. Southern Illinois School of Medicine.
Lange, R., Greyson, B. and Houran, J. (2004). A Rasch scaling validation of a ‘core’ near-death experience. British Journal of Psychology, 95, 161-177.
Linacre, J. M. (2013). Facets computer program for many-facet Rasch measurement, version 3.71.2. Beaverton, Oregon: Winsteps.com.
Lukoff, D. (1998). From Spiritual Emergency to Spiritual Problem: The Transpersonal Roots of the New DSM-IV Category. Journal of Humanistic Psychology, 38, 21-50.
Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational measurement (3rd ed., pp. 13– 103). Washington, DC: The American Council on Education and the National Council on Measurement in Education.
Michell, J. (2003) Measurement: A beginner’s guide. Journal of Applied Measurement, 4(4), 298-308.
Minichiello, V. (1990). In-Depth Interviewing: Researching People. Longman Cheshire.
Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, La Valetta, Malta, p. 45–50.