A composite score for a measuring instrument utilising re-scaled Likert values and item weights from matrices of pairwise ratios

Authors:
Piet J. Becker[1,2]
J.S. Wolvaardt[3]
Angie Hennessy[4]
Carin Maree[5]

Affiliations:
[1]Biostatistics Unit, Medical Research Council, South Africa

[2]Department of Clinical Epidemiology, University of Pretoria, South Africa

[3]Department of Decision Sciences, University of South Africa, South Africa

[4]Tygerberg Hospital, South Africa

[5]Department of Nursing Science, University of Pretoria, South Africa

Correspondence to:
Piet J. Becker

e-mail:
pbecker@mrc.ac.za

Postal address:
Division of Clinical Epidemiology, University of Pretoria, P.O. Box 667, Pretoria, 0001

Keywords:
subjective judgement theory; Likert scale; weights; measuring instrument; pairwise comparison

Dates:
Received: 29 July 2008
Accepted: 27 Nov. 2008
Published: 18 June 2009

How to cite this article:
Becker, P.J., Wolvaardt, J.S., Hennessy, A. & Maree, C. 2009. A composite score for a measuring instrument utilising re-scaled Likert values and item weights from matrices of pair wise ratios. *Health SA Gesondheid* 14(1), Art. #412, 4 pages. DOI: 10.4102/hsag.v14i1.412

This article is available at: www.hsag.co.za

ABSTRACT

A methodology is proposed to develop a measuring instrument (metric) for evaluating subjects from a population that cannot provide data to facilitate the development of such a metric (e.g. pre-term infants in the neonatal intensive care unit). Central to this methodology is the employment of an expert group that decides on the items to be included in the metric, the weights assigned to these items, and an index associated with the Likert scale points for each item. The experts supply pairwise ratios of an importance between items, and the geometric mean method is applied to these to establish the item weights – a well-established procedure in multi-criteria decision analysis. The ratios are found by having a managed discussion before asking the members of the expert panel to mark a visual analogue scale for each item.

OPSOMMING

'n Metode word aangebied waarmee 'n meetinstrument (metriek) ontwikkel kan word vir die evaluering van persone uit 'n populasie wat nie self die data vir die ontwikkeling van die metriek kan voorsien nie (bv. vroeggebore babas in die neonatale intensiewe sorgeenheid). Die kern van hierdie werkswyse is die gebruik van 'n deskundige groep wat die items vir die meetinstrument kies, gewigte aan die items toeken, en vir elke item 'n indeks opstel wat met die Likert-skaal punte geassosieer word. Die deskundiges het paarsgewyse verhoudings tussen items verskaf en die meetkundig-gemiddelde metode is hierop toegepas om die itemgewigte te verkry – 'n goedgevestigde gebruik in meerdoelwitbesluitkunde. Die paarsgewyse verhoudings is gewerf deur die deskundiges, na 'n bestuurde bespreking, vir elke item 'n visuele analoogskaal te laat invul.

INTRODUCTION

In this article, a methodology is proposed to develop a measuring instrument (metric) for evaluating a variety of conditions and situations. It is particularly valuable when the population to be evaluated cannot participate in the construction of the metric, e.g. for item reduction. It is generally useful in its use of weights to put the appropriate emphasis on the items included, and in putting values on the equidistant Likert scale points. The methodology is presented in the context of a case study in which the stress levels in pre-term infants are to be measured.

The problem originated during a research study that aimed to measure pre-term infants' stress levels before and after developmentally supportive positioning, but an appropriate measuring instrument was not available (Hennessy, Maree & Becker 2007:3-11). Contrary to situations such as psychometric testing, in which the subjects (from the population for which a metric needs to be developed) participate, usually by way of a self-administered questionnaire, it is impossible for pre-term infants in the neonatal intensive care unit (NICU) to furnish data that can assist with the development of an instrument capable of measuring stress levels in the pre-term infant.

In this setting, the development of a suitable metric relied on inputs from two expert panels that contributed in three ways: to determine the items that needed to go into such an instrument (first panel), and to allocate weights to the items and to re-scale the Likert scale points for each item (second panel). The need for weights is because items very seldom contribute equally to the composite score of a metric and weights put emphasis on items according to their contribution.

We refer to Likert scale points (rather than values) and reserve the term values for numerical values associated with the Likert scale points. For a particular item one may associate with the five Likert scale points 0, 1, 2, 3, 4 used here the values 0,00; 0,12; 0,35; 0,72; 1,00 indicating that the condition associated with Likert point 2 is 0,35 on a scale from 0 to 1. The values are not equidistant, but attempt to represent the severity of the condition associated with the particular Likert point. In this example, the value associated with Likert point 3 is six times that of Likert point 1 (0,72 versus 0,12).

METHODS

Input by expert groups

Initially, a group of experts (first panel) was chosen and consulted individually, by written correspondence and telephonic conversation, to determine the items that went into the metric (see Table 1, which illustrates the Hennessy Stress Scale for the pre-term infant (HSSPI)). The $n=15$ items included in the HSSPI were decided on by consensus among the members of the first panel. Subsequently, a second group of experts, hereafter referred to as the panel, was used to provide the information needed to estimate the required item weights and numerical values to be associated with the Likert scale points.

The panel originally consisted of ten members, but three members were excluded due to collaboration and untrustworthiness of their inputs, which showed external influences. The remaining expert group, consisting of m=7 members, were provided with the 15 item HSSPI and, for each item individually, every point on the five-point Likert scale (0, 1, 2, 3, 4) was clearly described. To start with, panel members

Journal of Interdisciplinary Health Sciences
Article #412

were expected to indicate for each item where the Likert scale points (1, 2, 3) lie on a visual analogue scale (VAS) beginning at 0 and ending at 4. The values for these points were set equal to their distances from 0 and were then re-scaled to fall between 0 and 1. For each of the 15 items, this resulted in the five Likert scale points (0, . . . , 4) being well described in terms of a medical condition, and a numerical value of between 0 and 1 associated with each. The numerical values were supplied by the panel and correspond to the severity of the conditions.

The weights of the items remained to be found. A well-established method in multi-criteria decision analysis (MCDA) was employed – the use of pairwise ratios of importance (Belton & Steward 2002:132; Lootsma 1999:53). Each of the panel members contributed to an own $n \times n$ judgement matrix $A = (a_{ij})$ with the ij-th element a_{ij} denoting the ratio between w_i and w_j where w_i is the importance of item *I* and a_{ij} therefore represents the importance of item *i* relative to item *j*. The value for a_{ij} follows when the panel member denotes with a single marking on a VAS the importance of item *i* relative to item *j*. The mark divides a bar of standard length into a left part, denoting the importance of item *i*, and the (remaining) right part indicating the importance of item *j*. The length of the left divided by that of the right assigns a numerical value of a_{ij}.

Estimation of distances between Likert scale values

For item *i* (*i* = 1, 2, . . . , *n*) denote the distance on the VAS between the Likert scale value *j* (*j* = 1,2,3) and the left-most value 0, as suggested by panel member k, with $d_{ij}^{(k)}$, then for item *i* the distance to *j* is estimated as the mean value for the *m* panel members.

i.e. $$\hat{d}_{ij} = \frac{1}{4m}\sum_{k=1}^{m} d_{ij}^{(k)}$$

The Likert scale value 4 is situated at the right-most end of the VAS. Replace the original Likert scale values with weights equal to the latter distances bounded by 0 and 4, re-scale these weights to range from 0 to 1, and denote them by $w_{i,j}$ (i = 1, 2, . . . , n ; j = *0,1,…, 4*).

Estimation of item weights

The problem of calculating a preference vector from the ratios entered in a judgement matrix can be presented as a general linear model (Crawford & Williams 1985:393-4). To find the weights of the *n* items of the metric, defined as the sum of the weighted scores for the *n* items, each of the *m* panel members compiled a $n \times n$ judgement matrix $A = (a_{ij})$.

Assume an underlying vector $w^t = (w_1,...,w_n)$. The a_{ij} are estimates of ratios of the elements of w with random error. Panel member k (via the VAS) supplied estimates $a_{ij}^{(k)}$ to estimate w_i/w_j.

Random errors $f_{ij}^{(k)}$ are introduced and the elements of $A^{(k)}$, supplied by panel member k, are

$$a_{ij}^{(k)} = \frac{w_i}{w_j} f_{ij}^{(k)}, \text{ for } i = 1,...,n-1, \; j = i+1,...,n$$

The indices point to the upper diagonal of $A^{(k)}$ only. This is because $a_{ii}^{(k)} = 1$, and $a_{ji}^{(k)}$ is the reciprocal of $a_{ij}^{(k)}$.

Taking logs of $a_{ij}^{(k)} = \frac{w_i}{w_j} f_{ij}^{(k)}$, it follows that $\ln a_{ij}^{(k)} = \ln w_i - \ln w_j + \ln f_{ij}^{(k)}$

Set $e_{ij}^{(k)} @ \ln(f_{ij}^{(k)})$ then

$$\ln a_{ij}^{(k)} = \ln w_i - \ln w_j + e_{ij}^{(k)} \quad (1)$$

In matrix notation, (1) can now be expressed in the general linear model form $y = Xb + e$ with $\ln a_{ij}^{(k)}$ the general element of the observation vector y, $\ln w_i$ the general element of the coefficient vector b, and $e_{ij}^{(k)}$ the general element of the vector e, where $e_{ij}^{(k)} \sim N(0, \sigma_e^2)$, and the different elements of the design matrix X,

$$X = \begin{bmatrix} 1 & -1 & 0 & 0 & \cdots & 0 & 0 \\ 1 & 0 & -1 & 0 & \cdots & 0 & 0 \\ \vdots & & & & & & \vdots \\ 1 & 0 & 0 & 0 & \cdots & 0 & -1 \\ 0 & 1 & -1 & 0 & \cdots & 0 & 0 \\ \vdots & & & & & & \vdots \\ 0 & 0 & 0 & 0 & \cdots & 1 & -1 \end{bmatrix}$$

with a column for each item, can take the values –1, 0 or 1, e.g.

From (1) the first line of X would give $\ln a_{12}^{(1)} = \ln w_1 - \ln w_2 + e_{12}^{(1)}$. The m panel members supplied replicates and this was accommodated by the number of equations in $y = Xb + e$, which is equal to the total number of entries in all the matrices $A^{(k)}$.

The item weights $w_1,...,w_n$ with $w_i = \exp(\ln w_i)$ follow by determining the vector b that minimises the sum of squares

$$\sum_{i=1}^{n-1}\sum_{j>i}^{n}\sum_{k=1}^{m} [\ln a_{ij}^{(k)} - (\ln w_i - \ln w_j)]^2 \quad (2)$$

using ordinary least-squares regression software and not fitting the constant. Generally it is convenient to re-scale the weights w_i so that they add up to 100. As n increases, the $n \times n$ matrix A^k is enlarged and it becomes unfeasible for each panel member to assess every position in the matrix.

Since $a_{ij} = \frac{1}{a_{ji}}$ it is only necessary to consider the $\frac{1}{2}n(n-1)$ entries for which $i < j$ in order to estimate the w_i. However, if this number is still too large, the estimation of the w_i can be based on a sample of the a_{ij}, selected in such a way that all *i* and *j* are connected. For the current problem, 61 of the possible $105 = \frac{1}{2}15 \times 14$ a_{ij} were considered.

A further generalisation of (2) is to allow the m panel members to fill in different cells in their $A^{(k)}$. The same comparisons by different panel members are regarded as replicates. With or without this generalisation, the assumption is that the panel members are able to supply replicates. In general, even the handpicked members of an expert panel are not that similar, and the effect of panel members should be accounted for in the analysis. In this study, the latter was done by adding 'panel member' as a fixed effect to our model.

The metric or composite score, out of 100, for the HSSPI is

$$\sum_{i=1}^{15}\sum_{j=0}^{4} w_i w_{ij} I_j \quad (3)$$

where $I_j = 1$ if item *i* takes the point *j* on the Likert scale and = 0 otherwise. The items are taken from Table 1, the values w_{ij} from Table 2 and the weights w_i from Table 3.

RESULTS

The values w_{ij} assigned to the Likert scale points by the expert panel are given in Table 2, and the item weights w_i determined using Stata Statistical Software Release 8.0 (Statacorp 2003), both when assuming panel members are replicates (Crawford & Williams 1985:395) and when including panel members as a fixed effect in the regression model, are given in Table 3.

An example of how (3) is computed: Suppose the Likert values for items 1 to 15 for a given infant were 1; 0; 3; 2; 4; 1; 1; 3; 3; 2; 0; 1; 3; 2; 1, then the composite score when panel member is a fixed effect is

(7.19)(0.28) + (5.51)(0) + (5.56)(0.76) + (5.06)(0.51) + (4.79)(1)
+ (7.01)(0.21) + (6.86)(0.25) + (7.85)(0.76) + (9.69)(0.78) + (8.72)(0.57)
+ (7.92)(0) + (7.27)(0.22) + (6.90)(0.78) + (4.58)(0.49) + (5.09)(0.16)
= 45.53 (score out of 100)

The HSSPI, now weighted, was then used in a pre-test/post-test design to observe infants prior to a specific positioning intervention, and again after the intervention (Hennessy *et*

TABLE 1
Hennessy Stress Scale for the Preterm Infant - Items

		4	3	2	1	0
Neurological System	**1. Neck & back**	severe jitteriness hyperextension hypertonia hypotonia	mild jitters	sensitivity to moro reflex stimulus fright	neck stiffness	reflexes correct for GA flexed head
	2. Extremities, fingers & toes	flaccid flexor or extensor spasm hypertonic white knuckles	flying movements extension kicks frantically poor attempt at movement grasping hands poor grip	increased muscle tone in hands hands clasped together pushing with feet	lower than normal muscle tone attempts flexion holds on to object fists	relaxed midline flexion self-regulating behaviour hands to mouth co-ordination feet flexed
	3. Crying	hysterical crying high pitched never cries	very upset crying constant moaning poor attempt to cry	upset crying irritability moaning	cries for attention	not crying and relaxed mouth
	4. Face	gaping mouth drifting eyes half-open no focus no response	grimace frown tonguing continuously squeezes eyes closed	tonguing at times yawn sneeze blinks eyes tightly	pulls faces Visual aversion	relaxed face relaxed mouth rooting reflex sucks pacifier eyes fixed on an object
	5. Sleep – wake cycle	no sleep	periods of light sleep no deep sleep	at times establishing	periods of deep sleep but wakes easily	>30 min periods of deep sleep
Respiratory System	**6. Respiration rate (/min)**	<20 >120	20-30 100-119	30-35 81-99	35-40 61-80	40-60
	7. Respiratory sounds	gasping	severe expiratory grunting	moderate expiratory grunting	mild expiratory grunting	no expiratory grunting
	8. SpO_2 (%)	<85	85-87	88-90	91-93	94-100 or apnoea monitor
Cardiovascular System	**9. Heart rate *(beats/min)***	<80 >200	80-89 181-200	90-99 161-180	100-119 141-160	120-140
	10. Heart rhythm	irregular		irregular at times		regular
	11. Blood pressure *(mmHg)* Gestational age = GA	BP mean 2 or more > GA or > 6 > GA	BP mean 1 < GA	BP mean equal to GA	BP mean 2 > GA	BP mean 2 - 6 > GA
	12. Skin colour	blue or grey mottled purple	pale grey areas turns red when crying	ashen	pale pink	pink
	13. Perfusion	central and peripheral cyanosis	centrally pink peripheral cyanosis	cold extremities	cool extremities centrally pink peripherally pink	pink tongue no cyanosis warm extremities
GIT – System	**14. Nutrition per 3 hour period**	no absorption	absorbs < 25%	absorbs 25-50%	absorbs 50-75%	absorbs >75%
	15. GIT related responses	abdominal distension vomiting	signs of nausea visible gaga reflex	hiccups cramps (cries and pulls legs up)	breaking wind	peaceful after feed

al. 2007:3-11). The composite score expressed stress level as a percentage. The pre-test stress scale was performed as the pre-term infant was waking up for the three-hourly routine before care commenced. Routine care was then done and, once routine care was completed, the pre-term infant was positioned according to specific principles with the use of the positioning aids. The infant was left for three hours without unnecessary disturbance. Before the following routine care commenced, the post-test stress scale was performed to determine whether the intervention had been successful. This would be confirmed if the stress levels measured lower on the post-test than the pre-test stress scale. The results of the study were published in a previous article by Hennessy *et al*. (2007:3-11).

DISCUSSION

A wide variety of conditions and phenomena manifest themselves not as a single measurable attribute, but as a phenomenon that is intuitively understood but not easily measured, mainly because there is no single attribute. In these many cases, researchers resort to constructing a composite score that takes into account a number of aspects (called items) of the phenomenon considered. The process starts with the nomination of (supposedly) all the relevant items and proceeds to cull them in what is called item reduction. Item reduction removes the items that are not contributing to the measurement because they are not relevant, or because they coincide with one or more of the other items and are made redundant by the presence of these items. The item reduction is frequently done by the statistical analysis of large samples of questionnaires. We propose the use of a panel of experts both for the nomination and the reduction of the items.

A panel of experts would then consist of a group of expert individuals with special knowledge or skill in a subject (The Concise Oxford Dictionary 1990:411), working together to produce a desired result As discussed by De Vos (1998:180), literature that exists in any discipline usually represents only a section of the knowledge of people involved in a specialised field on a daily basis. An expert panel can contribute the knowledge relevant to the metric to be constructed. They bring explicit knowledge and a wealth of experience that cannot be gleaned from any number of questionnaires filled in by the subjects of the study. In particular, they contribute to the clarity and relevance of the selected items (Gauthier & Froman 2001:301).

Once the items have been identified, the metric is constructed as a composite score, where the items (frequently) have a number of possible outcomes. Most metrics do not weigh the items, meaning that every item is as important as any other. We agree with Lynn (1986:382), who argues that different items contribute at different levels and have different content validity ratings or weights, and propose that an expert panel be asked for inputs into a process for the calculation of the weights. The process

Journal of Interdisciplinary Health Sciences Article #412

TABLE 2
Expert panel weights for the Likert scale points

	LIKERT SCALE POINTS				
	0	1	2	3	4
ITEM	Values ($w_{i,l}$) for Likert Scale Points				
1	0	0.28	0.57	0.78	1
2	0	0.21	0.52	0.78	1
3	0	0.17	0.47	0.76	1
4	0	0.23	0.51	0.76	1
5	0	0.20	0.44	0.74	1
6	0	0.21	0.47	0.75	1
7	0	0.25	0.51	0.75	1
8	0	0.19	0.51	0.76	1
9	0	0.23	0.51	0.78	1
10*	0	n/a	0.57	n/a	1
11	0	0.23	0.47	0.74	1
12	0	0.22	0.48	0.75	1
13	0	0.24	0.52	0.78	1
14	0	0.24	0.49	0.74	1
15	0	0.16	0.43	0.73	1

* Heart rhythm was assessed regular (0), irregular at times (2) or irregular (4)

TABLE 3
Item weights following analysis of panel members' Subjective Judgement Matrices

	Item Weights (w_i) when panel member is a	
ITEM	Fixed effect	Replicate
1	7.19	7.78
2	5.51	5.88
3	5.56	5.86
4	5.06	5.30
5	4.79	4.95
6	7.01	7.18
7	6.86	6.95
8	7.85	7.85
9	9.69	9.57
10	8.72	8.51
11	7.92	7.67
12	7.27	6.94
13	6.90	6.54
14	4.58	4.30
15	5.09	4.70
Total	100.00	100.00

is based on a well-known and frequently used technique from multi-criteria decision analysis. We propose that the eliciting of answers be based on a VAS because of its direct appeal to the mental model being queried, thus avoiding various well-documented problems with semantic scales (Belton & Steward 2002:132; Lootsma 1999:53).

Many metrics allocate equidistant values to the different outcomes of an item, e.g. most metrics that use the Likert scale. Considering Table 2 above, it is clear that this is not necessarily correct and can even be seen as mostly wrong. It is possible to construct cases where this practice gives excessively wrong answers, but it is always preferable to eliminate even the small measuring errors. For this purpose we propose that the Likert scale be seen as consisting of an index (the Likert scale point) that takes on values $0,\ldots,m$, a description (in this case study a clinical description) associated with each Likert scale point, and a value associated with the description and associated Likert scale point. We also propose that the values be found by using a VAS. Finally, the metric is the weighted average of the Likert values of the items.

For the case study, the compilation of the panel was based on an identification of the disciplines relevant to the problem. This increased the reliability and validity of the study, as a wider perspective was achieved. Triangulation of all the main sources for the accumulation of the stress scale content was done to enhance the reliability and validity of the stress scale, including clinical observation, expert opinion, theory and empirical research. The stress scale was based on conceptual definitions and concepts of the research to ensure that one base of knowledge was used for research and instrument development. The stress scale was also pilot-tested to allow for revision and alteration before data collection commenced.

A possible shortcoming involving incongruence between study conceptualisations and scale content was reduced by providing the expert panel with the research proposal and stress scale for feedback prior to the meeting of the expert panel.

CONCLUSION

A methodology was discussed according to which the employment of an expert group is central in deciding on the items to be included in the metric, the weights assigned to these items and, for each item, an index associated to the Likert scale points. The experts supply pairwise ratios of an importance between items, and the geometric mean method is applied to these to establish the item weights – a well-established procedure in multi-criteria decision analysis. The ratios are found by having a managed discussion before asking the members of the expert panel to mark a visual analogue scale for each item.

This methodology is proposed to develop a measuring instrument (metric) for evaluating subjects in a population that cannot provide data to facilitate the development of such a metric (e.g. pre-term infants in the neonatal intensive care unit).

REFERENCES

Allen, R.E. (ed.). 1990. *The concise Oxford dictionary of current English*. 8th edn. Oxford: Clarendon Press.

Belton, V. & Steward, T.J. 2002. M*ultiple criteria decision analysis: an integrated approach.* Boston: Kluwer Academic.

Crawford, G. & Williams, C. 1985. A note on the analysis of subjective judgement matrices. *Journal of Mathematical Psychology* 29, 387-405.

De Vos, A.S. 1998. *Research at grass roots: A primer for the caring professions*. Pretoria: Van Schaik.

Gauthier, D.M. & Froman, R.D. 2001. Preferences for care near the end of life: Scale development and validation. *Research in Nursing & Health* 24, 298-306.

Hennessy, A., Maree, C. & Becker, P. 2007. The effects of developmentally supportive positioning (DSP) on preterm infants' stress levels. *Health SA Gesondheid* 12(1), 3-11.

Lootsma, F.A. 1999. Multi-criteria decision analysis via ratio and difference judgement. Dordrecht: Kluwer Academic.

Lynn, M.R. 1986. Determination and quantification of content validity. *Nursing Research* Nov/Dec 34(6), 382-385.

Statacorp. 2003. *Stata statistical software: release 8.0.* College Station: Stata Corporation.