Some fundamental issues in oral health-related quality of life research

Kieffer, J.M.

CHAPTER 7
Model Specification in Oral Health-Related Quality of Life Research

INTRODUCTION

Because current dental practice increasingly recognizes the psychosocial impact of oral diseases, the patient's view and, by implication, the use of patient-reported outcomes, have become imperative. Among other things, this understanding has led to the development of various multi-item questionnaires measuring oral health-related quality of life (OHRQoL) (Gift et al. 1995, McGrath et al. 1999). Because the items in these questionnaires are presumed to tap into a distinct domain, namely OHRQoL, it is also assumed that items are positively correlated. This assumption underlies the common use of statistical methods such as factor analysis for the construction and validation of questionnaires or of Cronbach's alpha for estimating the reliability of questionnaires, in the construction and analyses of OHRQoL questionnaires. This study assesses the tenability of the hypothesized measurement model of OHRQoL, and also determines whether the methods used for developing and analyzing OHRQoL questionnaires are appropriate.

Formative and reflective measurement models

In the social sciences, theory-building is particularly important as a theory explains how concepts such as oral health and quality of life (QoL) are related. Hypotheses on the nature of these relationships between concepts steer the selection of the measurement model that underlies the methods and techniques to be used for developing and analyzing questionnaires (Edwards et al. 2000). These measurement models represent connections between a latent variable and its observable variables. There are many possible measurement models, two of which are discussed in this section: the reflective measurement model (also referred to as an indicator model); and the formative measurement model (also referred to as a causal model).

In a reflective measurement model, a change in the variables observed is caused by a change in the latent variable (Bollen 1984). This can best be explained by using the latent variable of intelligence as an example. Intelligence is an unobservable variable that can be measured and tested only through observable variables such as verbal fluency or mathematical ability. Because the variable intelligence is reflected in the observable variables verbal fluency or mathematical ability, the underlying measurement model is reflective. In a reflective measurement model, the latent variable is therefore explanatory, and the variables observed are dependent (Bollen et al. 1991), as illustrated in Fig. 1A.

Conversely, in a formative measurement model a change in the latent variable is caused by a change in the variables observed. For instance, the concept of socioeconomic status (SES) can be measured by observable variables such as education and income. These variables must be considered determinants, rather than effects, of SES, and fit a formative measurement model instead of a reflective one (Bollen et al. 1991). By contrast to the reflective measurement model, the variables observed are explanatory and the latent variable is dependent (Bollen et al. 1991), as shown in Fig. 1B.

The assumption of Local Independence

The previous section outlines the difference between the reflective and the formative measurement models. However, there is also a statistical difference, in that only the reflective measurement model assumes local independence (i.e. the assumption that the variables observed are statistically or conditionally independent given an individual score

on the latent variable) (Borsboom et al. 2003). Thus, the latent variable explains why the observed variables are correlated. In the example of intelligence, local independence would mean that the observable variables – verbal fluency and mathematical ability – are correlated only because each is related to the latent variable, intelligence. Consequently, high intelligence affects not only verbal fluency, but also mathematical ability. This is not to say that a person is equally good at both, but consistently better at both than someone with a lower intelligence.

In the social sciences, many, if not most psychological and social concepts are presumed to have an underlying reflective measurement model and many questionnaires are developed in such a way that the answers to the questions reflect an unobservable latent variable. Hence, a reflective measurement model is assumed, implying local independence of the variables observed.

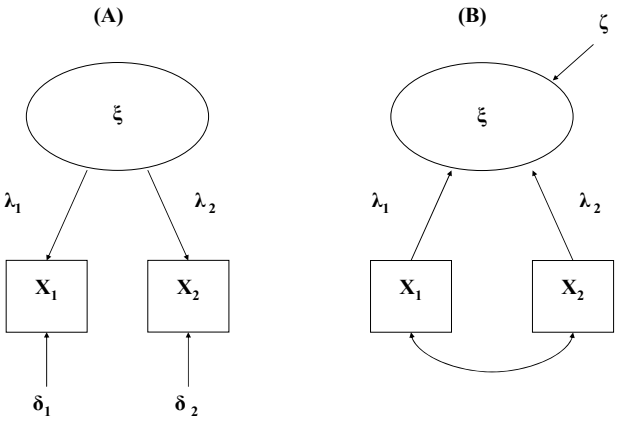

Fig 1. Path diagram of reflective (A) and formative (B) measurement models (Bollen, 1984). (A) λ_1 and λ_2 give the expected effect of the latent variable ξ on the observable variables X_1 and X_2 and δ_1 and δ_2 are random errors of measurement. (B) X_1 and X_2 are the variables observed which predict the latent variable ξ, and ζ is a residual term.

APPLICATION TO OHRQOL

Predicament: model specification

It has been argued that health-related quality of life (HRQoL) questionnaires typically consist of two types of items (Fayers et al. 1997a, Fayers et al. 1997b, Fayers et al. 2002). The same can be said about OHRQoL questionnaires.

The first type concerns psychosocial impacts. These items, which are reminiscent of the observables in the example of intelligence, are merely reflective of OHRQoL. They are expected to be correlated, because any covariance between these items can be ascribed only to their relationship with the common cause (Fayers et al. 1997a, Fayers et al. 2002). Thus, for example, people reporting poor OHRQoL would report that they felt both upset and tense. Accordingly, these items behave consistently with a reflective measurement model.

The second type of items describes symptoms or side effects. If one could imagine that the occurrence of a symptom or side-effect would diminish one's OHRQoL, these

types of items are determinants of OHRQoL, rather than reflections. Therefore, these types of items could covary, regardless of their relationship with the common cause (Fayers et al. 2002). Likewise, while a symptom is sufficient to diminish one's OHRQoL, it need not be necessary. Accordingly, a person does not have to suffer from all, or even any, symptoms to experience a change in his or her OHRQoL (Fayers et al. 1997a), because a person reporting poor OHRQoL does not necessarily exhibit symptoms such as bad breath or toothache. Hence, there need not be a correlation between these items, which shows these items to behave consistently with a formative measurement model. Although this has a vast impact on their construction and evaluation, the fact that many OHRQoL questionnaires consist of both formative and reflective items is generally ignored.

Practical dilemma

The traditional psychometric approach for developing and analyzing questionnaires emanates from the idea that items in a questionnaire reflect variation in the latent variables (Diamantopoulos 2008), thus assuming local independence. Consequently, methods of item-selection, validation, reliability assessment, scaling, and analysis are based on reflective measurement models (Fayers et al. 2002). Perhaps the most central criterion in measurement practice to date is internal consistency, both in constructing and validating questionnaires. This is characterized by the fact that items related to the measured construct should be positively correlated with each other. Thus, it has become common practice in many measurement endeavours to screen correlation matrixes for items that cluster together and to discard items that do not (Bollen et al. 1991). Therefore, we now address several practical concerns, attributable to the coincidence of both formative and reflective items, pertaining to some conventional statistical methods used in OHRQoL research.

Reliability and validity

Two highly valued concepts in measurement practice are reliability and validity. Reliability refers to the extent to which scores on a questionnaire are reproducible, all other things being equal (Hays et al. 1993). Here, only the internal reliability aspect is considered. The most frequently used method for assessing the internal reliability of a questionnaire is Cronbach's alpha. In statistical terms this is the ratio of the inter-item covariance to the total variance. Accordingly, as an estimate of the test variance caused by a common factor among the items, Cronbach's alpha reports how much a test score depends upon that factor (Cronbach 1951). A questionnaire is therefore considered to be a reliable measure of a latent variable if that variable underlies all items (Huysamen 2006). Thus, the degree to which a questionnaire is reliable depends strongly on positive correlations among items. However, because formative items can be positively or negatively correlated, or not correlated at all (Bollen et al. 1991), assessing the internal reliability of questionnaires containing formative items would give misleading results. To illustrate, in an article on item-order effects (Kieffer et al. 2008a) the Oral Health Impact Profile (OHIP) (Slade et al. 1994) was used to measure OHRQoL. The OHIP contains items such as 'Have you had painful aching in your mouth?' Intuitively, one would identify 'painful aching in your mouth' as a determinant of OHRQoL, and classify this item as formative. Nevertheless, this study reported an internal reliability of the total scale of $\alpha = 0.95$. Normally, one would be happy to find such an estimate, were it not for the presence of formative items in the

OHIP. Then, how could Cronbach's alpha be so high? The logical explanation would be that correlations between the formative items in this questionnaire were a result of other factors, perhaps more situational in nature, than the intended construct, OHRQoL (Fayers et al. 2002). In any case, the estimate says nothing about the internal reliability of this questionnaire, because not all correlations between the items originate from their relationship with the construct to be measured. Tragically, it has become commonplace to report Cronbach's alpha scores.

Validity is commonly referred to as the extent to which an instrument measures what it is supposed to measure. Exploratory Factor Analysis (EFA) is frequently used to confirm if a questionnaire has the intended underlying structure (i.e. construct validity). EFA can also be used for item selection to further the development of a questionnaire, by analyzing whether or not items contribute to the measured construct (Fayers et al. 1997a). In essence, it is a method to see which items hang together or cluster, in order to form a questionnaire measuring the same underlying construct. As with Cronbach's alpha, it is a method based on evaluating the correlation (or covariance) structure between items that make up the questionnaire (Fayers et al. 1997a). Therefore, also for this method positive correlations between items are a required condition. However, formative items are not necessarily correlated.

Consequently, the use of EFA as evidence for the validity of a questionnaire or as an item-selection procedure has serious implications when a questionnaire contains both formative and reflective items. This is demonstrated in a study on the use of EFA in QoL research (Fayers et al. 1997a). Because large differences were found among other studies concerning the factor structure (number of factors ranged from two to nine factors) of a frequently used QoL questionnaire in cancer trials, factor analysis was applied to data from a clinical trial for colorectal cancer and four factors were found. One factor found -a psychological distress factor- was also found in the other studies. The other three factors were more symptom-like factors and loaded on items describing symptoms typical for colorectal cancer, an outcome completely different from that found in the other studies. This demonstrates that the factor solution was situational or group specific, which is not a preferable quality of a questionnaire.

Besides the fact that this QoL questionnaire is prone to certain situational factors, these factors are also causal to the measured construct. Moreover, through our own research efforts on the OHIP, whilst still unaware of the presence of formative items, we established that EFA would not lead to a meaningful factor solution. In one of our studies we did, however, suggest a more parsimonious factor structure on the basis of high correlations between subscales (Kieffer et al. 2008a), and in the same breath we pointed to two studies that did use factor analysis on the OHIP and found a three- and a four-factor structures, respectively (John et al. 2004, Mumcu et al. 2007). Both solutions came from two completely different samples and once more suggest that situational factors play a role. This comes to show that knowing the processes underlying the correlations, rather than the correlations themselves, are a crucial ingredient in the validity process and as Bollen concludes by citing Blalock; 'one should be especially on guard against procedures that supposedly permit one to appraise the validity of an indicator (item) on the basis of the magnitude of correlation coefficients, without the benefit of a specific theoretical model' (Bollen 1984, p383-384).

DISCUSSION

It has been argued that OHRQoL questionnaires are a mix of both a formative and a reflective measurement model; this is consistent with other studies which suggest that reflective models are hardly ever feasible in the social sciences because they suppose one common cause, whereas it is in fact more likely that social constructs are part of a causal network of observable variables (McGrath 2005, Borsboom 2006). This view proposes that a reflective measurement model is not sufficient to underpin the complexity of many social constructs (Borsboom 2006), including OHRQoL, making conventional methods that focus on homogeneous scales and high inter-item correlations (classic of a reflective measurement models), fall short.

The question is; 'where did we go wrong?' The foregoing has touched on the subject of theory building, which is central to the social sciences and is, in fact, the answer. Hence, 'nobody starts constructing measurement instruments without the finest idea of the processes that lead to the measurement outcome' (Borsboom et al. 2004, p1067). OHRQoL research has done just that, for it is lacking a substantiated theory of OHRQoL. Even definition-wise no consensus is reached, and we refer to an article that gives five different definitions of OHRQoL (Locker et al. 2007a). Most OHRQoL questionnaires are rather based on classification schemes, such as the International Classification of Impairments, Disabilities and Handicaps, developed by the World Health Organization (World Health Organization 1980), on which actually no inferences can be made as to how constructs are related to each other. In addition, the traditional psychometric approach, which has been around for many generations, relentlessly predominates in contemporary scientific education and thus in current research efforts; hence, the readily acceptance of a reflective measurement model.

This article does not mean to demoralize researchers in this field. However much vigilance is needed for the issues described above. In the short term, we would recommend that researchers separate the reflective from the formative component in OHRQoL questionnaires and apply the aforementioned statistical methods only to the reflective component. The simplest way of distinguishing between formative and reflective items is by doing a so-called thought experiment (Fayers et al. 1997b). In designing such a thought experiment, Jarvis et al. (Jarvis et al. 2003) mentioned four sets of criteria relating to the direction of causality, to the interchangeability of items, to the inter-item correlations, and to the question of whether or not items have the same antecedents and consequences. Interestingly, Fayers et al. (Fayers et al. 1997b) proposed a statistical approach for distinguishing between formative item and reflective items. This approach holds the idea that formative items have an asymmetric relationship with QoL, an idea quite similar to the criteria suggested by Jarvis et al. (Jarvis et al. 2003) and that is based on frequency analyses using chi-square statistics; this is an attractive approach and worth looking into.

For future reference, however, we would like to challenge researchers to make themselves familiar with that side of measurement practice that can deal with both formative and reflective items. However, as many studies demonstrate different ways of dealing with formative measurement (Diamantopoulos et al. 2001, Jarvis et al. 2003, Howell et al. 2007), and give ample recommendation on how to model formative constructs, there is a catch. Structural equation models (SEM), used to test theory,

represent theoretical constructs that have substantive meaning apart from the estimation of parameters linking them to their indicators (Howell et al. 2007), implying that the latent variable exists independently from its measures. This is assumed only in reflective measurement, not in formative measurement (Borsboom et al. 2003). We therefore again stress the importance of the theoretical interpretation of the construct OHRQoL.