ARTICLE

Validity of a scale for consumer evaluation of mental health service delivery

Grenville Rose BA PhD\textsuperscript{a}, Ivan Beale MSc PhD\textsuperscript{b}, John Malone BA MA\textsuperscript{c}, Stephen Kinkead\textsuperscript{d} and Judi Higgin BSocSc MBA\textsuperscript{e}

\textsuperscript{a} Registered Psychologist and Research Manager, Aftercare, Rozelle Hospital Site, NSW, Australia
\textsuperscript{b} Registered Psychologist and Senior Visiting Research Fellow, University of New South Wales, Sydney, NSW, Australia
\textsuperscript{c} Registered Psychologist and Executive Director, Aftercare, Rozelle Hospital Site, NSW, Australia
\textsuperscript{d} Immediate Past CEO, New Horizons and Chartered Accountant, New Horizons, Ryde, NSW, Australia
\textsuperscript{e} Chief Executive Officer, New Horizons, Ryde, NSW, Australia

Abstract

Objective: The factor structure of a previously developed consumer constructed scale is based on whether the questions are positively or negatively worded and extant literature suggests that this may reflect the wording rather than the genuine structure of the questionnaire. Given the potential usefulness of this type of scale, further investigation of its structure and applicability to the development of person-centered mental health service organisation and delivery is warranted.

Method: The questionnaire, with consumer suggested modifications, was delivered twice across two Australian non-government mental health services. The two data sets were thus independent. The factor structure was investigated using the group 1 data and then applied to both groups using Confirmatory Factor Analysis (CFA) and multiple regression.

Results: The factor structure was affected by the phrasing of the questions. It was not possible to derive a model using the full set of questions and a reduced set was derived. Using the reduced set, valid models were produced in both the 2 and 1 factor CFAs, but only the model based on 1 factor was robust when applied to an independent data set.

Conclusions: The 1 factor model with a reduced question set offers the advantages of being quicker to deliver, not use negative phrasing and produces a more robust CFA. Further work investigating the delivery of such a questionnaire may contribute usefully to the development of person-centered mental health service organisation and delivery.

Keywords
Consumer evaluation, health, mental health services, outcome, patient satisfaction, process, scales, survey, validation, validity

Correspondence address
Dr. Grenville Rose, P.O. Box 82, Rozelle, 2039, NSW, Australia. E-mail: grenville.rose@aftercare.com.au

Accepted for publication: 1 November 2011

Introduction

Consumer satisfaction is an increasingly important concept in mental health services [1-4] and one which puts consumer perceptions at the forefront of service provision. Feedback from the consumer’s perspective is a good measure of the effectiveness of a service and provides an evidential basis for improvement of services and guidelines for the creation of new services [5-7]. Studies of consumer satisfaction with community-based services and other general medical health services that have used quantitative measures, have generally yielded high levels of satisfaction; usually around the 90% level [2,8-11]. Further, satisfaction with the treatment processes, treatment compliance and positive treatment outcomes have been shown to be inter-related [12]. However, most of these satisfaction studies were designed, conducted and administered by mental health professionals or the service providers and it has been argued that staff administered questionnaires can introduce response bias which may affect the validity of the studies [1].

Some studies have demonstrated that consumers surveyed by consumer interviewers reported significantly less satisfaction with the effectiveness of services and gave more extreme negative responses that those interviewed by staff interviewers as respondents feel safer and more willing to disclose true feelings which minimize social desirability effects or response bias [13,14]. Best practice, therefore, demands an increase in the involvement of
mental health service consumers in the development of outcome measures that assess the quality and effectiveness of mental health services [13,15-19]. The national standard for mental health services in Australia includes consumer participation as its third standard, encouraging collaboration between mental health consumers and professionals [20]. Nevertheless, there are only a handful of scales that have been developed with consumer involvement as a primary concern [1,8,21,22]. The use of consumers as researchers has resulted in an increase in the reliability and validity of the outcome measures, as well as empowering the consumers involved [13,14,23-28].

In response to these issues, Law and Oades [1] produced the “Consumer Evaluation of Mental Health Services” (CEO-MHS) after a three year research project in which the consumers were trained as researchers who then developed a service evaluation instrument [29]. In this project, 14 mental health consumers were trained in the research skills required to conduct interviews and run focus groups [30,31]. Utilising the information gained from in depth interviews and focus groups with other consumers, 11 themes were developed and from these themes, with further input from the consumer researchers, a model of consumer evaluation of mental health services and outcome measures was developed [1]. The result is a genuinely consumer-based survey that can be used to evaluate services from a more consumer directed point of view.

It has been argued that research into consumer satisfaction in mental health is a well-researched area, but its focus has traditionally been directed towards government provided services with little or no attention to the possible merits of private and NGO providers [32]. Additionally, two questions in the Oades [1] survey are not applicable to NGO services in Australia at this time, as they directly reference medication information and diagnosis, which are not the province of NGOs in Australia. It would be useful, then, to build on the consumer-centred survey development of Oades [1] and to develop further its applicability to NGO mental health services in Australia, while keeping the focus on the consumer involvement.

It is possible that methodological difficulties can be encountered when using positive and negatively worded questions in a survey [33]. This has been attributed to a number of causes, including careless responding [34], response style [35] and cognitive processing difficulties [33], which may make the survey used here problematic in its present form. Furthermore, it has been pointed out that the finding of two factors based on negatively and positively worded questions occurs in a number of different fields of research and that researchers reaching such an outcome “should be highly suspicious of factors loaded primarily with negatively keyed items” [34]. It has been noted that while it is not the case that a two factor solution based on negatively or positively keyed items is necessarily artifactual, it is better to investigate early in the development process, rather than having alternative interpretations to contend with, perhaps decades after development [36].

It is also possible that questions that are phrased in a negative way and that therefore relate negative affect may have a detrimental effect on vulnerable populations, such as those with severe mental illness. While we could find no studies that directly addressed the question of whether negatively phrased questions distress survey participants, a number of studies have addressed concerns that people may be affected by participation in survey research [37]. Although there seems to be little evidence that there is a risk of harm from participation in survey based research, [37,38] any risk should be minimised when working with vulnerable populations. In the two factor model previously reported [1], all positively worded questions loaded on factor 1, empowerment and all negatively worded questions loaded on factor 2, dehumanisation and this potential difficulty was not strongly addressed in that paper. Thus, an investigation of the effect of the negatively phrased questions on the factor structure would appear to be warranted at this point in development. Informal conversations with staff and consumers of the NGO over the past 4 years (G. Rose, Observation) also suggest that positively worded questions are preferred. Positively worded statements are also a better fit with the strengths-based approach that is being adopted by Australian government and non-government mental health agencies as a part of the focus on consumer recovery from mental illness.

Additionally, survey completion, however good its intentions, is not the principal focus of staff or consumers at the point of service delivery. Requirements of funding bodies mean that there is a plethora of forms and data to collect and additional burden is sometimes met with hostility or disinterest from staff and consumers. A shorter questionnaire that is easier to understand and complete would therefore be useful to services.

**Method**

The surveys were conducted face to face by consumer researchers who had previously been trained and involved in the development of the survey at the University of Wollongong. Survey data were collected across two Australian non-government mental health organisations that provide supported accommodation and outreach for people with mental illness. The original purpose of the survey was to give information that would be useful in developing staff training and to test the effect on consumers of that training. Thus, the data were collected in two waves, 6 months apart. Consumers in the outreach branches of the organisations were approached via their usual care worker and invited to take part in the study. The participants completed the survey either singly, for low literacy participants, or up to three at a time, with a consumer researcher. Ethics approval was granted through

**Consumer evaluation of mental health services**
the University of Wollongong. Two questions were removed from the survey as they did not apply to non-government mental health organisations supplying psychosocial support and would have confused the participants and introduced variance into the data: “I would like to have received my diagnosis from this service earlier” and “I have not been given adequate information about my medication and its side effects”. The consumer advisory committee at one organisation was consulted about the questionnaire and it suggested the inclusion of two questions on rights and responsibilities. To reinforce both the perception and reality of anonymity and as it was not germane to the purpose of the study, minimal demographic information was collected. There were 26 questions overall in the modified survey and it took between 5 and 30 minutes to deliver depending on literacy level of the participant. Average time to deliver the questionnaire was approximately 10 minutes.

Statistical analysis was conducted using SPSS 18 and Amos 18. Cronbach’s $\alpha$ was used to assess the reliability of the survey and Principal Components Analysis (PCA), Exploratory and Confirmatory Factor Analysis (EFA, CFA) methods were used to test the validity of one factor and two factor questionnaire models, guided by previously published methodology [33]. Fit measures used to assess the CFA were: $\chi^2$ (a useful global, but flawed, measure of the fit of the model); Root Mean Square Error of Association (RMSEA) (recommended as an adequate measure of misspecification) [39]; Adjusted Goodness of Fit Index (AGFI) (an absolute index of fit [40]) and the Akaike Information Criterion (AIC), (a measure of parsimony). Test of factorial equivalence of nested models was performed using the above, plus change in Cumulative Fit Index $\Delta$CFI [40]. Automatic backward stepwise ordinary least squares multiple regression was used to determine the most important factors that drove satisfaction with services. In tandem with CFA, this was used to reduce the length of the questionnaire and to develop a valid model of consumer satisfaction. Age and gender ratios were tested by $\chi^2$. Variance differences between positively and negatively worded questions were assessed using the $\chi^2$ Median test.

Part 1: Full question set

Demographics and satisfaction with services

Four hundred and eighty-one mental health service consumers of two non-government mental health and disability services completed the survey in two groups approximately 6 months apart. Sixty-nine of the consumers completed the survey at both collection occasions and those participants were removed from the second data set so as to make the two data sets independent. There were 266 participants in the first data set, which was used as a calibration data set for model testing and 215 in the second data set. Participation rate was greater than 50%. The age breakdown was almost identical at times 1 and 2, ($\chi^2$df1=0.004, $p>0.05$). Age information was collected in four ranges, as shown in Table 1. One participants’ data was found to have no variance and was excluded.

Table 1 Age breakdown

<table>
<thead>
<tr>
<th>Group</th>
<th>18-25</th>
<th>26-40</th>
<th>41-65</th>
<th>&gt;65</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>4.90%</td>
<td>32.50%</td>
<td>48.30%</td>
<td>14.30%</td>
</tr>
<tr>
<td>Group 2</td>
<td>4.20%</td>
<td>32.10%</td>
<td>49.30%</td>
<td>14.40%</td>
</tr>
</tbody>
</table>

Table 2 Gender breakdown

<table>
<thead>
<tr>
<th>Collection group</th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Group 1</td>
<td>57.6%</td>
<td>42.4%</td>
</tr>
<tr>
<td>Group 2</td>
<td>57.9%</td>
<td>42.1%</td>
</tr>
</tbody>
</table>

The gender balance was also nearly identical across the two collection groups ($\chi^2$df1=0.004, $p>0.05$, Table 2). The overall willingness of people to recommend the service to others, a proxy for service satisfaction, was 79% for group 1 and 82% for group 2. The number of people who would not recommend the service was 14% for group 1 and 11% for group 2.

Comparison of models using exploratory factor analysis for group 1 (calibration data set)

Cronbachs $\alpha$ for all questions for group 1 was 0.894 and was only improved marginally by the deletion of one of the negatively worded items, ‘I experience stigma from staff’. All variables in the group 1 data set had a strong ($>0.3$) correlation with at least one other variable and Kaiser-Meyer-Olin statistic was 0.914. Principal Components Analysis replicated the two factor solution found previously [1]. The first factor contained only positively worded questions, including the new questions relating to legal rights and all reverse coded negatively worded questions loaded on the second factor. These two factors accounted for 31% and 12% of the variance respectively.

Using the group 1 data set, a $\chi^2$ test showed that the variance for the negatively phrased items was significantly higher than that for the positively phrased items (Yates Corrected $\chi^2$df1= 9.58, $p<0.01$).

In the analyses presented in Table 3, the fit statistics for the single and two factor solutions with no correlated residuals, models 1 and 2 respectively, were the weakest. This is indicated by the high RMSEA (RMSEA >0.08 an adequate measure of misspecification ([40] p80), low AGFI (< 0.9 ([40] p78) and high AIC.
Correlating the residuals of the negative items on the single factor model tests for the effect of negative wording and resulted in a significantly lower $\chi^2$ value when compared with the uncorrelated single factor model, but a non-significantly different value when compared to the model with all positive items correlated. This is shown by significance testing on $\chi^2$ change $\Delta\chi^2$, ($\Delta\chi^2$df 21, 330.8, $p<0.01$, $\Delta\chi^2$df 265, 274.8, $p>0.05$ respectively). Model fitting on this calibration data set guided by modification indices and theory, did not result in models with acceptable AGFI.

**Table 3 Exploratory factor models of consumer satisfaction**

<table>
<thead>
<tr>
<th>Description</th>
<th>CMin $\chi^2$</th>
<th>df</th>
<th>$p$</th>
<th>RMSEA</th>
<th>AGFI</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single factor uncorrelated residuals</td>
<td>1004.3</td>
<td>299</td>
<td>$&lt;0.01$</td>
<td>0.094</td>
<td>0.696</td>
<td>1108</td>
</tr>
<tr>
<td>Two factor uncorrelated residuals</td>
<td>731.19</td>
<td>298</td>
<td>$&lt;0.01$</td>
<td>0.091</td>
<td>0.74</td>
<td>831.9</td>
</tr>
<tr>
<td>Correlated negative items single factor</td>
<td>673.5</td>
<td>278</td>
<td>$&lt;0.01$</td>
<td>0.073</td>
<td>0.79</td>
<td>819.5</td>
</tr>
<tr>
<td>Correlated positive items</td>
<td>398.7</td>
<td>13</td>
<td>$&lt;0.01$</td>
<td>0.084</td>
<td>0.73</td>
<td>822.7</td>
</tr>
<tr>
<td>Single factor Fitted</td>
<td>411</td>
<td>228</td>
<td>$&lt;0.01$</td>
<td>0.057</td>
<td>0.839</td>
<td>559.6</td>
</tr>
<tr>
<td>Two factor fitted</td>
<td>489.2</td>
<td>291</td>
<td>$&lt;0.01$</td>
<td>0.052</td>
<td>0.842</td>
<td>609.2</td>
</tr>
</tbody>
</table>

**Discussion of exploratory factor analysis**

Though the greater number of parameters estimated in the positively correlated model would normally be expected to lead to an improved $\chi^2$, there was no significant difference between that model and the negatively correlated model. Additionally, Marsh [33] states that relative measures of fit are more useful when comparing models and all fit indices in the current models suggest that the negatively correlated model has better fit. There would thus appear to be a weak method effect, where the factor structure of the model is being affected, or perhaps effected, by the wording of the questions rather than the perceptions of those who completed the survey. Correlating variables in either the single or two factor model, guided by modification indices and model theoretical considerations (for example, “supports my legal rights” and “tells me of my legal rights” were correlated), resulted in a model with acceptable fit statistics. In the attempt to generate a well fitted model, there were 9 modifications made to correlations of residual covariances in the two factor model and 16 modifications made to the single factor model, 9 of which in the latter case correlated the negatively worded questions, which further suggests a method effect in the data from the negatively worded items. Though no adequate fit was found for any model, the fit statistics are not entirely without merit, the RMSEA is acceptable for both the single and two factor models, for example. The question therefore arises as to whether a subset of questions might yield a valid model.

**Part 2: Confirmatory factor analysis of a reduced question set**

The data were originally gathered as part of a project to develop a training program for the organisations. As a part of the development process, an automatic backward stepwise regression analysis was conducted to identify those themes and questions that were important in predicting consumer satisfaction and which would therefore be important for staff training. Based on the data for group 1 this analysis identified 5 questions that were significantly related to the satisfaction question in the survey for both groups and between them accounted for 30% of the variance in the measure of satisfaction used in the survey “I would recommend this service to my family and friends”. Tolerances for each of the 5 questions were 0.80 or above. They included two negatively worded questions, ‘staff are ignorant to my needs’ and ‘I am not taken seriously by staff’ and 3 positively worded questions; ‘learn more about myself’, ‘feel safe and secure’, ‘my choice of treatment is respected’. These 5 variables, which represented 4 of the themes that formed the basis of the questionnaire development (two questions related to service responsiveness), were matched to one question representing each of the other 11 themes that formed the basis for questionnaire development [1], with the exception of diagnosis theme which does not apply to non-government psychosocial support mental health organisations. In the complete questionnaire, there was more than 1 question on each theme and therefore the question within each theme that had the strongest standardised beta weight in the group 1 data set was chosen for inclusion (Table 4).

Cronbach’s $\alpha$ for this set of questions on the group 1 data set was 0.84. The $\alpha$ was improved by deleting the negatively phrased questions. Calibration models were
Table 4 Questions in reduced data set and their variable number in final CFA

<table>
<thead>
<tr>
<th>Variable number in CFA</th>
<th>Question</th>
</tr>
</thead>
<tbody>
<tr>
<td>V1</td>
<td>This service is responsive when my symptoms change</td>
</tr>
<tr>
<td>V2</td>
<td>I think the consumer involvement in this service is excellent.</td>
</tr>
<tr>
<td>V3</td>
<td>This service has provided me with excellent information on mental health.</td>
</tr>
<tr>
<td>V4</td>
<td>This service encourages consumers to support each other.</td>
</tr>
<tr>
<td>V5</td>
<td>This service has assisted me to learn more about myself.</td>
</tr>
<tr>
<td>V6</td>
<td>This service has empowered me.</td>
</tr>
<tr>
<td>V7</td>
<td>This service helps consumers overcome any stigma they hold about themselves.</td>
</tr>
<tr>
<td>V8</td>
<td>I feel safe and secure when being looked after by this service.</td>
</tr>
<tr>
<td>V9</td>
<td>This service has told me about my legal rights</td>
</tr>
<tr>
<td>V10</td>
<td>This service gives me my legal rights</td>
</tr>
<tr>
<td>V11</td>
<td>I am not taken seriously by service staff.</td>
</tr>
<tr>
<td>V12</td>
<td>The staff at this service understand me.</td>
</tr>
<tr>
<td>V13</td>
<td>Service staff are ignorant to my needs.</td>
</tr>
</tbody>
</table>

Table 5 Model fit of 1 and 2 factor models with reduced question set.

<table>
<thead>
<tr>
<th>Description</th>
<th>CMin $\chi^2$</th>
<th>df</th>
<th>$p$</th>
<th>RMSEA</th>
<th>AGFI</th>
<th>AIC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Single factor</td>
<td>90.5</td>
<td>61</td>
<td>&lt;0.01</td>
<td>0.043</td>
<td>0.93</td>
<td>150.5</td>
</tr>
<tr>
<td>Two factor</td>
<td>63.7</td>
<td>50</td>
<td>0.092</td>
<td>0.032</td>
<td>0.94</td>
<td>119.7</td>
</tr>
</tbody>
</table>

Figure 1 Final 1 factor CFA model

Figure 1 Final 1 factor CFA model

The modifications to the residual correlations were made on the calibration data sets, based on both the indices and theoretical considerations (Figure 1). The modifications were identical in both the single and two factor models. When the model was subjected to a multigroup analysis, applying the result from group 1 to both groups simultaneously, the fit statistics were almost identical for both models, with the exception that the change in the cumulative fit index ($\Delta$CFI) for the covariances, assuming the measurement weights to be correct in the 2 factor model, was unacceptable [41]. All of the indices for the single factor model were within acceptable limits and all regression paths are significant at $p<0.01$. The AIC is similar for the two models and points to no advantage of parsimony. These solutions, however, are both more parsimonious than those for the models which used the full questionnaire (Table 6).
Discussion of confirmatory factor analysis of the reduced set

The models derived from the reduced question set require few modifications to derive well fitting models. Additionally, the single factor model fitted well to a validation data set, which suggests that the model is robust, whereas the two factor model failed on one criterion. These results suggest that the reduced data set yields a better fit to the data and also that the single factor model is the more robust solution.

Conclusions

This study re-affirms the high reliability of this scale, however, although it has been suggested that it was unlikely that the two factor structure previously found for this scale was a result of artefact [1], extant literature shows that a two factor structure based on positive or negative phrasing can be problematic [33]. However, the current study reaches a relatively weak conclusion regarding the advantage of using a single or two factor model. Correlating the negative items’ error terms yielded a non-significantly smaller \( \chi^2 \) compared to correlating the larger number of positive items, yet the modification indices suggested that the negatively correlated model was superior. The non significant \( \Delta \chi^2 \) does not argue for a method effect, yet the fit indices, which are the stronger indicator, suggest that the factor result may be due to the negative phrasing. As it was not possible to develop adequately fitted 1 factor or 2 factor solutions for the full set of questions, a reduced set was developed based on analyses conducted independently for staff training. When data from this reduced set were tested, both models fitted acceptably to the data of group 1, but when the model was applied to a validation data set (group2), only the 1 factor model had an acceptable fit. However, the fit statistics were reasonably good for both models and were superior to those of the full data set. Although analysis of models where the positively and negatively worded questions were allowed to correlate showed that there was not a strong method effect from the wording, the higher variance that was found for the negatively worded questions and the improvements in Cronbach’s \( \alpha \) when some negatively worded questions were removed suggests that positively wording the questions might yield a stronger result.

The current results suggest that, in one form or another, this questionnaire is a valid instrument and would be useful to Australian NGOs. The results are not unequivocal, yet they do suggest that there is a problem arising from negative wording in this questionnaire. Given that the extant literature suggests that cognitive and linguistic ability can affect responses [33,42], that there is the (small) possibility of harm being done in survey research and that the target group for this survey is people with a mental illness many of who will be from a multicultural non-English speaking background, the current analysis indicates the use of only positively phrased questions for this type of survey, if possible.

The reduced set offers a better fitting CFA model compared to that with the full questionnaire set and one which offers the advantage of brevity, aiding administration in a mental health service setting. Though there does not appear to be any strong advantage to using either the single or two factor model, if questions were to be worded positively then a single factor model becomes a possible outcome, as the PCA models developed for this study and those derived by the survey authors [1] are based on the positive or negative wording of the questions. Administration of the same questionnaire with all questions worded positively would be the most direct test of whether the two factors are based on the method or on factors implicit in the questions.

The results indicate that this instrument is a useful measure of consumer satisfaction with mental health services and thus an important tool for the development of person-centered mental health services organisation and delivery. The Cronbach’s \( \alpha \) of both the full and reduced questionnaire sets was still more than adequate for this reduced data set. It is suggested then that a reduced length questionnaire with only positively worded questions may be the best solution, one which keeps the intent of the questionnaire development intact, that it be consumer constructed and one which keeps the equanimity of the consumer at the forefront by only having positively phrased questions and thus minimising the risk of a negative effect as the person completes the survey.
Acknowledgements

We thank service participants who put their thoughts, time and effort into improving this survey and to the Mental Health Coordinating Council who funded the study.

References


