Training in Routine Mental Health Outcome Assessment: an evaluation of the Victorian experience

TIM COOMBS, TOM TRAUER, AND KATHY EAGAR

Tim Coombs is Co-ordinator of Mental Health – Outcomes and Assessment Training in New South Wales,
Tom Trauer is Associate Professor in the Department of Psychiatry, University of Melbourne and
Kathy Eagar is Professor and Director of the Centre for Health Service Development, Faculty of Commerce, University of Wollongong.

Abstract

This paper evaluates training in the use of measures of outcomes and casemix provided to four pilot agencies in Victoria, Australia. The training program is outlined along with key evaluation findings. The knowledge and skills of participants developed during training is described. Deficiencies in the training program are identified and opportunities for improvement outlined.

Background

The introduction of the Health of the Nation Outcome Scales (HoNOS) (Wing et al 1998), the short 16 item version of the Life Skills Profile (LSP-16) (Rosen et al 1989) and the Focus of Care (FoC) (Buckingham et al 1998) into routine clinical practice is an essential component of the information development strategy for Australian mental health services (Department of Health and Aged Care 1999). These measures of outcomes and casemix are seen as useful in routine clinical practice for monitoring the progress of the consumer (Stedman et al 1997) and at a broader service level (Buckingham et al 1998).

This paper evaluates the training process undertaken in the first four agencies to introduce these measures into routine clinical practice in Victoria. The development of the training materials used has been described elsewhere (Trauer, Coombs and Eagar 2002). The evaluation of the training focused on three broad areas. These were a general evaluation of the participant's experience of training; the knowledge gained by participants of the process for implementing these measures into routine use, along with the data collection protocol; and the skills developed during training in the completion of the standard measures.

Approach to Training

Four agencies in Victoria were identified as pilot sites for the implementation of these measures into routine clinical practice. These agencies are referred to here as Agency A, B, C and D. These agencies undertook two different types of training. Agencies A and B participated in 'Train the Trainer' workshops. These trainers would then go on to train the remainder of the workforce in the area. Agencies C and D received direct training ('Train the Troops') consisting of several $3^{1/2}$-hour clinician training sessions delivered by experienced trainers.

Both models involved a multi-modal approach including written case studies, video presentations, practice ratings of the outcome measures and group discussion (Eagar et al 2000).

The direct clinician training was delivered around three core 1-hour modules. Module 1 gave an overview to the background of the Victorian strategy and the data collection protocol. Module 2 focused on training in the HoNOS. The session began with a brief presentation on the background of the instrument and how it was developed, followed by a general outline of the rating rules or guidelines for the instrument. Trainees then read case vignettes that gave relevant clinical details and watched a video vignette that gave more information on the clinical presentation of the consumer to be rated. Trainees used the information from this vignette to practice rating the HoNOS, LSP-16 and the FoC. Trainees then declared their ratings and discussion of the rating rules were encouraged and elaborated. Misconceptions around the ratings were corrected at this time. Module 3 gave an overview of the consumer self-rating instrument being implemented in Victoria (the BASIS-32). The LSP-16 and the FoC were described and participants had the opportunity to rate a different vignette using a similar method as that outlined for Module 2.

The 'Train the Trainer' program included not only skills development in the completion of the HoNOS, the LSP–16 and the FoC, but also training in how to deliver a $3^{1/2}$ hour training session, and more detailed discussion of the data collection protocol along with local implementation issues. The 'Train the Trainer' sessions also included more detailed discussion of the background to the Victorian strategy.

At the completion of training, each program was evaluated. The evaluation measured knowledge and skills and sought feedback on the training program in general.

Evaluation method

Overall evaluation of training

Sessions were evaluated in general by use of a session evaluation form. This evaluation was in two sections. The first section involved participants self-assessing their knowledge of the outcome measures strategy, the tools that were to be used and the role that the tools play in the outcome measures strategy. Participants responded on a 5-point scale from 1 (I don't understand) to 5 (I understand most of it). Participants were also asked what they thought of the presenters, again on a 5-point scale from 1 (Helpful and approachable) to 5 (Not helpful or approachable at all).

The second section of the evaluation gave participants an opportunity to comment on the program, on ways in which the program could be improved, and on their further training needs. This section also asked two additional questions. The first asked participants if they felt their service should collect outcome measures. The second asked if participants felt that they should share their clinician ratings on outcome measures with consumers.

Knowledge test

A ten-item true/false knowledge test was developed by the training team that focused on the knowledge gained of the three clinician-rated measures. An initial form of this test was amended following its use in the first agency. This modified version was used in the other three agencies. In order to maximise participation rates, no demographic information was collected. The test was offered along with the general evaluation at the completion of the training session.

Skills assessment

Skills in completion of the measures were measured during the training. Participants rated a vignette and shared these ratings with other participants during the training session. This sharing of ratings served two purposes. First, it allowed for discussion around the rating rules for the particular measures and was an aid to training. Second, it provided an opportunity to compare the ratings of participants with ratings of the vignette previously made by an 'expert panel'.

During the course of training, the trainers gave a general indication that ratings one point either side of the recommended rating were usually acceptable, at least for training purposes. Skills in completion of the measures could therefore be measured based on the degree of agreement between the participants' ratings and those of the expert panel. Participants were assessed as having developed skills in the completion of the measures if their ratings fell one point either side of that given by the expert panel. For the purposes of calculating scores in relation to these skills, half of those who scored one point higher or lower than the expert panel were added to the number who scored the same as the expert panel. This gave a measure of the number of participants who achieved the expert panel criterion in the completion of the measures.

Results

200 staff were trained in the four pilot agencies, 17 and 9 in A and B and 101 and 73 in C and D. Evaluation data were collected from all four agencies. However, due to an administrative error, Agency C did not complete the general evaluation component.

General Evaluation

Evaluation questionnaires were completed in agencies A, B and D. Partial or complete forms were returned by 17, 9 and 72 respondents from these agencies respectively. Table 1 presents the aggregated results.

Table 1: Aggregated responses to the general evaluation questionnaire

Question number	Evaluation question	Mean Score
1	I gained a general understanding of the outcome strategy	4.2
2	I understand why outcome measurement is being introduced	4.2
3	I understand the assessment measures and how they are to be collected	4.0
4	I understand my role in the collection of the outcome measures	4.4
5	I found the presenter to be helpful and approachable (lower scores indicate more helpful and approachable)	1.7 % of participants responding yes
8	I think I need further training	44.6%
9	I think that my service should routinely collect consumer outcome measurement	92.8%
10	In general, I think that clinicians should discuss the ratings they make with the consumer	83.1%

Most respondents agreed with question 1, although 20% to 25% of respondents in D and A expressed some reservations. A similar pattern was observed with question 2, with 12% to 21% at D and A expressing some reservations. There was general agreement with questions 3 and 4, with no variation between agencies. In all three agencies, the majority of participants found the presenters to be helpful and approachable, but in two (D and A) 20% to 25% expressed reservations. None of the differences on these questions was statistically significant.

Forty-five percent of respondents at D and 64% at A expressed a need for further training, compared to none at B. This difference was close to reaching statistical significance ($\chi^2 = 5.64$, df = 2, p = .06).

This difference may be the result of the length of time taken to train in each agency. As a result of staffing issues, the total training time in site B was $1^{1/2}$ days, at A one day, and at C and D, $3^{1/2}$ hours. Differences may also be the result of the amount of experience agencies had with the use of these measures. Site B had extensive prior experience in the use of the HoNOS and LSP-16.

There was near-unanimous support for the collection of these measures as part of clinical practice, with only 6 respondents at D disagreeing. Over 80% of respondents in all three agencies believed that clinicians should share their ratings of the standard measures with consumers.

Evaluation comments

The evaluation form also encouraged participants to make comments on the training program. When asked what was the most difficult part of the training session, the majority of respondents who offered comments (n = 37) indicated the amount of information to be assimilated in such a short session *"to learn so much in such a small amount of time."* However, this was thought to be not insuperable because, *"with further use, my understanding will improve."* In terms of further training, a participant *"identified the need for 'reinforcement"* and the *"need to reflect on it"* including completing *"more case studies*". In the longer term, the need for ongoing training was also identified "*maybe after the introduction in 3 months time*".

In terms of whether participants felt that their services should collect routine outcome measures, one respondent caught the flavour of comments made at several agencies, *"as long as it is fed back in a meaningful, useful and positive way and not just added to the paper collected work*", while another participant indicated some concern about the introduction of outcome measurement by stating that "*depends on how routinely, still unsure on real benefits to clients*."

Although the majority of respondents indicated that sharing ratings with consumers was regarded as good clinical practice, a degree of concern remained. One participant indicated that *"I'm not sure, think it is OK as long as it does not have a detrimental effect on the therapeutic relationship*" while another participant saw it as *"not necessary, at times not possible when acutely unwell*".

Knowledge

Table 2 shows the aggregated responses to the knowledge test. The correct responses are in bold. The 186 respondents comprised 97 and 73 from C and D, and 16 from A. The results for Agency B are excluded.

Table 2: Responses to knowledge test (correct response in bold)

		True	False	Missing	% correct
HoNOS					
1.	You should never need to rate a 9 when completing the HoNOS	61	**122**	3	65.6
2.	Quality of relationships are not considered when rating the HoNOS	54	**128**	4	68.8
3.	Psychotic depression is only rated on the Hallucination Scale of the HoNOS.	18	**160**	8	86.0
4.	If the consumer has been extremely suicidal during the rating period, but is not suicidal at the time of rating, the HoNOS can be rated either a 3 or a 4	95	**81**	10	43.5
LSP-16					
5.	You should rate the Life Skills Profile on the basis of the person's worst level of functioning over the past 3 months	20	**166**	0	89.2
6.	The Life Skills Profile 16 is a clinical interview	19	**167**	0	89.8
7.	The five subscales of the Life Skills Profile are accommodation, self care, aggression, compliance and withdrawal	76	**104**	6	55.9
Focus of Care					
8.	For the Focus of Care, Acute is the highest rating and Maintenance is the lowest	38	**148**	0	79.6
9.	Functional gain represents a transition stage between Acute and Maintenance Focus of Care	102	**83**	1	44.6
10.	A patient on an acute ward need not be rated Acute Focus of Care	**114**	70	2	61.3

With the exceptions of questions 4 and 9, the majority of respondents gave the correct answer, but in certain cases the proportion giving the correct answer was not much greater than 50%. Across the test as a whole, 68.4% of answers were correct.

Differences between the three agencies were non-significant for seven questions, near the 0.05 criterion of significance on two questions, and highly significant for one question (question 9). Given that multiple significance tests were conducted, only this last finding will be discussed. Question 9 states that "Functional gain represents a transition stage between Acute and Maintenance Focus of Care". Sixty-nine percent of the respondents at A gave the correct response of False, compared with 53% at D and 34% at C. Agency A respondents had been trained to become trainers while those at D and C were clinicians who had received the shorter 'Train the Troops' program.

On average for the knowledge test, the correct response rate for the three LSP questions was 78%, for the four HoNOS questions it was 66%, and for the three FoC questions it was 62%. These results are consistent with the other indications during training that there are some specific issues and problems, especially with the FoC. Feedback from participants suggested that certain questions could be reworded to improve clarity.

Skill in the HoNOS

Two vignettes, which we shall call P and Q, were used for HoNOS training. One hundred participants were trained in sessions that used P, and 96 in sessions that used Q. Occasionally items were omitted by participants, resulting in 99 to 100 ratings across the 12 HoNOS items for P, and 90 to 96 ratings for Q.

For these aggregated data, correct ratings were computed as all recommended ratings plus half of all ratings that were one point higher or lower than the recommended rating. Table 3 shows the percentage accuracy for the 12 HoNOS items for the two vignettes.

Table 3: Percentage correct ratings to two HoNOS vignettes

	P	Q
Item 1	94.0	57.9
Item 2	97.5	97.9
Item 3	99.0	80.2
Item 4	78.5	62.0
Item 5	71.5	93.2
Item 6	88.0	80.2
Item 7	43.9	97.3
Item 8	44.9	60.3
Item 9	50.5	70.7
Item 10	78.8	80.5
Item 11	98.0	66.5
Item 12	82.0	48.3
All items	77.3	74.8

Accuracy levels range from near perfect to below 50%, averaging around 75%. There is little consistency between the item accuracy levels between the two vignettes (the correlation is –0.06), indicating that items achieved different accuracy levels, presumably on the basis of differences in the vignettes themselves.

Skill in the LSP-16

Two vignettes, which we shall call R and S, were used for LSP-16 training. One hundred participants were trained in sessions that used R, and 94 in sessions that used S. Occasionally items were omitted by participants, resulting in 92 to 100 ratings across the 16 LSP-16 items for R, and 90 to 94 ratings for S.

As with the HoNOS, an expert panel prior to the training itself determined recommended or "correct" ratings, and the same procedure for computing accuracy was used. Table 4 shows the percentage accuracy for the 16 LSP items for the two vignettes.

Table 4: Percentage correct ratings to two LSP-16 vignettes

	R	S
Item 1	78.1	80.9
Item 2	73.2	85.1
Item 3	36.4	88.3
Item 4	50.0	94.1
Item 5	50.0	87.8
Item 6	79.5	46.8
Item 7	89.7	100.0
Item 8	63.5	69.7
Item 9	60.3	59.0
Item 10	58.8	83.2
Item 11	58.4	71.1
Item 12	93.1	73.1
Item 13	59.9	65.6
Item 14	81.5	93.5
Item 15	53.7	53.9
Item 16	70.7	86.0
All items	66.1	77.4

As with the HoNOS, there was little consistency between the accuracy levels between the two vignettes (the correlation is 0.04) once again indicating that items achieved different accuracy levels, presumably on the basis of differences in the vignettes themselves.

Skill in the Focus of Care

The same vignettes used for LSP-16 training were used to train the FoC. Sixty-three FoC ratings were obtained using the R vignette, and 93 using S. The distribution of ratings is displayed in Table 5.

Table 5: Focus of Care ratings of two vignettes

	Vignette R (number and %)	Vignette S (number and %)
Acute	2 (3.18%)	0 (0.00%)
Functional Gain	36 (57.14%)	52 (55.91%)
Intensive Extended	8 (12.70%)	31 (33.33%)
Maintenance	17 (26.98%)	10 (10.75%)

The vignettes had been designed such that the correct response for R was Functional Gain and for S, Maintenance. Thus the accuracy was 57.1% for the R vignette and 10.8% for the S vignette. Over a quarter of participants rated R as Maintenance instead of the recommended Functional Gain and over a half rated S as Functional Gain instead of the recommended Maintenance. Overall, the accuracy levels for the FoC were weak. The two vignettes only cover the Functional Gain and Maintenance categories, so we can say nothing about the other two categories. Nevertheless, these results are indicative of a problem in either the vignette material itself, or the conceptual basis of the FoC categories, or both.

Discussion

Overall, the evaluation of training was positive, with the majority of participants indicating that they had developed a general understanding of the strategy, the instruments being used and the role that they had to play. The overwhelming support for the collection of these measures within mental health services is a positive indicator for the introduction of outcome measures into the remaining Victorian Mental Health Services. There was also support (although with reservations) for staff to share their ratings with consumers. The willingness of clinicians to share their ratings with consumers and to discuss the responses to the consumer-rated measure will be a valuable adjunct to clinical practice by promoting and supporting dialogue between the consumer and clinician. The clinical utility of the measures is essential to the successful introduction of outcome measures into routine clinical practice (Callaly et al 1998).

While the overall findings were positive, participants in the training sessions express a note of caution. They indicated that a significant amount of new information must be assimilated during a short training session and suggested that further training will be required. However, the need for further training depended on the experience of the agencies with the collection of outcome measures, with the more experienced agency indicating less need for further training.

Irrespective of the level of previous experience in outcome measurement, there is a clear need for further training and retraining. Those agencies that received "Train the Troops" direct clinician training sessions identified the need for individuals who could act as resource persons in the future. If these resource people could be accredited, or be required to complete regular repeat training sessions, ongoing data quality would be enhanced and a system would be established that could manage issues such as changes to instruments and protocols (Trauer, Coombs and Eagar 2002).

The results of the knowledge test in relation to the HoNOS and LSP-16 indicate that the majority of participants could offer the correct response to questions following training. The FoC achieved the poorest results. This may suggest problems in the design of the FoC.

However, since we have no independent indication of the difficulty of the test items, or of the quality of the teaching of the various instruments, it is not possible to say whether these results are particularly good or bad. It may be that the results simply indicate poorly written questions. Some participants indicated that the wording of some of the questions in the knowledge test was ambiguous.

In terms of skills developed during training, accuracy levels in the HoNOS averaged around 75%. However, certain items differed in their degree of difficulty between the vignettes. Likewise, the two vignettes for training in the LSP produced quite different accuracy levels, which appeared to be related to the actual content of the

vignettes themselves. This variation in rating the measures could be the result of either difficulties with the instruments themselves or difficulties with the training materials. The HoNOS has been criticised because of the poor interrater reliability with regard to certain scales, in particular scale 4, 7, 8 and 12 (Bebbington et al 1999). This pattern of poor interrater reliability was not demonstrated in the ratings of the video vignettes used in training. The conclusion that can be drawn is that variations in ratings by participants are related to variations in the information provided either in the written case notes or in the video or in both.

Video material has been used for a variety of purposes in mental health, from improving treatment acceptability for consumers (Foxx et al 1996) to improving the counselling skills of clinicians (Caris-Verhallen et al 2000). However, this is the first time to our knowledge that video vignettes have been used to support the rating of outcome measures. It may be that variation in participants' ratings is the result of visual cues in the material producing a response bias. Discussion of one vignette in particular often elicited the response from participants that "… she didn't move around enough to be manic". Although this response bias is not a significant issue during training (because it prompts discussion around the measures), it does highlight the difficulty of attempting to include sufficient material in one vignette to allow for the accurate completion of a suite of outcome measures.

Even after training in the FoC, the results of both the knowledge and skill tests indicated a general lack of agreement between ratings offered by participants and those identified by trainers as the correct response. The FoC is a single item that requires clinicians to identify retrospectively the main goals of care for the consumer during the preceding period of care. Training in the FoC was limited with participants simply being given a brief didactic overview of the item and the rating rules. Discussion of this measure indicated either an inability to understand the concept or a philosophical disagreement with particular terminology used by the measure such as the word 'maintenance'. This suggests that the item may not be a measure of treatment intention and provider activity but an idiosyncratic indicator of the philosophy of certain units. The FoC requires far more training than a cursory introduction. The quality of the information gathered requires local monitoring to avoid the development of idiosyncratic rating rules.

As we have shown, even after relatively short training sessions, staff can demonstrate sufficient knowledge and skill in the completion of the HoNOS and LSP-16 to begin data collection using these measures. More problematic is the quality of information generated through completion of the FoC. The evaluation of the training undertaken in the four pilot agencies in Victoria indicates the need for ongoing training and retraining to support the implementation of these measures into routine clinical practice. The evaluation also indicates the need to establish local quality improvement processes that monitor the quality of data collected.

References

Buckingham W, Burgess P, Solomon S, Pirkis J & Eagar K 1998, *'Casemix Classification for Mental Health Services'*, Commonwealth Department of Health and Family Services, Canberra.

Bebbington P, Brugha T, Hill T, Marsden L, & Window S 1999, 'Validation of the health of the Nation Outcome Scales', *British Journal of Psychiatry*, vol 174, pp389-394.

Callaly T, Trauer T, & Hantz P 1998, 'Integration of outcome measures into clinical practice', *Australasian Psychiatry*, vol 6, no 4, pp 188-190.

Caris-Verhallen W, Kerkstra A, Bensing J, & Grypdonck M 2000, 'Effects of video interaction analysis training on nurse-patient communication in the care of the elderly', *Patient Education & Counselling*, vol 39, no 1, pp 91-103.

Department of Health and Aged Care 1999, *'Mental Health Information Development: National Information Priorities and Strategies under the Second National Mental Health Plan'* 1998-2003, Commonwealth Department of Health and Family Services, Canberra.

Eagar K, Buckingham B, Coombs T, Trauer T, Graham C, Eagar L & Callaly T 2000, *'Outcome Measurement in Adult Area Mental Health Services: Implementation Resource Manual'*, Department of Human Services, Victoria.

Foxx R, McHenry W & Bremer B 1996, 'The effects of a video vignette on increasing treatment acceptability', *Behavioral Interventions*, vol 11, no 3, pp 131-140.

Rosen A., Hadzi-Pavlovic D & Parker G 1989, 'The Life Skills Profile: a measure assessing function and disability in schizophrenia', *Schizophrenia Bulletin*, vol 15, no 2, pp 325-337.

Stedman T., Yellowlees P, Mellsop G, Clarke R & Drake S 1997, *'Measuring Consumer Outcomes in Mental Health'*, Department of Health and Family Services, Canberra.

Trauer T, Coombs T & Eagar K 2002, 'Training in Routine Mental Health Outcome Assessment: The Victorian Experience', *Australian Health Review*, vol 25, No 2, 122-128.

Wing J, Beevor A, Curtis R, Park S, Hadden S & Burns, A 1998, 'Health of the Nation Outcome Scales (HoNOS)', *British Journal of Psychiatry*, vol 172, pp 11-18.