4 Testing methods of assessment

The assessment method used in the appraisal process may influence the reliability and validity of the measure of service quality produced by the Instrument. In this section the assessment methods trialled in the pilot are compared in terms of their reliability, validity and appropriateness.

As outlined in Chapter 2, there were five methods for the assessment of agencies against the HACC National Service Standards. The most commonly used methods were self-assessment, self-assessment with verification and joint assessment. Some assessments were carried out by a peer review method and a few assessments used an independent or external assessor. Each of these methods involved a different level of interchange between the assessor and the agency. They also involved differences in the timing and context of data collection/review and the processes used to arrive at ratings.

The various methods for the assessment of agencies against the standards are tested by four means in this chapter. The face validity of each method is examined by reports on the experiences of assessors. The extent to which concurrent validity is dependent on the type of assessment is tested by comparison of agency and assessor ratings and comparison of global performance appraisals and Instrument Scores. Statistical tests of inference were conducted to analyse the degree to which the assessment method affects the assessment outcome. Finally, comparisons are made of variations in rater reliability across the methods of assessment.

4.1 Face validity

4.1.1 General observations

The following comments on the apparent validity of the methods of assessment were made by assessors after they had completed the assessment process, either via the Assessment of Review Process questionnaires (described in Chapter 2, Section 2.1.2) or during debriefing sessions.

A number of observations were common across the assessment methods. Most particularly, assessors participating in all assessment methods concurred that the assessment interview was a crucial step in conducting an accurate appraisal. Assessors expressed the opinion that an in-house visit gave agencies the opportunity to challenge the process, to disagree or concur with the reviewers, to substantiate claims, and to display their efforts and achievements. During a visit, the assessor had the opportunity to pick up on agency culture and attitudes that may either confirm or belie a purely

paper-based assessment. Although it was acknowledged that agency visits were time consuming, and the need to find a balance between documentation-based and interactive assessment was recognised, assessors generally felt that the effort of visiting agencies was worthwhile and resulted in much more balanced and appropriate assessments.

In general, assessors believed that the assessment process trialled in the pilot was not sufficient to detect all difficulties in service provision. Assessors expressed the opinion that flaws in service quality may go undetected by an assessment method limited to viewing an agency's completed Instrument and documentation and discussing agency practice in an interview. This was particularly seen to be the case where an agency might not wish to disclose flaws in service quality. It was considered that obtaining data about agency performance from other sources was essential to the process. Assessors believed that a more thorough assessment would require observation of agency practice and, perhaps more importantly, that consumer input should have a place in the measurement of service quality.

Assessors expressed the opinion that the assessment process would have been made easier if agencies had received the same information as assessors regarding how ratings against the standards were derived. The Instrument which agencies received did not contain information on how to rate answers against the standards, whereas the assessor guidelines were intended to assist in rating answers. The Instrument, itself, was generally seen as too long and too repetitive but, nevertheless, was generally seen as being a well-constructed and helpful tool. Indeed, on balance, the assessment process was believed to be a positive experience for raising agency awareness of service quality issues. It was seen as important and beneficial for agencies in terms of establishing their own accountability and being able to document and substantiate how they met the standards. It was also seen to be helpful in encouraging agencies to review their own progress and evaluate the framework within which they have planned and developed their services.

The time frame of the pilot allowed agencies a minimum time of one week to complete the Instrument, although many agencies had longer than this. A common complaint to assessors was that this was not enough time, particularly where agency staff were volunteers and/or worked only part-time hours. In New South Wales, where agencies undertook the assessment more formally,[1] service providers considered that a more appropriate time to complete the Instrument with the approval of their management committees was three months.

4.1.2 Observations on individual assessment methods

The following discussion summarises assessors' comments on each of the assessment methods they that participated in. A full account of each assessment approach was given in Section 2.2.

1. Agencies participating in the pilot in New South Wales were advised that, should the Instrument remain substantially the same after the pilot, they would not have to complete the Instrument again. They were also required to discuss the responses to the Instrument with their management committees.

Self-assessment

In this method, agencies were required to fully complete the Instrument without the assistance of an assessor, but knowing that a proportion of agencies would be randomly selected for a verification interview.

The majority of agencies which undertook a self-assessment did not participate in an assessment interview. Assessors who conducted interviews with the 14% of the self-assessed agencies in New South Wales and 50% of the self-assessed agencies in the Australian Capital Territory agreed that verification visits were important to clarify and explore issues, to provide support and education and to cooperatively develop an effective forward action plan.

Agencies undertaking self-assessment received no assistance in completing the Instrument other than the short guide to scoring outlined on page 12. Assessors who conducted the random verification interviews found that agencies would have benefited from more detailed information both prior to and in the process of undertaking assessment. In at least one case, a verification interview was conducted as a joint assessment because of the agency's difficulty in completing the Instrument. Conversely, some agencies indicated to assessors that the verification interview was unnecessary since it effectively required them to undertake assessment twice—once in completing the Instrument and once in reviewing it with the assessor. This type of assessment may be most effective when tailored to the capacity of the agency to complete the Instrument.

Self-assessment with verification

The self-assessment with verification method required agencies to complete the Instrument prior to the verification visit, but the ratings were considered to be draft ratings. Final ratings were then reached after discussion with the assessor. Assessors involved in self-assessment with verification did not receive completed agency Instruments before attending the assessment interview. Assessors generally believed that assessment visits would have been more constructive and efficient if they had the opportunity to view completed Instruments prior to this meeting. An advantage of the self-assessment with verification method was that agencies were extremely well prepared and had thoroughly completed the review, making the process less time-consuming for the assessor.

Joint assessment

In joint assessments, agency staff completed the required performance information and assembled relevant documentation prior to the visit, but did not rate their performance against the standards prior to the assessment visit.

Joint assessments were found to be very useful for service providers who were unfamiliar with the requirements of service appraisals. New agencies and those operated by staff with limited experience of such procedures were believed to have benefited most from the intensive and collaborative contact with a government officer that a joint assessment offered.

A disadvantage of joint assessment included the length of time required for the whole assessment process, particularly where there were time constraints on both parties. Estimates of the time required to complete a joint assessment varied from 3 hours to 7.5 hours, and in one case, 15 hours including travel time. Joint assessments were more time consuming for assessors because agencies had done less preparation work. In some cases this was understandable because agencies were unsure of how to prepare, but where agencies were well able to complete the Instrument a joint assessment was viewed as unnecessarily time consuming for assessors.

The majority of assessors using joint assessment were government officers whose responsibility it was to liaise regularly with the agencies they assessed. Some concern was raised by these officers that they might have been too close to the agencies to be sufficiently objective. Officers expressed some difficulty in coming down hard on agencies with good intentions who may have failed to fully satisfy standards through inexperience rather then through neglect or ill-intent. As stated elsewhere in this report, however, understanding the context in which an agency is operating and the history and circumstances of its development can be an advantage in assessing how a service is most appropriately delivered and improved.

Independent or external assessor

The independent or external assessors who participated in the pilot were State government officers with extensive experience in quality measures and appraisal, but external to the HACC program. They undertook either self-assessments with verification or joint appraisals. These assessors undoubtedly benefited the assessment process by bringing their experience to the quality appraisal interview. The small number of appraisals undertaken by independent assessors did not, however, allow for any viable statistical testing of their effect on the appraisal process.

According to these assessors, their lack of familiarity with the services they appraised was a disadvantage. Some familiarity was seen to be necessary so as to not to be too prescriptive. For example, knowledge of local service provider networks is of particular relevance to the assessment of HACC agency performance against standards relating to advocacy, referral, or provision of services to special needs groups. An independent or external assessor may not possess such knowledge.

One of the independent assessors was from the State head office. This officer speculated that the anxiety or defensiveness of the agencies may have been increased by being visited by someone from head office rather than a familiar regional representative. Counterbalancing this was the advantage of objectivity that an external assessor brought to the assessment interview. A suggested alternative was that an external evaluation officer could assist a regional government officer in conducting assessments. Such a procedure may be particularly beneficial in circumstances where the appraisal process is expected to be unusually complex or problematic.

Peer review assessments

Peer review assessments were conducted by staff of other HACC agencies, employing either self-assessments with verification or joint appraisals (in the Northern Territory) or desk audits (in South Australia).

There was general agreement that the consultative and interactive aspects of the peer review process were worthwhile and positive. Assessors agreed that the process of peer review had the added value of promoting sharing of information, expertise and practices between agencies and encouraging agencies to learn from the experience of their peers. It also assisted in the development of stronger service networks through awareness of other service providers and participation in a cooperative process of service quality improvement. The agencies represented at the debriefing session stated that they were more comfortable with one of their industry colleagues evaluating them, rather than an outside consultant, because of the possibilities for sharing information and maintaining the positive, collaborative and helpful feel of the exercise.

On the other hand, the resources needed to complete an assessment and to participate in the review process were of concern to agencies participating in peer review. Small agencies and agencies in rural and remote areas were seen to be at a particular disadvantage in this regard. It was noted that, regardless of enthusiasm, small agencies do not have the capacity (in terms of either time or people) to participate in the process at the same level as do larger agencies. The ability to gain from the feedback and quality development aspects of the review was also seen as difficult for rural and remote agencies. Not only was it very time consuming to conduct peer reviews of these agencies (and vice versa), but the lack of other agencies in the area would mean that there would be little available support to improve service provision.

Service providers who had acted as assessors agreed that it was very difficult to evaluate the performance and quality of another agency without conducting a visit to the agency. In conducting the peer review process, agencies concurred that it would have been most effective to use the following method:

- review documentation which had been supplied by the agency with the Instrument;
- have a telephone conversation with the agency to discuss issues and make a time for a face-to-face interview; and
- conduct an agency visit of some duration to clarify points, substantiate claims and to get a general impression of the service.

4.2 Concurrent validity

The concurrent validity of the Instrument was examined in Chapter 3, Section 4. On average, exact agreement between assessors and agencies regarding ratings against each standard was 76%. The concurrence between the assessor's overall appraisal of agency service quality and the Instrument Score was measured as a correlation of 0.74. The equivalent correlation for agency-determined scores was 0.64. In this section, further statistical comparisons are made to test whether the assessment method affects the concurrent validity of the Instrument. Firstly, comparisons are made of the convergence of global appraisals and Instrument Scores across assessment methods. Secondly, comparisons are made of agency and assessor agreement on ratings across assessment methods.

4.2.1 Validation against a global assessment of quality

As described in Chapter 3, assessors and agencies were asked to give the agency they assessed an overall appraisal of service quality according to a four-point scale from 1 (*fails to meet the HACC National Service Standards*) to 4 (*exemplary*).

Table 4.1 shows the correlations between global assessments and Instrument Scores according to both assessors and agencies for the five assessment methods trialled in the pilot. The numbers of cases used to calculate each correlation are listed. These numbers are very small for some assessment methods, indicating that some degree of caution should be exercised when interpreting the statistics. Numbers are limited by the size of the sample of agencies in each assessment method and by the failure of some agencies or assessors to complete the global appraisal question. The number of cases available for the independent/external assessor method are too small to consider in comparison to the others and will not be discussed.

Table 4.1: Correlations between global assessments and Instrument Scores by method of assessment

Informant	Self-assessment	Self-assessment with verification	Joint assessment	Peer review	Independent/ external rater
Assessor	0.88** (n = 13)	0.59** (n = 30)	0.47 (n = 9)	0.70* (n = 7)	1.00 (n = 2)
Agency	0.68** (n = 41)	0.62** (n = 17)	0.61 (n = 10)	0.70 (n = 11)	0.28 (n = 4)

*p<0.05, ** p<0.01

The highest correlations between assessors' Instrument Scores and global assessment scores occurred for peer review (r = 0.70) and self-assessment (r = 0.88). Indeed, the correlation between the assessors' global scores and Instrument Scores for self-assessment is very close to a one-to-one correspondence. These results suggest that, for these methods, the same or similar appraisal occurs whether assessors rigorously score each standard or simply form an overall opinion of the agency based on their replies. This result throws some doubt on the validity of these assessment methods, since it suggests that ratings against standards may be heavily influenced by the assessors' general impression. According to the peer review method, ratings were determined without visiting the agency and were based solely on written responses to the Instrument and provided documentation. According to the self-assessment method, agencies had fully completed the Instrument on their own without foreknowledge of whether they would receive a verification visit, and assessors received Instruments prior to their visit to the agency, allowing them to form an opinion of the agency prior to the verification.

The lowest correlation between assessors' global assessments of agency quality and Instrument Scores occurred for the joint assessment method (r = 0.47). This method involved the most collaborative process for reaching decisions about ratings. The assessors' overall opinion of the agency's service quality has least relation to the ratings the agency received in the Instrument. This may suggest that in this collaborative process assessors were most likely to have been affected by their communication with the agency. They may have been more likely to take into account the special and

individual circumstances of the agencies. As assessors expressed in debriefing sessions (see Section 4.1.2), they sometimes found it difficult to be as objective as the Instrument required, making allowances for agencies such that the ratings they gave were less representative of how they actually felt the agency was doing.

The correlation between assessor-determined Instrument Scores and global assessments was $r = 0.59$ for the self-assessment with verification method. The size of this correlation indicates concurrent validity between the two scores without being so high as to suggest that assessor opinion of the agency is synonymous with the scores that they achieve for the standards. It is also not so low as to suggest that the Instrument Scores do not represent a reasonable measure of assessors' opinion of agency service quality.

The correlation between agency-determined Instrument Scores and agency global appraisal scores fell between 0.6 and 0.7 for all assessment methods (with the exception of independent/external assessor where the number in the sample was too low to obtain a reliable statistic for this comparison). The size of these correlations suggests that agency ratings may be influenced by their own opinions of the general quality of their service. This relative consistency across methods and the high correlations are to be expected since, initially, all agencies received an equivalent minimal amount of information regarding how to score themselves against the standards. Hence there is no appreciable reason why the correspondence between these scores would differ greatly across methods.

There is some evidence, however, that discussion with assessors that occurred during the assessment process was associated with some difference between agencies' individual ratings and agencies' global appraisals. The correlations for self-assessment with verification and joint assessment using agency-determined ratings ($r = 0.62$ and $r = 0.61$ respectively) were slightly lower than those for agencies which experienced peer review or self-assessment ($r = 0.70$ and $r = 0.68$ respectively). It may be surmised that the agency's own opinion of the quality of its service was somewhat less influential in the decision about appropriate self-ratings when they expected to have to take into account the views of an assessor. In both peer review and self-assessments, agencies had no contact with assessors when fully completing the Instrument and, furthermore, did not believe they would necessarily have verification interviews.

4.2.2 Agency and assessor agreement on ratings

In this section individual standard ratings given to agencies by assessors are compared to agency self-ratings for each method of assessment. As described in Chapter 3, Section 3.4.1, agreement means that if the assessor rated their agency 'met', the agency also rated it as 'met'; or if the assessor rating was 'partly met', the agency rating was also 'partly met'; or if the assessor rating was 'not met' the agency rating was also 'not met'. Table 4.1 lists the average percentage of exact agreement between agencies and assessors for each assessment method. Again, the independent/external assessor method is excluded from consideration due to the low sample size.

Table 4.2: Average proportion of exact agreement on ratings between the assessor and the agency for each assessment method

Self-assessment	Self-assessment with verification	Joint assessment	Peer review	Independent/ external rater
71.4% (n = 15)	81.3% (n = 23)	92.6% (n = 10)	59.8% (n = 18)	94.0% (n = 4)

Joint assessment

In a joint assessment both the agency and the assessor (government project officer or equivalent) completed their ratings against the standards in the Instrument and on Ratings Summary Forms during their meeting together. Where possible they were asked to come to an agreement about ratings. The highest proportion of agreement between agencies and assessors occurred for this assessment method. On average, agencies and assessors agreed on the ratings for standards 93% of the time (see Table 4.2). Table 4.3 shows the mean rating given to agencies for standards for each assessment method. The average rating across all standards according to agencies was 1.71 (out of a maximum possible score of 2); the average rating across all standards according to assessors was 1.73 (see Table 4.1). A test of the significance of this result indicated that there was not a significant difference between agency and assessor average ratings ($t = -0.31$). This result is to be expected given the aims of joint assessment.

Table 4.3: Average agency rating, average assessor rating by assessment type, and significance test of the difference between average scores

Assessment type	Agency rating (mean)	Assessor rating (mean)	Significance test	
			t-test	probability
Joint assessment	1.71	1.73	–0.31	NS
Self-assessment with verification	1.64	1.68	–1.48	NS
Self-assessment	1.55**	1.43	1.85	0.01
Peer review	1.64	1.20	3.21	0.01
Independent/external rater	1.67	1.64	0.14	NS
Total sample*	1.63	1.50		

NS = Not Significant.
* Where both Instrument and Ratings Summary Form were received.
** Agency mean for the sample of self-assessed agencies which received a random verification visit. The mean for all self-assessment agencies was 1.64.

Self-assessment with verification

In a self-assessment with verification the agency completed their ratings against the standards in the Instrument prior to meeting with the assessor, and the assessor completed their ratings against the standards in the Ratings Summary Form during the meeting with the agency. Self-assessment with verification, while less collaborative than the joint assessment method, also showed a high level of agreement between agencies

and assessors. These two parties agreed on ratings an average in 81% of cases. The average rating across all standards according to agencies was 1.64; the average rating across all standards according to assessors was 1.68. This difference was not significant (t = –1.48).

Self-assessment

In the self-assessment method, the agency completed the Instrument, including their ratings, and sent it to the Institute. Agencies were informed that a small proportion of Instruments would be verified by a visiting government officer but were not informed whether they would be part of this sample until after they had sent in their Instrument. The average rating across standards for all agencies (i.e. with and without verification visits) which completed the Instrument as a self-assessment was 1.64.

The statistics presented in Tables 4.2 and 4.3 are for the sample of agencies which received verification visits. For this sample, government officers from the Australian Capital Territory and New South Wales were sent completed agency Instruments and supporting documentation. They subsequently completed the Ratings Summary Forms during meetings with the agencies. In the self-assessment method, agencies and assessors agreed on ratings an average of less than three-quarters of the time (71%). Table 4.3 shows that self-assessing agencies (with random verification visits) were less likely to agree with their assessors, rating themselves higher than did their assessors. The average rating across all standards according to agencies was 1.55; the average rating across all standards according to assessors was 1.43. This difference was significant (t = 1.85, p<0.01).

Peer review

Agencies assessed under the peer review method completed the Instrument, including their ratings, and sent it to a peer review team. This review team completed the Ratings Summary Form on the basis of written responses and documentation supplied with the Instrument. The least agreement occurred for this assessment method, in which agencies and assessors did not work together at any stage in determining ratings. Agreement occurred on average only 60% of the time. Table 4.3 shows that peer-reviewed agencies rated themselves higher than did their assessors. The average rating across all standards according to agencies was 1.64; the average rating across all standards according to assessors was 1.20. This difference was significant (t = 3.21, p<0.01).

Independent or external assessor

A small number of assessments were undertaken by independent assessors or external assessors. Table 4.3 shows that agencies gave themselves similar ratings to those the assessor gave them. The average rating across all standards according to agencies was 1.66; the average rating across all standards according to assessors was 1.64. This difference was not significant (t = 0.14).

4.3 Inferential tests of the difference between methods

If the assessment methods are each equally effective at reflecting the true service quality of agencies then the Instrument should produce the same average performance scores across each assessment method (within a degree of error to be expected by chance). Conversely, if differences greater than chance occur between the average performance of the agencies in each assessment type, then this difference could be proposed to be the result of factors associated with the assessment method. In this section significance tests are conducted of the differences between the means scores for each assessment method.

The rating for agencies in each assessment method, averaged over standards, is presented in Table 4.3. For all of the agencies involved in the pilot, for whom both Instruments and Ratings Summary Forms were received ($n = 74$), the average rating across standards according to agencies was 1.63 whereas the average rating across standards according to assessors was 1.50 (out of a maximum possible score of 2).

As Table 4.3 indicates, both agencies and assessors rated agencies in the joint assessment method higher than agencies overall. This difference was significant for both the agency joint assessment rating compared with the average agency rating ($t = 1.80$, $p < 0.05$) and for the assessor joint assessment rating compared with the average assessor rating ($t = 4.09$, $p < 0.005$). It is possible that the collaborative approach of this method may have resulted in agencies achieving scores higher than average and, considering the results of Section 4.2, possibly higher than their actual performance warranted.

States which carried out joint assessments and self-assessments with verification were asked to select agencies for each assessment type such that agencies were equally likely to have a joint assessment as a self-assessment with verification. It can be assumed, then, that there is no reason for joint-assessed agencies to have significantly better performance against the standards than agencies undergoing self-assessment with verification. The information sharing and cooperative process of the joint assessment may have influenced both agencies and assessors to take a more lenient approach.

Agencies which were assessed by self-assessment with verification did not rate themselves significantly higher or lower than the average for agencies in the total pilot sample. However, the assessors who conducted the verifications rated these agencies significantly higher than the average for all agencies (assessor ratings) in the pilot sample ($t = 3.71$, $p < 0.005$). The opportunity for information exchange between agency and assessor may have been responsible for the higher than average ratings against the standards by assessors.

Agencies undertaking self-assessment (with random verification visits) did not rate themselves significantly differently from other agencies in the pilot (the average ratings were statistically equivalent for the self-assessment group and the total sample ($t = 1.38$, $p < 0.10$). Likewise, the assessors of these agencies also rated them similarly to other agencies in the pilot ($t = 1.37$, $p < 0.10$).

The average rating given by peer reviewers was significantly lower than the average rating given by assessors in general ($t = 5.17$, $p < 0.005$). The self-ratings of agencies in the peer review sample were not significantly different from the average for all agencies

in the pilot. There is no reason to believe that the performance of agencies which underwent a peer review would be substantially poorer against the standards. However, there is reason to believe that the paper-based reviews conducted by peers would not be as accurate as other methods of assessment that involved more direct contact with agencies.

The agency assessments performed by independent/external assessors were not found to be significantly different from average agency or assessor ratings. The small number of cases in this assessment method precludes drawing firm conclusions from this data.

This analysis reveals that there are significant differences in the appraisal outcomes when different assessment methods are used. Where no interview with the agency occurs in the process of assessment, assessors rate agencies lowest against the standards. Agencies obtain significantly higher ratings from assessors when assessment incorporates an interview with the agency and assessment is undertaken collaboratively, as in a joint assessment.

A number of questions are raised by the finding that the appraisal outcomes for self-assessment with verification are higher, on average, than the appraisal outcomes for the randomly verified self-assessments. For the random verification visits, assessors had received Instruments prior to their visit, allowing them to form a preliminary opinion of the agencies' performance. This was not the case for the self-assessment with verification method. This preliminary opinion may have made assessors less inclined to make allowances for the individual circumstances of agencies. There may also have been effects that flowed from the method of informing agencies that only a sample would be verified. In this case, agencies could not count on receiving the assistance of an assessor if they could not complete the Instrument. They were required to sort through all difficulties unaided. The verification process was thus less collaborative and one that may have seemed to some to be more like having an assessment twice rather than receiving valued assistance and cooperation in achieving performance goals. From the assessor's position, the random verification method may not have placed them in a position to cooperatively assist agencies, but rather required them to check ratings in a more 'police-like' manner.

4.4 The effect of assessment type on rater reliability

The method of assessment may have affected the reliability of the Instrument. An examination of rater reliability is undertaken for joint assessments, self-assessments with verification, and peer review, although results should be interpreted with some caution since only five assessments were available for analysis in each assessment type.

The reliability of ratings given by independent/external assessors is not examined due to the low number of these assessments undertaken and the difficulty of performing reliability interviews in a sufficient number of these to enable comparison.

The inter-rater reliability for agencies which undertook self-assessment but received a verification visit was examined in the previous chapter. Desk audits of these

Instruments were compared with assessor ratings, and agreement occurred less than half the time.

When inter-rater reliability is assessed by a visiting reliability rater, the highest reliability between raters occurred for those agencies which had undertaken a self-assessment with verification. Agreement between raters averaged 86% across standards. (For this method, two Ratings Summary Forms were not returned by assessors, leaving a sample of only three. Conclusions must therefore be tentative only.)

For joint assessments, the reliability between the visiting rater and the assessor, as indicated by the average agreement, was moderately high (79% averaged over all standards, with agreement as low as 40% on some standards).

Of the reliability assessments made by visiting agencies, the agreement between the reliability rater and the assessor was lowest for agencies assessed by peer review (agreement averaged 62% across standards).

4.5 Summary

4.5.1 Findings

- Assessors involved in all assessment methods highly recommend the inclusion of an agency visit in the appraisal process. Analyses of the peer review method indicate that a purely paper-based review is not as reliable, and does not produce as much agreement between assessors and agencies, as one involving more direct contact.
- Assessors believed that, while the assessments undertaken using the Instrument and visiting agencies were very beneficial, they would not necessarily reveal all problems in service quality, particularly where agencies did not want to reveal such flaws. Information about service quality from other sources was also necessary, with consumer input being one very important source.
- If there is to be a choice between assessing an individual agency via the joint assessment or self-assessment with verification method, the decision should take into account the ability of an agency to adequately complete the Instrument without assistance.
- Assessors considered that agencies would have benefited from additional information (such as that contained in the assessor guidelines) to help them decide on their own ratings.
- Agencies undertaking self-assessment would have benefited from a 'preparation' period, to provide education and training about the standards, the Instrument and the assessment process.
- The joint assessment method was seen as particularly beneficial to new or small agencies—but was seen as unnecessarily time consuming for assessors in most cases.
- Assessors considered that the self-assessment with verification method would have been less time consuming and more productive if assessors had received the agencies' completed documentation prior to the visit.

- The peer review process was seen to have great potential benefit to agencies by encouraging closer service-provider networking and information sharing. However, the resources needed to complete a peer review process were considered critical, and sometimes prohibitive, for some agencies—particularly small agencies and agencies in remote areas operating under adverse conditions.
- While the use of an external or independent assessor was seen to offer objectivity to the assessment process, familiarity with services and their environments was also seen to be important in determining agency performance.
- Assessors' overall appraisals of agency performance were in closest correspondence to the Instrument Scores for the self-assessment and peer review methods, where assessors and agencies engaged in minimal dialogue in the determination of ratings. The correspondence was such that individual standards ratings may have little validity if too heavily influenced by the assessor's general view of the quality of the agency's service. Ratings by these methods were also harshest. In joint assessments, where assessment dialogue was at a maximum, correspondence between overall appraisals and Instrument Scores was at its lowest, and assessors reported difficulties in being objective. Instrument ratings were highest by the joint assessment method. Self-assessment with verification was found to have the most acceptable level of concurrent validity between assessors' overall appraisals and Instrument Scores.
- Agreement on ratings between agencies and assessors was highest for agencies in which a visit to the agency by the assessor was an integral part of the assessment. The greatest agreement between assessors and agencies regarding ratings occurred for joint assessments. There was no significant difference between the ratings of agencies and assessors in the self-assessment with verification sample. Agencies which undertook self-assessment (with random verification visits) disagreed significantly with their assessors. The lowest agreement occurred between agency ratings and reviews done by documentation only (with no assessor visit).
- Although limited by a small sample size, a comparison of rater reliability between the joint assessment, self-assessment with verification and peer review methods revealed that the self-assessment with verification method had the highest rater reliability and peer review had the lowest.

4.5.2 Recommendations

- The Instrument should be revised to incorporate appropriate information from the assessor guidelines. This revised version can be found in Appendix A.
- It is recommended that the assessment process include a visit to the agency. In particular, the self-assessment with verification model presents as the method most likely to produce valid and reliable Instrument Scores.
- Peer review is not recommended unless the method is modified to include face-to-face contact with agencies.
- Joint assessment, while of considerable value to those agencies experiencing difficulty interpreting and completing the Instrument for the particular circumstances of their agency, should not be considered as optimal for providing reliable or valid Instrument Scores.

- A flexible approach to the selection of the most appropriate appraisal process can be adopted where necessary, with assessment methods requiring more intensive resource allocation being targeted on those agencies identified as needing most attention.
- Training for agencies in completing the Instrument and the appraisal process should be given to agencies before assessment begins.
- Wherever possible, assessors should receive copies of the agency's documentation (i.e. completed Instrument, ratings and supporting documentation) before making a visit to the agency. The Instrument received prior to the visit should not, however, be taken as the basis for determining the agency's performance appraisal. Ratings should be determined during the assessment visit with the agency's input so that the assessment process is one of consultation and education rather than 'police-like' auditing and double-checking.