



Reliability of the Australian National Aged Care Classification shadow assessments

Results from dual assessments conducted in
the AN-ACC shadow period

23 March 2022



ACN 087 047 809
ABN 29 087 047 809
www.taylorfry.com.au

Taylor Fry Pty Ltd



Table of contents

1	Executive Summary	4
2	Introduction	5
2.1	Background	5
2.2	Reliability of AN-ACC	5
3	Methods	7
3.1	Dual assessment process	7
3.2	Inter-rater reliability test statistics	8
3.3	Hypothesis testing.....	9
4	Results	10
4.1	Dual assessment summary	10
4.2	Reliability analysis	10
5	Discussion	17
	Appendix A FIM Cognition breakdown by Assessor Occupation	18

1 Executive Summary

The Australian National Aged Care Classification (AN-ACC) assessment tool will be core to the new AN-ACC funding model for residential aged care. So it is important that AN-ACC assessments are reliable, such that equivalent AN-ACC assessments of the same individual provide consistent results.

A robust process was followed to obtain 533 dual AN-ACC assessments for a representative sample of the residential aged care population. In each dual assessment the same resident was independently assessed by two assessors. This has enabled thorough investigation of the (inter-rater) reliability of the AN-ACC assessment tool. Reliability was tested at an overall level and for cohorts of assessors grouped by their Assessment Management Organisation (AMO) and their professional training.

The AN-ACC assessment tool had excellent reliability, overall and within almost all cohorts. Agreement on the de Morton Mobility Index (DEMMI) was excellent across all AMOs and professional cohorts. This is important because DEMMI plays a critical role in the overall determination of AN-ACC class.

Some instruments and cohorts, while having very good reliability, had weaker levels of agreement relative to others. These points are noted for completeness but do not call into question reliability of AN-ACC:

- Behaviour Resource Utilisation Assessment (BRUA) Disruptive scores had the lowest reliability of all instruments within AN-ACC. This is consistent with assessors' feedback on the BRUA Disruptive, that emotional dependence is very subjective and quite difficult to score for most residents.
- Assessor profession did not adversely impact reliability of the AN-ACC assessment overall or on specific instruments.

2 Introduction

2.1 Background

The AN-ACC will replace the Aged Care Funding Instrument (ACFI) on 1 October 2022. AN-ACC assessments will be the foundation of the new funding model that will apply to all permanent aged care residents in Australia.

To prepare for the AN-ACC transition, from April 2021 onwards all residents of an Australian Government-funded residential aged care facility (except residents who are nearing end of life) will receive an the AN-ACC assessment in addition to current assessment for ACFI funding arrangements. This period is referred to as the “shadow assessment period”.

Six independent Assessment Management Organisations (AMOs) located around Australia have been contracted to conduct AN-ACC assessments on behalf of the Australian Government. All AN-ACC assessors have clinical qualifications as registered nurses, occupational therapists, or physiotherapists, and have considerable experience working in aged care settings. They have also undergone comprehensive training on the AN-ACC assessment process.

2.2 Reliability of AN-ACC

Validity and reliability are two of the most important characteristics of a well-designed assessment. In this context, validity refers to the degree to which the AN-ACC assessment tool assesses what it intends to assess i.e. cost drivers in residential aged care including frailty, mobility, activities of daily living (ADL) function, cognition, communication and behaviour. Validity of the AN-ACC assessment tool has previously been tested in two ways.

- First, the tool has been developed using a series of pre-existing clinical instruments that have been validated in other populations e.g. the de Morton Mobility Index (DEMMI) has been validated in many different cohorts^{1,2}.
- Second, the Resource Utilisation and Classification Study (RUCS) assessed correlations within and between the related clinical tools in AN-ACC. There were strong relationships between various items and sub-scales that were consistent with what would be expected clinically. These relationships were used as the basis for a series of validation tests which are automatically programmed into the shadow data assessment tool. Validation error rates are closely monitored as part of quality assurance.

It is important that AN-ACC assessments produce **reliable** results. That is, that repeated or equivalent AN-ACC assessments of the same individual provide consistent results. As the AN-ACC is a relatively new instrument little is known about its reliability and the factors that influence inter-rater agreement. The AN-ACC Trial Report concluded that the training and clinical supports developed to equip assessors to undertake assessments were effective, however, the following areas of improvement were identified to optimise assessment consistency and accuracy:

- Increasing the length of training from two to three days, with additional time spent on providing training on assessing residents with complex cognitive variables and those from diverse linguistic and cultural backgrounds.
- Post-training assessor clinical support structures.

These changes were both implemented early 2021, and prior to commencement of dual assessments in the shadow assessment period.

¹ <https://onlinelibrary.wiley.com/doi/10.1111/j.1741-6612.2010.00497.x>

² <https://doi.org/10.1080/09638288.2018.1430176>

During the shadow assessment period, reliability of the shadow assessments was further tested via dual assessment and inter-rater reliability testing. **The purpose of this report is to summarise the dual assessment process and results on the reliability of AN-ACC shadow assessments.**

3 Methods

3.1 Dual assessment process

The primary method to test reliability was dual assessment (DA). In a dual assessment, a single resident was independently assessed by two different assessors, at the same time. All conditions in the DAs were the same as assessment by a single assessor, i.e. the same information and instructions were provided to the assessors and AMOs at the beginning of the assessment process.

In the shadow roll-out, each AMO was assigned responsibility for completing the assessments in specific facilities. For this reason, all dual assessments were completed by two assessors from the same AMO.

The target number of dual assessments was chosen to achieve a representative sample for each of the strata in Table 3.1. The actual number of dual assessments completed (533) was lower than target due to COVID-related restrictions on assessment activity during the dual assessment period (September 2021 to December 2021).

The dual assessors for each DA were randomly selected depending on who was on site at a facility at the time scheduled for dual assessments³. Consideration was given to ensure there was a mix of professions across dual assessors (i.e. a mix of registered nurses, occupational therapists, and physiotherapists).

The residents who were assessed by two assessors were also selected randomly on site (by name or ID number prior to viewing the resident). That is, residents were not selected to display particular traits or diagnoses. No instructions were given to the resident regarding the dual assessment.

Table 3.1 - Split of dual assessments (by no. of residents) across different types of aged care facilities

Remoteness	Type	No. beds	Population size	Dual assessments
MMM1	For profit	0 to 59	6,715	23
		60 to 119	31,559	68
		120 to 179	19,588	72
		180+	1,814	12
	Other	0 to 59	12,987	29
		60 to 119	31,655	80
		120 to 179	17,873	64
		180+	3,933	4
MMM2	For profit	Any	5,500	27
	Other	Any	10,743	44
MMM3	Any	Any	15,807	34
Regional (MMM4-5)	Any	Any	21,656	71
Remote (MMM6-7)	Any	Any	943	5
Homeless Specialised	Any	Any	1,712	9

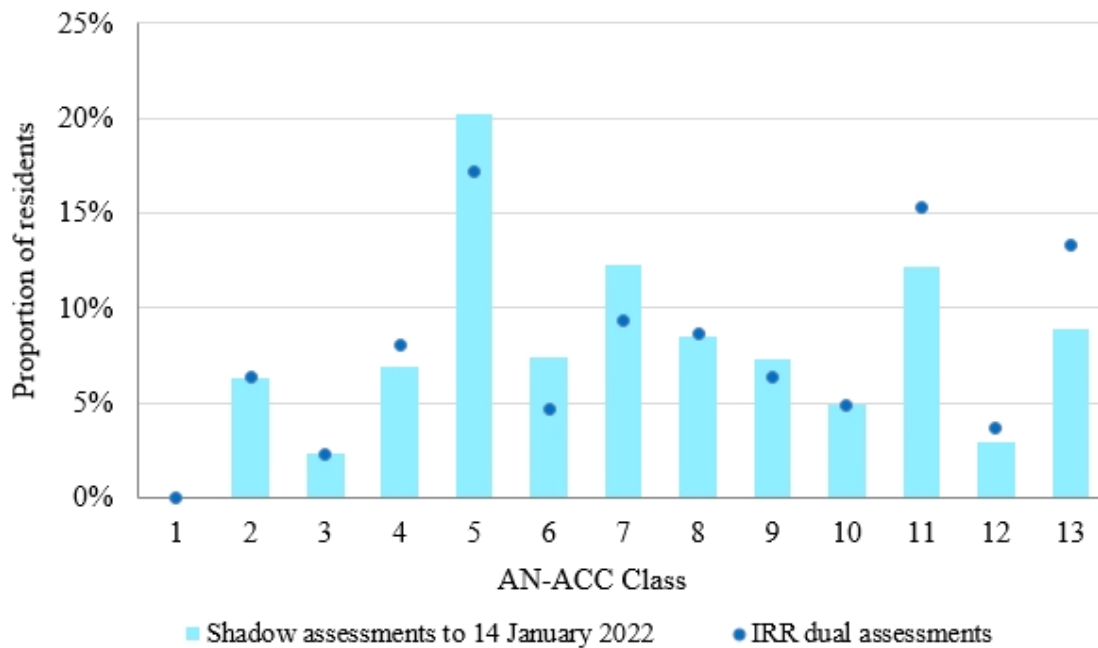
³ Of the 533 dual assessments completed around 5% (or 25 out of 533) were done by a senior assessor – assessor pair rather than an assessor – assessor pair. We can see no reason why this would bias the results.

3.1.1 Comments on dual assessment data

The actual number of dual assessments completed (by number of residents assessed by two assessors) is shown in Table 3.1. A reasonable spread of dual assessments across different types of facilities (defined by groups of facilities in which a different size, ownership type and remoteness) was achieved.

The dual assessment data also represents a good spread across different types of residents, as illustrated in Figure 3.1. The casemix of dual assessments is broadly comparable to the casemix of the cumulative shadow assessments to 14 January 2022.

Figure 3.1 - Casemix distribution of dual assessments by AN-ACC class versus cumulative shadow assessments to 14 January 2022



3.2 Inter-rater reliability test statistics

Cohen's kappa (Kappa) measures the degree of agreement between a pair of variables and is frequently used as a metric of inter-rater agreement for categorical variables. Kappa compares the probability of agreement to that expected if the ratings are independent. The values of range lie in $[-1, 1]$ with

- Kappa = 1 representing perfect agreement
- Kappa = 0 representing no agreement
- Kappa < 0 implies that the agreement is worse than random.

The standard for a “good” or “acceptable” Kappa value is not clearly specified. Commonly cited is Fleiss (1981) guidelines that Kappa greater than or equal to 0.75 is excellent⁴, whereas other papers use lower thresholds (e.g. Flack et al. (1988) use a guide of values of kappa above 0.60 show good to excellent agreement between the two raters' scores⁵). We use the more stringent threshold of Kappa greater than or equal to 0.75.

⁴ Fleiss, J.L. (1981). *Statistical methods for rates and proportions* (2nd ed.). New York: John Wiley. [ISBN 978-0-471-26370-8](#).

⁵ Flack, V., Afifi, A., Lachenbruch, P. and Schouten, H. (1988). *Sample size determinations for the two-rater kappa statistic*, *Psychometrika*, 53, issue 3, p. 321-325.

Cohen’s weighted kappa is typically used for categorical data with an ordinal structure. Weighted kappa allows for the fact that some disagreements are more material than others (or that not all differences between raters should be treated as equally important). E.g. two assessors who assess the same resident as Class 2 versus Class 3 have a lower level of disagreement than two assessors who assess the same resident as Class 2 versus Class 13; weighted kappa allows for this. Weights used in testing are shown in Table 3.2.

Table 3.2 – Variables assessed for reliability, with variable type and test statistics used

Variable	Type	Test statistic
AN-ACC Class	Categorical, with close to ordinal structure ⁶	Cohen’s Weighted Kappa (weight by RVU)/ Correlation co-efficient
DEMMI Total	Ordinal	Cohen’s Weighted Kappa (weight by total score)/ Correlation co-efficient
FIM Cognition Total	Ordinal	Cohen’s Weighted Kappa (weight by total score)/ Correlation co-efficient
RUG-ADL Total	Ordinal	Cohen’s Weighted Kappa (weight by total score)/ Correlation co-efficient
Braden Scale Total	Ordinal	Cohen’s Weighted Kappa (weight by total score)/ Correlation co-efficient
BRUA Disruptive Score	Ordinal	Cohen’s Weighted Kappa (weight by total score)/ Correlation co-efficient

Both Kappa and correlation coefficients can be used to assess reliability of ordinal rating scales⁷. For ordinal rating scales, the same conclusions about inter-rater reliability are generally arrived at whether Kappa or correlation coefficients are used⁵.

As shown in Table 3.2, the instruments that make up AN-ACC typically have ordinal scales. For this reason, we use Pearson’s Correlation coefficient as a secondary measure of reliability.

3.3 Hypothesis testing

We test the null hypothesis that $Kappa > 0.75$ i.e. that inter-rater agreement is “excellent”. To do this we calculate a 95% confidence interval for Kappa⁸. The standard error formula incorporates the weights and the number of dual assessments completed. For segments in which fewer dual assessments are completed, the confidence interval is wider.

If the lower bound of the 95% confidence interval is above 0.75, we can conclude that inter-rater agreement is excellent in that group. Alternatively, the null hypothesis is rejected if the lower bound of the 95% confidence interval is below 0.75. If lower bound of the 95% confidence interval is below 0.75, but above 0.60, we can conclude that inter-rater agreement is “very good” in that group⁹. As these formulae are based on a normal approximation, the adequacy of this approach depends on sample sizes being “large” (around 100 or more) and the approximate normality of the distribution of Kappa¹⁰.

⁶ In most (but not all) cases Relative Value Units (RVUs) increase as AN-ACC class increases.

⁷ <https://link.springer.com/article/10.1007/s00357-021-09386-5>

⁸ See formulae here <https://www.real-statistics.com/reliability/interrater-reliability/weighted-cohens-kappa/>

⁹ Flack, V., Afifi, A., Lachenbruch, P. and Schouten, H. (1988). Sample size determinations for the two-rater kappa statistic, *Psychometrika*, 53, issue 3, p. 321-325.

¹⁰ Using method derived here <https://doi.org/10.1007/BF02294215>.

4 Results

4.1 Dual assessment summary

Overall, the assessment data featured:

- 533 residents who had two separate assessments conducted
- 1,066 assessments overall
- 217 assessors who completed the dual assessments.

The dual assessments were completed by assessors of varying occupations as broken down in Table 4.1.

Table 4.1 – Distribution of assessor occupation within dual assessments

Assessor Occupation	Number of Assessors	Proportion of Assessors
Registered Nurse	163	75%
Occupational Therapist	14	6%
Physiotherapist	40	18%
Total	217	100%

We note that all assessors who completed a dual assessment have conducted at least one AN-ACC assessment prior to the dual assessment, although they had a range of experience levels (measured by the number of days between the earliest dual assessment, and the earliest assessment completed in the AN-ACC shadow assessments). As Table 4.2 shows, most assessors had over 180 days between their first dual assessment and first AN-ACC shadow assessments, with only 8% of assessors having less than 90 days of experience. The average level of assessor experience is 181 days between the first dual assessment and first AN-ACC shadow assessment.

Table 4.2 – Distribution of dual assessment assessor experience, measured by the number of days between the earliest dual assessment, and the earliest assessment completed in the shadow assessments

Assessor experience	Number of Assessors	Proportion of Assessors
0 – 90 days	17	8%
90 – 180 days	70	32%
180 – 270 days	130	60%
Total	217	100%

4.2 Reliability analysis

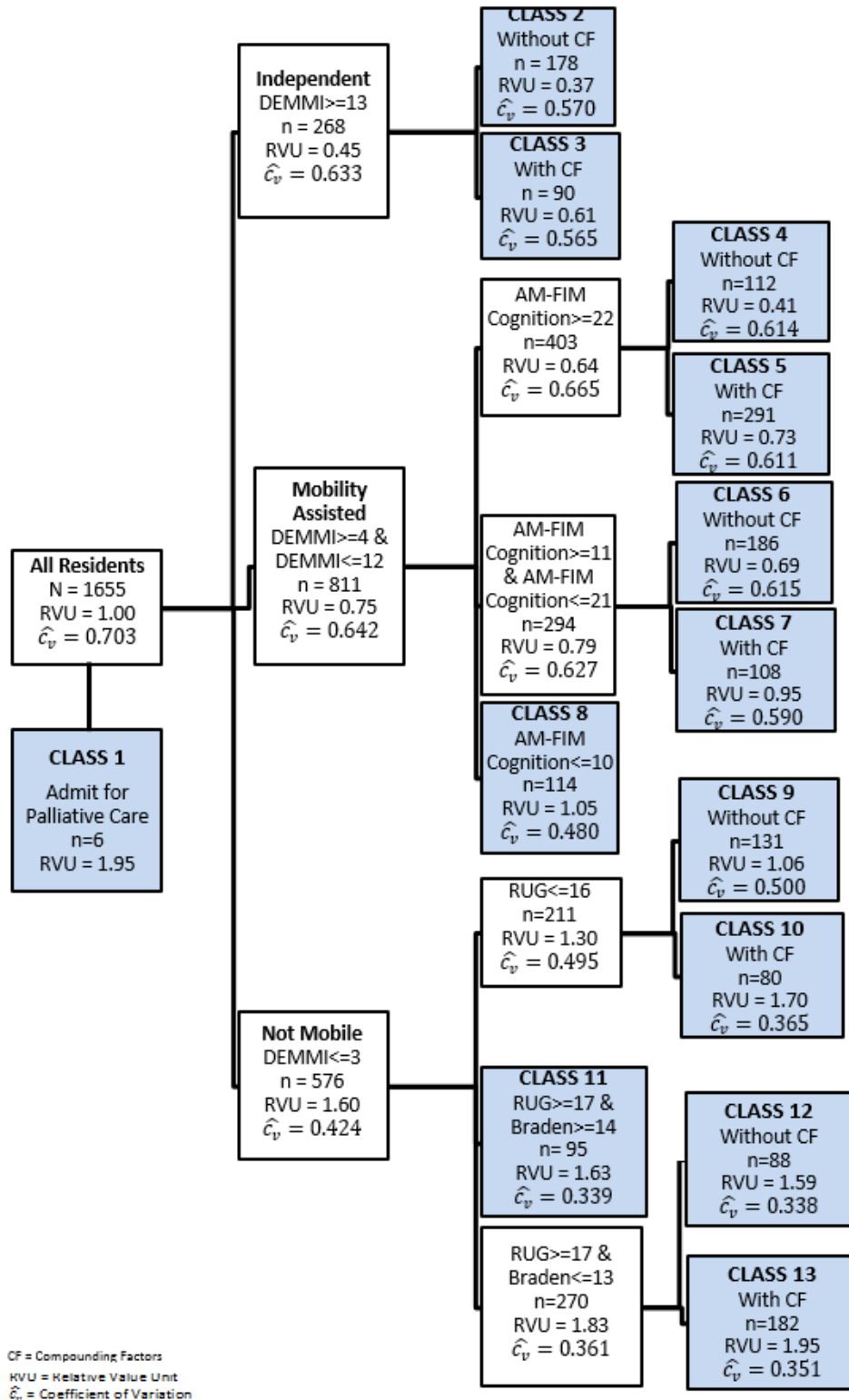
The reliability analysis looks at the calculated Weighted Kappa, 95% confidence interval of Weighted Kappa, and correlation coefficient of varying metrics for dual assessments, split into different cohorts.

The analysis begins by looking at the overall agreement of AN-ACC ratings for all dual assessments, and then splits into different assessor cohorts (AMOs and assessor occupation groups). To understand which parts of the assessment may be causing stronger and weaker agreement, we also look at the agreement of

underlying instrument scores which are used when calculating AN-ACC class. These include the DEMMI Total, FIM Cognition Total, RUG-ADL Total, Braden Scale Total, and BRUA Disruptive Score.

The structure of the AN-ACC branches is summarised in Figure 4.1.

Figure 4.1 – Structure of AN-ACC branches and calculations



Source: Eagar K, McNamee J, Gordon R et al. (2019) The Australian National Aged Care Classification (AN-ACC). *The Resource Utilisation and Classification Study: Report 1*. Australian Health Services Research Institute, University of Wollongong. ISBN: 978-1-74128-295-5.

One of the most relevant takeaways from Figure 4.1 is that there is a hierarchy of instruments within the AN-ACC. The DEMMI Total determines the *first* branch: Independent Mobility, Mobility Assisted or Not Mobile. Residents assessed as Independent Mobility must be in Class 2 or 3, residents assessed as Mobility Assisted must be in Class 4 to 8 and residents assessed as Not Mobile must be in Class 9 to 13. Following this, FIM-Cognition Total, RUG-ADL Total, Braden Scale Total are the next instruments used to determine AN-ACC Class. Thirdly, compounding factors (CF) may be used to determine the final AN-ACC Class, which the BRUA Disruptive score is used to calculate.

This means that the reliability of some instruments may be more important than others, when assessing the overall reliability of AN-ACC. For example, as DEMMI Total determines which branch of AN-ACC a resident is in, the reliability of DEMMI would have a greater impact on the reliability of AN-ACC than the reliability of BRUA Disruptive.

4.2.1 Overall analysis

The overall reliability of the AN-ACC assessment, as measured by Weighted Kappa, is excellent (as the Weighted Kappa is 0.90, well above the 0.75 threshold with a 95% confidence interval of (0.88,0.92)). The overall reliability for individual instrument in Table 4.3 is also excellent.

Table 4.3 – Agreement statistics for all dual assessments, by metric

Metric	Weighted Kappa	95% Confidence – Lower bound	95% Confidence – Upper bound	Pearson’s Correlation
AN-ACC	0.90	0.88	0.92	0.96
DEMMI Total	0.91	0.90	0.93	0.98
FIM Cognition Total	0.86	0.84	0.88	0.96
RUG-ADL Total	0.92	0.91	0.94	0.98
Braden Scale Total	0.87	0.85	0.89	0.96
BRUA Disruptive Score	0.86	0.83	0.90	0.90

In summary, Table 4.3 shows that:

- There is excellent reliability across each metric (as measured by Weighted Kappa and the secondary check of high Pearson’s correlation values).
- Scorings for DEMMI Total and RUG-ADL Total have excellent agreement and correlation. So the reliability of DEMMI is contributing to the excellent overall reliability of AN-ACC.
- Scorings for FIM Cognition Total and BRUA Disruptive Scores have excellent reliability, but reliability is weaker when compared to the other instruments.
 - We note that assessors have provided feedback around the BRUA Disruptive score, saying that emotional dependence is very subjective and quite difficult to score for most residents.
 - Similarly, issues may arise with FIM Cognition which requires assessors to measure comprehension, expression, social interaction, problem solving and memory.

The following sections analyse the agreement of the chosen metrics within specific cohorts, to investigate if there may be groups of assessors displaying stronger or weaker agreement, hence affecting the reliability of assessments.

4.2.2 Analysis by Assessor Occupation

This section investigates the agreement of assessments based on assessor occupation. As discussed by Westera et al. (2019), AN-ACC assessors should be credentialed individuals from appropriate professional groups whose undergraduate degree includes function and mobility as a core component, i.e. nursing (registered), occupational therapy and physiotherapy¹¹. These three undergraduate degrees have a different focus and emphasis on mobility and other elements of functioning, so it is possible that this may influence how an assessor rates residents on components of the AN-ACC assessment.

Also the quality assurance of AN-ACC shadow assessments has found that, in some weeks, registered nurses had a different propensity to assess a resident as ‘Not Mobile’ than non-registered nurses, indicating that the assessor occupation group may influence the reliability of assessments.

To test whether reliability differs depending on the assessor’s occupation, we consider two assessor occupation groups – Registered Nurses, and Non-Registered Nurses (comprising of Occupational Therapists and Physiotherapists). A resident with a dual assessment can fall into one of three categories:

- Both assessments completed by assessors in the Registered Nurses group,
- Both assessments completed by assessors in the Non-Registered Nurses group, or
- One assessment completed by an assessor in the Registered Nurse group, and the other completed by an assessor in the Non-Registered Nurse group (‘Mixed’).

Overall, assessments from each assessor occupation group showed excellent reliability (with a strong Weighted Kappa and correlation). Table 4.4 shows the results by each assessment occupation group.

Table 4.4 - Agreement statistics for AN-ACC, split by assessor occupation group

Assessor Occupation Group	Weighted Kappa	95% Confidence – Lower bound	95% Confidence – Upper bound	Pearson’s Correlation	Number of dual assessments
Registered Nurses	0.90	0.88	0.93	0.96	341
Non-Registered Nurses	0.92	0.85	1.00	0.98	45
Mixed	0.89	0.86	0.93	0.97	147

Our key observations from Table 4.4 are:

- There is excellent reliability of AN-ACC across all assessor occupation groups.
- There difference in agreement between each assessor occupation group is very small, indicating that assessor occupation does not have a material influence on the overall AN-ACC casemix.

DEMMI

The agreement statistics for DEMMI Total are consistent with the agreement statistics for AN-ACC, as seen in Table 4.5.

¹¹ Westera A et al. (2019) *The AN-ACC assessment model. The Resource Utilisation and Classification Study: Report 2*. Australian Health Services Research Institute, University of Wollongong. ISBN 978-74128-296-2

Table 4.5 - Agreement statistics for DEMMI Total, split by assessor occupation group

Assessor Occupation Group	Weighted Kappa	95% Confidence – Lower bound	95% Confidence – Upper bound	Pearson’s Correlation	Number of dual assessments
Registered Nurses	0.92	0.90	0.94	0.98	341
Non-Registered Nurses	0.92	0.86	0.98	0.97	45
Mixed	0.90	0.87	0.93	0.97	147

In summary, Table 4.5 shows:

- There is excellent reliability of DEMMI Total (measured by Weighted Kappa and Pearson’s correlation values).
- There difference in agreement between each assessor occupation group is very small, indicating that assessor occupation does not have a material influence on the overall AN-ACC casemix.

FIM Cognition

Unlike the agreement statistics for AN-ACC and DEMMI Total, the agreement statistics for FIM Cognition Total show a material difference based on the assessor occupation group of the dual assessors, as seen in Table 4.6.

Table 4.6 - Agreement statistics for FIM Cognition Total, split by assessor occupation group

Assessor Occupation Group	Weighted Kappa	95% Confidence – Lower bound	95% Confidence – Upper bound	Pearson’s Correlation	Number of dual assessments
Registered Nurses	0.87	0.85	0.90	0.96	341
Non-Registered Nurses	0.90	0.84	0.95	0.97	45
Mixed	0.81	0.76	0.85	0.93	147

Table 4.6 illustrates that:

- There is an overall excellent agreement of FIM Cognition Total measured by Weighted Kappa for assessments who were either only assessed by Registered Nurses, or only assessed by Non-Registered Nurses, also supported by strong Pearson’s correlation values.
- Assessments completed by dual assessors from varying occupation groups showed a materially lower level of agreement measured by Weighted Kappa, when compared to assessments completed by assessors from the same occupation group. However, the assessment agreement still remained above the 0.75 ‘excellent’ threshold.
- The 95% confidence interval of Weighted Kappa for the “Mixed” group does not overlap with the 95% confidence interval of Weighted Kappa for assessments by Registered Nurses. This suggests a statistically significant occupational effect on the reliability of this measure. This effect is highly unlikely to influence overall AN-ACC results, however, because agreement across all groups is excellent.

Detailed results for each of the instruments within FIM Cognition can be found in Appendix A.

RUG-ADL

Dual assessments showed strong agreement in RUG-ADL Total when split by assessor occupation group, although there was a material difference in agreement between each split.

Table 4.7 - Agreement statistics for RUG-ADL Total, split by assessor occupation group

Assessor Occupation Group	Weighted Kappa	95% Confidence – Lower bound	95% Confidence – Upper bound	Pearson's Correlation	Number of dual assessments
Registered Nurses	0.93	0.91	0.94	0.98	341
Non-Registered Nurses	0.96	0.94	0.99	0.99	45
Mixed	0.91	0.88	0.94	0.97	147

In summary, Table 4.7 shows:

- There is an overall excellent agreement of RUG-ADL Total measured by Weighted Kappa across each assessor occupation group, also supported by strong Pearson's correlation values.
- Assessments completed by Non-Registered Nurse assessors had the strongest level of agreement of RUG-ADL Total, with a Weighted Kappa of 0.96 and correlation of 0.99.
- Mixed assessments had strong, but the weakest level of agreement of RUG-ADL Total, with a Weighted Kappa of 0.91.

Braden Scale

Agreement levels for Braden Scale Total were excellent, despite being a little weaker than some of the previously analysed metrics.

Table 4.8 – Agreement statistics for Braden Scale Total, split by assessor occupation group

Assessor Occupation Group	Weighted Kappa	95% Confidence – Lower bound	95% Confidence – Upper bound	Pearson's Correlation	Number of dual assessments
Registered Nurses	0.87	0.85	0.89	0.96	341
Non-Registered Nurses	0.89	0.84	0.94	0.98	45
Mixed	0.85	0.81	0.89	0.95	147

Table 4.8 shows:

- There is an overall strong agreement of Braden Scale Total measured by Weighted Kappa across each assessor occupation group, also supported by strong Pearson's correlation values.
- Similar to agreement in RUG-ADL Total, the strongest levels of agreement for Braden Scale total were found in Non-Registered Nurses, and the weakest was found in Mixed assessments.

BRUA Disruptive

Agreement levels for BRUA Disruptive scores were strong, with all point estimates of Weighted Kappa being above 0.75. However, the lower bound of the Weighted Kappa 95% confidence interval fell below 0.75 for assessments by Non-Registered Nurses, and Mixed assessments.

Table 4.9 – Agreement statistics for BRUA Disruptive, split by assessor occupation group

Assessor Occupation Group	Weighted Kappa	95% Confidence – Lower bound	95% Confidence – Upper bound	Pearson’s Correlation	Number of dual assessments
Registered Nurses	0.88	0.84	0.92	0.92	341
Non-Registered Nurses	0.85	0.73 ¹²	0.97	0.90	45
Mixed	0.83	0.75	0.90	0.87	147

Our key observations from Table 4.9 are:

- Consistent with previous results, there is a strong, but weaker level of agreement of BRUA Disruptive scores for dual assessments where the assessors come from Mixed occupation groups.
- Dual assessments completed by Registered Nurses had the strongest level of agreement of BRUA Disruptive scores, with a Weighted Kappa of 0.88.
- The lower bounds of the Weighted Kappa 95% confidence interval for Non-Registered Nurses and assessors from Mixed occupation groups are quite close to our set threshold of 0.75 (0.73 and 0.75, respectively), whereas for Registered Nurses this is higher (0.84). This indicates that assessments performed by Non-Registered Nurses may tend to show less agreement for BRUA Disruptive scores.

Overall, the agreement of AN-ACC and selected instruments was very good or excellent across each assessor occupation group, although dual assessments performed by a mixed set of assessor occupations tended to show a weaker level of agreement.

¹² As the lower bound of the 95% confidence interval is below 0.75, we cannot say the reliability is excellent, but it is very good.

5 Discussion

Overall, the Department of Health have followed a rigorous process to obtain a representative sample of the aged-care population, as reflected in Figure 3.1. This has enabled thorough investigation of the level of reliability of the AN-ACC assessment tool.

Overall, across 533 dual assessments, AN-ACC showed excellent reliability, with a Weighted Kappa of 0.9 (with a 95% confidence interval ranging from 0.88 to 0.92). With this, we can say with high confidence that there is excellent reliability in AN-ACC shadow assessments.

There were instruments and cohorts which tended to have stronger levels of agreement than others. Importantly, agreement of DEMMI Total and RUG-ADL Total was excellent across all AMOs and assessor occupation cohorts. This is encouraging because DEMMI plays a critical role in the overall determination of AN-ACC class.

Some instruments and cohorts, while having strong reliability in absolute terms, had weaker levels of agreement relative to other instruments and cohorts. These minor points are noted for completeness but do not call into question reliability of AN-ACC:

- Agreement on the FIM Cognition instrument was weaker than for other instruments. FIM Cognition Total also tended to have weaker levels of agreement in dual assessments that were completed by a mixed group of assessor occupations, compared to assessments completed by the same group of assessor occupation.
- BRUA Disruptive scores showed a statistically significant difference in agreement between assessments completed by assessor from different occupation groups, when compared to assessments completed by two registered nurses.

Appendix A FIM Cognition breakdown by Assessor Occupation

Table B.1 – Weighted Kappa of instruments used to calculate FIM Cognition Total, split by assessor occupation

Assessor Occupation Group	FM Comprehension	FM Expression	FM Memory	FM Problem Solving	FM Social Interaction
Registered Nurses	0.85	0.87	0.84	0.85	0.82
Non-Registered Nurses	0.86	0.82	0.85	0.90	0.85
Mixed	0.76	0.75	0.82	0.82	0.75



'F
TAYLOR
FRY

www.taylorfry.com.au