AN-ACC Inter-rater Reliability Analysis

Department of Health and Aged Care

13 December 2024

Final report

Building a better working world

RELEASE NOTICE

Ernst & Young ("EY") was engaged on the instructions of the Commonwealth of Australia as represented by the Department of Health and Aged Care ("the Client") to perform an inter-rater reliability analysis ("the Project"), in accordance with the engagement agreement dated 11 June 2024 ("the Engagement Agreement").

The results of EY's work, including the assumptions and qualifications made in preparing the report, are set out in EY's report dated 13 December 2024 ("Report"). You should read the Report in its entirety including any disclaimers and attachments. A reference to the Report includes any part of the Report. No further work has been undertaken by EY since the date of the Report to update it.

Unless otherwise agreed in writing with EY, any party accessing the Report or obtaining a copy of the Report ("Recipient") agrees that its access to the Report is provided by EY subject to the following terms:

1. The Report cannot be altered.

2. The Recipient acknowledges that the Report has been prepared for the Client and may not be disclosed to any other party or used by any other party or relied upon by any other party without the prior written consent of EY.

3. EY disclaims all liability in relation to any party other than the Client who seeks to rely upon the Report or any of its contents.

4. EY has acted in accordance with the instructions of the Client in conducting its work and preparing the Report, and, in doing so, has prepared the Report for the benefit of the Client, and has considered only the interests of the Client. EY has not been engaged to act, and has not acted, as advisor to any other party. Accordingly, EY makes no representations as to the appropriateness, accuracy or completeness of the Report for any other party's purposes.

5. No reliance may be placed upon the Report or any of its contents by any party other than the Client. A Recipient must make and rely on their own enquiries in relation to the issues to which the Report relates, the contents of the Report and all matters arising from or relating to or in any way connected with the Report or its contents.

6. EY have consented to the Report being published electronically on the Department of Health and Aged Care website for informational purposes only. EY have not consented to distribution or disclosure of the Report beyond this.

7. No duty of care is owed by EY to any Recipient in respect of any use that the Recipient may make of the Report.

8. EY disclaims all liability, and takes no responsibility, for any document issued by any other party in connection with the Project.

9. A Recipient must not name EY in any report or document which will be publicly available or lodged or filed with any regulator without EY's prior written consent, which may be granted at EY's absolute discretion.

10. A Recipient:

    a) may not make any claim or demand or bring any action or proceedings against EY or any of its partners, principals, directors, officers or employees or any other Ernst & Young firm which is a member of the global network of Ernst & Young firms or any of their partners, principals, directors, officers or employees ("EY Parties") arising from or connected with the contents of the Report or the provision of the Report to the recipient; and

    b) must release and forever discharge the EY Parties from any such claim, demand, action or proceedings.

11. If a Recipient discloses the Report to a third party in breach of this notice, it will be liable for all claims, demands, actions, proceedings, costs, expenses, loss, damage and liability made or brought against or incurred by the EY Parties, arising from or connected with such disclosure.

12. If a Recipient wishes to rely upon the Report that party must inform EY and, if EY agrees, sign and return to EY a standard form of EY's reliance letter. A copy of the reliance letter can be obtained from EY. The Recipient's reliance upon the Report will be governed by the terms of that reliance letter.

Ernst & Young's liability is limited by a scheme approved under Professional Standards Legislation.

Glossary

The table below details the list of acronyms used throughout this report.

Acronym	Definition
AFM/FIM	Australian Functional Measure, based on the Functional Independence Measure
AMO	Assessment Management Organisation
AN-ACC	Australian National Aged Care Classification
BS	Braden Scale
BRUA	Behaviour Resource Utilisation Assessment
DEMMI	De Morton Mobility Index-Modified
FIM.Cognition	Cognition component of the Australian Modified Functional Independence Measure
FIM.Motor	Motor component of the Australian Modified Functional Independence Measure
IRR	Inter-Rater Reliability
IRR 1	Review of IRR of dual assessments conducted over November 2022 to December 2022
IRR 2	Review of IRR of dual assessments conducted over September 2023 to November 2023
IRR 3	Review of IRR of dual assessments conducted over June 2024 to August 2024
MMM	Modified Monash Model
NWAU	National Weighted Activity Unit
OT	Occupational Therapist
PA	Pure Agreement
PT	Physiotherapist
QA	Quality Assurance
RN	Registered Nurse
RVU	Relative Value Units
RUG-ADL	Resource Utilisation Groups—Activities of Daily Living

Contents

1. Executive summary

1.1 Background

EY has been working with the Department of Health and Aged Care (the Department) to identify and analyse trends, anomalies and patterns in Australian National Aged Care Classification (AN-ACC) assessments which may be of concern in the assessment process and to support ongoing quality assurance of AN-ACC assessments.

This report presents the results of inter-rater reliability (IRR) analysis conducted on 1,098 IRR assessments (i.e. a total of 2,196 individual assessments) occurring between 3 June and 2 August 2024. This represents the third IRR round ("IRR 3"), with the first round ("IRR 1") representing samples of assessments taken from November to December 2022 and the second round ("IRR 2") representing samples of assessments taken from September to November 2023.

The IRR analysis considers the rate at which assessors agree on the outcomes of the assessment, either by assigning the resident to the same AN-ACC classification or assigning the same score on an instrument within the AN-ACC assessment. Statistical adjustments are also performed to allow for the probability of assessors agreeing by chance and for a tolerance for small discrepancies in scoring between assessors.

1.2 Key findings

The key findings from the IRR analysis have been outlined below:

- **Consistently excellent agreement** - High rates of agreement were observed for the AN-ACC classification and underlying assessment instruments. Agreement was also highly consistent between initial assessments, reclassifications and reconsiderations, which may imply a broader improvement in the consistency of IRR assessments conducted in comparison to previous IRR rounds.

- **Improved agreement since the previous IRR round** – There were improvements in agreement from IRR 2 to IRR 3 for most Assessment Management Organisations (AMO), following an initial overall drop in agreement from IRR 1 to IRR 2. AMOs 2 and 5 continued to have the greatest agreement rates, while AMOs 3 and 6 maintained the lowest levels of agreement.
- **Agreement has declined for *"moderate only"* assessor experience pairs** – Agreement for assessor pairs where both assessors had moderate experience decreased in IRR 3 and was generally lower than for assessor pairings including at least one assessor with extensive experience. Assessor experience levels vary between 'limited', 'moderate' and 'extensive'.
- **Agreement for assessor pairs where both assessors had extensive experience, or included at least one physiotherapist (PT) or occupational therapist (OT), were relatively weaker** – Compared to *"extensive only"* assessor experience pairs, agreement was higher for *"extensive – moderate"* and *"extensive – limited"* pairs for the AN-ACC classification and all instruments. In addition, pairs of registered nurse (RN) assessors had the highest agreement for the AN-ACC classification and all instruments, followed by *"mixed"* pairs, containing one RN and either one PT or OT, and then *"other"* pairs, containing no RNs.

1.3 Limitations

The analysis of IRR results for the purposes of monitoring assessment standards and driving improvements in quality comes with inherent limitations:

- This is a data analysis exercise to assess the level of consistency in the application of AN-ACC and its components by the assessors. However, the analysis itself cannot differentiate the underlying quality of trainings, assessment standards or improvement processes for assessments of excellent agreement, nor can it associate any agreement in class/score as being the 'correct' class/score.

- The results from the analysis are observational and any differences observed are not necessarily causal in nature. Further investigation would be needed to establish causal relationships.

- The analysis in this report is limited in scope and any results are limited to what can be explained by the data only. Furthermore, in undertaking the investigations related to this report, EY has relied on the accuracy and completeness of information supplied by the Department and has not performed any quality assurance or validation on the data provided by the Department back to its source.

2. Background and Approach

2.1 Background

EY has been analysing AN-ACC assessment and re-classification data to identify trends, anomalies and patterns which may be of concern in the assessment process and to support ongoing quality assurance of AN-ACC assessments.

This report presents the results of IRR (inter-rater reliability) analysis conducted on 1,098 IRR assessments occurring from 3 June to 2 August 2024[1]. As in previous IRR reviews (in March 2022, April 2023 and April 2024), IRR was tested through dual assessments, which saw the same resident independently assessed by two different assessors on the same day. In this report, we refer to these dual assessments as "IRR assessments". To strengthen the quality of the IRR assessments, both assessments were completed under the same environment conditions by assessors who had access to the same information and were provided the same instructions.

2.2 Data Sources

The following data sets and information were provided by the Department of Health and Aged Care (Department) through the Health Data Portal:

Table 1: Data sources

#	File	Description of file
Inter-rater reliability		
1	IRR assessment Data	Excel data containing IRR assessment data from 3 June 2024 to 2 August 2024.
Other reference data sets		
2	All AN-ACC assessments	Excel data files containing Qlik extracts of AN-ACC assessment data from April 2021 to 2 August 2024.
3	Provider and facility details by NAPS ID	Excel file containing details of active and closed facilities and providers as of 30 August 2024.

2.3 Analysis

This section details the analysis techniques employed in this report. We have assessed inter-rater reliability of the June to August 2024 IRR assessments at both a final AN-ACC classification level and individual assessment instrument level using several summary statistics:

- Pure rates of agreement – The simple proportion of assessments where assessors gave the same class, score or range of scores.
- Correlation – The Pearson correlation coefficient calculated from pairs of National Weighted Activity Units (NWAUs) or total scores on each instrument[2].

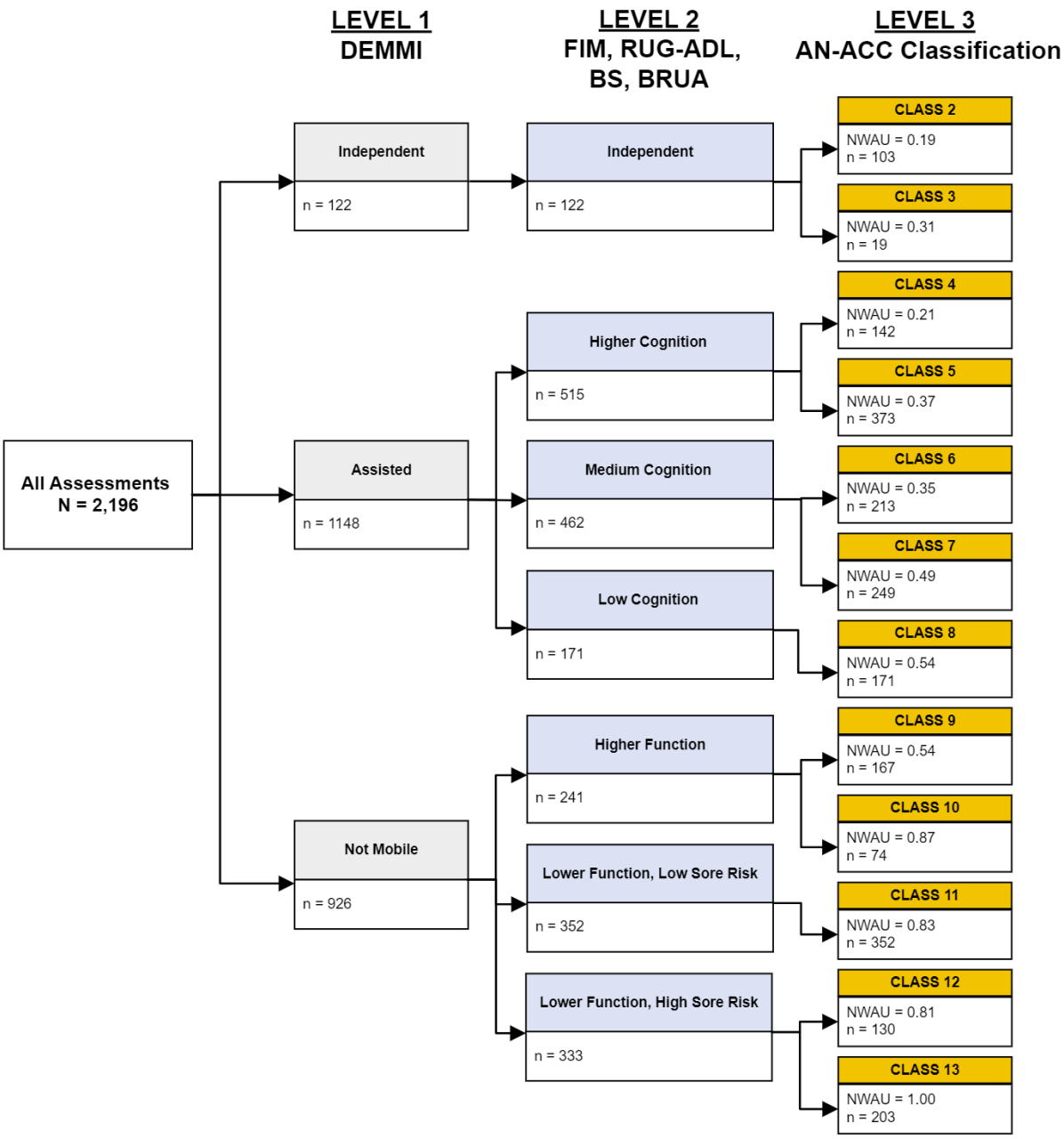
[1] In a dual assessment, two assessors simultaneously conduct an assessment of a resident (scoring that resident independently) and the results of both assessments are uploaded to Department servers for comparison.
[2] (2008). Pearson's Correlation Coefficient. In: Kirch, W. (eds) Encyclopedia of Public Health. Springer, Dordrecht. https://doi.org/10.1007/978-1-4020-5614-7_2569

- Kappa statistics – More complex but robust measures of agreement which adjust for expected agreement by chance between assessors (discussed further in Section 2.3.1).

In addition to agreement between assessors on final AN-ACC classification and the underlying assessment instruments in the AN-ACC assessment, we have also considered agreement at different levels of the AN-ACC classification algorithm and in score groupings on specific instruments (e.g., the total DEMMI score is grouped into three mobility categories). This is shown in the Figure 1 representation of AN-ACC classifications.

These measures, calculated across all assessments and split by various assessor characteristics, provide a set of key summary statistics to form a view on the inter-rater reliability of AN-ACC assessments.

Figure 1: AN-ACC decision tree with assessment counts[3]



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

[3] All individual assessments are counted, such that there exist two assessments for each resident.

2.3.1 Inter-rater reliability statistics

This section details the analysis techniques employed in this report. We have assessed inter-rater reliability of IRR assessments at both a final AN-ACC classification level and individual assessment instrument level using several summary statistics:

1. **Pure agreement** – The proportion of assessments where assessors gave the same class or score. This statistic does not consider a "degree of agreement" or allow for "partial agreement" and only considers cases where assessors agree perfectly.

2. **Same or adjacent class/score** – The proportion of assessments where assessors are within one class or score (i.e., treating an adjacent score or class as an agreement), this allows for a level of "partial agreement".[4]

3. **Fleiss' kappa** – A measure of agreement between pairs of assessors who assess different residents, with an adjustment for the probability those two assessors will agree by chance. Similar to the pure agreement, this statistic does not consider a "degree of agreement" or allow for "partial agreement" and only considers cases where assessors agree perfectly.

4. **Weighted Fleiss' kappa** – A measure of agreement that allows for partial agreement between assessors using a set of weights which vary depending on the difference in score between assessors. The selected weightings used in this report are discussed further in Appendix A.2.

5. **Pearson's correlation coefficient** – Correlation measures the general direction of agreement only and should be interpreted with caution since correlation is not a reliable measure of agreement.

This report primarily uses Fleiss' kappa to represent "exact agreement" and weighted Fleiss' kappa to represent a "tolerance for partial agreement".

We have interpreted agreement rates with a tolerance for partial agreement with respect to reference kappa values in Table 2. As shown, *"excellent agreement"* is achieved where a kappa higher than 0.75 is observed.

Table 2: Kappa interpretation

Kappa Value	Interpretation
[-1.0, 0.0]	No agreement
(0.00, 0.20]	Poor agreement
(0.20, 0.40]	Fair agreement
(0.40, 0.75]	Moderate agreement
(0.75, 1.0]	Excellent agreement

**Kappa bands were chosen to align with the 2022 reliability assessment, the original Cohen report[5] and previous IRR reviews.*

[4] We note that adjacent ordering at Level 1 and Level 2 of the AN-ACC decision tree in Figure 1 may occur between broader branches, e.g. 'Assisted – Low Cognition' to 'Not Mobile – Higher Function', as well as within branches, e.g. 'Assisted – Medium Cognition' to 'Assisted – Low Cognition'. This may represent differing aspects of adjacency and is a limitation of the approach taken.

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

[5] McHugh ML. Interrater reliability: the kappa statistic. Biochem Med (Zagreb). 2012;22(3):276-82. PMID: 23092060; PMCID: PMC3900052.

2.3.2 IRR assessment process

The IRR assessments are a sample from the population of all assessments conducted during the same period and were chosen by the Department to be representative of the broader assessment distribution as outlined in Table 3.

- Assessor pairs for each resident were assigned by the AMO based on the group of assessors present at a facility.
- Special considerations were taken so that each AMO and each assessor profession performed a sufficient sample of IRR assessments for analysis. This may impact the representation in IRR assessments due to the characteristics of facilities within each AMO's locality.
- Residents were assigned by name or ID to the assessors prior to any contact and no notices were given regarding the IRR assessment. No considerations were taken to select residents with distinct traits or conditions.

Table 3: Split of IRR assessments (by no. of residents) across different types of aged care facilities

Remoteness	Type	No. Beds	6-Month Assessment Volume*	Proportion	IRR Assessments	Proportion
MMM 1	For Profit	0 to 59	1,944	2.1%	18	1.6%
		60 to 119	14,952	16.4%	173	15.8%
		120 to 179	14,090	15.5%	180	16.4%
		180+	2,454	2.7%	57	5.2%
	Not for Profit	0 to 59	3,770	4.1%	30	2.7%
		60 to 119	14,656	16.1%	161	14.7%
		120 to 179	10,887	12.0%	143	13.0%
		180+	2,182	2.4%	30	2.7%
	Government	Any	352	0.4%	2	0.2%
MMM 2	For Profit	Any	3,211	3.5%	43	3.9%
	Not for Profit	Any	4,516	5.0%	77	7.0%
	Government	Any	516	0.6%	4	0.4%
MMM 3	Any	Any	7,634	8.4%	88	8.0%
Regional (MMM 4-5)	Any	Any	9,458	10.4%	92	8.4%
Remote (MMM 6-7)	Any	Any	342	0.4%	0	0.0%
Total			90,964	100.0%	1,098	100.0%

** Calculated based on total AN-ACC assessments completed 1 December 2023 to 1 June 2024.*

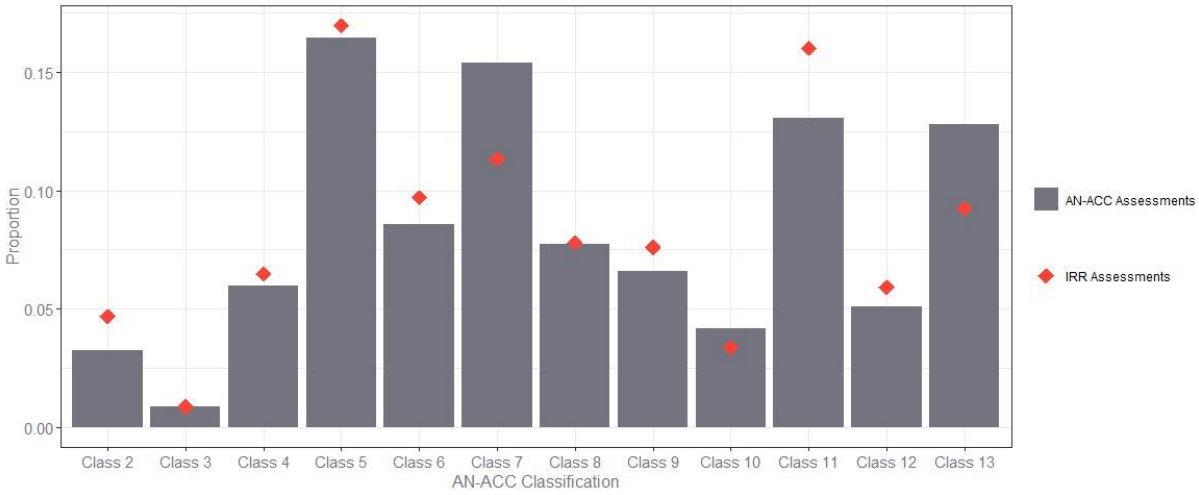
Overall, the distribution of IRR assessments by remoteness, facility type and facility size aligns closely to that of the previous 6 months of AN-ACC assessments.

3. Results

From 3 June 2024 to 2 August 2024, a total of 2,212 raw assessments were completed by 272 assessors for the purposes of the IRR analysis. 1,098 unique residents from 444 facilities were assessed with the average assessor completing approximately eight dual assessments. 2,196 (99.3%) of the assessments were submitted to the Department on the same day by both assessors, indicating good compliance with IRR protocols.

The dual assessment sample was broadly representative of the AN-ACC assessment casemix over the previous six months as shown in Figure 2.

Figure 2: AN-ACC classification distribution of assessments: IRR assessments vs AN-ACC assessments (1 Dec 23 – 1 Jun 24)



3.1 Overall agreement results

As discussed in Section 2.3.1, kappa statistics aim to remove agreement by chance from a simple rate of agreement measure, allowing us to gain a better understanding of the true agreement rates between assessors. Table 4 shows the Fleiss' kappa and the Weighted Fleiss' kappa test statistics alongside rates of agreement, calculated across all IRR assessments by AN-ACC decision tree splits as provided in Figure 1.

Table 4: Agreement rates by AN-ACC decision tree level

Level	Category	Fleiss' Kappa		Agreement Type	
		Exact (unweighted)	Tolerance for partial (linearly weighted)	Pure Agreement[6]	Same or Adjacent Class/Score[7]
1	Mobility Branch	0.94	0.96	96.9%	100.0%
2	Mobility: Cognition/ Function/Pressure Sores	0.88	0.94	89.8%	97.8%
3	AN-ACC Class	0.84	0.93	86.0%	94.1%

Key observations from Table 4 are:

- Assessors assigned residents into the same mobility category in 96.9% of assessments, the same level 2 category in 89.8% of assessments and the same final AN-ACC classification in 86.0% of assessments. This implies excellent pure agreement (exact agreement) at each level of the AN-ACC decision tree.

- When also treating classification to adjacent categories as an agreement, the agreement rates increased to 100%, 97.8% and 94.1% respectively. In the case of level 1 mobility, this implies that there were no situations where the two assessors separately classified a resident as "independent mobility" and "not mobile".

  These high rates of agreement when allowing for adjacent classifications also indicate that, in cases where results differed between assessors, the results tended to be similar. In the case of level 1 mobility, we note that there are only three possible outcomes and that this result was not unexpected.

- Both the weighted and unweighted Fleiss' kappa statistics imply excellent agreement between assessors at all levels of the decision tree.

  The unweighted Fleiss' kappa, relying on exact agreement only, decreased slightly at each level (or branch) of the decision tree as the range of potential classifications available increases, therefore increasing the chances of small discrepancies between assessments which result in a disagreement. Introducing a tolerance for partial agreement, the weighted Fleiss' kappa was higher at all levels of the decision tree. This again indicates that, in circumstances where assessors did not assign the exact same AN-ACC classification, they more often than not assigned AN-ACC classifications which were similar or adjacent (i.e., most disagreements were relatively small).

[6] Note that throughout this report, the tables containing 'pure agreement' and 'same or adjacent class/score' statistics have been shaded according to the colour scheme in Table 3, which refers to interpretation of a kappa statistic. Technically, the pure agreement rate contains a component of "agreement by chance", which is removed in calculating a kappa statistic. However, this shading scheme has been adopted for consistency.
[7] AN-ACC classifications were not ordered by resource utilisation/funding

Table 5 below contains the average raw score differences between assessors on final AN-ACC classification (by NWAU) for each underlying assessment instrument in the AN-ACC assessment.

Table 5: Average score difference by metric

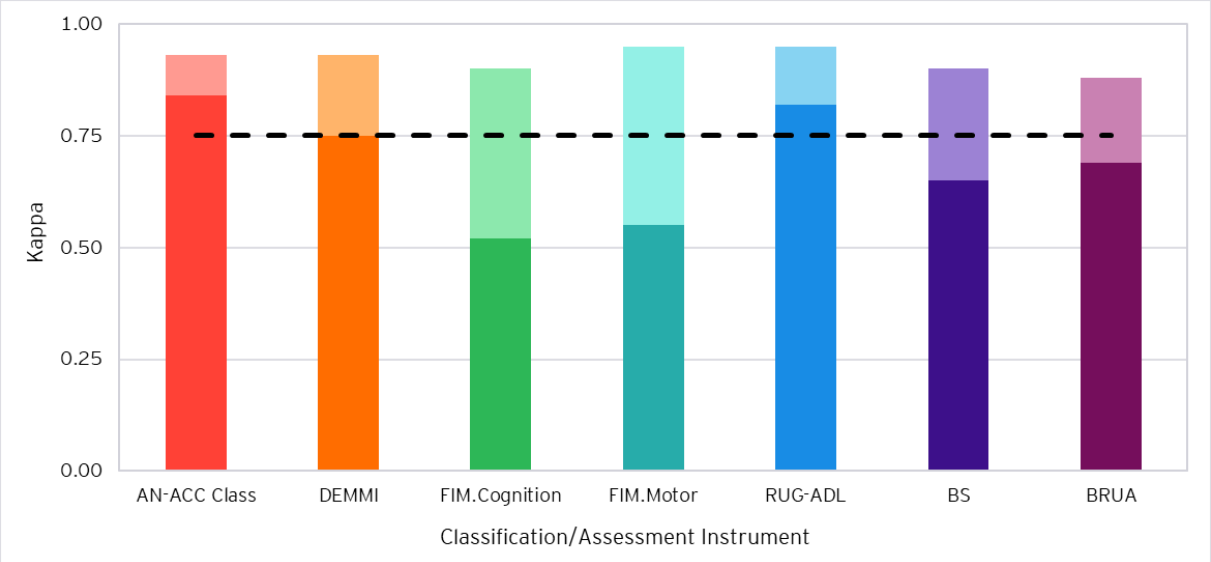
Metric	Score Range	Number of Possible Scores	Average Score Difference*	Average Score Difference Divided by Score Range
AN-ACC Class (NWAU)	[0.19, 1]	11	0.019	0.023
DEMMI	[0,15]	16	0.335	0.022
FIM.Cognition	[5, 35]	31	1.006	0.034
FIM.Motor	[12, 84]	73	1.243	0.017
RUG-ADL	[4, 18]	15	0.332	0.024
BS	[8, 23]	16	0.443	0.030
BRUA	[5, 20]	16	0.476	0.032

**The absolute value of score differences were taken prior to averaging*

The score differences are, by construction, closely related to the possible range of scores available. Standardising the average score difference by the score range shows that discrepancies are generally within 2% to 3% of the total range of possible scores.

The Fleiss' kappa and the weighted Fleiss' kappa test statistics for final AN-ACC classification and each underlying assessment instrument across all IRR assessments are shown in Figure 3 below, with values of each statistic presented in Table 6.

Figure 3: Agreement rates by metric using Fleiss' kappa and Weighted Fleiss' kappa



i. *Excellent agreement is represented by the dashed line (0.75)*
ii. *Exact agreement (unweighted kappa) is shown in darker shading for each classification/assessment instrument.*
iii. *Agreement with tolerance for partial agreement (weighted kappa) above exact agreement is shown in lighter shading.*

Table 6: Agreement rates by metric

Category	Fleiss' Kappa		Agreement Type		Pearson's Correlation
	Exact (unweighted)	Tolerance for partial (linearly weighted)	Pure Agreement	Same or Adjacent Class/Score	
AN-ACC Class	0.84	0.93	86.0%	94.1%	0.97
DEMMI	0.75*	0.93	77.7%	92.8%	0.98
FIM.Cognition	0.52	0.90	54.5%	74.9%	0.98
FIM.Motor	0.55	0.95	56.4%	74.2%	0.99
RUG-ADL	0.82	0.95	84.2%	90.8%	0.98
BS	0.65	0.90	67.9%	91.1%	0.97
BRUA	0.69	0.88	71.6%	89.9%	0.95

** Excellent agreement is observed when using exact values.*

Key observations from Figure 3 and Table 6 are:

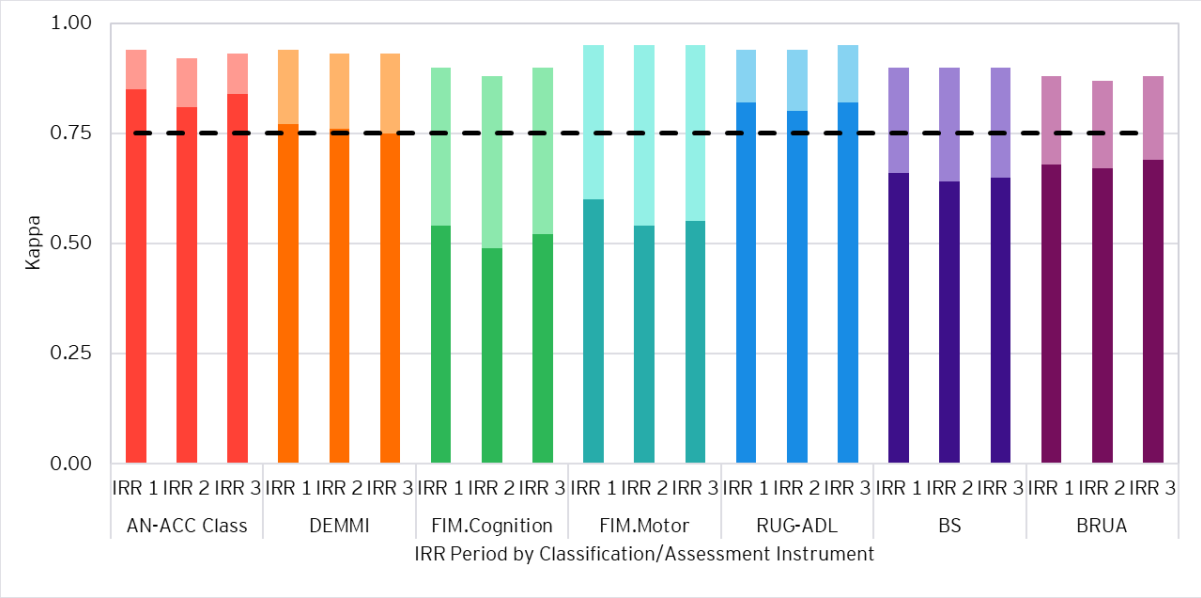
- Excellent agreement between assessors was observed for all instruments when measured with a tolerance for partial agreement (linearly weighted Fleiss' kappa).
- The level of agreement was broadly consistent (linearly weighted kappas within 0.07 of each other) across all instruments when measured with a tolerance for partial agreement.
- When examining exact agreement (unweighted Fleiss' kappa), only the AN-ACC classification, DEMMI and RUG-ADL had excellent exact agreement, and the FIM.Cognition, FIM.Motor, BS and BRUA were observed to have moderate agreement.
    - For the FIM.Cognition and FIM.Motor, this is expected due to the reduced likelihood of exact agreement where there is a larger range of possible scores available (see Table 5 for the full list of score ranges). The FIM.Cognition has a larger range of scores than the FIM.Motor, as observed in Table 5.
    - The BS and BRUA have a similar range of scores to the DEMMI and RUG-ADL but were observed to have lower rates of exact agreement.
- When treating same or adjacent scores (instrument scores within 1) as an agreement, the FIM.Cognition and FIM.Motor were slightly below excellent agreement (within 1 percentage point).
- Strong correlation was observed across all instruments. As discussed in Section 2.3.1, this should be interpreted with caution since correlation is not a reliable measure of agreement as it only measures the general direction of agreement.

3.1.1 Comparing agreement levels against IRR 1 and IRR 2

The three IRR periods represent samples of assessments taken over just three discrete periods of time: November to December 2022 ("IRR 1"), September to November 2023 ("IRR 2") and June to August 2024 ("IRR 3"). Their results are potentially impacted by seasonality, influences from the nature of the assessments included within the samples (such as assessment type, AMO and other characteristics discussed in this report) and randomness. Further analysis on additional IRR data would need to be conducted to conclude whether there are any underlying trends.

Figure 4 shows the comparison in Fleiss' kappa statistics between the IRR rounds.

Figure 4: Agreement statistics by IRR period and classification/assessment instrument



i. *Excellent agreement is represented by the dashed line (0.75)*
ii. *Exact agreement (unweighted kappa) is shown in darker shading for each classification/assessment instrument.*
iii. *Agreement with tolerance for partial agreement (weighted kappa) above exact agreement is shown in lighter shading.*

Key observations from Figure 4 are:

- Excellent agreement between assessors was observed for all instruments when measured with a tolerance for partial agreement (linearly weighted Fleiss' kappa) in all IRR periods.
- The FIM.Motor and BS had no change to weighted agreement across all IRR periods.
- AN-ACC classification and FIM.Cognition increased in weighted agreement by 0.01 and 0.02 respectively from IRR 2 to IRR 3, following a decrease of 0.02 from IRR 1 to IRR 2.
- The DEMMI had no change in weighted agreement from IRR 2 to IRR 3, while the BRUA increased in weighted agreement by 0.01. Both instruments had decreased in weighted agreement by 0.01 from IRR 1 to IRR 2.
- The RUG-ADL increased in weighted agreement by 0.01 from IRR 2 to IRR 3, following no change from IRR 1 to IRR 2.
- Exact agreement (unweighted Fleiss' kappa) increased by up to 0.03 for all instruments, except DEMMI, from IRR 2 to IRR 3, following a decrease by up to 0.06 for all instruments from IRR 1 to IRR 2. IRR 1 exact agreement was generally the highest out of the three IRR rounds.

It is possible that changes in the underlying assessment casemix for each IRR period may drive changes in agreement rates instead of genuine changes to assessor agreement. To eliminate the potential that the distribution of IRR assessments performed by AMO and State has influenced changes in the overall agreement rates, we have constructed re-weighted pure agreements for the IRR 1 and IRR 2 assessments. The re-weighted pure agreement is a measure of the average pure agreement in IRR 1 and IRR 2 assessments with each observation weighted by the distribution of AMOs and states within the IRR 3 assessment sample. This enables a 'like-for-like' comparison between the IRR 1 re-weighted pure agreement rates, the IRR 2 re-weighted pure agreement rates and the IRR 3 pure agreement rates.

Table 7 shows the comparison in IRR 3 pure agreement and the re-weighted IRR 1 and IRR 2 pure agreements.

Table 7: IRR 3 agreement rates and re-weighted IRR 1 and IRR 2 agreement rates by AN-ACC classification/assessment instrument

Category	Pure Agreement		
	Re-weighted IRR 1*	Re-weighted IRR 2*	IRR 3
AN-ACC Class	84.4%	81.9%	86.0%

** Agreement for each AMO and State combination are weighted by representation across IRR 3 assessments*

Key observations from Table 7 are:

- The improvement from IRR 2 to IRR 3 is more pronounced following re-weighting of the IRR 2 agreement rates.
- Due to the reduction in IRR 1 agreement when re-weighted, the reduction in agreement from IRR 1 to IRR 2 previously observed across all instruments is less pronounced on this comparable basis.
- This is reflective of changes in the distribution of IRR assessments by AMO and State from IRR 1 to IRR 3, particularly where increases in proportions of AMO-State combinations occurred where pure agreement rates during IRR 1 and IRR 2 were relatively lower. This is less impactful for IRR 2 due to the shorter time difference from the IRR 3 round.

Overall, there was an increase in agreement observed between the IRR 2 and IRR 3 assessments. This may be representative of refinement in standards and developing assessor experience over time as the AN-ACC assessment process matures, however further IRR rounds are required to assess this trend.

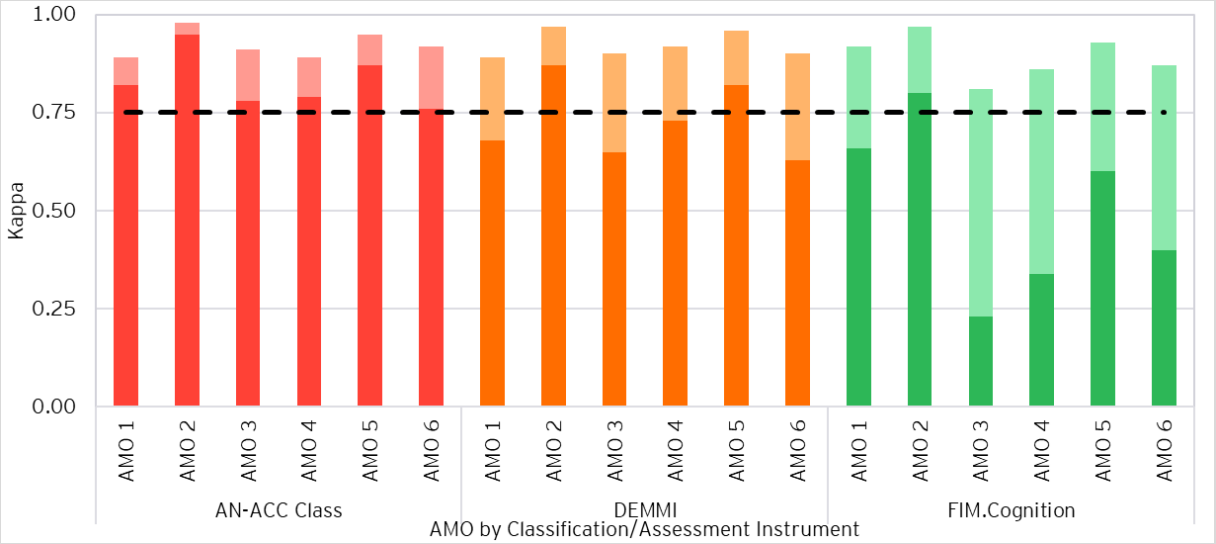
Changes in overall agreement can be further explained by changes within subgroups of assessors. Assessors are explored further by AMO, experience and profession in the following sections.

3.2 Segmentation by AMO

This section examines the agreement of assessors by their AMO. It is important to note that all IRR assessments taking place through June to August 2024 and analysed in this report were performed by pairs of assessors from the same AMO. This practice promoted consistent scheduling of IRR assessments and allowed for two assessors to observe the resident under the same conditions. For this reason, caution should be applied when comparing IRR statistics between AMOs.

There are six independent AMOs currently conducting AN-ACC assessments on behalf of the Department. The Fleiss' kappa and the weighted Fleiss' kappa test statistics for final AN-ACC classification and each underlying assessment instrument by AMO are shown in Figure 5 and Figure 6 below, with values of each statistic presented in Appendix B.1.

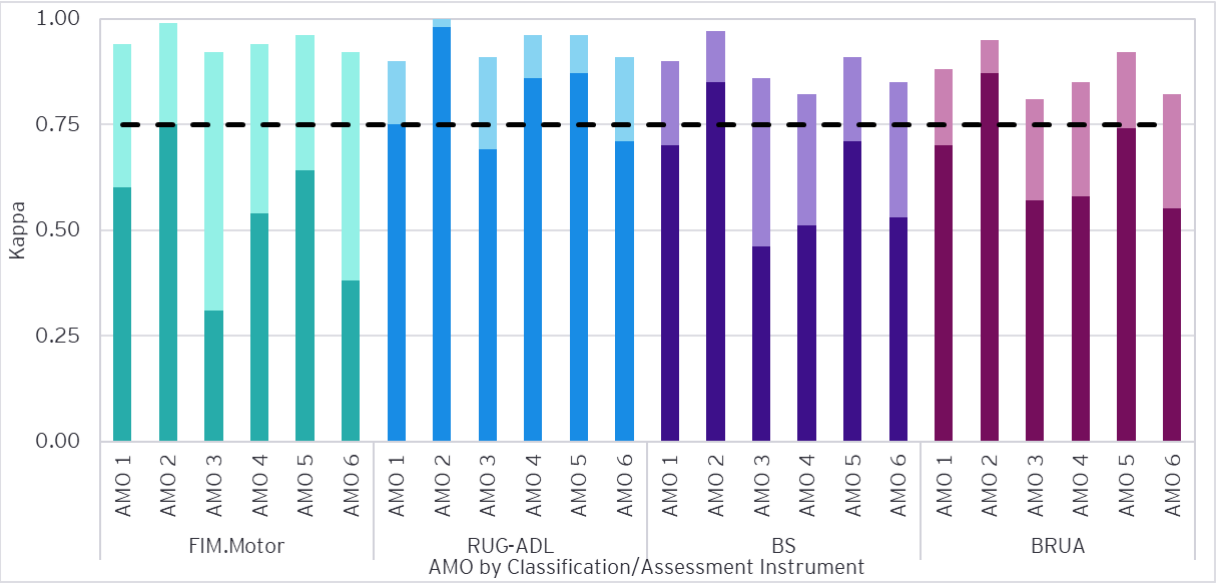
Figure 5: Agreement statistics by AMO (AN-ACC Class, DEMMI and FIM.Cognition)



Excellent agreement is represented by the dashed line (0.75)

i. *Exact agreement (unweighted kappa) is shown in darker shading for each classification/assessment instrument.*
ii. *Agreement with tolerance for partial agreement (weighted kappa) above exact agreement is shown in lighter shading.*

Figure 6: Agreement statistics by AMO (FIM.Motor, RUG-ADL, BS and BRUA)



i. *Excellent agreement is represented by the dashed line (0.75)*
ii. *Exact agreement (unweighted kappa) is shown in darker shading for each classification/assessment instrument.*
iii. *Agreement with tolerance for partial agreement (weighted kappa) above exact agreement is shown in lighter shading.*

Key observations from Figure 5 and Figure 6 are:

- Excellent agreement between assessors for all AMOs was observed for all instruments, when measured with a tolerance for partial agreement.

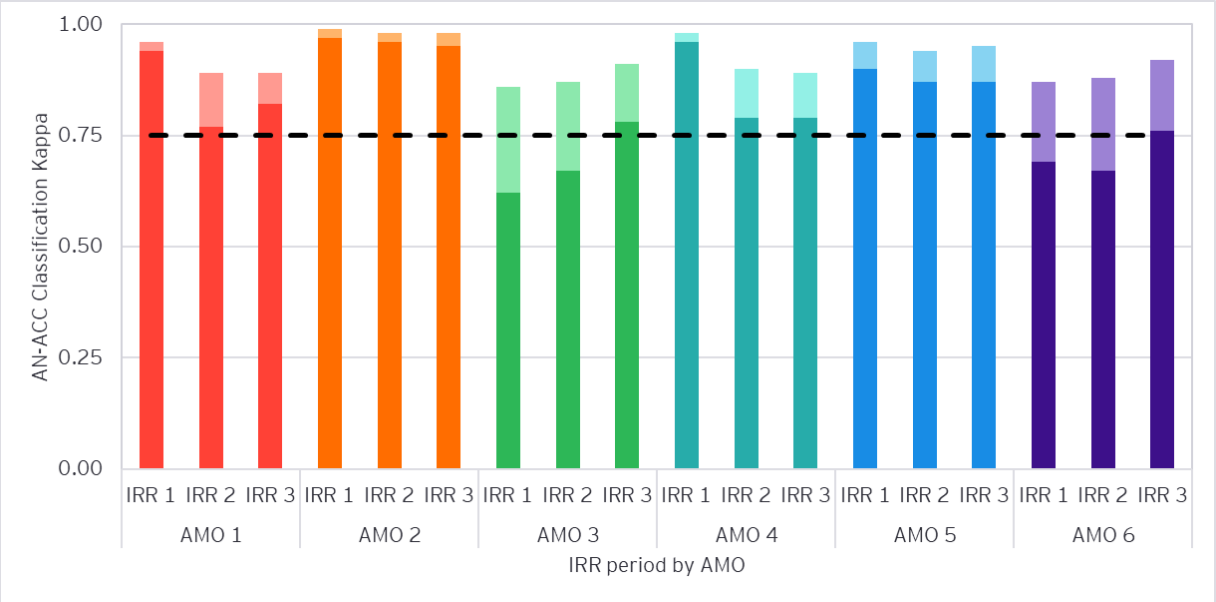
- AMOs 2 and 5 generally had higher rates of agreement than other AMOs, with weighted agreement statistics greater than 0.91 across all instruments.
    - AMO 2 had the highest agreement rates of all the AMOs and pure agreement above the "excellent" threshold for each instrument. This resulted in near perfect agreement on the final AN-ACC classification (96%), corresponding to only 6 disagreements on 149 IRR assessments.
- AMO 1 had weighted agreement similar to AMO 5 (within 0.05) on FIM.Cognition, FIM.Motor, BS and BRUA, had weighted agreement similar to AMO 2 (within 0.05) on FIM.Cognition and FIM.Motor, but had lower weighted agreement on the DEMMI (0.89) and RUG-ADL (0.90).
- Whilst having excellent agreement measured with a tolerance for partial agreement across all instruments, AMOs 3, 4 and 6 were generally observed to have agreement levels lower than other AMOs.
    - AMO 6 had fair pure agreement (below 40%) on FIM.Motor and AMO 4 had fair pure agreement on FIM.Cognition. Meanwhile, AMO 3 had fair pure agreement on both instruments.
- The FIM.Cognition and FIM.Motor, with the largest scoring ranges, had below 75% pure agreement across all AMOs except AMO 2, but performed much better with the linearly weighted Fleiss' kappa. The measured agreement on the FIM.Cognition and FIM.Motor showed the most improvement between the pure agreement rate and the linearly weighted Fleiss' kappa.

We note the absence of cross AMO IRR assessments introduces the risk that IRR values between AMOs may not be directly comparable.

3.2.1 Comparing agreement levels against IRR 1 and IRR 2

Figure 7 shows the comparison in Fleiss' kappa statistics for the final AN-ACC classification between the IRR 1, IRR 2 and IRR 3 assessments.

Figure 7: AN-ACC classification agreement statistics by IRR period and AMO



i. *Excellent agreement is represented by the dashed line (0.75)*
ii. *Exact agreement (unweighted kappa) is shown in darker shading for each classification/assessment instrument.*
iii. *Agreement with tolerance for partial agreement (weighted kappa) above exact agreement is shown in lighter shading.*

The key observations from Figure 7 are:

- Excellent agreement measured with a tolerance for partial agreement was observed at the AN-ACC classification level for all AMO's in each IRR period.
- For the AN-ACC classification,
    - AMO 2 has continued to have near perfect weighted agreement above that of all other AMOs across IRR 1 (0.99), IRR 2 (0.98) and IRR 3 (0.98). Exact agreement rates have generally been higher than weighted agreement across other AMOs throughout each IRR round.
    - AMO 5 has continued to have very high weighted agreement across IRR 1 (0.96), IRR 2 (0.94) and IRR 3 (0.95).
    - AMOs 1 and 4's weighted agreements have remained similar from IRR 2 to IRR 3 with small decreases of 0-0.01, following their relatively large decreases in agreement of up to 0.08 from IRR 1 to IRR 2. Exact agreement had increased by 0.05 for AMO 1 in IRR 3.
    - AMOs 3 and 6's weighted agreements have increased from IRR 1 to IRR 3, reaching similar levels to AMOs 1 and 4 in IRR 3. Previously, AMOs 3 and 6 had agreement rates lower than other AMOs.

Figures showing the comparison of Fleiss' kappa statistics for underlying assessment instruments for each AMO between IRR periods are shown in Appendix B.4. Changes in AMO agreement on

assessment instruments generally follow the trends in AN-ACC classification agreement shown in Figure 7.

AMO and State contributions towards IRR 2 to IRR 3 changes in agreement

It is relatively straightforward to compare the change between periods in the rate of pure agreement (as opposed to kappa statistics). This provides a good sense of the contributors to the change between periods.

As shown in Section 3.1.1 above, "re-weighting" the IRR 1 and IRR 2 assessments to the IRR 3 distribution of AMO and State enables a 'like-for-like' comparison between pure agreement rates. To identify the AMO and State pairs that had the greatest impact on the overall change in AN-ACC classification pure agreement, we have broken down the sources of change in pure agreement by AMO and State.

Table 8 below shows the AN-ACC classification pure agreement rate for IRR 1, IRR 2 and IRR 3, and the change in percentage points attributable to each AMO and State pair. The change for each AMO and State pair accounts for both the observed change in agreement rates and the representation amongst IRR assessments. Additionally shown is the change in pure agreement that is attributable to the changing AMO and State mix.

Table 8: AN-ACC classification pure agreement for each IRR period and the change in agreement attributable to AMO and State pairs

IRR period	AN-ACC Classification Pure Agreement	Source of change from IRR 2 to IRR 3	Change (percentage points)
IRR 2 actual	82.9%		
IRR 2 (with IRR 3 weights*)	81.9%	Driven by change in AMO and State sample mix	-1.0
		Four AMO and State combinations	+3.4
		All other AMO and State combinations	+0.7
IRR 3 actual	86.0%	Total like-for-like sample change	+4.1

** Weights are applied by AMO and State representation across IRR 3 assessments*

Key observations from Table 8 are:

- Although there was a 3.1 percentage point increase in AN-ACC classification pure agreement, changes in the AMO and State sample mix alone would have led to a 1.0 percentage point reduction in pure agreement.
- Agreement rates at the AMO and State combination level increased by 4.1 percentage points on average. Of this amount, four specific AMO and state combinations contributed to 3.4 percentage points.

3.3 Segmentation by assessor experience

This section examines the agreement of assessors by their experience. The number of assessments completed by the assessor between 1 October 2022 and 2 August 2024 was used as a proxy to indicate experience. This measure was adopted to recognise that assessment volumes are likely a better indicator of experience than elapsed time since the first assessment completed by an assessor to allow for extended periods without completing assessments.

Assessor experience was segmented into three categories; those with 100 or less assessments were classified as having "*limited*" experience, those with 350 or less assessments "*moderate*" experience, and those with over 350 were classified as having "*extensive*" experience. Of the 272 assessors involved in IRR assessments, 4 (1.5%) had 0 payment impacting AN-ACC assessments performed since 1 October 2022. These assessors were therefore also classified as having "*limited*" experience.

Table 9 shows the distribution of assessors and assessments by experience, while

Table 10 shows the distribution of IRR assessments for each combination of assessors by experience.

Table 9: Distribution of assessments by assessor experience

Assessor Experience	Number of Assessments Competed	Number of Assessors	Number of Assessments	Proportion of Assessments
Extensive	(350, Max]	207	1,688	76.9%
Moderate	(100, 350]	44	372	16.9%
Limited	[0, 100]	21	136	6.2%
Total		272	2,196	100.0%

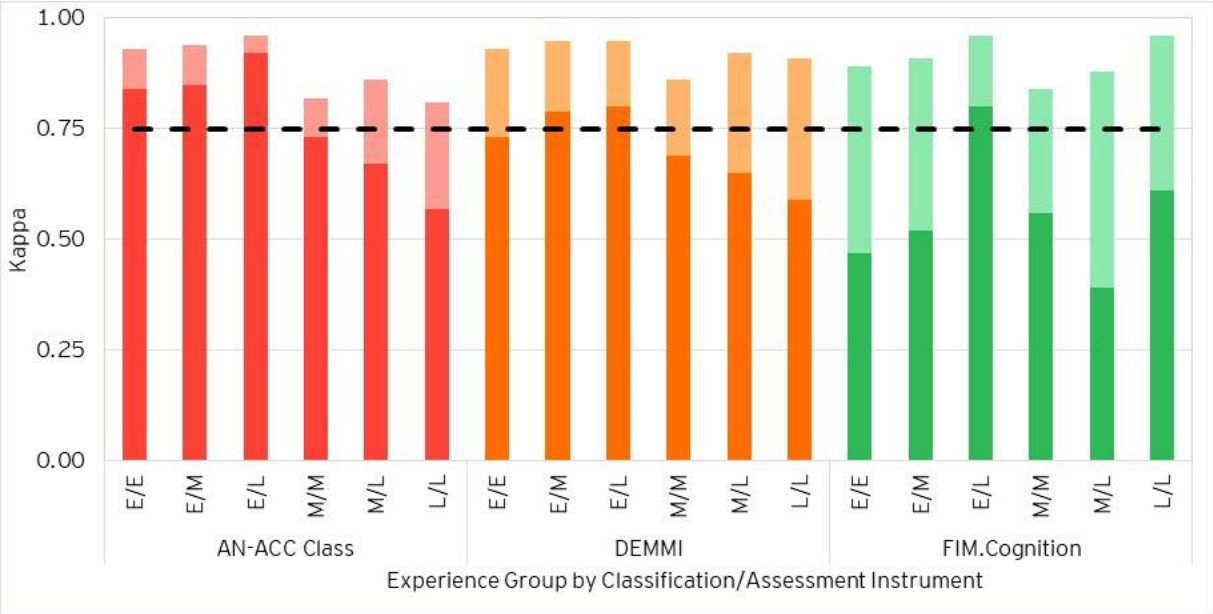
Table 10: Distribution of IRR assessments by assessor experience combination

Assessor Experience Combination		Mapping	Number of IRR Assessments	Proportion of IRR Assessments
Extensive	Extensive	Extensive only (E/E)	662	60.3%
Extensive	Moderate	Extensive - moderate (E/M)	251	22.9%
Extensive	Limited	Extensive - limited (E/L)	113	10.3%
Moderate	Moderate	Moderate only (M/M)	55	5.0%
Moderate	Limited	Moderate - limited (M/L)	11	1.0%
Limited	Limited	Limited only (L/L)	6	0.5%
Total			1,098	100%

The most common combination of assessor experience in IRR assessments was "*extensive only*", with 1,324 (60.3%) assessments by this combination. We note that 12 (0.5%) assessments were completed by a combination of two assessors with limited experience.

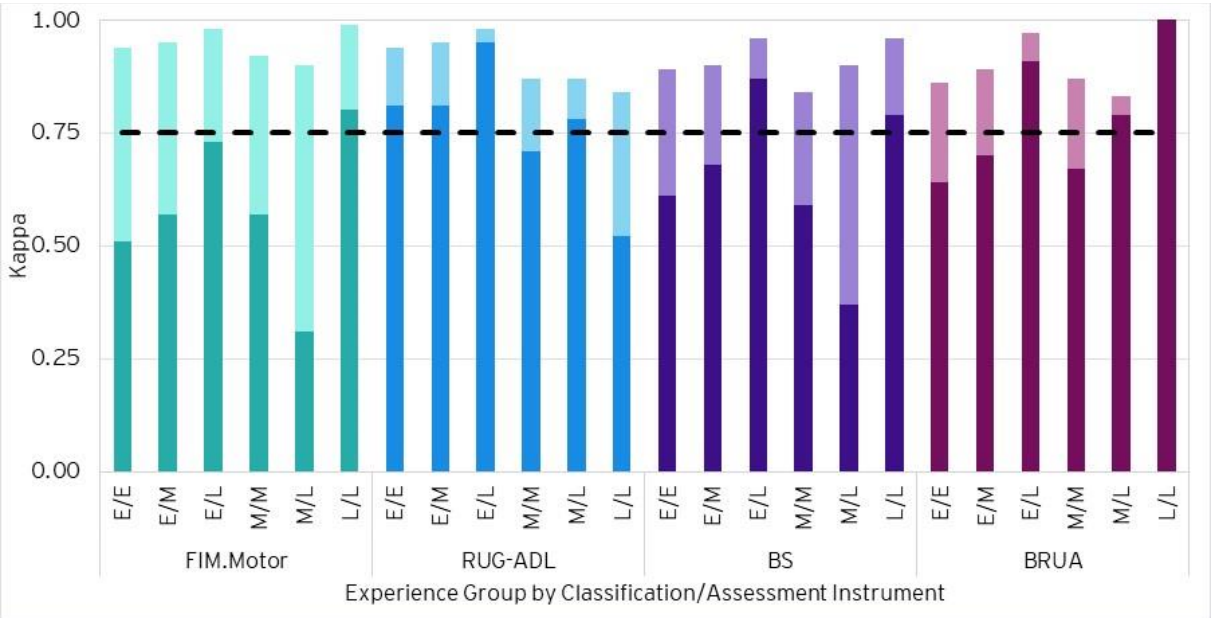
The Fleiss' kappa and the weighted Fleiss' kappa test statistics for final AN-ACC classification and each underlying assessment instrument by assessor experience group are shown in Figure 8 and Figure 9 below, with values of each statistic presented in Appendix B.2.

Figure 8: Agreement statistics by assessor experience (AN-ACC Class, DEMMI and FIM.Cognition)



i. *Excellent agreement is represented by the dashed line (0.75)*
ii. *Exact agreement (unweighted kappa) is shown in darker shading for each classification/assessment instrument.*
iii. *Agreement with tolerance for partial agreement (weighted kappa) above exact agreement is shown in lighter shading.*

Figure 9: Agreement statistics by assessor experience (FIM.Motor, RUG-ADL, BS and BRUA)



i. *Excellent agreement is represented by the dashed line (0.75)*
ii. *Exact agreement (unweighted kappa) is shown in darker shading for each classification/assessment instrument.*
iii. *Agreement with tolerance for partial agreement (weighted kappa) above exact agreement is shown in lighter shading.*

Key observations from the above figures are:

- With a tolerance for partial agreement (weighted kappa statistics), excellent agreement was observed across all instruments for all pairs of assessors by experience.
- Higher agreement rates were generally observed for assessor pairs including at least one assessor with extensive experience.

- It was expected that agreement rates would be highest for *"extensive only"* pairs. However, agreement rates were higher for *"extensive – moderate"* and *"extensive – limited"* pairs compared to *"extensive only"* pairs across the AN-ACC classification and all underlying assessment instruments. Exact agreement for *"extensive – limited"* pairs was sharply higher across all instruments except the DEMMI.

- As expected, *"moderate only"* pairs generally had lower agreement than assessor pairings including at least one assessor with extensive experience.

- Agreement rates for a pair of assessors with limited experience and *"moderate – limited"* pairs were more volatile between instruments, particularly for the FIM.Cognition, FIM.Motor, BS and BRUA.

3.3.1 Comparison against IRR 1 and IRR 2

Assessor experience has increased from IRR 1 to IRR 2 to IRR 3. This is expected as more assessments continue to be performed in the AN-ACC system and assessor tenure develops.

Table 11 shows the number of assessors by experience in each IRR period and Table 12 shows the number and proportion of IRR assessments completed by each pair of assessors by experience.

Table 11: Distribution of IRR assessment by assessor experience and IRR period

Assessor Experience	Number of Assessments Competed	IRR 1 Assessors		IRR 2 Assessors		IRR 3 Assessors	
		No.	Prop.	No.	Prop.	No.	Prop.
Extensive	> 350	71	33%	179	66%	207	77%
Moderate	> 100, <= 350	67	31%	50	19%	44	17%
Limited	<= 100	79	36%	40	15%	21	6%
Total		217	100%	269	100%	272	100%

The proportion of assessors with extensive experience (more than 350 AN-ACC assessments performed) has grown from 33% to 77%, with the increase of 33% between IRR 1 and IRR 2 being much more substantial than the increase of 11% between IRR 2 and IRR 3. The proportions of assessors with moderate and limited experience subsequently decreased, particularly for assessors with limited experience from IRR 2 to IRR 3.

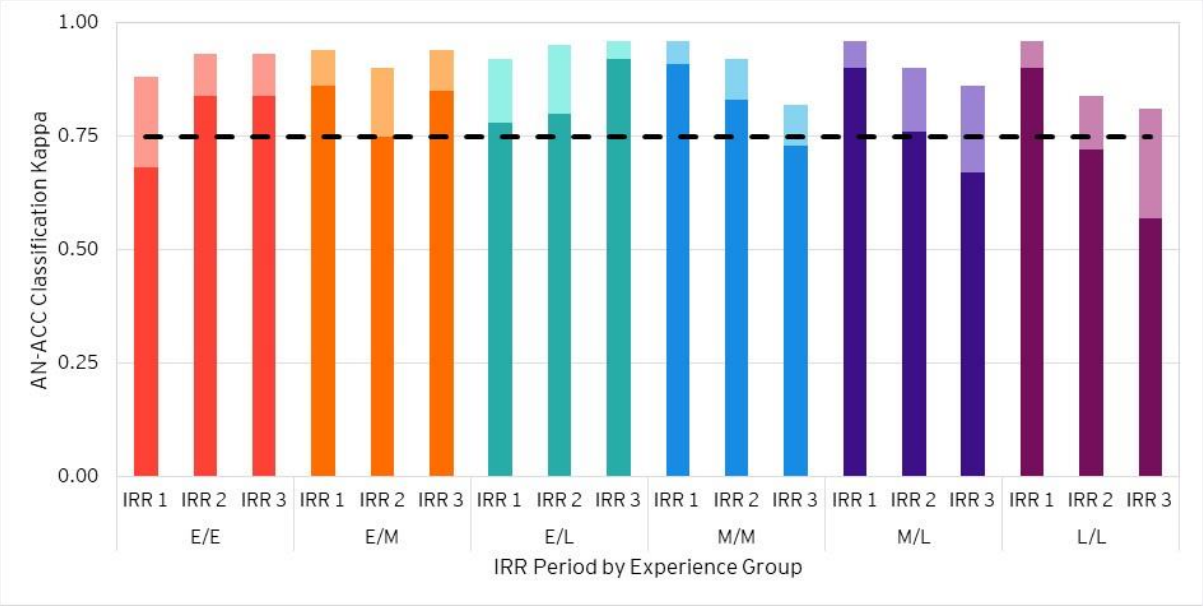
Table 12: Distribution of IRR assessments by assessor experience pairing and IRR period

Assessor Experience Combination		Mapping	IRR 1 Assessments		IRR 2 Assessments		IRR 3 Assessments	
			No.	Prop.	No.	Prop.	No.	Prop.
Extensive	Extensive	Extensive only (E/E)	90	14%	474	51%	662	60%
Extensive	Moderate	Extensive - moderate (E/M)	176	27%	240	26%	251	23%
Extensive	Limited	Extensive - limited (E/L)	81	12%	108	12%	113	10%
Moderate	Moderate	Moderate only (M/M)	136	21%	34	4%	55	5%
Moderate	Limited	Moderate - limited (M/L)	124	19%	34	4%	11	1%
Limited	Limited	Limited only (L/L)	49	7%	29	3%	6	1%
Total			656	100%	919	100%	1,098	100%

The proportion of IRR assessments completed by a pair of *"extensive only"* assessors increased from 14% to 60%, with subsequent decreases to the proportions of assessments completed by all other assessor experience combinations, particularly for *"moderate only"*, *"limited only"* and *"moderate – limited"* pairs. This is reflective of the overall change in assessor experience mix as assessors perform more assessments and gain more experience.

Figure 10 shows the comparison in Fleiss' kappa statistics for AN-ACC classification for pairs of assessors by experience level in each IRR period.

Figure 10: AN-ACC classification agreement statistics by IRR period and experience group



i. *Excellent agreement is represented by the dashed line (0.75)*
ii. *Exact agreement (unweighted kappa) is shown in darker shading for each classification/assessment instrument.*
iii. *Agreement with tolerance for partial agreement (weighted kappa) above exact agreement is shown in lighter shading.*

Key observations from Figure 10 are:

- Excellent agreement was observed when measured with a tolerance for partial agreement at the AN-ACC classification level for all assessor experience pairs in each IRR period.
- Agreement rates for pairs of *"extensive - moderate"* and *"extensive – limited"* assessors increased from IRR 2 to IRR 3. *"Extensive only"* pairs maintained the same high agreement from IRR 2 to IRR 3.
- Agreement rates for pairs of *"moderate only"*, *"moderate - limited" and "limited only"* assessors decreased in IRR 3 and formed decreasing trends from IRR 1 to IRR 3, with the largest decrease of 10 percentage points for *"moderate only"* pairs of assessors in IRR 3.
    - 53% of *"moderate only"* assessor pairs were from AMO 5, and 42% were from AMO 1.
    - The decrease in agreement rates for pairs of *"moderate – limited"* and *"limited only"* assessors in IRR 3 are influenced by volatility and lower reliability arising from low IRR assessment counts, as shown in Section 3.3.

An improvement in overall agreement rates between assessors in IRR 3 and future IRR rounds is expected as the proportion of assessors with extensive experience and agreement of these assessor pairs increases. This was shown through the increase in the proportion of assessments completed by *"extensive only"* pairs, despite no change in agreement on the AN-ACC classification from IRR 2 to IRR 3, alongside the decrease in the proportions of assessments completed by *"moderate – limited"* and *"limited only"* pairs.

3.3.2 Assessor experience - Adjusted thresholds

To investigate maturing assessor experience over time and the associated impacts on agreement rates for experienced assessors, higher thresholds for the number of assessments completed have additionally been developed and splits the IRR assessor cohort into Levels A, B and C.

Table 13 shows the distribution of IRR assessors by assessor experience using alternative experience level thresholds, while Table 14 shows the distribution of IRR assessments by assessor experience pairing using these alternative thresholds.

Table 13: Distribution of IRR assessors by assessor experience, with new thresholds

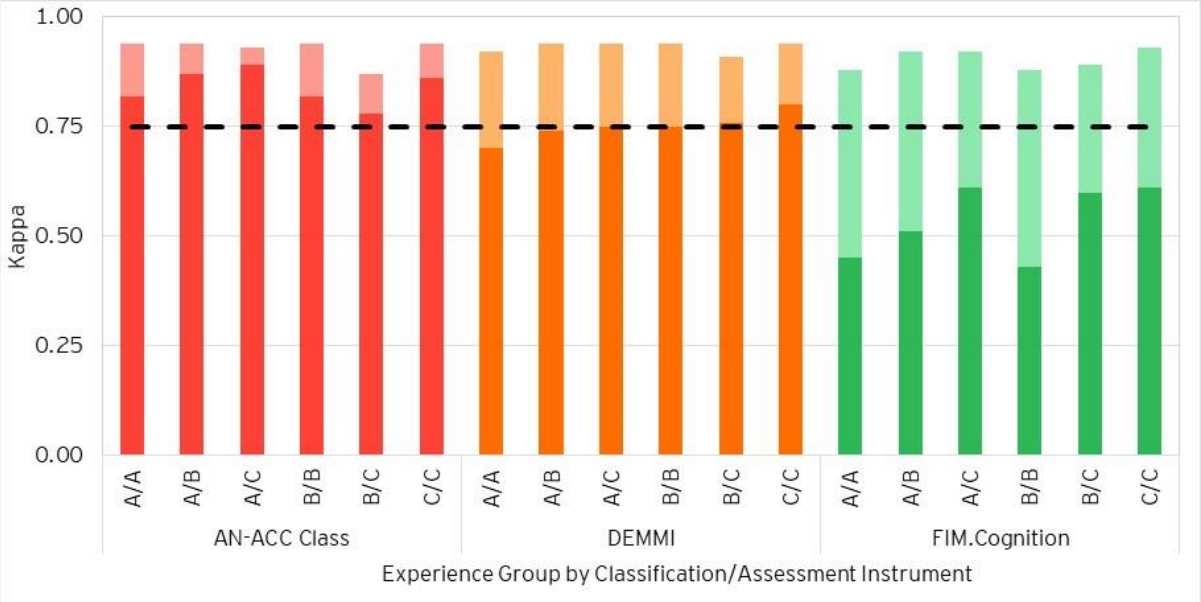
Assessor Experience Level	Number of Assessments Competed	IRR 3 Assessors	
		No.	Prop.
A	(1300, Inf)	91	34%
B	(600, 1300]	92	34%
C	[0, 600]	89	33%
Total		272	100%

Table 14: Distribution of IRR assessments by assessor experience pairing, with new thresholds

Assessor Experience Combination		Mapping	IRR 3 Assessments	
			No.	Prop.
A	A	A only (A/A)	175	16%
A	B	A - B (A/B)	229	21%
A	C	A - C (A/C)	263	24%
B	B	B only (B/B)	104	9%
B	C	B - C (B/C)	213	19%
C	C	C only (C/C)	114	10%
Total			1,098	100%

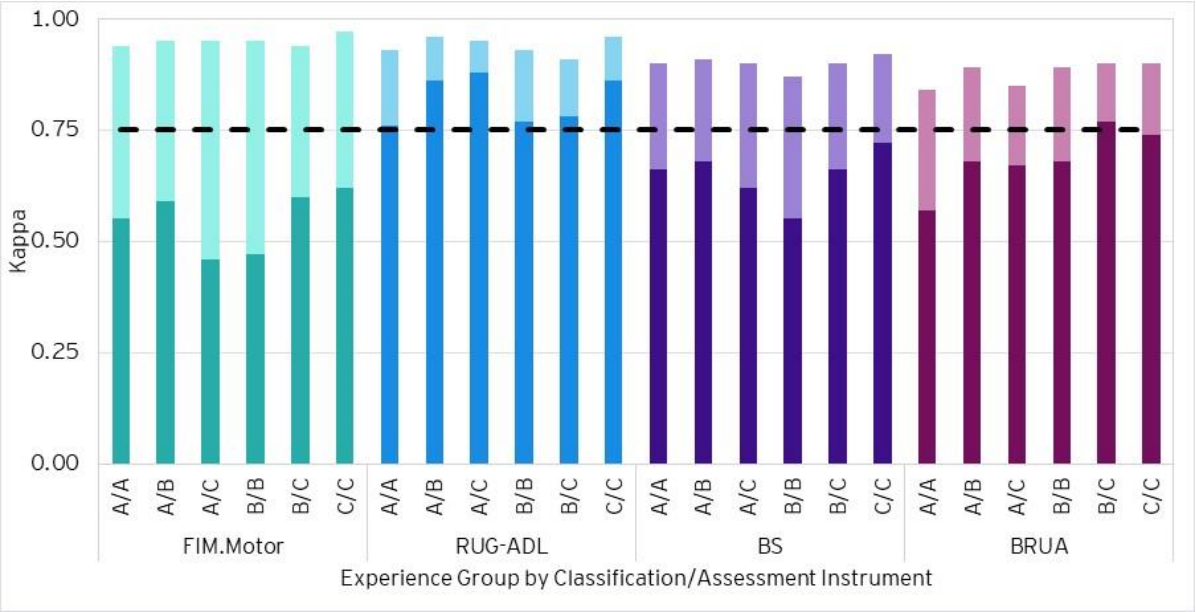
As observed in Table 13 and Table 14, a roughly even distribution of assessors between assessor experience levels was produced after adjusting the thresholds. This resulted in a greater representation across all assessor experience pairings containing no assessor with extensive experience, particularly in those pairings containing at least one assessor with limited experience.

Figure 11: Agreement statistics by assessor experience under new thresholds (AN-ACC Class, DEMMI and FIM.Cognition)



i. *Excellent agreement is represented by the dashed line (0.75)*
ii. *Exact agreement (unweighted kappa) is shown in darker shading for each classification/assessment instrument.*
iii. *Agreement with tolerance for partial agreement (weighted kappa) above exact agreement is shown in lighter shading.*

Figure 12: Agreement statistics by assessor experience under new thresholds (FIM.Motor, RUG-ADL, BS and BRUA)



i. *Excellent agreement is represented by the dashed line (0.75)*
ii. *Exact agreement (unweighted kappa) is shown in darker shading for each classification/assessment instrument.*
iii. *Agreement with tolerance for partial agreement (weighted kappa) above exact agreement is shown in lighter shading.*

Key observations from Figure 11 and Figure 12 are:

- Excellent agreement was observed for all assessor experience pairs when measured with a tolerance for partial agreement across the AN-ACC classification and all underlying instruments.
- Higher agreement rates were generally observed for assessor pairs including at least one assessor with experience level A, with the exception of the BRUA where *"A only"* and *"A – C"* pairs had the lowest weighted agreement rates.
- Similarly to under the existing thresholds, it was expected that agreement rates would be highest for *"A only"* pairs. However, agreement rates between *"A only"* pairs were equal to or lower than *"A – B"* and *"A – C"* pairs across all instruments, suggesting that the most experienced assessors do not have higher agreement over pairs involving a highly experienced assessor with an assessor with lower experience. However, using the alternative thresholds reduces the difference in agreement observed compared to when using the old thresholds, and suggests that overall assessment quality becomes more consistent once assessors gain enough experience.
- *"B – C"* pairs generally had among the lowest agreement across the assessor pairs for most instruments, with the exception of the BS and BRUA.
- Agreement rates for *"C only"* pairs were generally high compared to all other assessor pairs.

3.4 Segmentation by assessor profession

This section examines the agreement of assessors by their profession.

AN-ACC assessors completing IRR assessments are associated with 3 different professions: registered nurses (RN), occupational therapists (OT), and physiotherapists (PT). Due to the limited number of assessments completed by occupational therapists and physiotherapists, these professions have been grouped as "*Other*" for this analysis.

Table 15 and Table 16 below show the distributions of assessors and IRR assessments by assessor profession and assessor profession pairing respectively.

Table 15: Distribution of assessors in IRR 3 by assessor profession

Assessor Profession	No. Assessors	Prop. of assessor population
RN	210	77.2%
OT	23	8.5%
PT	39	14.3%
Total	272	100%

Table 16: Assessor profession mapping and distribution of IRR assessments by pairing

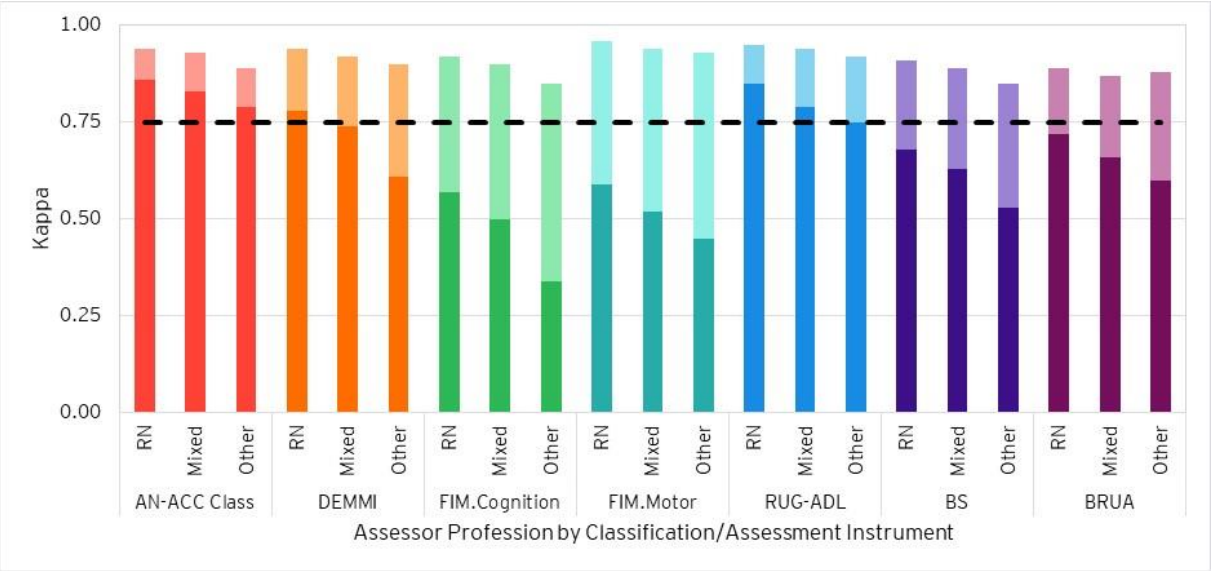
Assessor Profession Pairings		Mapping	No. IRR Assessments	Prop. in IRR 3
RN	RN	RN Only	617	56.2%
RN	Other	Mixed	389	35.4%
Other	Other	Other	92	8.4%
Total			1,098	100%

"Other" assessor profession pairings were primarily made up of AMO 5 and 6 assessments, comprising 90% of all other pairings. *"Mixed"* assessor profession pairings were primarily made up of AMO 5 assessments, making up 43% of all mixed pairings. *"RN Only"* assessor profession pairings were made up of a wider range of AMOs, with no AMO comprising more than 36% of dual assessments for any other assessor profession pairing.

As such, we note that agreement rates for "Other" and "Mixed" profession pairings are heavily biased towards certain AMO assessors and not necessarily representative of all "Other" and "Mixed" profession pairings.

Figure 13 contains agreement statistics for the IRR assessments conducted by different profession pairs, with values of each statistic presented in Appendix B.3.

Figure 13: Agreement statistics by assessor profession and classification/assessment instrument



i. *Excellent agreement is represented by the dashed line (0.75)*
ii. *Exact agreement (unweighted kappa) is shown in darker shading for each classification/assessment instrument.*
iii. *Agreement with tolerance for partial agreement (weighted kappa) above exact agreement is shown in lighter shading.*

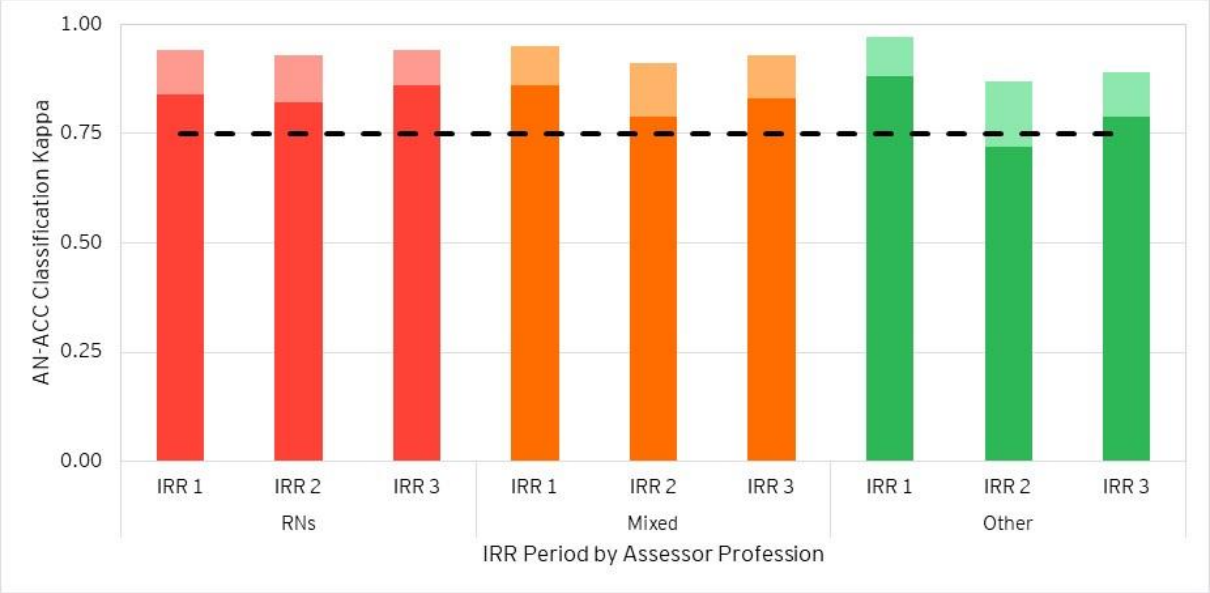
Key observations from Figure 13 are:

- Excellent agreement was observed for all assessor profession pairings when measured with a tolerance for partial agreement.
- At the AN-ACC classification level and across all instruments, agreement statistics were highest for a pair of RNs, followed by an RN and an OT or PT, and lastly a pair compromised of PTs or OTs, except for the weighted agreement for the BRUA where "*other*" pairs had a slightly higher weighted agreement but lower exact agreement than "*mixed*" pairs.

3.4.1 Comparison against IRR 1 and IRR 2

Figure 14 shows the comparison in Fleiss' kappa statistics for AN-ACC classification for pairs of assessors by profession in each IRR period.

Figure 14: AN-ACC classification agreement statistics by IRR period and assessor profession



i. *Excellent agreement is represented by the dashed line (0.75)*
ii. *Exact agreement (unweighted kappa) is shown in darker shading for each classification/assessment instrument.*
iii. *Agreement with tolerance for partial agreement (weighted kappa) above exact agreement is shown in lighter shading.*

Key observations from Figure 14 are:

- Excellent agreement was observed when measured with a tolerance for partial agreement at the AN-ACC classification level for all assessor profession pairs in each IRR period.
- Agreement measured with a tolerance for partial agreement slightly increased from IRR 2 to IRR 3 for all pairs of assessors by profession, by 0.02 for "*mixed*" and "*other*" assessor pairings and by 0.01 for pairs of RNs. Following the decrease in agreement from IRR 1 to IRR 2, IRR 1 maintains the highest agreement of the three IRR rounds when measured with a tolerance for partial agreement, noting that IRR 1 and IRR 3 weighted agreement are equal for RNs.

3.5 Segmentation by assessment type

Each IRR assessment consists of one *"payment impacting"* assessment performed alongside a *"no-payment impact"* quality assurance assessment. This section analyses agreement rates segmented by the assessment type of the *"payment impacting"* assessment.

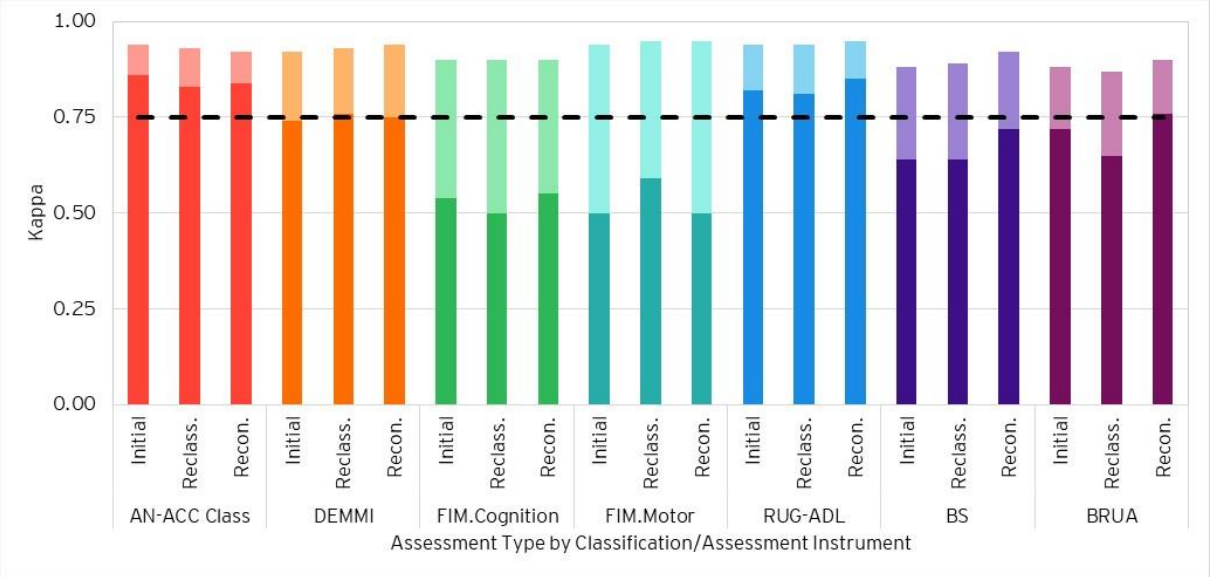
Table 17 shows the number and proportion of IRR assessments completed by assessment type.

Table 17: IRR assessments by *"payment impacting"* assessment type

Assessment Type	No. IRR Assessments	Prop. IRR Assessments
Initial	375	34.2%
Reclassification	624	56.8%
Reconsideration	99	9.0%
Total	1,098	100%

The Fleiss' kappa and weighted Fleiss' kappa test statistics for the final AN-ACC classification and each underlying instrument segmented by assessment type is shown in Figure 15.

Figure 15: Agreement statistics by assessment type and classification/assessment instrument



i. *Excellent agreement is represented by the dashed line (0.75)*
ii. *Exact agreement (unweighted kappa) is shown in darker shading for each classification/assessment instrument.*
iii. *Agreement with tolerance for partial agreement (weighted kappa) above exact agreement is shown in lighter shading.*

Key observations from Figure 15 are:

- Excellent agreement between assessors for all assessment types was observed for the AN-ACC classification and all underlying assessment instruments when measured with a tolerance for partial agreement.
- Weighted agreement was consistent (within 0.04) across assessment types for the AN-ACC classification and all the underlying instruments.

- Whilst levels of agreement were consistent, weighted agreement on initial assessments was observed to be lower than other assessment types for the DEMMI, FIM.Motor and BS, and higher than other assessment types for the AN-ACC classification.

  Weighted agreement on reconsiderations was observed to be generally slightly higher than other assessment types, particularly for the BS and BRUA. Exceptions include the AN-ACC classification, where the other assessment types have higher agreement, the FIM.Cognition, where weighted agreement was equal across all assessment types, and FIM.Motor, where weighted agreement for reclassifications and reconsiderations was equal and above that of initial assessments.

3.5.1 Comparing agreement levels against IRR 1 and IRR 2

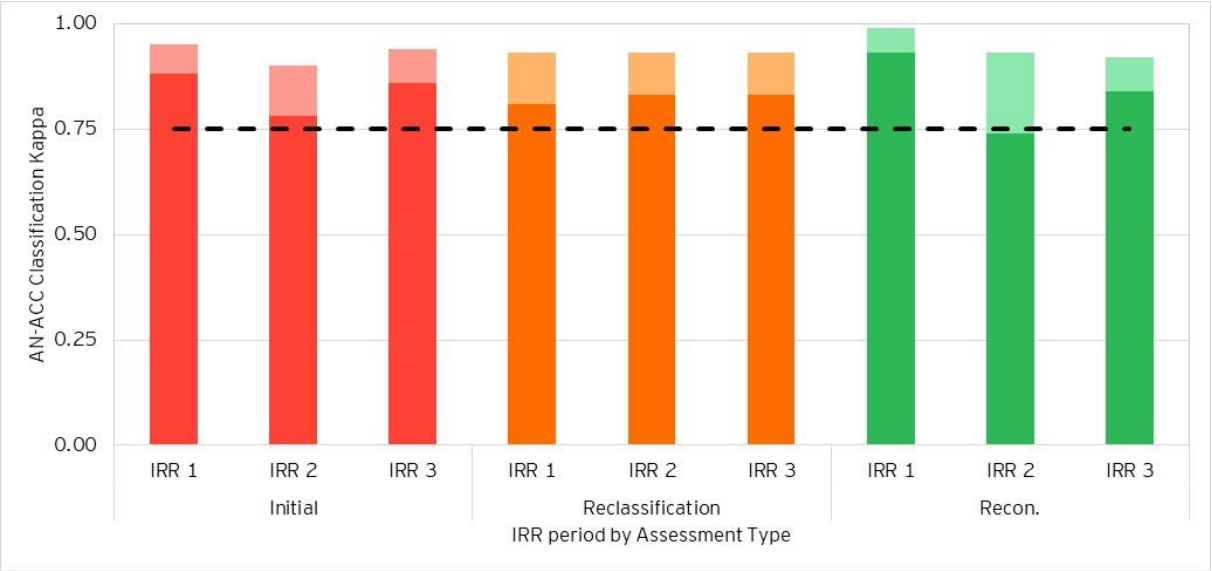
The mix of AN-ACC assessments by assessment type has changed since the beginning of the AN-ACC period on 1 October 2022, with an increasing proportion of assessments being reconsideration assessments and a decreasing proportion of assessments being initial assessments, as observed through regular monitoring for the Department. Meanwhile, the proportion of reclassification assessments increased from the beginning of the AN-ACC period to approximately March 2023, where it then stabilised moving forward. These trends in assessments by type was particularly reflected in the changing mix of assessments in the samples chosen for IRR 1 and IRR 2, as shown in Table 18 below, with some allowance for volatility in sample selection between IRR 2 and IRR 3 causing smaller changes in proportions.

Table 18: Distribution of IRR assessments by *"payment impacting"* assessment type and IRR period

Assessment Type	IRR 1 Assessments		IRR 2 Assessments		IRR 3 Assessments	
	No.	Prop.	No.	Prop.	No.	Prop.
Initial	311	47.4%	300	32.7%	375	34.2%
Reclassification	328	50.0%	541	59.1%	624	56.8%
Reconsideration	17	2.6%	75	8.2%	99	9.0%
Total	656	100%	916	100%	1,098	100%

Figure 16 shows the comparison in Fleiss' kappa statistics for each assessment type by AN-ACC classification and assessment instrument between the IRR 1, IRR 2 and IRR 3 assessments.

Figure 16: AN-ACC classification agreement statistics by IRR period and assessment type



i. *Excellent agreement is represented by the dashed line (0.75)*
ii. *Exact agreement (unweighted kappa) is shown in darker shading for each classification/assessment instrument.*
iii. *Agreement with tolerance for partial agreement (weighted kappa) above exact agreement is shown in lighter shading.*

The key observations from Figure 16 are:

- Excellent agreement when measured with a tolerance for partial agreement was observed for the AN-ACC classification for all assessment types across each IRR period.
- Weighted agreement was overall more consistent across assessment types when compared to IRR 1 and IRR 2. This may imply a broader improvement in the consistency of assessments in general regardless of assessment type which may extend beyond agreement rates.
- Weighted agreement for the AN-ACC classification on reclassification assessments remained constant at 0.93 across each of the three IRR rounds.
- Weighted agreement for the AN-ACC classification on initial assessments increased by 0.04 from IRR 2 to IRR 3, following a decrease of 0.05 from IRR 1 to IRR 2.
- Weighted agreement for the AN-ACC classification on reconsideration assessments decreased by 0.01 from IRR 2 to IRR 3, following a decrease of 0.06 from IRR 1 to IRR 2 and forming a declining trend.

The shift in the mix of IRR assessments towards initial assessments, and the associated increase in weighted agreement, were key drivers towards the overall increase in agreement between IRR 2 and IRR 3 (as observed in Section 3.1.1). Meanwhile, the smaller increase in the proportion of reconsiderations, and the associated slight decrease in weighted agreement, did not have a large impact on the overall agreement in IRR 3.

Appendix A

A.1 Kappa statistics

A kappa statistic measures the level of agreement between assessors with an adjustment for the probability that those assessors will agree purely by chance. For example, if two assessors each place 20% of residents into class 5, then there is a 4% chance that they will both place the same resident into class 5 before allowing for any actual agreement or disagreement between them (i.e. purely by chance).

This expected level of agreement by chance (P_E) is removed from the observed level of agreement (P_O) between assessors when calculating kappa (κ) to provide a more robust measure of agreement for inter-rater reliability testing. This is expressed as:

$$\kappa = \frac{P_O - P_E}{1 - P_E}$$

Where κ = kappa, P_O = probability of agreement observed and P_E = probability of agreement expected. The calculation of the kappa statistic can be broken down into two components:

- $1 - P_E$, the potential additional agreement which can be achieved above chance (a value of 1 represents perfect agreement and we subtract the expected level of agreement by chance)
- $P_O - P_E$, the level of agreement actually achieved above chance

It is noted that P_E will equal to 1 when all assessors give the exact same score, thereby causing an erroneous kappa value of NA. However, this is expected to be an extremely rare scenario, particularly at higher assessor counts, and therefore is only noted as a limitation.

There are several formulations of the kappa statistic, each taking a slightly different approach to calculating P_O and P_E. The kappa statistics used in this analysis are summarised in Table 19.

Table 19: Summary of kappa statistics

#	Test statistic	Application	Limitations
1	Fleiss' kappa	Any number of assessors (sampled randomly) classify a fixed number of items into mutually exclusive (non-ordinal) categories.	AN-ACC classifications and the underlying assessment instruments are fundamentally ordinal by RVU or score. All disagreements are considered equal (e.g. classes 2 and 13 have the same influence on kappa as classes 2 and 3). Combinations of assessors completing IRR assessments are influenced by geographical and logistical constraints (i.e. assessors are not sampled randomly).
2	Weighted Fleiss' kappa	When classification categories are ordinal, the level of disagreement can be weighted to produce both weighted observed agreement and weighted expected agreement by chance.	Judgement is required in selecting the weighting structure and interpreting the kappa statistic.

As detailed in the '*Limitations*' column above, there are challenges with using each individual kappa statistic for the purpose of evaluating agreement in AN-ACC IRR assessments. Therefore, a level of judgement is required when interpreting the calculated kappa statistics as well as considering the rate of agreement and correlation results.

Guidance on interpreting values of kappa is most readily available for Cohen's kappa and is detailed in Table 2. This interpretation aligns with the guidelines adopted in the "*Reliability of the Australian National Aged Care Classification shadow assessments*" report on AN-ACC assessment IRR published in March 2022.[8]

[8] Available from: https://www.health.gov.au/resources/publications/reliability-of-the-australian-national-aged-care-classification-shadow-assessments

A.2 Kappa weightings

Before explaining weights in the kappa calculation, we will first reference the unweighted kappa. In this case the "*weights*" applied are essentially values of 1, for agreement, and 0, for disagreement between assessors.

This is shown in Table 20, with the example of two assessors making a classification between 3 categories. If in this example the assessors also make the classification at random (i.e., there is equal chance of each assessor scoring 1, 2 or 3) then the expected rate of agreement by chance would be 1 in 3.

Table 20: Unweighted kappa - Weights

Weights		Assessor 1 Score		
		1	2	3
Assessor 2 Score	1	1	0	0
	2	0	1	0
	3	0	0	1

Table 21 introduces linear weights based on the difference in scores compared to the difference between the maximum and minimum scores available. In this example, perfect agreement is still assigned a weight of 1, however, a score difference of 1, is assigned a weight of 0.5, because it is half of the full range of possible disagreement (2).

Table 21: Linearly weighted kappa - Weights

Weights		Assessor 1 Score		
		1	2	3
Assessor 2 Score	1	1	0.5	0
	2	0.5	1	0.5
	3	0	0.5	1

There is no definitive or prescriptive method for determining the weights which should be used in the calculation of a weighted kappa, other than the general principle in all modelling that they should be chosen to best represent the system or situation which is being analysed.

For the AN-ACC classification the NWAU for each class is used to define weights in weighted Fleiss' kappa.

Appendix B

B.1 Assessment Management Organisation

B.1.1 AMO – Fleiss' kappa

Table 22 contains kappa statistics by AMO. Exact agreement (unweighted Fleiss' kappa) is labelled *"UN"* and agreement with a tolerance for partial agreement (weighted kappa) is labelled *"W"*.

Table 22: Fleiss' kappa by AMO

Metric	AMO 1		AMO 2		AMO 3		AMO 4		AMO 5		AMO 6	
	W	UN	W	UN	W	UN	W	UN	W	UN	W	UN
AN-ACC Class	0.89	0.82	0.98	0.95	0.91	0.78	0.89	0.79	0.95	0.87	0.92	0.76
DEMMI	0.89	0.68	0.97	0.87	0.90	0.65	0.92	0.73	0.96	0.82	0.90	0.63
FIM.Cognition	0.92	0.66	0.97	0.80	0.81	0.23	0.86	0.34	0.93	0.60	0.87	0.40
FIM.Motor	0.94	0.60	0.99	0.75	0.92	0.31	0.94	0.54	0.96	0.64	0.92	0.38
RUG-ADL	0.90	0.75	1.00	0.98	0.91	0.69	0.96	0.86	0.96	0.87	0.91	0.71
BS	0.90	0.70	0.97	0.85	0.86	0.46	0.82	0.51	0.91	0.71	0.85	0.53
BRUA	0.88	0.70	0.95	0.87	0.81	0.57	0.85	0.58	0.92	0.74	0.82	0.55

B.1.2 AMO – Correlation coefficient

Table 23 contains Pearson correlation coefficients by AMO.

Table 23: Correlation by AMO

Metric	AMO 5	AMO 2	AMO 1	AMO 3	AMO 4	AMO 6
AN-ACC Class	0.98	1.00	0.92	0.96	0.96	0.97
DEMMI	0.99	1.00	0.95	0.97	0.98	0.98
FIM.Cognition	0.99	1.00	0.98	0.95	0.97	0.97
FIM.Motor	0.99	1.00	0.99	0.99	0.99	0.99
RUG-ADL	0.99	1.00	0.95	0.97	0.99	0.97
BS	0.98	1.00	0.96	0.97	0.95	0.96
BRUA	0.97	0.98	0.96	0.90	0.96	0.93

B.2 Assessor experience

B.2.1 Assessor experience – Fleiss' kappa

Table 24 contains kappa statistics by assessor experience. Exact agreement (unweighted Fleiss' kappa) is labelled "*UN*" and agreement with a tolerance for partial agreement (weighted kappa) is labelled "*W*".

Table 24: Fleiss' kappa by assessor experience pairs

Metric	Both Extensive		Extensive/ Moderate		Extensive/ Limited		Both Moderate		Moderate/ Limited		Both Limited	
	W	UN	W	UN	W	UN	W	UN	W	UN	W	UN
AN-ACC Class	0.93	0.84	0.94	0.85	0.96	0.96	0.82	0.73	0.86	0.67	0.81	0.57
DEMMI	0.93	0.73	0.95	0.79	0.95	0.95	0.86	0.69	0.92	0.65	0.91	0.59
FIM.Cognition	0.89	0.47	0.91	0.52	0.96	0.96	0.84	0.56	0.88	0.39	0.96	0.61
FIM.Motor	0.94	0.51	0.95	0.57	0.98	0.98	0.92	0.57	0.90	0.31	0.99	0.80
RUG-ADL	0.94	0.81	0.95	0.81	0.98	0.98	0.87	0.71	0.87	0.78	0.84	0.52
BS	0.89	0.61	0.90	0.68	0.96	0.96	0.84	0.59	0.90	0.37	0.96	0.79
BRUA	0.86	0.64	0.89	0.70	0.97	0.97	0.87	0.67	0.83	0.79	1.00	1.00

B.2.2 Assessor experience – Correlation coefficient

Table 25 contains Pearson correlation coefficients by assessor experience.

Table 25: Correlation coefficient by assessor experience pairs

Metric	E/E	E/M	E/L	M/M	M/L	L/L
AN-ACC Class	0.97	0.99	0.98	0.85	0.95	0.98
DEMMI	0.98	0.99	0.99	0.93	0.99	1.00
FIM.Cognition	0.98	0.98	0.99	0.93	0.98	1.00
FIM.Motor	0.99	0.99	1.00	0.98	0.98	1.00
RUG-ADL	0.98	0.99	1.00	0.92	0.94	0.98
BS	0.97	0.98	0.99	0.93	0.99	1.00
BRUA	0.95	0.96	0.99	0.96	0.86	1.00

B.3 Assessor profession

B.3.1 Assessor profession – Fleiss' kappa

Table 26 contains unweighted Fleiss' kappa statistics by assessor profession. Exact agreement (unweighted Fleiss' kappa) is labelled "*UN*" and agreement with a tolerance for partial agreement (weighted kappa) is labelled "*W*".

Table 26: Fleiss' kappa across different assessor profession pairs

Metric	RN		Mixed		Other	
	W	UN	W	UN	W	UN
AN-ACC Class	0.94	0.86	0.93	0.83	0.89	0.79
DEMMI	0.94	0.78	0.92	0.74	0.90	0.61
FIM.Cognition	0.92	0.57	0.90	0.50	0.85	0.34
FIM.Motor	0.96	0.59	0.94	0.52	0.93	0.45
RUG-ADL	0.95	0.85	0.94	0.79	0.92	0.75
BS	0.91	0.68	0.89	0.63	0.85	0.53
BRUA	0.89	0.72	0.87	0.66	0.88	0.60

B.3.2 Assessor profession – Correlation coefficient

Table 27 contains Pearson correlation coefficients by assessor profession.

Table 27: Correlation coefficient across difference assessor profession pairs

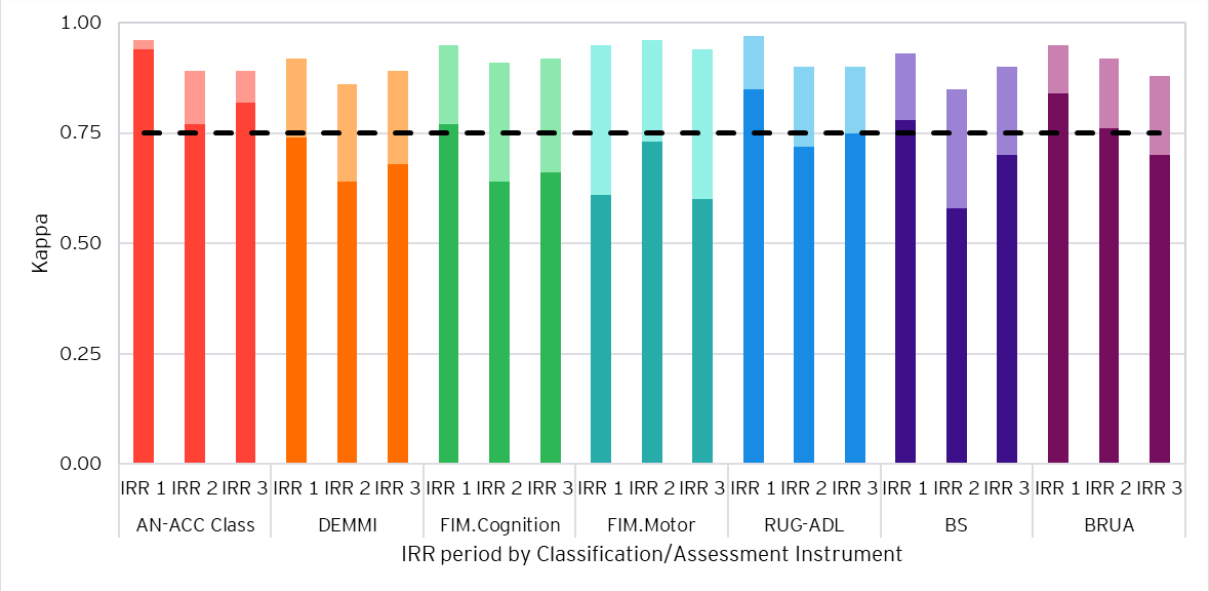
Metric	RN	Mixed	Other
AN-ACC Class	0.98	0.97	0.95
DEMMI	0.99	0.98	0.98
FIM.Cognition	0.98	0.98	0.96
FIM.Motor	0.99	0.99	0.99
RUG-ADL	0.98	0.98	0.98
BS	0.98	0.97	0.96
BRUA	0.95	0.95	0.97

B.4 AMO agreement over time

Figure 17 to

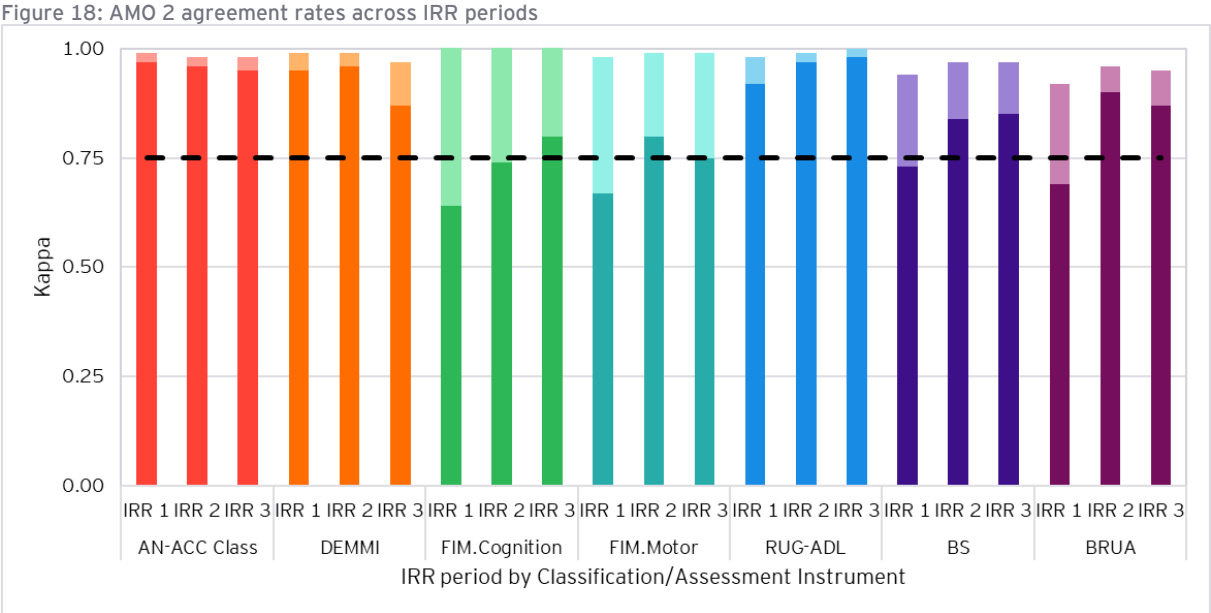
Figure 22 show the kappa statistics for the AN-ACC classification and each of the underlying assessment instruments in IRR 1, IRR 2 and IRR 3.

Figure 17: AMO 1 agreement rates across IRR periods



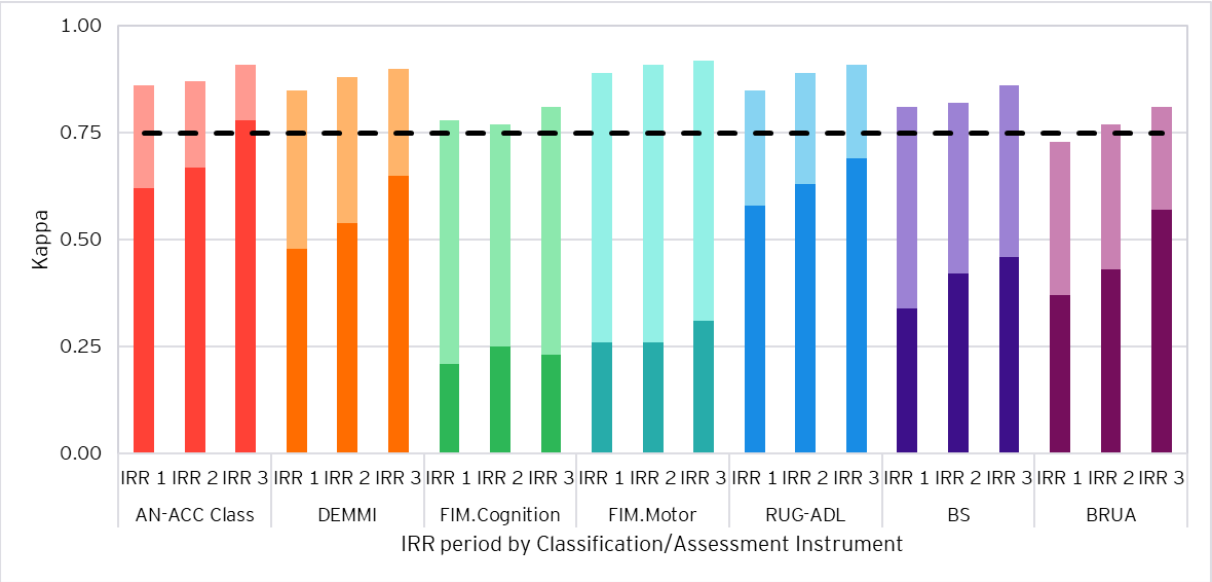
i. *Excellent agreement is represented by the dashed line (0.75)*
ii. *Exact agreement (unweighted kappa) is shown in darker shading for each classification/assessment instrument.*
iii. *Agreement with tolerance for partial agreement (weighted kappa) above exact agreement is shown in lighter shading.*

Figure 18: AMO 2 agreement rates across IRR periods



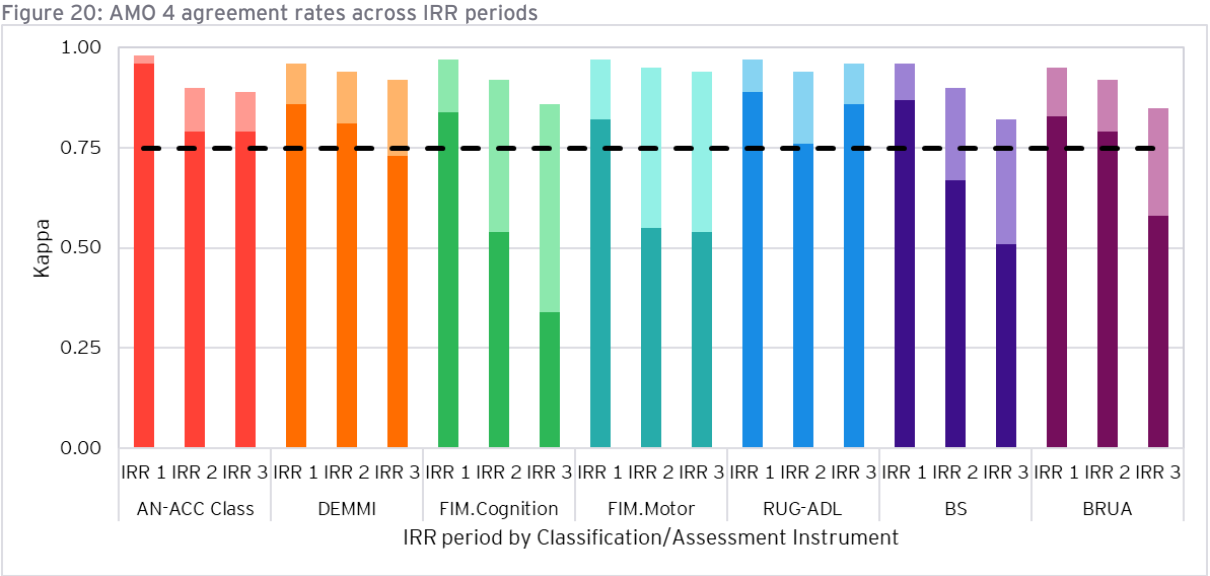
i. *Excellent agreement is represented by the dashed line (0.75)*
ii. *Exact agreement (unweighted kappa) is shown in darker shading for each classification/assessment instrument.*
iii. *Agreement with tolerance for partial agreement (weighted kappa) above exact agreement is shown in lighter shading.*

Figure 19: AMO 3 agreement rates across IRR periods



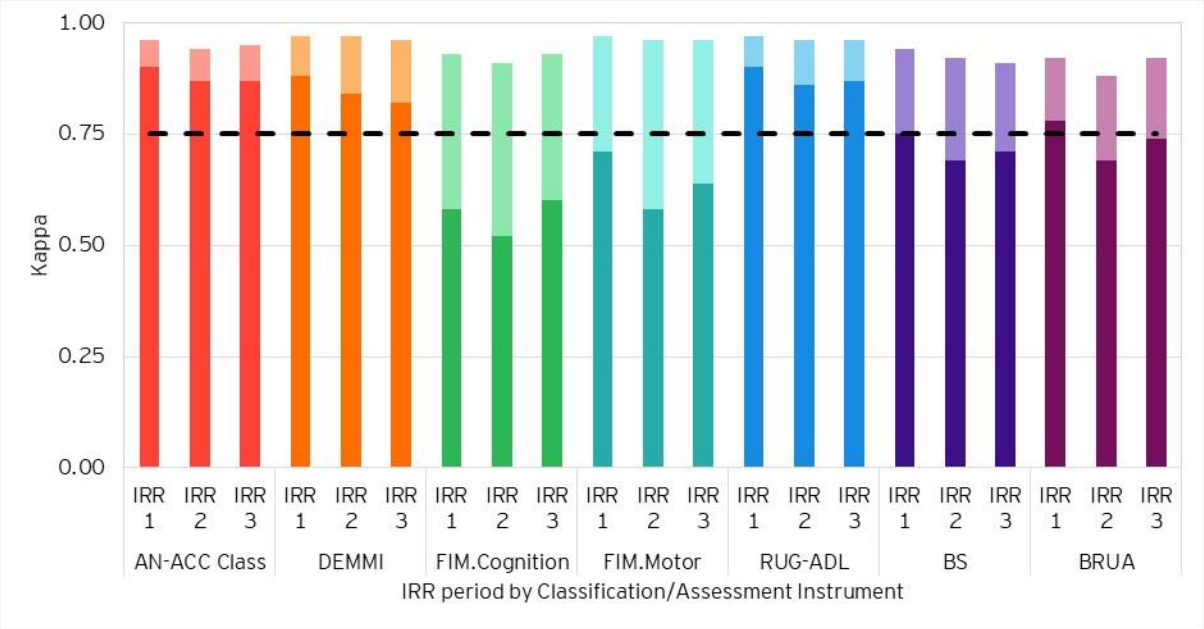
i. *Excellent agreement is represented by the dashed line (0.75)*
ii. *Exact agreement (unweighted kappa) is shown in darker shading for each classification/assessment instrument.*
iii. *Agreement with tolerance for partial agreement (weighted kappa) above exact agreement is shown in lighter shading.*

Figure 20: AMO 4 agreement rates across IRR periods



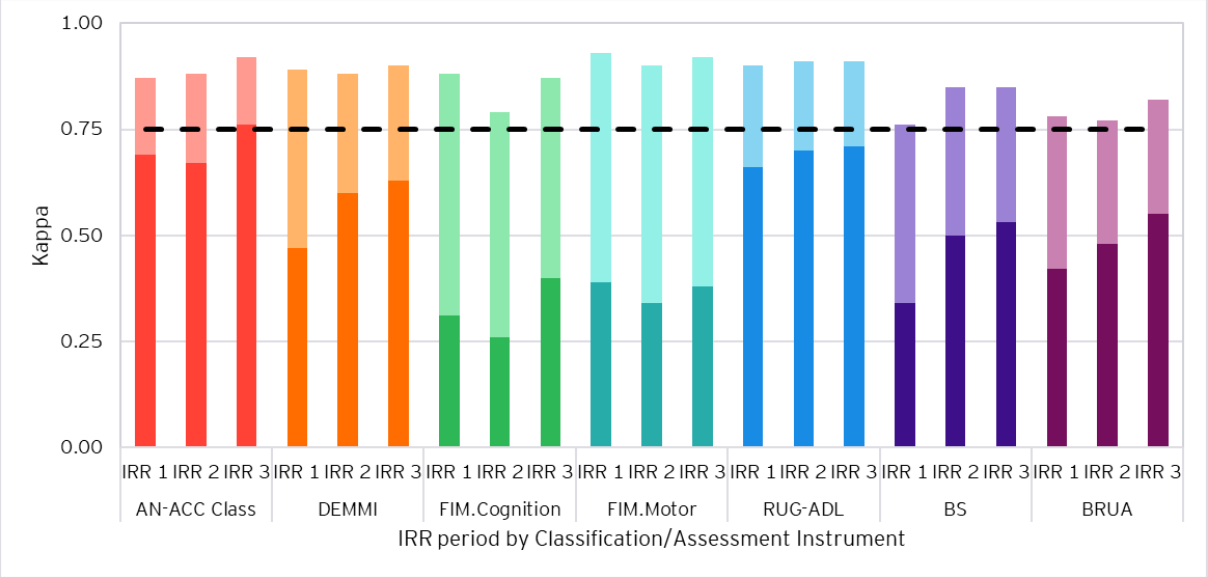
i. *Excellent agreement is represented by the dashed line (0.75)*
ii. *Exact agreement (unweighted kappa) is shown in darker shading for each classification/assessment instrument.*
iii. *Agreement with tolerance for partial agreement (weighted kappa) above exact agreement is shown in lighter shading.*

Figure 21: AMO 5 agreement rates across IRR periods



i. *Excellent agreement is represented by the dashed line (0.75)*
ii. *Exact agreement (unweighted kappa) is shown in darker shading for each classification/assessment instrument.*
iii. *Agreement with tolerance for partial agreement (weighted kappa) above exact agreement is shown in lighter shading.*

Figure 22: AMO 6 agreement rates across IRR periods



i. *Excellent agreement is represented by the dashed line (0.75)*
ii. *Exact agreement (unweighted kappa) is shown in darker shading for each classification/assessment instrument.*
iii. *Agreement with tolerance for partial agreement (weighted kappa) above exact agreement is shown in lighter shading.*

B.5 Assessor pairs

There were 353 unique assessor pairs in the IRR assessments. Of those pairs, ten assessed ten or more residents together. We conducted reliability testing on each of these pairs to measure agreement between specific assessor pairs.

All pairs of assessors had excellent agreement on the AN-ACC classification when measured with a tolerance for partial agreement, reflective of the overall IRR results presented in this report. Three pairs of assessors had moderate pure agreement on the AN-ACC classification.

In the April 2023 IRR report, nine (different) pairs of assessors had 10 or more IRR assessments performed together. These pairs of assessors collectively completed 122 IRR assessments, for which excellent agreement was also observed between all nine pairs of assessors, when measured with a tolerance for partial agreement.

B.6 Assessors in IRR 2 and IRR 3

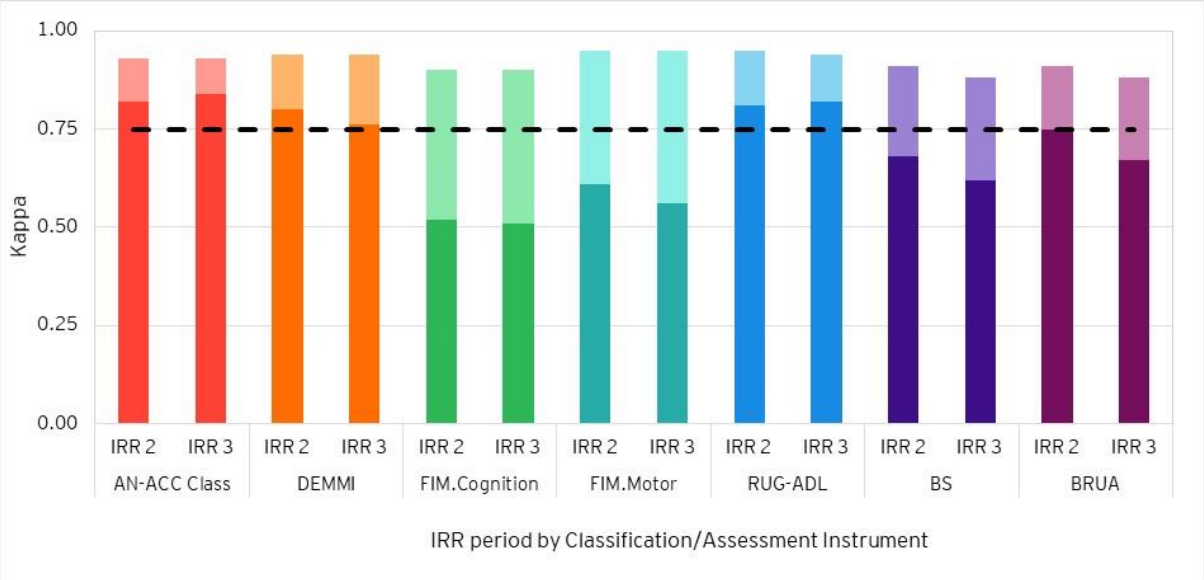
93 pairs of assessors performed IRR assessments together in both periods. These 93 pairs of assessors completed a total of 348 IRR assessments in each of IRR 2 and IRR 3.

Table 28 shows agreement rates by AN-ACC decision tree level across IRR 2 and IRR 3, and Figure 23 shows the comparison in kappa statistics at the AN-ACC and instrument level for these assessors, split by IRR period.

Table 28: Agreement rates by AN-ACC decision tree level for pairs of assessors in IRR 1 and 2 by IRR period

Level	Category	Fleiss' Kappa			
		Exact (unweighted)		Tolerance for partial (linearly weighted)	
		IRR 2	IRR 3	IRR 2	IRR 3
1	Mobility Branch	0.96	0.95	0.98	0.96
2	Mobility: Cognition/ Function/ Pressure Sores	0.86	0.89	0.95	0.95
3	AN-ACC Class	0.82	0.84	0.93	0.93

Figure 23: Agreement rates by AN-ACC classification/assessment instrument for pairs of assessors in IRR 1 and 2 by IRR period



i. *Excellent agreement is represented by the dashed line (0.75)*
ii. *Exact agreement (unweighted kappa) is shown in darker shading for each classification/assessment instrument.*
iii. *Agreement with tolerance for partial agreement (weighted kappa) above exact agreement is shown in lighter shading.*

The key observations from Table 28 and Figure 23 are:

- At each level of the AN-ACC decision tree weighted agreement statistics are within 0.02.
- For each of the underlying assessment instruments weighted agreement statistics are within 0.03.

EY | Building a better working world

EY exists to build a better working world, helping to create long-term value for clients, people and society and build trust in the capital markets.

Enabled by data and technology, diverse EY teams in over 150 countries provide trust through assurance and help clients grow, transform and operate.

Working across assurance, consulting, law, strategy, tax and transactions, EY teams ask better questions to find new answers for the complex issues facing our world today.

Our Report may be relied upon by the Department of Health and Aged Care for the supporting quality assurance of AN-ACC assessments through performing an inter-rater reliability analysis ("the Project") pursuant to the terms of our engagement agreement dated 11 June 2024. We disclaim all responsibility to any other party for any loss or liability that the other party may suffer or incur arising from or relating to or in any way connected with the contents of our report, the provision of our report to the other party or the reliance upon our report by the other party.

ey.com