AN-ACC Inter-rater Reliability Analysis

Department of Health, Disability and Ageing

Final report
10 February 2026

EY

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, Disability and Ageing ("**Client**") to perform an inter-rater reliability analysis ("**Project**"), in accordance with the contract signed on 11 June 2024 under Official Order number E24-209018 and deed of variation signed on 26 June 2025 under Official Order number E25-313508, including the General Terms and Conditions ("the **Engagement Agreement**").

The results of EY's work, including the assumptions and qualifications made in preparing EY's report dated 10 February 2026, ("**Report**") are set out in the 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, access to the Report by any party other than the Client (the Recipient) is made only on the following basis and in either accessing the Report or obtaining a copy of the Report the Recipient agrees to the following terms.

1. Subject to the provisions of this notice, 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.

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

3. 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.

4. No reliance may be placed upon the Report or any of its contents by any party other than the Client. Any party receiving a copy of the Report 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. No duty of care is owed by EY to any Recipient of the Report in respect of any use that the Recipient may make of the Report. EY disclaims all liability, and takes no responsibility, for any document issued by any other party in connection with the Project.

5. Subject to clause 6, the Report is confidential and must be maintained in the strictest confidence and must not be disclosed to any party for any purpose without the prior written consent of EY.

6. All tax advice, tax opinions, tax returns or advice relating to the tax treatment or tax structure of any transaction to which EY's services relate ("**Tax Advice**") is provided solely for the information and internal use of the Client and may not be relied upon by anyone else (other than tax authorities who may rely on the information provided to them) for any purpose without EY's prior written consent. If the Recipient wishes to disclose Tax Advice (or a portion or summary thereof) to any other third party, they shall first obtain the written consent of the Client before making such disclosure. The Recipient must also inform the third party that it cannot rely on the Tax Advice (or a portion or summary thereof) for any purpose whatsoever without EY's prior written consent.

7. The Recipient must not name EY in any document prepared by the Recipient that is to be lodged or filed by you with any regulator or is to be made publicly available ("Public Document") without EY's prior written consent, which may be granted at EY's absolute discretion. If the Recipient is required by applicable laws or regulations to name EY in any Public Document, the Recipient may do so provided it has given EY prior written notice, if such prior written notice is permitted by applicable laws or regulations.

8. The Recipient of the Report:

    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

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

9. In the event that the Recipient discloses the Report to a third party in breach of the terms of this notice ("**Breach Event**"), the Recipient 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 the Breach Event.

Glossary

Acronym	Definition
AFM	Australian Functional Measure
AN-ACC	Australian National Aged Care Classification
AO	Assessment Organisation
Braden	Braden Scale
BRUA	Behaviour Resource Utilisation Assessment
DEMMI-modified	De Morton Mobility Index-Modified
AFM Cognition	Cognition component of the Australian Functional Measure
AFM Motor	Motor component of the Australian Functional 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
IRR 4	Review of IRR of dual assessments conducted over August 2025 to September 2025
MMM	Modified Monash Model
NWAU	National Weighted Activity Unit
PA	Pure Agreement
QA	Quality Assurance
RAC	Residential Aged Care
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, Disability and Ageing (the Department) to identify and analyse trends, anomalies and patterns in Residential Aged Care (RAC) funding assessments which may be of concern in the assessment process and to support ongoing quality assurance of RAC funding assessments.

This report presents the results of inter-rater reliability (IRR) analysis conducted on 1,693 IRR assessments (i.e. a total of 3,386 individual assessments) occurring between 4 August and 30 September 2025. This represents the fourth IRR round ("IRR 4"), with the first round ("IRR 1") representing samples of assessments taken from November to December 2022, the second round ("IRR 2") representing samples of assessments taken from September to November 2023 and the third round ("IRR 3") representing samples of assessments taken from June to August 2024.

The IRR analysis considers the rate at which assessors agree on the outcomes of the assessment, either by assigning the resident to the same Australian National Aged Care Classification (AN-ACC) classification or assigning the same score on an instrument within the RAC funding 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. To assist in interpreting the below results, analysis techniques and terminology are detailed in Section 0.

▸ **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 implies 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 3 to IRR 4 across the AN-ACC classification and all assessment instruments, following from a generally increasing trend in agreement rates observed between IRR 2 and IRR 3. The greatest improvements in IRR 4 were observed for the Behaviour Resource Utilisation Assessment (BRUA), followed by the Resource Utilisation Groups—Activities of Daily Living (RUG-ADL) and the De Morton Mobility Index-Modified (DEMMI-modified).

▸ **Two assessment organisations (AOs) demonstrated agreement rates that were noticeably higher than other AOs, while one AO demonstrated noticeably lower rates of agreement** – AO 4 and AO 10 both experienced high rates of agreement in IRR 4, particularly on the AN-ACC classification where both AOs demonstrated perfect pure agreement. In contrast, AO 7 continued to have the lowest agreement rates across the AOs, specifically on the AN-ACC classification, Australian Functional Measure Cognitive items (AFM Cognition), Australian Functional Measure Motor items (AFM Motor) and BRUA.

▸ **Assessor pairs consisting of two Category B[1] assessors experienced the lowest agreement rates in IRR 4** – When segmenting by assessor category, agreement rates for pairs of only

---

[1] Assessor categories are assigned by the number of AN-ACC funding assessments completed by an assessor. Assessors who have completed the greatest number of assessments are classified as Category A, followed by Category B and Category C. An assessor pair conducting an IRR assessment may be classified as, for example, A/B, where the pair consists of a Category A and Category B assessor.

Category B assessors were generally lower than any assessor pair consisting of at least one Category A or C assessor across the AN-ACC classification and all assessment instruments.

1.3   Limitations

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

▸   This is a data analysis exercise to assess the level of consistency in the application of RAC funding assessments by 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 classification data to identify trends, anomalies and patterns which may be of concern in the assessment process and to support ongoing quality assurance of RAC funding assessments.

This report presents the results of IRR analysis conducted on 1,693 IRR assessments occurring from 4 August to 30 September 2025[2]. As in previous IRR reviews (in March 2022, April 2023, April 2024 and August 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 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 4 August 2025 to 30 September 2025.
Other reference data sets		
2	All RAC funding assessments	Excel data files containing extracts of RAC funding assessment data from April 2021 to 3 October 2025.
3	Provider and facility details by NAPS ID	Excel file containing details of active and closed facilities and providers as of 16 October 2025.
4	Matching assessor IDs	Excel file mapping past and updated assessor IDs as of 24 October 2025.

---

[2] 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.

2.3   Analysis

This section details the analysis techniques employed in this report. We have assessed inter-rater reliability of the August to September 2025 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.

- ▸ **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).

- ▸ **Correlation** – The Pearson correlation coefficient calculated from pairs of National Weighted Activity Units (NWAUs) or total scores on each instrument.

In addition to agreement between assessors on the final AN-ACC classification and the underlying assessment instruments in the RAC funding 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-modified 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 RAC funding assessments.

Figure 1: AN-ACC decision tree with assessment counts and NWAUs as at October 2024[3]
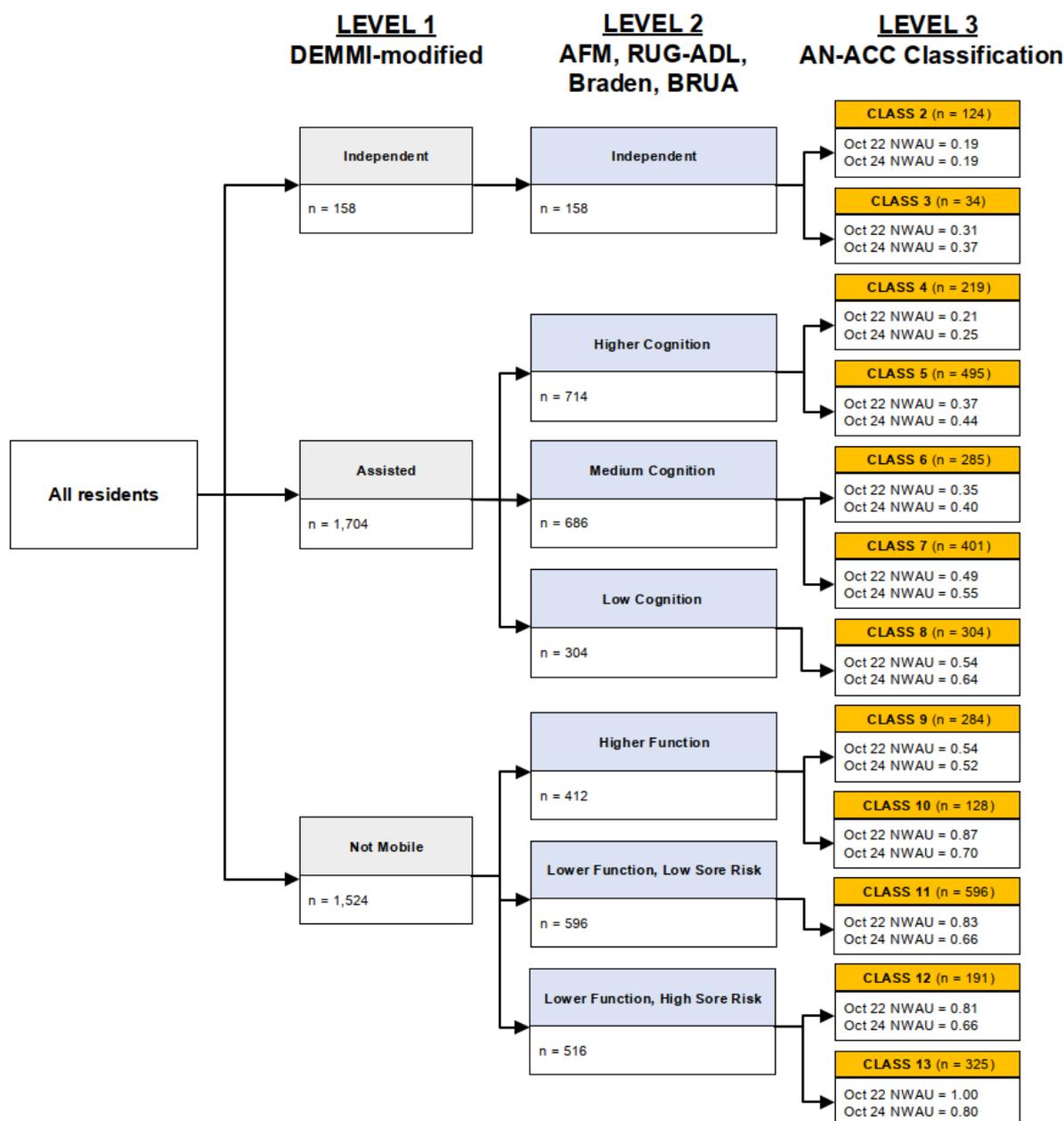
Decision tree source: Eagar K, McNamee J, Gordon R et al. (2019) The Australian National Aged Care Classification (AN-ACC). The Resource Utilisation and Classification Study: Report 1. Australian Health Services Research Institute, University of Wollongong. ISBN: 978-1-74128-295-5

NWAU (October 2022 basis) source: Australian Government Department of Health, Disability and Ageing, Schedule of subsidies and supplements for aged care, March 2023.

NWAU (October 2024 basis) source: Australian Government Department of Health, Disability and Ageing, Schedule of subsidies and supplements for aged care, October 2024.

---

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

Department of Health, Disability and Ageing – RAC funding assessment Anomaly Analysis – Inter-rater reliability – Final report

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 assessor outcomes are within one class or score of each other (i.e., treating an adjacent score or class as an agreement). This allows for a level of "partial agreement".[4]

3. **Fleiss' kappa (exact agreement)** – 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 (weighted agreement)** – 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. This measure uses weights based on NWAUs for the AN-ACC classification. The selected weightings used in this report are discussed further in Appendix A.1.1.

5. **Pearson's correlation coefficient** – Correlation measures the general direction of agreement only, for example the extent to which assessors scored towards the higher end of the range of scores, and should be interpreted with caution since correlation is not a reliable measure of agreement.

6. **Root Mean Squared Error (RMSE)** – A measure used to quantify the typical range of differences in funding outcomes due to disagreement on the AN-ACC classification in IRR assessments.

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.

---

[4] Same or adjacent class/score agreement on the AN-ACC classification assumes that adjacent classes have the closest classification likelihood, which in some cases may not accurately reflect realistic variations in outcomes and care needs. This therefore limits the reliability of this measure for AN-ACC classification. Kappa statistics remain the most robust measure of agreement.

Table 2: Kappa interpretation

Kappa Value	Interpretation
$-1.00 \leq kappa \leq 0.00$	No agreement
$0.00 < kappa \leq 0.20$	Poor agreement
$0.20 < kappa \leq 0.40$	Fair agreement
$0.40 < kappa \leq 0.75$	Moderate agreement
$0.75 < kappa \leq 1.00$	Excellent agreement

i. *Kappa bands were chosen to align with the initial reliability assessment conducted in 2022[5] as well as subsequent reviews and are considered to be consistent with academic literature.[6,7]*

ii. *Note that throughout this report, 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.*

2.3.2 IRR assessment process

The IRR process checks RAC funding assessments to provide assurance to RAC providers and the sector that RAC funding assessments conducted by independent assessors are consistent regardless of the assessor. The portion of IRR assessments conducted during the set period is taken by the Department to be representative of the broader assessment distribution as outlined in Table 3 in the next section.

▸ Assessor pairs for each resident were assigned by the AO, with both assessors needing to have an unrestricted status and a good mix of cross disciplines used.

▸ Each AO was given a minimum number of assessments that they were required to undertake during the period.

▸ Residents were assigned by name or ID to the assessors prior to any contact regarding the IRR assessment. Both assessors undertaking the assessment were required to spend equal time at the facility, interact with the resident and staff together, where appropriate, and review the same medical records and care documentation.

---

[5] Taylor Fry, Quality Assurance of the Australian National Aged Care Classification – Final Report, July 2022

[6] Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. Biometrics, 33(1), 159-174. https://doi.org/10.2307/2529310

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

3. Results

From 4 August 2025 to 30 September 2025, 3,392 raw assessments were completed for the purposes of the IRR analysis. 3,386 (99.8%) of the assessments were submitted on the same day by both assessors, indicating good compliance with IRR protocols. However, 3 IRR assessments were excluded from the analysis as the assessment pairs either were uploaded three or more days apart from each other or appeared erroneous (see Appendix A.1 for details). Therefore, 3,386 assessments were included in the final IRR analysis, performed by 292 assessors across 490 facilities and 1,674 unique residents. Each assessor completed approximately 12 IRR assessments on average.

Distribution of IRR assessments

Overall, the distribution of IRR assessments by remoteness, facility type and facility size aligns reasonably closely to that of the previous 6 months of RAC funding assessments. However, there has been increased skewness towards metropolitan facilities resulting from larger absolute volumes of assessments conducted in Modified Monash Model (MMM) category 1-2 and slightly lower volumes of IRR assessments in MMM 3-5 compared to the previous IRR period. In addition, there were no IRR assessments in this round for MMM 6-7; however, this is broadly consistent with the low proportion of RAC funding assessments across the previous 6 months..

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

Remoteness	Type	No. Beds	6-Month Assessment Volume*	Proportion	IRR Assessments	Proportion
MMM 1	For profit	0 to 59	1,845	1.9%	43	2.5%
		60 to 119	15,565	16.2%	277	16.4%
		120 to 179	14,106	14.6%	330	19.5%
		180+	2,651	2.8%	64	3.8%
	Not for profit	0 to 59	3,813	4.0%	40	2.4%
		60 to 119	16,213	16.8%	253	14.9%
		120 to 179	12,335	12.8%	294	17.4%
		180+	2,546	2.6%	40	2.4%
	Government	Any	298	0.3%	8	0.5%
MMM 2	For profit	Any	3,006	3.1%	45	2.7%
	Not for profit	Any	4,896	5.1%	132	7.8%
	Government	Any	578	0.6%	12	0.7%
MMM 3	Any	Any	8,228	8.5%	79	4.7%
Regional (MMM 4-5)	Any	Any	9,889	10.3%	76	4.5%
Remote (MMM 6-7)	Any	Any	364	0.1%	0	0.0%
Total			96,333	100.0%	1,693	100.0%

*   *Calculated based on total RAC funding assessments completed 1 February 2025 to 31 July 2025.*

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

Figure 2: AN-ACC classification distribution of assessments: IRR assessments vs RAC funding assessments (1 Feb 25 – 31 Jul 25)
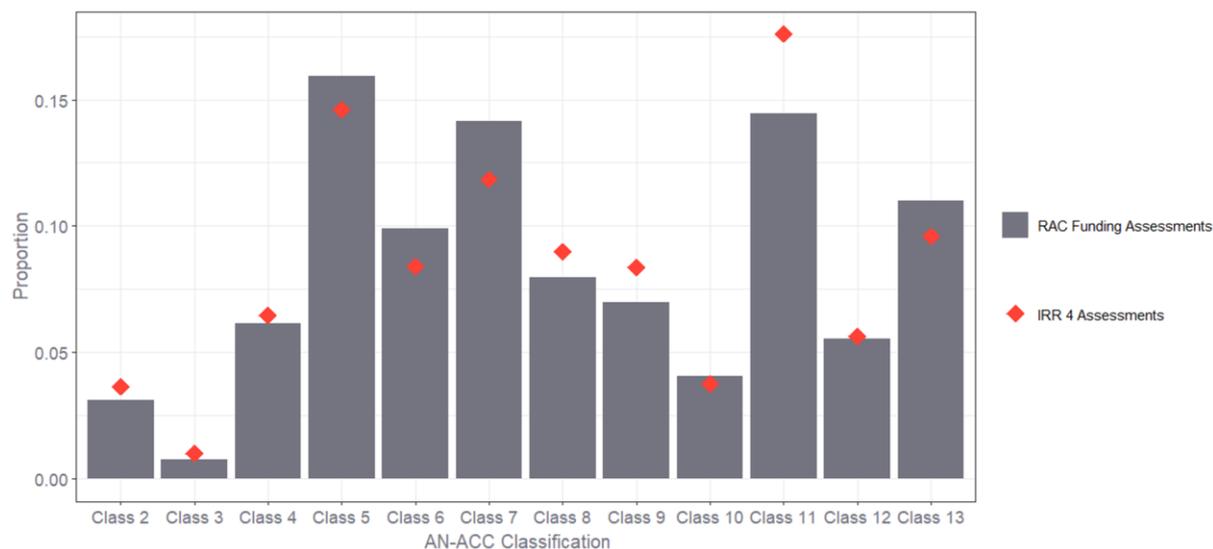
3.1   Overall agreement results

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.

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

Level	Category	Fleiss' Kappa		Agreement Type	
		Exact (unweighted)	Tolerance for partial agreement (weighted)	Pure Agreement	Same or Adjacent Class/Score[8]
1	Mobility Branch	0.96	0.97	97.8%	100.0%
2	Mobility: Cognition/ Function/Pressure Sores	0.89	0.95	90.7%	98.6%
3	AN-ACC Class	0.86	0.92	87.5%	94.7%

Key observations from Table 4 include:

- ▸ Assessors assigned residents into the same mobility category in 97.8% of assessments, the same Level 2 category in 90.7% of assessments and the same final AN-ACC classification in 87.5% 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%, 98.6% and 94.7% respectively. In the case of Level 1 mobility, this

---

[8] AN-ACC classifications were not ordered by resource utilisation/funding for simplicity. For example, using October 2024 NWAUs, Class 3 (NWAU = 0.37) would be considered "adjacent" to Class 2 (NWAU = 0.19), while Class 4 (NWAU = 0.25) would not be considered "adjacent" to Class 2 even though the NWAU difference between Class 2 and 4 is smaller. This therefore limits the reliability of this measure for the AN-ACC classification. Kappa statistics remain the most robust measure of agreement.

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.

▸ 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.

▸ The weighted Fleiss' kappa statistics at Levels 2 and 3 of the decision tree were higher than the unweighted Fleiss' kappa statistics. This 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).

Table 5 below contains the average raw score differences between assessors on final AN-ACC classification (by NWAU) and for each underlying assessment instrument in the RAC funding 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, 0.80]	11	0.015	0.013
DEMMI-modified	[0,16]	17	0.233	0.014
AFM Cognition	[5, 35]	31	0.871	0.028
AFM Motor	[12, 84]	73	0.944	0.013
RUG-ADL	[4, 18]	15	0.210	0.014
Braden	[6, 23]	18	0.366	0.020
BRUA	[5, 20]	16	0.320	0.020

    i.   *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 1% 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
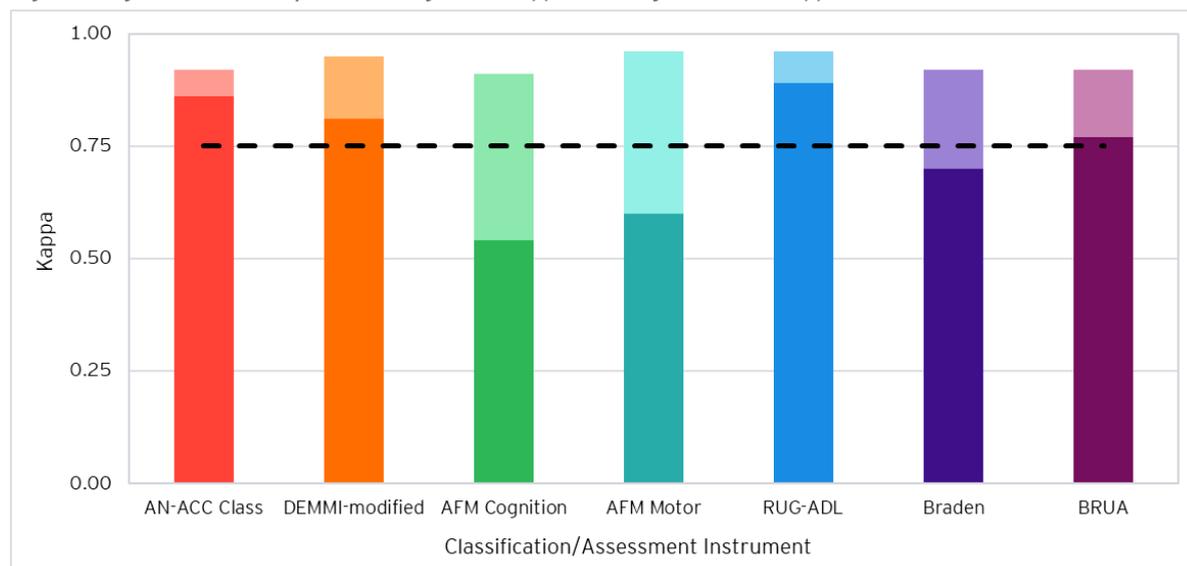
i.   Excellent agreement is represented by the dashed line (0.75).
ii. Exact agreement (unweighted kappa) is shown in darker shading for each classification/assessment instrument.
iii. Agreement with tolerance for partial agreement (weighted kappa) above exact agreement is shown in lighter shading.
iv. The weighted Fleiss' kappa for the AN-ACC classification uses the October 2024 NWAU basis.

Table 6: Agreement rates by metric

Category	Fleiss' Kappa		Agreement Type		Pearson's Correlation
	Exact (unweighted)	Tolerance for partial agreement (weighted)	Pure Agreement	Same or Adjacent Class/Score	
AN-ACC Class	0.86	0.92	87.5%	94.7%	0.95
DEMMI-modified	0.81	0.95	82.6%	95.7%	0.99
AFM Cognition	0.54	0.91	56.5%	79.0%	0.98
AFM Motor	0.60	0.96	61.8%	79.9%	0.99
RUG-ADL	0.89	0.96	90.0%	95.5%	0.99
Braden	0.70	0.92	72.5%	93.0%	0.98
BRUA	0.77	0.92	79.1%	93.9%	0.97

i.   The weighted Fleiss' kappa for the AN-ACC classification uses the October 2024 NWAU basis.

Key observations from Figure 3 and Table 6 include:

‣   Excellent agreement between assessors was observed for all instruments when measured with a tolerance for partial agreement (weighted Fleiss' kappa).

‣   When examining exact agreement (unweighted Fleiss' kappa), the AN-ACC classification, DEMMI-modified, RUG-ADL and BRUA had excellent exact agreement, while the AFM Cognition, AFM Motor and Braden were observed to have moderate agreement.

    o   For the AFM Cognition and AFM Motor, this was not unexpected 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). However, these instruments showed the greatest improvements with the weighted Fleiss' kappa. This indicates that although disagreements on these instruments were more frequent, they tended to reflect small differences in outcomes.

- o The Braden has a similar range of scores to the DEMMI-modified, RUG-ADL and BRUA, but was observed to have a lower rate of exact agreement.

▸ Excellent agreement between assessors was observed for all instruments when treating same or adjacent scores (instrument scores within 1) as an agreement.

▸ 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 Comparison of agreement levels across IRR studies

The four IRR periods represent samples of assessments taken over four discrete periods of time: November to December 2022 ("IRR 1"), September to November 2023 ("IRR 2"), June to August 2024 ("IRR 3") and August to September 2025 ("IRR 4"). Results are potentially impacted by seasonality, influences from the nature of the assessments included within the samples (such as assessment type, AO 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.

Considerations in comparing results across IRR studies

When comparing the weighted Fleiss' kappa statistics for the AN-ACC classification between prior IRR periods (IRR 1 to IRR 3) and IRR 4, it is important that the comparison is conducted on a like-for-like basis. As the weighted Fleiss' kappa statistic for AN-ACC classifications is calculated using weights based on NWAUs (see Appendix A.1.1), the results differ when using the October 2022 NWAU basis (applicable for IRR 1 to IRR 3) and the October 2024 NWAU basis (applicable for IRR 4). For IRR 4, this metric has been calculated on both bases, whereby the weighted Fleiss' kappa calculated using the October 2022 NWAU basis should be used to compare with prior periods.
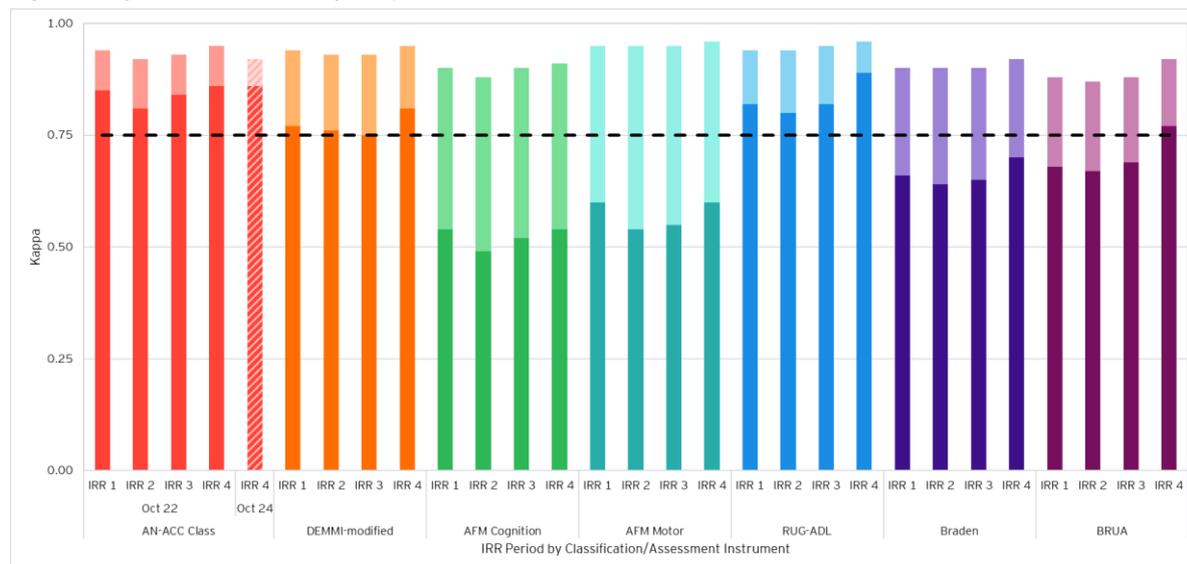
Under the October 2022 NWAU basis, weighted agreement for the AN-ACC classification in IRR 4 is higher than that on the October 2024 NWAU basis. This is primarily driven by a reduction in the overall spread of NWAUs across AN-ACC classifications on the October 2024 basis. When assessing partial agreement between assessors, the narrower range of NWAUs across AN-ACC classifications means that they can be considered more "adjacent" to each other. This increases the expected level of partial agreement between assessors by chance to a greater extent than the corresponding increase in observed partial agreement. This results in reduced weighted agreement (after adjusting for chance) on the October 2024 basis. This is shown in the results presented throughout the remainder of this report. Refer to Appendix A.1 for more technical detail.

Results

Overall, there was an increase in agreement observed between the IRR 3 and IRR 4 assessments. This is reasonable given the underlying expectation that inter-rater reliability should increase over time as the system and assessor workforce matures. Changes in overall agreement can be further explained by changes within subgroups of assessors.

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.*
iv.   *Agreement for the AN-ACC classification using the October 2022 NWAU basis is shown in the solid-coloured columns, enabling a like-for-like comparison between IRR periods.*
v.   *Agreement for the AN-ACC classification using the October 2024 NWAU basis in IRR 4 is shown in the striped column. The NWAU basis only affects the calculation of weighted agreement (weighted Fleiss' kappa).*

Key observations from Figure 4 include:

▸   Excellent agreement between assessors was observed for all instruments when measured with a tolerance for partial agreement (weighted Fleiss' kappa) across all IRR periods.

▸   Agreement rates in IRR 4 were generally the highest across the four IRR periods. Exact agreement (unweighted Fleiss' kappa) increased by up to 0.08 for all instruments from IRR 3 to IRR 4, reflecting the largest improvements since IRR 1. Similarly, weighted agreement rates have improved to a lower extent for all assessment instruments.

▸   Although exact agreement for the AN-ACC classification has slightly increased in IRR 4 in comparison to IRR 3, weighted agreement has slightly decreased when the IRR 4 agreement is calculated based on the October 2024 NWAU basis. However, when adjusting to the October 2022 NWAU basis (which was applicable from IRR 1 to IRR 3), the weighted agreement has instead increased by 0.02 from IRR 3 to IRR 4, as shown by the grey dot.

▸   The greatest improvement in exact and weighted agreement rates in IRR 4 compared to IRR 3 is observed for the BRUA, with a 0.08 and 0.04 increase respectively. The RUG-ADL and DEMMI-modified have also greatly improved, with 0.07 and 0.06 increases in exact agreement respectively.

▸   Exact agreement rates for the AFM Cognition and AFM Motor in IRR 4 have returned to IRR 1 rates.

Adjusting for change in mix of AOs and states in IRR 4

Comparisons of results between different IRR periods are limited by changes in the makeup of the underlying IRR assessment samples. To explore the sensitivity of agreement rates to sample representation and adjust for the impact of changing sample bias, we have produced 're-weighted' results which apply the IRR 4 assessment sample's mix of AO and state representation across all IRR periods. This enables a more 'like-for-like' comparison of agreement rates between IRR periods.

This approach has been chosen because IRR assessments are scheduled at the AO level and are limited by geographic restrictions in practice. Additionally, results from previous IRR studies have shown substantial variation in agreement by AO and state. We note however that this analysis is intended to provide indicative insights only.

Further, following the implementation of the Single Assessment System in December 2024, the mix of assessments between AOs changed significantly in the IRR 4 period compared to previous IRR periods. To maintain comparability, the set of AOs considered when re-weighting agreement rates for each IRR period has been restricted to only those that conducted assessments across all periods. This excludes 263 IRR assessments in IRR 4, or approximately 16% of total IRR assessments.

Table 7 compares re-weighted pure agreement rates across each IRR period. Due to the exclusion of assessments and AOs described above, these re-weighted results do not align to others presented in this report and should be interpreted with caution.

Table 7: IRRs 1 to 4 re-weighted agreement rates by AN-ACC classification/assessment instrument

Category	Pure Agreement (re-weighted)			
	IRR 1	IRR 2	IRR 3	IRR 4
AN-ACC Class	86.3%	82.7%	86.4%	87.1%

i.   *Agreement rates are weighted by AO and state representation in IRR 4. Only AOs which conducted assessments across all IRR periods are included.*

Key observations from Table 7 include:

▸   The improvement in pure agreement from IRR 3 to IRR 4 is less pronounced following re-weighting of the previous IRR agreement rates.

▸   Due to the generally large reduction in IRR 1 agreement when re-weighted, the reduction in pure 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 AO and state from IRR 1 to IRR 4. For instance, increases in the proportions of AO-state combinations generally occurred where agreement rates during IRR 1 and IRR 2 were relatively lower. Meanwhile, increases in the proportions of AO-state combinations generally occurred where agreement rates during IRR 3 were relatively higher.

Root Mean Squared Error

The Root Mean Squared Error (RMSE) has been calculated for each IRR period as an indicator of the variability (or the typical range of error) in funding due to disagreements on the AN-ACC classification. These results are shown in Table 8 and presented in both NWAUs and dollars.

To allow for a consistent basis, the set of NWAUs relevant as at both October 2024 and October 2022 were applied separately across all IRR rounds, with the AN-ACC price of $282.44 as at October 2024 applied to both sets.

The RMSE can be interpreted similarly to a standard deviation and is calculated as the square root of the average squared differences in NWAU from IRR assessments, including those where assessors agreed (i.e. a difference of zero).

Table 8: RMSEs for IRR 1 to IRR 4 by NWAU basis

NWAU basis	RMSE (NWAUs)				RMSE ($)			
	IRR 1	IRR 2	IRR 3	IRR 4	IRR 1	IRR 2	IRR 3	IRR 4
October 2024	0.0465	0.0536	0.0499	0.0477	$13.13	$15.14	$14.09	$13.47
October 2022	0.0543	0.0680	0.0619	0.0527	$15.34	$19.21	$17.48	$14.88

i.   The RMSE metrics presented do not represent the average funding impact of a single disagreement. Rather, they provide a comparative view of the overall level and variability of funding impacts due to assessor disagreement.
ii. The AN-ACC price used for RMSE ($) was $282.44, applicable as at October 2024.

Key observations from Table 8 include:

▸ On the October 2024 basis, IRR 1 results implied the lowest impact on funding resulting from assessor disagreement across the IRR periods, followed by relatively high funding impacts in IRR 2 and IRR 3. This has since reduced in IRR 4, almost returning to the IRR 1 level.

▸ Meanwhile, on the October 2022 basis, IRR 4 results implied the lowest funding impact, followed by IRR 1. This difference in trend to the October 2024 basis purely reflects variations to the NWAU differences associated with each possible disagreement on the AN-ACC classification. The extent of the NWAU differences has reduced under the more recent October 2024 basis, resulting in a narrower $2.01 range in implied funding impacts between IRR rounds compared to $3.32 under the October 2022 basis.

▸ Although the differences in trend between the NWAU bases may suggest secondary drivers in relation to the mix of disagreements, this has not been explored in this report, with the primary purpose focused on reflecting overall funding impacts due to disagreements.

3.2   Segmentation by AO

This section examines the agreement of assessors, segmented by AO. It is important to note that all IRR assessments taking place between August to September 2025 and analysed in this report were performed by pairs of assessors from the same AO. This practice supported the consistent scheduling of IRR assessments to allow each pair of assessors to observe the resident under the same conditions. For this reason, caution should be applied when comparing IRR statistics between AOs.

In addition, AOs were de-identified for reporting purposes, with each AO randomly assigned a number. While the AO numbering is consistent throughout this report, it is not consistent with the numbering within previous IRR reports.

Following the implementation of the Single Assessment System in December 2024, the cohort of AOs conducting assessments in IRR 4 has changed significantly compared to previous IRR rounds. There are now ten independent AOs currently conducting RAC funding assessments on behalf of the Department, compared to six AOs during the previous IRR rounds.

The Fleiss' kappa and the weighted Fleiss' kappa test statistics for final AN-ACC classification and each underlying assessment instrument by AO are shown in Figure 5, Figure 6 and Figure 7 below, with values of each statistic presented in Appendix B.1.

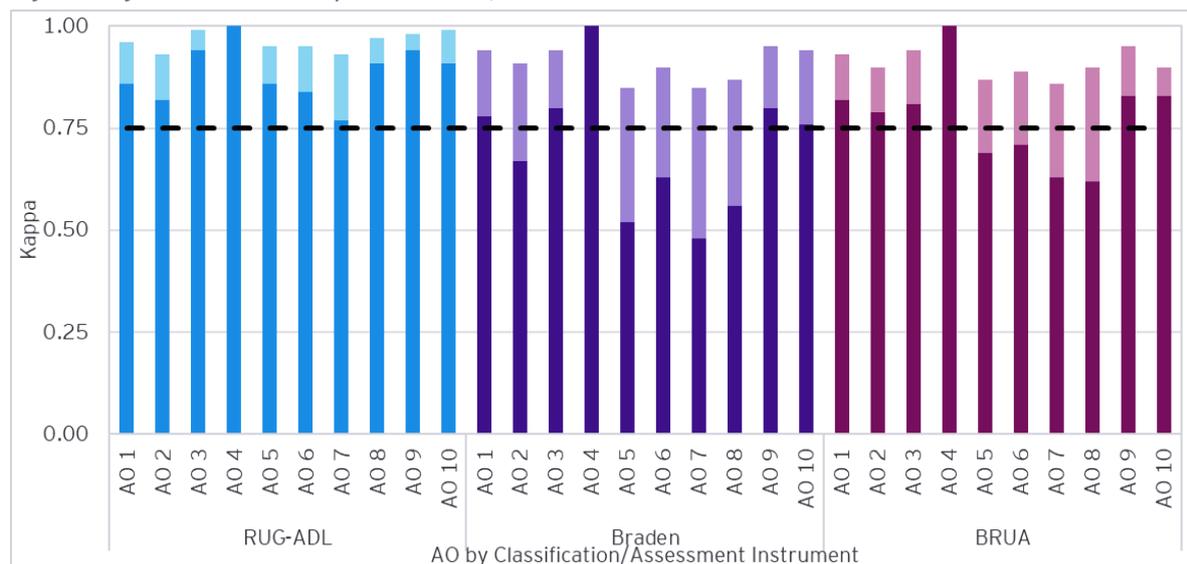Figure 5: Agreement statistics by AO (AN-ACC Class and DEMMI-modified)
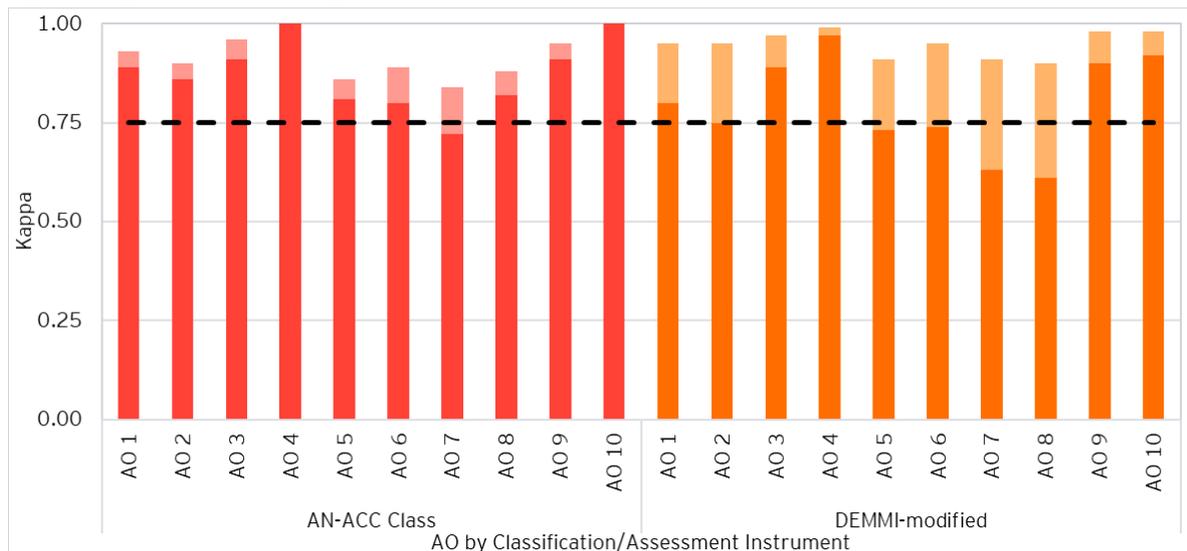
i.   *Excellent agreement is represented by the dashed line (0.75).*
ii. *Exact agreement (unweighted kappa) is shown in darker shading for each classification/assessment instrument.*
iii. *Agreement with tolerance for partial agreement (weighted kappa) above exact agreement is shown in lighter shading.*
iv. *The weighted Fleiss' kappa for the AN-ACC classification uses the October 2024 NWAU basis.*

Figure 6: Agreement statistics by AO (AFM Cognition and AFM Motor)
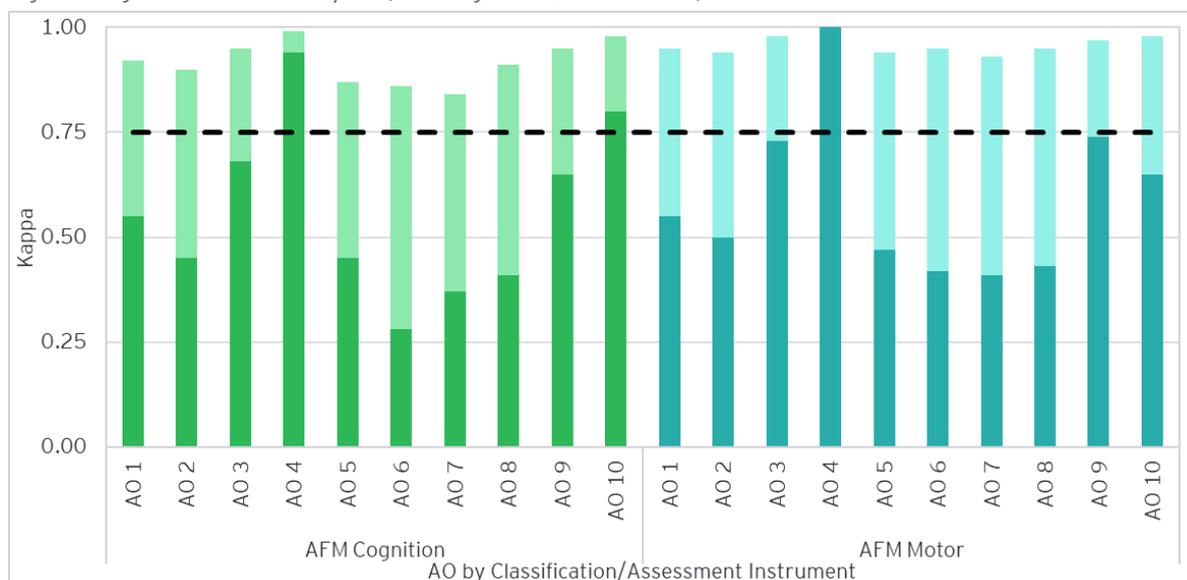
i.   *Excellent agreement is represented by the dashed line (0.75).*
ii. *Exact agreement (unweighted kappa) is shown in darker shading for each classification/assessment instrument.*
iii. *Agreement with tolerance for partial agreement (weighted kappa) above exact agreement is shown in lighter shading.*

Figure 7: Agreement statistics by AO (RUG-ADL, Braden 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, Figure 6 and Figure 7 include:

▸ Excellent agreement between assessors of all AOs was observed for all instruments when measured with a tolerance for partial agreement.

▸ AO 4 had the highest rates of agreement, with weighted agreement above 0.99 across all instruments. Perfect agreement of 1.00 was observed for the AN-ACC classification, AFM Motor, RUG-ADL, Braden and BRUA, meaning that assessor scores were exactly aligned across all IRR assessments performed.

▸ AO 10 also had amongst the highest rates of agreements across the AOs, with perfect exact agreement on the AN-ACC classification and weighted agreement on the RUG-ADL, DEMMI-modified, AFM Cognition and AFM Motor above 0.98.

▸ AO 3 and AO 9, which had amongst the highest assessor and assessment volumes of the AOs, generally had high rates of agreement, with weighted agreement greater than 0.94 across all instruments.

▸ AO 1 had weighted agreement similar to, but slightly lower than, AO 3 and AO 9 (within 0.03) across all instruments.

▸ AO 2 and AO 8 generally had mid-range agreement relative to the other AOs.

  o AO 2 had weighted agreement slightly lower than AO 3 and AO 9 (within 0.05) across the DEMMI-modified, AFM Cognition, AFM Motor, Braden and BRUA, but had poorer agreement on the AN-ACC classification and RUG-ADL.

  o AO 8 had weighted agreement slightly lower than AO 3 and AO 9 (within 0.05) across the AFM Cognition, AFM Motor, RUG-ADL and BRUA, but had poorer agreement on the AN-ACC classification, DEMMI-modified and Braden.
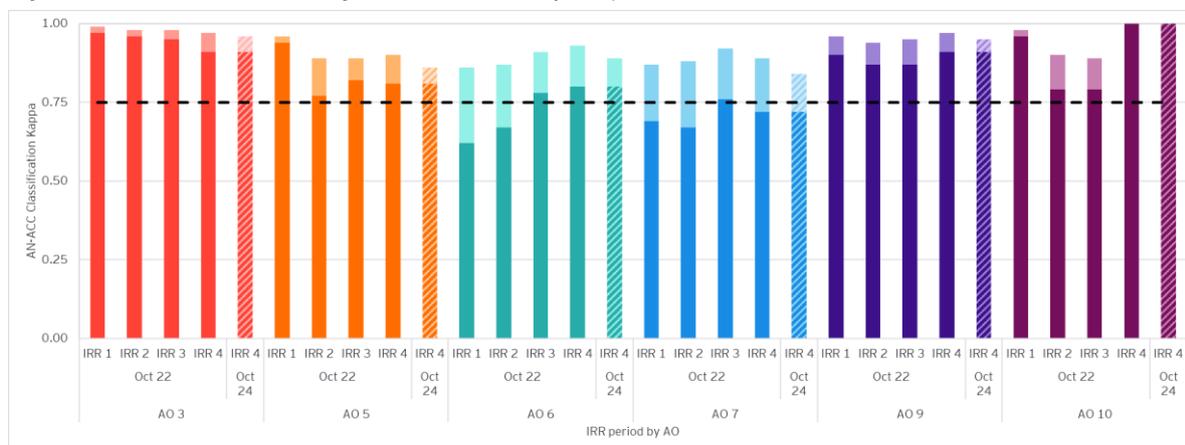
- AO 5 and AO 6 were generally observed to have agreement levels lower than the other AOs.

  - AO 5 had relatively low pure agreement and weighted agreement across most instruments compared to other AOs, particularly for the Braden.

  - Although AO 6 had mid-range weighted agreement in comparison to other AOs, fair pure agreement and exact agreement (20% to 40%) were observed on the AFM Cognition, which was the lowest across the AOs and assessment instruments.

- Whilst having excellent agreement when measured with a tolerance for partial agreement across all instruments, AO 7 had the lowest agreement levels across the AOs, with the lowest weighted agreement on the AN-ACC classification, AFM Cognition, AFM Motor and BRUA. Fair pure agreement and exact agreement (20% to 40%) were observed on the AFM Cognition.

We note that the absence of cross-AO IRR assessments could mean that agreement results between AOs may not be directly comparable.

3.2.1   Comparison of agreement levels across IRR studies

Figure 8 shows the comparison in Fleiss' kappa statistics for the AN-ACC classification across the IRR periods. We note that these changes may have also been partially driven by the changes in NWAU as at October 2024, explained in further detail in Section 3.1.1.

Figure 8: AN-ACC classification agreement statistics by IRR period and AO

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.
iv. Agreement for the AN-ACC classification using the October 2022 NWAU basis is shown in the solid-coloured columns, enabling a like-for-like comparison between IRR periods.
v.   Agreement for the AN-ACC classification using the October 2024 NWAU basis in IRR 4 is shown in the striped column. The NWAU basis only affects the calculation of weighted agreement (weighted Fleiss' kappa).

The key observations from Figure 8 include:

- Excellent agreement measured with a tolerance for partial agreement was observed at the AN-ACC classification level for all AOs in each IRR period. Further, excellent exact agreement was observed for AOs 3, 9 and 10 across each IRR period.

- Exact and weighted agreement for AO 3 has slightly decreased in IRR 4 after having near perfect weighted agreement above other AOs across IRR 1 to 3. When comparing weighted agreement on a like-for-like basis using the October 2022 NWAUs, only a slight decrease of 0.01 was observed between IRR 3 and IRR 4.

- o In previous IRR periods, exact agreement for AO 3 was generally higher than weighted agreement across other AOs.

▸ For AO 5, exact and weighted agreement both decreased in IRR 4 compared to IRR 3. However, after adjusting to the October 2022 NWAU basis, a slight increase in weighted agreement is observed. Overall, weighted agreement appears to be stable between IRR 2 and IRR 4, following a substantial decrease in agreement between IRR 1 and IRR 2.

▸ Exact agreement has increased for AO 6 in IRR 4, resulting in the highest exact agreement for AO 6 across all previous IRR periods. While weighted agreement decreased on the October 2024 NWAU basis, it increased above the levels observed in previous IRR periods after adjusting to the October 2022 NWAU basis.

▸ Exact and weighted agreement for AO 7 reduced in IRR 4 compared to IRR 3. However, exact agreement and weighted agreement (after adjusting to the October 2022 NWAU basis) both remain above the levels observed in IRRs 1 and 2.

▸ AO 9 has continued to have very high weighted agreement in IRR 4, with an increase in exact agreement to slightly over that of IRR 1 following reduced agreement during IRR 2 and 3.

▸ Perfect exact agreement was observed for AO 10 in IRR 4, following a decreased level of agreement over IRR 2 and IRR 3 compared to IRR 1. The low volume of assessments conducted in IRR 4 compared to IRRs 1 to 3 (27 versus over 59 respectively) should be considered when interpreting these results.

Contributions of specific AO and state combinations to changes in agreement from IRR 3 to IRR 4

As indicated in Section 3.1.1, comparisons of results between different IRR periods are limited by changes in the makeup of the underlying IRR assessment samples. To explore the sensitivity of agreement rates to sample representation and attempt to 'correct' for the impact of changing sample bias, we have produced 're-weighted' results which provide consistency in AO and state representation using the IRR 4 assessment sample as the basis. As the cohort of AOs conducting assessments changed in IRR 4 following the implementation of the Single Assessment System, we have restricted the mix of AOs considered to those which conducted assessments across all IRR periods.

Figure 9 below compares the AN-ACC classification pure agreement rate for IRR 3 and IRR 4, and the change in percentage points attributable to re-weighting for the IRR 4 AO and state mix, and to each AO and state pair. The change for each AO and state pair accounts for both the observed change in agreement rates and the representation amongst IRR assessments.
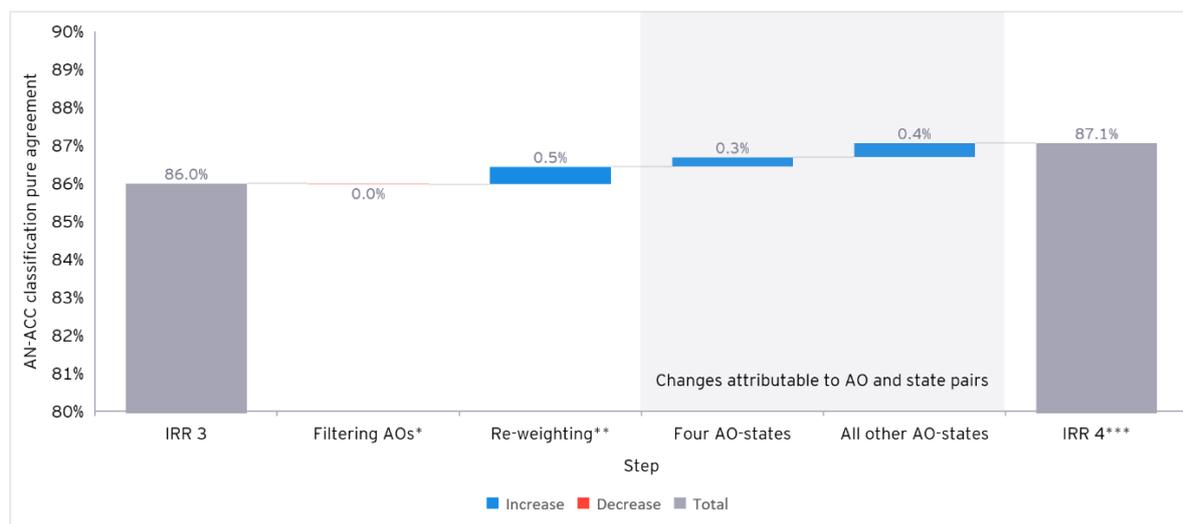
*   *Only AOs which conducted assessments across all IRR periods are included.*
** *Agreement rates are weighted by AO and state representation in IRR 4.*
*** *Due to the filtering of AOs and re-weighting, this IRR 4 pure agreement rate does not align to others presented in this report. See Section 3.1.1 for details.*

Key observations from Figure 9 include:

▸   Overall, there was 1.1 percentage point increase in the AN-ACC classification pure agreement between IRR 3 and IRR 4. Changes in the AO and state sample mix alone would have led to a 0.5 percentage point increase in pure agreement, as shown in the "Re-weighting" component of Figure 9.

▸   Agreement rates at the AO and state pair level were a positive source of change on the AN-ACC classification, contributing +0.7 percentage points to the overall change and driving the overall increase in the exact Fleiss' kappa agreement observed. Of this amount, four specific AO and state pairs contributed +0.3 percentage points.

3.3   Segmentation by assessor category

This section examines the agreement of assessors by their category. Assessor categories are assigned by the number of assessments completed by an assessor between 1 October 2022 and 30 September 2025. Assessors were segmented into three categories; those with over 350 assessments were classified as Category A, those with over 100 and less than or equal to 350 assessments as Category B, and those with 100 or less assessments were classified as Category C. Of the 292 assessors involved in IRR assessments, 5 (1.7%) had 0 payment impacting RAC funding assessments performed since 1 October 2022. These assessors were classified as part of Category C.

Table 9 shows the distribution of assessors and assessments by assessor category, while Table 10 shows the distribution of IRR assessments for each combination of assessor category.

Table 9: Distribution of assessments by assessor category

Assessor Category	Number of Assessments Completed	Number of Assessors	Number of Assessments*	Proportion of Assessments
A	(350, Max]	183	2,620	77.4%
B	(100, 350]	71	521	15.4%
C	[0, 100]	38	245	7.2%
Total		292	3,386	100.0%

* *Each IRR assessment consists of a "payment impacting" assessment performed alongside a "no-payment impact" quality assurance assessment. The number of assessments includes both "payment impacting" and "no-payment impacting" quality assurance assessments.*

Table 10: Distribution of IRR assessments by assessor category combination

Assessor Category Combination		Mapping	Number of IRR Assessments	Proportion of IRR Assessments
A	A	A/A	1,125	66.5%
A	B	A/B	255	15.1%
A	C	A/C	115	6.8%
B	B	B/B	98	5.8%
B	C	B/C	70	4.1%
C	C	C/C	30	1.8%
Total			1,693	100.0%

i. *Proportions may not add up exactly due to rounding.*

The most common combination of assessor category in IRR assessments was A/A, with 1,125 (66.5%) IRR assessments completed by this combination.

The Fleiss' kappa and the weighted Fleiss' kappa test statistics for the AN-ACC classification and each underlying assessment instrument by assessor category are shown in Figure 10 and Figure 11 below, with values of each statistic presented in Appendix B.3.

Figure 10: Agreement statistics by assessor category (AN-ACC Class, DEMMI-modified and AFM Cognition)
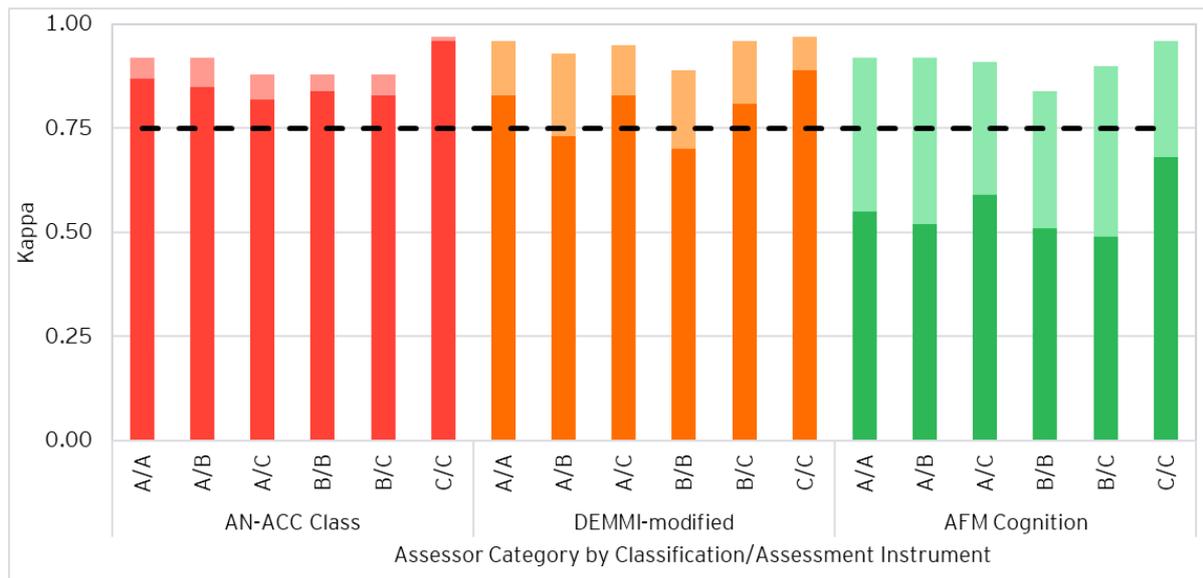
i. Excellent agreement is represented by the dashed line (0.75).
ii. Exact agreement (unweighted kappa) is shown in darker shading for each classification/assessment instrument.
iii. Agreement with tolerance for partial agreement (weighted kappa) above exact agreement is shown in lighter shading.
iv. The weighted Fleiss' kappa for the AN-ACC classification uses the October 2024 NWAU basis.

Figure 11: Agreement statistics by assessor category (AFM Motor, RUG-ADL, Braden and BRUA)
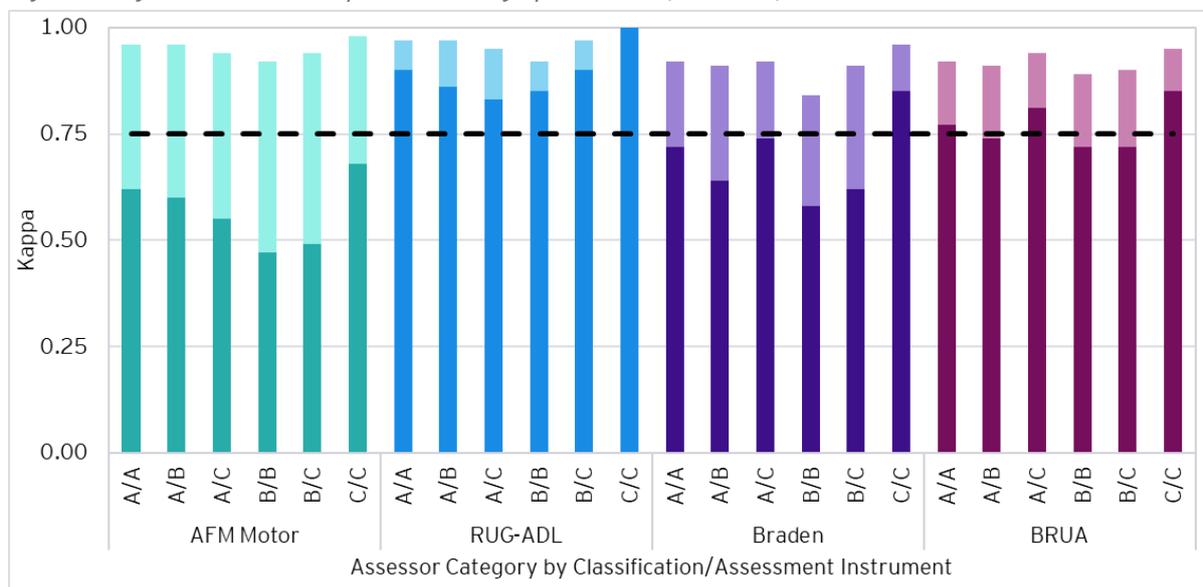
i. Excellent agreement is represented by the dashed line (0.75).
ii. Exact agreement (unweighted kappa) is shown in darker shading for each classification/assessment instrument.
iii. Agreement with tolerance for partial agreement (weighted kappa) above exact agreement is shown in lighter shading.

Key observations from the above figures include:

▸ With a tolerance for partial agreement (weighted kappa), excellent agreement was observed across all instruments for all pairs of assessor category.

▸ Agreement was consistently high across assessor pairs with at least one Category A assessor.

▸ Agreement for B/B assessor pairs was consistently lower than all other assessor pairs, particularly for the DEMMI-modified, AFM Motor and Braden.

Department of Health, Disability and Ageing – RAC funding assessment Anomaly Analysis – Inter-rater reliability – Final report

▸ Both exact (unweighted kappa) and weighted agreement rates were highest for C/C assessor pairs across the AN-ACC classification and all instruments. It is noted that results for the C/C group should be interpreted with care due to the low IRR assessment volume and limited AO representation.

3.3.1 Comparison of agreement levels across IRR studies

The proportion of assessors in Category A increased from IRR 1 to IRR 2 to IRR 3. This was expected as more assessments continued to be performed in the AN-ACC system and assessor experience developed. However, the proportion of assessors in Category A dropped from 77% in IRR 3 to 63% in IRR 4. Simultaneously, the proportions of assessors in Category B and Category C have both increased by 7%, over the same period. The composition of assessors participating in IRR 4 may reflect the change in the assessor workforce following the implementation of the Single Assessment System in December 2024 which introduced additional AOs and new assessors.

Table 11 shows the number of assessors by category in each IRR period and Table 12 shows the number and proportion of IRR assessments completed by each assessor category pairing. An assessor's category for each IRR period is defined based on the number of assessments they completed from 1 October 2022 to the end of the respective IRR period.

Table 11: Distribution of IRR assessments by assessor category and IRR period

Assessor Category	Number of Assessments Completed	IRR 1 Assessors		IRR 2 Assessors		IRR 3 Assessors		IRR 4 Assessors	
		No.	Prop.	No.	Prop.	No.	Prop.	No.	Prop.
A	> 350	71	33%	179	66%	207	77%	183	63%
B	> 100, <= 350	67	31%	50	19%	44	17%	71	24%
C	<= 100	79	36%	40	15%	21	6%	38	13%
Total		217	100%	269	100%	272	100%	292	100%

Table 12: Distribution of IRR assessments by assessor category pairing and IRR period
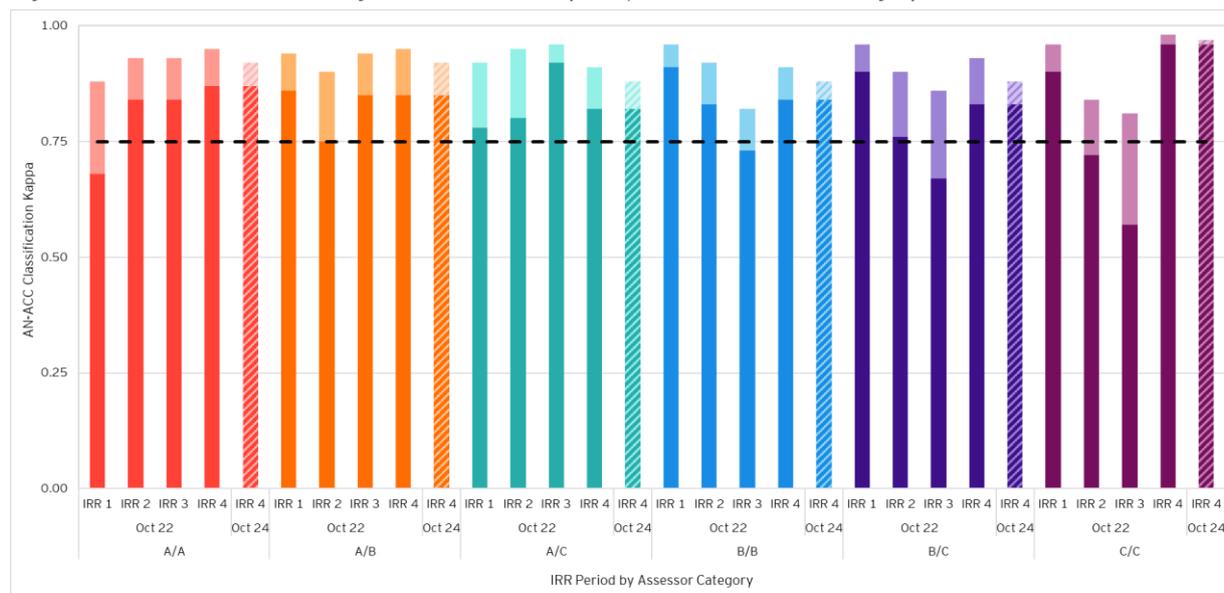
Assessor Category Combination		Mapping	IRR 1 Assessments		IRR 2 Assessments		IRR 3 Assessments		IRR 4 Assessments	
			No.	Prop.	No.	Prop.	No.	Prop.	No.	Prop.
A	A	A/A	90	14%	474	51%	662	60%	1,125	67%
A	B	A/B	176	27%	240	26%	251	23%	255	15%
A	C	A/C	81	12%	108	12%	113	10%	115	7%
B	B	B/B	136	21%	34	4%	55	5%	98	6%
B	C	B/C	124	19%	34	4%	11	1%	70	4%
C	C	C/C	49	7%	29	3%	6	1%	30	2%
Total			656	100%	919	100%	1,098	100%	1,693	100%

i.   *Proportions may not add up exactly due to rounding.*

The proportion of IRR assessments completed by a pair of Category A assessors increased from 14% to 67% between IRR 1 and IRR 4, with subsequent decreases to the proportions of assessments completed by all other assessor category combinations over the same period. This is reflective of the overall change in assessor category mix as assessors perform more assessments and gain more experience.

Figure 12 shows the comparison in Fleiss' kappa statistics for AN-ACC classification in each IRR period.

Figure 12: AN-ACC classification agreement statistics by IRR period and assessor category

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.*
iv. *Agreement for the AN-ACC classification using the October 2022 NWAU basis is shown in the solid-coloured columns, enabling a like-for-like comparison between IRR periods.*
v. *Agreement for the AN-ACC classification using the October 2024 NWAU basis in IRR 4 is shown in the striped column. The NWAU basis only affects the calculation of weighted agreement (weighted Fleiss' kappa).*

Key observations from Figure 12 include:

▸ Excellent agreement was observed when measured with a tolerance for partial agreement at the AN-ACC classification level for all assessor category pairs in each IRR period.

▸ Exact agreement rates increased for A/A assessor pairs and remained unchanged for A/B assessor pairs between IRR 3 and IRR 4.

▸ When adjusting to the October 2022 NWAU basis, weighted agreement increased between IRR 3 and IRR 4 for all assessor pairs, except A/C assessor pairs.

  o On the October 2022 NWAU basis, weighted agreement decreased by 0.05 for A/C assessor pairs between IRR 3 and IRR 4. AO 9 and AO 1 accounted for the largest proportions of A/C assessor pairs, representing 23% and 19% of total A/C assessor pairs in IRR 4 respectively.

▸ Agreement rates for pairs of C/C, B/C and B/B assessors increased materially between IRR 3 and IRR 4, with a further increase after adjusting to the October 2022 NWAU basis. The largest improvement was a 0.39 increase in exact agreement for C/C assessor pairs.

  o The increase in agreement rates for these pairs should be interpreted with care due to the volatility and lower reliability arising from lower IRR assessment counts and limited AO representation, as shown in Section 3.3.

The high proportion of Category A assessors and the increase in exact agreement for assessor pairs comprising only Category A assessor were key drivers towards the overall increase in exact agreement for the AN-ACC classification between IRR 3 and IRR 4 (as observed in Section 3.1.1).

Similarly, after adjusting to the October 2022 NWAU basis, the increase in weighted agreement for A/A and A/B assessor pairs were key drivers of the overall increase in weighted agreement for the AN-ACC classification, as observed in Section 3.1.1.

Further, adjusted weighted agreement improved for all assessor pairs, except A/C assessor pairs. As noted in the IRR 3 review in September 2024, an improvement in overall agreement rates between assessors in IRR 4 and future IRR rounds is expected as assessors gain more experience and agreement of these assessor pairs increases. This was also demonstrated through the increase in the proportion of assessments completed by A/A assessor pairs and the increase in exact agreement between IRR 3 and IRR 4.

3.4   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 13 shows the number and proportion of IRR assessments completed by assessment type.

Table 13: IRR assessments by *"payment impacting"* assessment type

Assessment Type	No. IRR Assessments	Prop. IRR Assessments
Initial	528	31.2%
Reclassification (Reclass.)	1,015	60.0%
Reconsideration (Recon.)	150	8.9%
Total	1,693	100%

i.   *Proportions may not add up exactly due to rounding.*

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 13.

Figure 13: Agreement statistics by assessment type and classification/assessment instrument
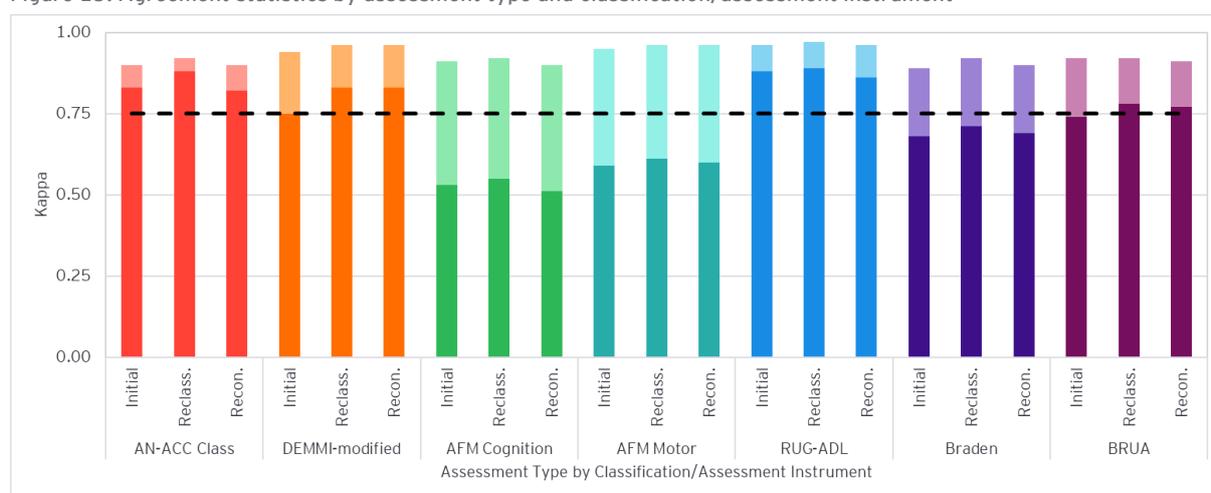
i.   *Excellent agreement is represented by the dashed line (0.75).*
ii.   *Exact agreement (unweighted kappa) is shown in darker shading for each classification/assessment instrument.*
iii.   *Agreement with tolerance for partial agreement (weighted kappa) above exact agreement is shown in lighter shading.*
iv.   *The weighted Fleiss' kappa for the AN-ACC classification uses the October 2024 NWAU basis.*

Key observations from Figure 13 include:

▸ 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 generally 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-modified, AFM Motor and Braden. This is consistent with our previous observations of IRR 3 data.

▸ Weighted agreement on reclassifications was observed to generally be slightly higher than other assessment types, particularly for the AN-ACC classification, AFM Cognition, RUG-ADL and Braden. Exceptions include the DEMMI-modified and AFM Motor, where weighted agreement for reclassifications and reconsiderations were exactly equal at 0.96 and above that of initial assessments.

3.4.1 Comparison of agreement levels against previous IRR studies
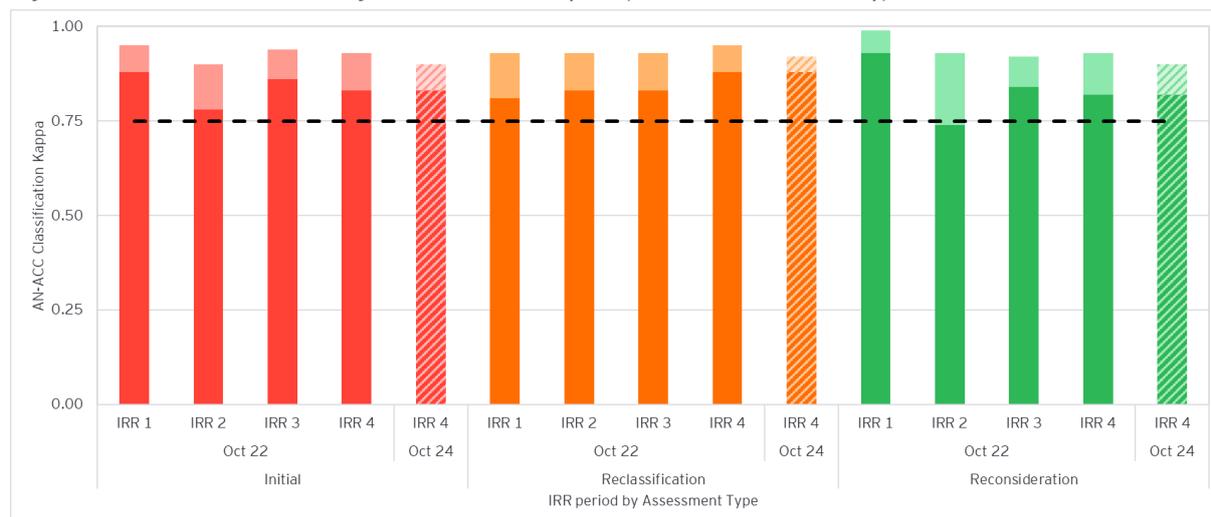
The mix of RAC funding assessments by assessment type has changed since the beginning of the AN-ACC period on 1 October 2022, with an increased proportion of assessments being reconsideration assessments and a decreased proportion of assessments being initial assessments. Meanwhile, the proportion of 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 were particularly reflected in the changing mix of assessments in the samples chosen for IRR 1 and IRR 2, as shown in Table 14 below, with some allowance for volatility in sample selection between IRR 2 and IRR 3 causing smaller changes in proportions. The mix of samples chosen for IRR3 and IRR 4 are broadly consistent, with an increase in the proportion of reclassifications and a corresponding decrease in the proportion of initial assessments for IRR 4.

Table 14: Distribution of IRR assessments by *"payment impacting"* assessment type and IRR period

Assessment Type	IRR 1 Assessments		IRR 2 Assessments		IRR 3 Assessments		IRR 4 Assessments	
	No.	Prop.	No.	Prop.	No.	Prop.	No.	Prop.
Initial	311	47.4%	300	32.7%	375	34.2%	528	31.2%
Reclassification	328	50.0%	541	59.1%	624	56.8%	1,015	60.0%
Reconsideration	17	2.6%	75	8.2%	99	9.0%	150	8.9%
Total	656	100%	916	100%	1,098	100%	1,693	100%

Figure 14 shows the comparison in Fleiss' kappa statistics for each assessment type by AN-ACC classification and assessment instrument between the IRR 1, IRR 2, IRR 3 and IRR 4 assessments.

Figure 14: 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.
iv. Agreement for the AN-ACC classification using the October 2022 NWAU basis is shown in the solid-coloured columns, enabling a like-for-like comparison between IRR periods.
v.   Agreement for the AN-ACC classification using the October 2024 NWAU basis in IRR 4 is shown in the striped column. The NWAU basis only affects the calculation of weighted agreement (weighted Fleiss' kappa).

The key observations from Figure 14 include:

▸ Excellent agreement was observed for the AN-ACC classification when measured with a tolerance for partial agreement 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.

▸ Exact agreement for the AN-ACC classification on reclassification assessments increased from 0.83 in IRR 3 to 0.88 in IRR4.

    o When using the October 2024 NWAU basis, weighted agreement decreased slightly between IRR 3 and IRR 4. However, when using the October 2022 NWAU basis, weighted agreement increased from 0.93 in IRR 3 to 0.95 in IRR 4.

▸ When comparing on a like-for-like basis using the October 2022 NWAUs, weighted agreement for the AN-ACC classification on initial assessments decreased slightly by 0.01 between IRR 3 and IRR 4, following an increase of 0.04 from IRR 2 to IRR 3. Similarly, exact agreement on initial assessments decreased by 0.03 between IRR 3 and IRR 4.

▸ Weighted agreement for the AN-ACC classification also decreased slightly for reconsideration assessments in IRR 4 under the October 2024 NWAU basis, following a declining trend from IRR 1 to IRR 3. However, after adjusting to the October 2022 NWAU basis, weighted agreement increased slightly by 0.01 between IRR 3 and IRR 4.

o   In addition, exact agreement on reconsiderations decreased slightly by 0.02 between IRR 3 and IRR 4, suggesting that assessors had slightly more frequent disagreements in IRR 4, but disagreed to a slightly lesser extent.

The shift in the mix of IRR assessments towards reclassifications and the increase in exact and adjusted weighted agreement were key drivers towards the overall increase in agreement between IRR 3 and IRR 4, particularly for the increase in pure and exact agreement for the AN-ACC classification (as observed in Section 3.1.1).

Department of Health, Disability and Ageing – RAC funding assessment Anomaly Analysis – Inter-rater reliability – Final report

Appendix A

A.1 Kappa statistics

A kappa statistic measures the level of agreement between assessors with an adjustment for the probability that assessors will agree purely by chance.

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

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

Such that:

$$P_O = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} p_{i,j}$$

$$P_E = \sum_{i=1}^{n} \sum_{j=1}^{n} w_{i,j} p_i p_j$$

Where:

- κ = kappa,

- n = number of classes or scores,

- P_O = probability of agreement <u>observed</u>,

- P_E = probability of agreement <u>expected</u> by chance,

- p_i = proportion of IRR assessments classified as Class/score i,

- $p_{i,j}$ = proportion of IRR assessments where the first assessor assigned Class/score i and the second assessor assigned Class/score j, and

- $w_{i,j}$ = weight assigned to IRR assessments where the first assessor assigned Class/score i and the second assessor assigned Class/score j. Further detail on weightings is provided in Appendix A.1.1.

P_O is calculated using $p_{i,j}$, which represents the observed proportion of IRR assessments where the first assessor assigned Class/score i and the second assessor assigned Class/score j. As an example for the AN-ACC classification, $p_{2,3} = 5\%$ indicates that in 5% of the observed IRR assessments, the first assessor assigned Class 2 while the second assessor assigned Class 3.

In contrast, P_E uses $p_i p_j$ which represents the expected probability that the first assessor will assign Class/score i and the second assessor will assign Class/score j by chance. As an example for the AN-ACC classification, if 10% of residents are placed into Class 2 ($p_2 = 10\%$) and 15% of residents are placed into Class 3 ($p_3 = 15\%$), then there is a $p_2 p_3 = 10\% \times 15\% = 1.5\%$ chance that the first assessor will assign Class 2 and the second assessor will assign Class 3 before allowing for any actual agreement or disagreement between them (i.e. purely by chance).

A.1.1. Kappa weightings

This section outlines how the weights are calculated for exact and weighted agreement.

Note that there is no definitive or prescriptive method for determining the weights used in the calculation of a weighted kappa. The calculation of weights follows the general principle that they should be chosen to best represent the system or situation which is being analysed. As such, the calculation of the weights differ when assessing agreement at an assessment instrument level compared to for the AN-ACC classification.

Kappa weightings for exact agreement

For the unweighted kappa, only exact agreement between assessors is considered. The "*weights*" applied are values of 1 for agreement and 0 for disagreement between assessors.

This is shown in Table 15, with an example of two assessors making a classification between 3 categories. In this example, if 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 15: 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

Kappa weightings for assessment instruments

Table 16 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 (3 – 1 = 2).

Table 16: 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

Linear weights are used to define weights in the weighted Fleiss' kappa for all assessment instruments.

Kappa weightings for the AN-ACC classification

A different approach is used for the AN-ACC classification, whereby the difference in NWAU between each pair of assigned AN-ACC classifications is used to define the weightings in the weighted Fleiss' kappa. Weights are expressed as:

$$w_{i,j} = 1 - |x_i - x_j|$$

Where:

- $w_{i,j}$ = weight assigned to IRR assessments where the first assessor assigned Class i and the second assessor assigned Class j, and

- x_i, x_j = NWAU for Class i and Class j, respectively.

To illustrate, let $w_{2,3}$ represent the weight for the difference between Class 2 and Class 3 and $w_{2,13}$ represent the weight for the difference between Class 2 and Class 13.

Under the October 2024 NWAU basis, the weights are calculated as follows:

$$w_{2,3} = 1 - |x_2 - x_3| = 1 - |0.19 - 0.37| = 0.82$$

$$w_{2,13} = 1 - |x_2 - x_{13}| = 1 - |0.19 - 0.80| = 0.39$$

As seen above, when classes are closer in NWAU, disagreements between them are considered less significant, resulting in higher corresponding weights. As such, disagreements between classes with similar NWAUs are allocated a larger positive contribution to the overall weighted Fleiss' kappa compared to disagreements between classes with substantial differences in NWAU.

Further, class agreement between an assessor pair, i.e. $x_i = x_j$, always results in a weighting of 1, aligned with the other approaches to determine kappa weightings above.

A.1.2 Interpretation of the Kappa statistic

As seen in Appendix A.1, the Kappa statistic can be expressed as:

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

where the expected level of agreement by chance (P_E) is removed from the observed level of agreement (P_O) between assessors to capture the level of actual agreement above chance.

The calculation of the kappa statistic can be broken down into two components:

- $P_O - P_E$, the level of actual agreement above chance, and

- $1 - P_E$, the <u>maximum</u> additional agreement above chance which can potentially be achieved (the value of 1 represents perfect agreement).

It is noted that P_O will equal to 1 when the exact same score across all IRR assessments is assigned by assessors, e.g. all residents are classified as Class 2 by all assessors. This results in a kappa value of 1.

Application of the Fleiss' kappa and weighted Fleiss' kappa statistics

The kappa statistics used in this analysis are summarised in Table 17.

Table 17: 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 categories.	Does not account for partial agreement. All disagreements are considered equal and not included in the calculation. The pairs of assessors determined to complete IRR assessments together are influenced by geographical and logistical constraints. Therefore, assessors are not sampled randomly, potentially introducing additional bias.
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. Weighted agreement accounts for the degree of disagreement between assessors. For example, disagreements between Classes 2 and 3 would have a greater positive contribution to weighted agreement than disagreements between Classes 2 and 13.	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 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 previous reports on RAC funding assessment IRR.

A.1.3 Impact of October 2024 NWAU changes on Kappa statistics and weightings

On 1 October 2024, NWAUs were updated for the first time since the commencement of AN-ACC. Overall, the revision in NWAUs reduced the spread of funding outcomes. This has no impact on IRR measures based on exact agreement, but does impact the weighted Fleiss' kappa calculation, specifically the calculation of weights (based on differences in NWAU between classes).

The weighted Fleiss' kappa for AN-ACC classification has reduced in IRR 4 under the October 2024 NWAUs, compared to applying the same set of disagreements using the October 2022 NWAU basis. This is primarily driven by a reduction in the overall spread of NWAUs across AN-ACC classifications on the October 2024 basis. When calculating partial agreement in the weighted Fleiss' kappa, the generally smaller differences in NWAU funding within AN-ACC classification pairs result in increased kappa weightings, particularly observed for cases of greater, but potentially less realistic, disagreement in AN-ACC classification outcome.

Appendix A.1 expressed the kappa statistic (κ) as:

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

P_O (probability of agreement observed) considers partial agreement for pairs of AN-ACC classifications which have actually been observed in the IRR sample. Meanwhile, P_E (probability of agreement expected by chance) considers all possible pairs of AN-ACC classifications, extending the range of partial agreement considered beyond what may be considered reasonable given practical considerations.

For example, when calculating P_O, there were zero cases of a Class 2 and a Class 13 outcome for an assessor pair observed in the IRR 4 sample. Meanwhile, when considering agreement by chance in the calculation of P_E, a non-zero but small chance of a Class 2 and a Class 13 outcome must be included. Alongside the increased weightings for cases of greater disagreement discussed above, this has resulted in an increase in P_E to a greater extent than the corresponding increase in P_O, when updating to the October 2024 NWAU basis.

When applying these impacts to the kappa statistic formula, the numerator has reduced to a lower extent compared to the denominator, resulting in a decreased weighted agreement result under the October 2024 NWAU basis.

Appendix B

B.1   Assessment Organisation

B.1.1 AO – Fleiss' kappa

Table 18 contains the rounded unweighted Fleiss' kappa agreement rates by AO, as previously observed in the figures provided in Section 3.2.

Table 18: Fleiss' kappa by AO

Metric	AO 1	AO 2	AO 3	AO 4	AO 5	AO 6	AO 7	AO 8	AO 9	AO 10
AN-ACC Class	0.89	0.86	0.91	1.00	0.81	0.80	0.72	0.82	0.91	1.00
DEMMI-modified	0.80	0.75*	0.89	0.97	0.73	0.74	0.63	0.61	0.90	0.92
AFM Cognition	0.55	0.45	0.68	0.94	0.45	0.28	0.37	0.41	0.65	0.80
AFM Motor	0.55	0.5	0.73	1.00	0.47	0.42	0.41	0.43	0.74	0.65
RUG-ADL	0.86	0.82	0.94	1.00	0.86	0.84	0.77	0.91	0.94	0.91
Braden	0.78	0.67	0.80	1.00	0.52	0.63	0.48	0.56	0.8	0.76
BRUA	0.82	0.79	0.81	1.00	0.69	0.71	0.63	0.62	0.83	0.83

*   *Moderate agreement is observed when using the exact, unrounded value.*

B.1.2 AO – Weighted Fleiss' kappa

Table 19 contains the rounded weighted Fleiss' kappa agreement rates by AO.

Table 19: Weighted Fleiss' kappa by AO

Metric	AO 1	AO 2	AO 3	AO 4	AO 5	AO 6	AO 7	AO 8	AO 9	AO 10
AN-ACC Class	0.93	0.90	0.96	1.00	0.86	0.89	0.84	0.88	0.95	1.00
DEMMI-modified	0.95	0.95	0.97	0.99	0.91	0.95	0.91	0.90	0.98	0.98
AFM Cognition	0.92	0.90	0.95	0.99	0.87	0.86	0.84	0.91	0.95	0.98
AFM Motor	0.95	0.94	0.98	1.00	0.94	0.95	0.93	0.95	0.97	0.98
RUG-ADL	0.96	0.93	0.99	1.00	0.95	0.95	0.93	0.97	0.98	0.99
Braden	0.94	0.91	0.94	1.00	0.85	0.90	0.85	0.87	0.95	0.94
BRUA	0.93	0.90	0.94	1.00	0.87	0.89	0.86	0.90	0.95	0.90

B.1.3 AO – Correlation coefficient

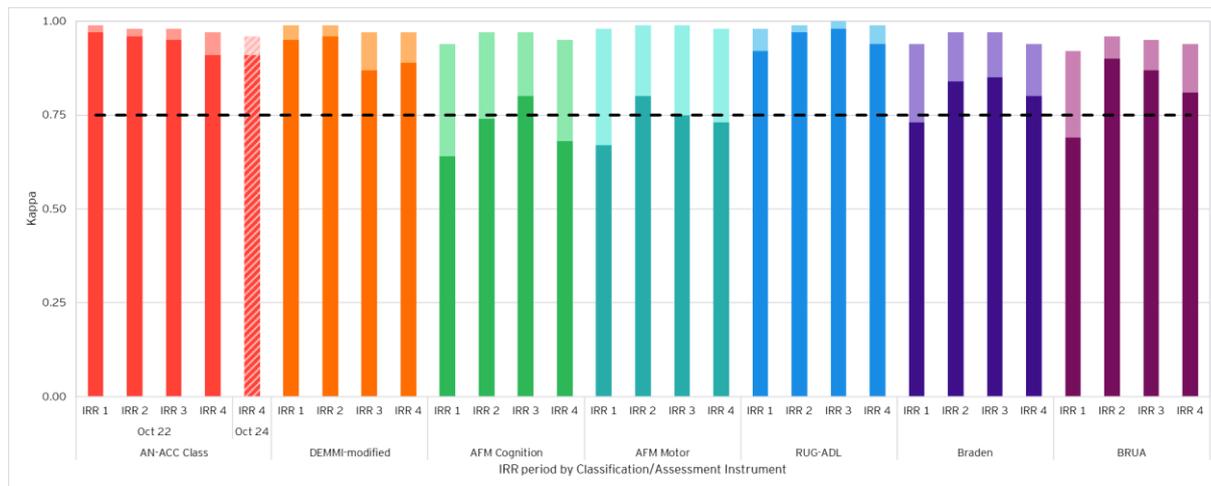Table 20 contains Pearson correlation coefficients by AO.

Table 20: Correlation by AO

Metric	AO 1	AO 2	AO 3	AO 4	AO 5	AO 6	AO 7	AO 8	AO 9	AO 10
AN-ACC Class	0.98	0.99	0.99	1.00	0.95	0.98	0.96	0.97	0.99	1.00
DEMMI-modified	0.99	0.99	0.99	1.00	0.97	0.99	0.98	0.98	1.00	0.99
AFM Cognition	0.99	0.98	0.99	1.00	0.96	0.98	0.96	0.99	0.99	1.00
AFM Motor	0.99	0.99	1.00	1.00	0.99	0.99	0.99	1.00	1.00	1.00
RUG-ADL	0.99	0.97	1.00	1.00	0.97	0.99	0.97	0.99	0.99	1.00
Braden	0.99	0.98	0.99	1.00	0.96	0.98	0.96	0.97	0.99	0.99
BRUA	0.98	0.96	0.98	1.00	0.93	0.96	0.94	0.98	0.99	0.95

i. *This table has been shaded according to the colour scheme in Table 3, which refers to interpretation of a kappa statistic. Technically, Pearson correlation coefficients and kappa statistics are not comparable, but this shading scheme has been adopted for consistency.*

B.2 AO agreement over time

Figure 15 to Figure 20 show the kappa statistics for the AN-ACC classification and each of the underlying assessment instruments across IRR periods.

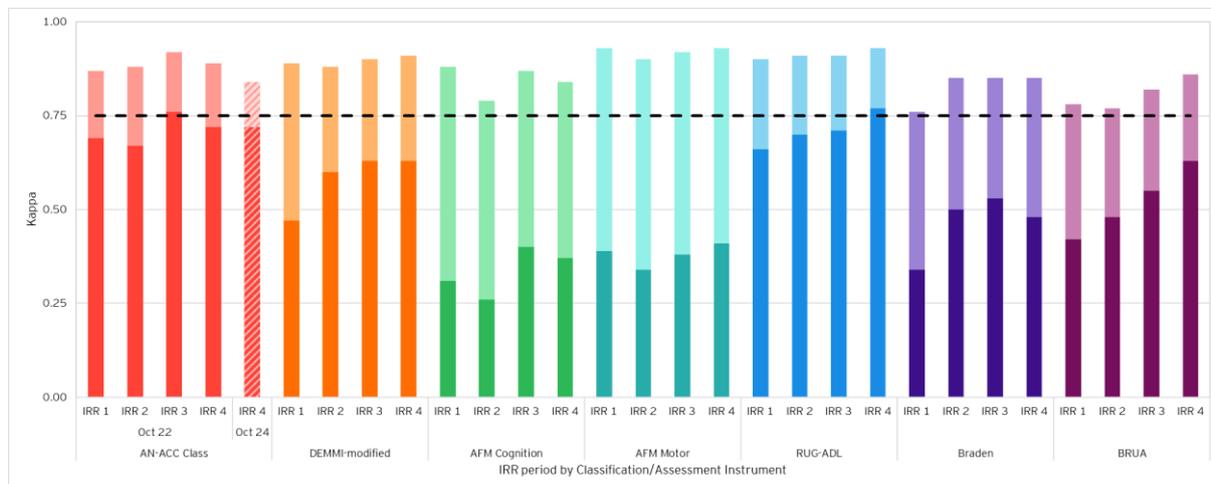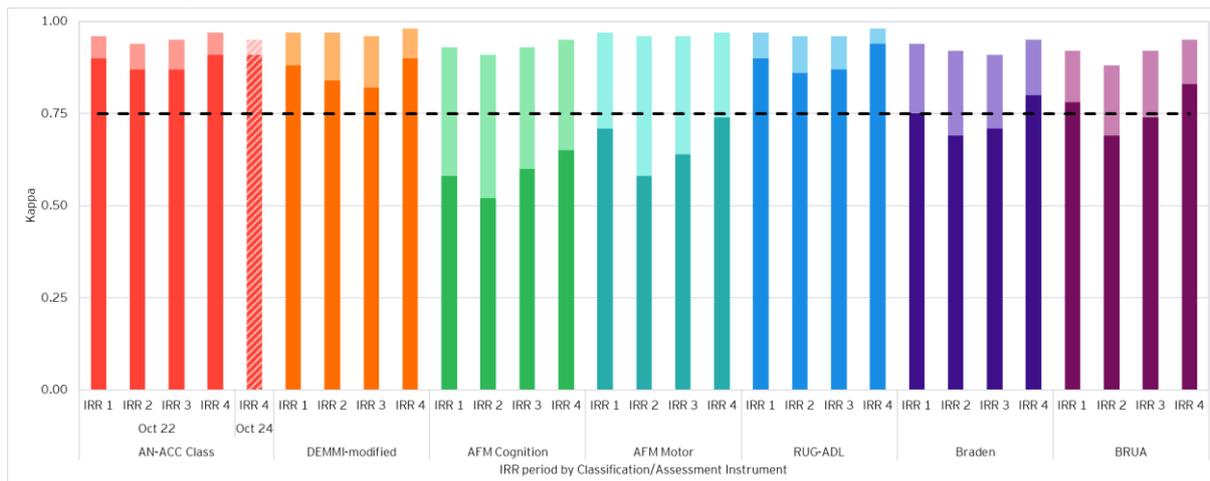Figure 15: AO 3 agreement rates across IRR periods

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.*
iv. *Agreement for the AN-ACC classification using the October 2022 NWAU basis is shown in the solid-coloured columns, enabling a like-for-like comparison between IRR periods.*
v. *Agreement for the AN-ACC classification using the October 2024 NWAU basis in IRR 4 is shown in the striped column. The NWAU basis only affects the calculation of weighted agreement (weighted Fleiss' kappa).*

Figure 16: AO 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.*
iv. *Agreement for the AN-ACC classification using the October 2022 NWAU basis is shown in the solid-coloured columns, enabling a like-for-like comparison between IRR periods.*
v. *Agreement for the AN-ACC classification using the October 2024 NWAU basis in IRR 4 is shown in the striped column. The NWAU basis only affects the calculation of weighted agreement (weighted Fleiss' kappa).*
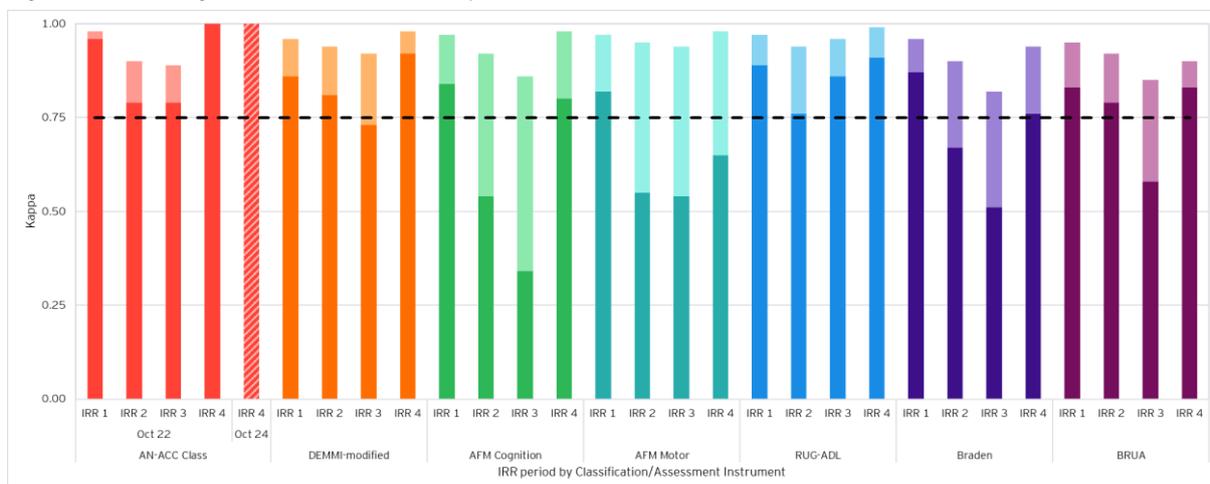
Figure 17: AO 6 agreement rates across IRR periods

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.
iv. Agreement for the AN-ACC classification using the October 2022 NWAU basis is shown in the solid-coloured columns, enabling a like-for-like comparison between IRR periods.
v.   Agreement for the AN-ACC classification using the October 2024 NWAU basis in IRR 4 is shown in the striped column. The NWAU basis only affects the calculation of weighted agreement (weighted Fleiss' kappa).

Figure 18: AO 7 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.
iv. Agreement for the AN-ACC classification using the October 2022 NWAU basis is shown in the solid-coloured columns, enabling a like-for-like comparison between IRR periods.
v.   Agreement for the AN-ACC classification using the October 2024 NWAU basis in IRR 4 is shown in the striped column. The NWAU basis only affects the calculation of weighted agreement (weighted Fleiss' kappa).

Figure 19: AO 9 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.*
iv. *Agreement for the AN-ACC classification using the October 2022 NWAU basis is shown in the solid-coloured columns, enabling a like-for-like comparison between IRR periods.*
v.   *Agreement for the AN-ACC classification using the October 2024 NWAU basis in IRR 4 is shown in the striped column. The NWAU basis only affects the calculation of weighted agreement (weighted Fleiss' kappa).*

Figure 20: AO 10 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.*
iv. *Agreement for the AN-ACC classification using the October 2022 NWAU basis is shown in the solid-coloured columns, enabling a like-for-like comparison between IRR periods.*
v.   *Agreement for the AN-ACC classification using the October 2024 NWAU basis in IRR 4 is shown in the striped column. The NWAU basis only affects the calculation of weighted agreement (weighted Fleiss' kappa).*

B.3   Assessor category

B.3.1 Assessor category – Fleiss' kappa

Table 21 contains the rounded unweighted Fleiss' kappa agreement rates by assessor category, as previously observed in the figures provided in Section 3.3.

Table 21: Fleiss' kappa by assessor category pairs

Metric	A/A	A/B	A/C	B/B	B/C	C/C
AN-ACC Class	0.87	0.85	0.82	0.84	0.83	0.96
DEMMI-modified	0.83	0.73	0.83	0.71	0.81	0.89
AFM Cognition	0.55	0.52	0.59	0.52	0.49	0.68
AFM Motor	0.62	0.60	0.55	0.47	0.49	0.68
RUG-ADL	0.90	0.86	0.83	0.85	0.90	1.00
Braden	0.72	0.64	0.74	0.58	0.62	0.85
BRUA	0.77	0.74	0.81	0.71	0.72	0.85

B.3.2 Assessor category – Weighted Fleiss' kappa

Table 22 contains the rounded weighted Fleiss' kappa agreement rates by assessor category.

Table 22: Weighted Fleiss' kappa by assessor category pairs

Metric	A/A	A/B	A/C	B/B	B/C	C/C
AN-ACC Class	0.92	0.92	0.88	0.88	0.88	0.97
DEMMI-modified	0.96	0.93	0.95	0.89	0.96	0.97
AFM Cognition	0.92	0.92	0.91	0.85	0.90	0.96
AFM Motor	0.96	0.96	0.94	0.92	0.94	0.98
RUG-ADL	0.97	0.97	0.95	0.92	0.97	1.00
Braden	0.92	0.91	0.92	0.84	0.91	0.96
BRUA	0.92	0.91	0.94	0.88	0.90	0.95

B.3.3 Assessor category – Correlation coefficient

Table 23 contains Pearson correlation coefficients by assessor category.

Table 23: Correlation coefficient by assessor category pairs

Metric	A/A	A/B	A/C	B/B	B/C	C/C
AN-ACC Class	0.98	0.98	0.96	0.95	0.98	1.00
DEMMI-modified	0.99	0.99	0.99	0.96	1.00	1.00
AFM Cognition	0.98	0.99	0.98	0.94	0.98	1.00
AFM Motor	1.00	1.00	0.99	0.98	0.99	1.00
RUG-ADL	0.99	0.99	0.98	0.95	0.99	1.00
Braden	0.98	0.98	0.98	0.94	0.99	1.00
BRUA	0.97	0.97	0.98	0.93	0.96	0.99

    i.   *This table has been shaded according to the colour scheme in Table 3, which refers to interpretation of a kappa statistic. Technically, Pearson correlation coefficients and kappa statistics are not comparable, but this shading scheme has been adopted for consistency.*

B.4   Assessor category

There were 346 unique assessor pairs in the IRR assessments. Of those pairs, 14 pairs assessed 15 or more residents together. We conducted reliability testing on each of these pairs to measure agreement between specific assessor pairs.

Most 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. Four pairs of assessors had perfect agreement on the AN-ACC classification.

B.5   Assessors in IRR 3 and IRR 4

35 pairs of assessors (10% of total assessor pairs in IRR 4) performed IRR assessments together across both IRR 3 and IRR 4. These assessor pairs completed a total of 320 IRR assessments across IRR 3 and IRR 4.

Table 24 shows agreement rates by AN-ACC decision tree level across IRR 3 and IRR 4, and Figure 21 shows the comparison in kappa statistics at the AN-ACC and instrument level for these assessors, split by IRR period.

Table 24: Agreement rates by AN-ACC decision tree level for pairs of assessors in IRR 3 and 4 by IRR period

Level	Category	Fleiss' Kappa			
		Exact (unweighted)		Tolerance for partial agreement (weighted)	
		IRR 3	IRR 4	IRR 3	IRR 4
1	Mobility Branch	0.95	1.00	0.96	1.00
2	Mobility: Cognition/ Function/ Pressure Sores	0.92	0.92	0.96	0.97
3	AN-ACC Class	0.91	0.86	0.97	0.94

Figure 21: Agreement rates by AN-ACC classification/assessment instrument for pairs of assessors in IRR 3 and 4 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.*
iv. *Agreement for the AN-ACC classification using the October 2022 NWAU basis is shown in the solid-coloured columns, enabling a like-for-like comparison between IRR periods.*
v. *Agreement for the AN-ACC classification using the October 2024 NWAU basis in IRR 4 is shown in the striped column. The NWAU basis only affects the calculation of weighted agreement (weighted Fleiss' kappa).*

The key observations from Table 24 and Figure 21 include:

▸ At Level 1 of the AN-ACC decision tree, exact (unweighted kappa) and weighted agreement improved from IRR 3 to IRR 4 within 5 percentage points.

▸ At Level 2 of the AN-ACC decision tree, agreement statistics were relatively consistent, with small improvements in weighted agreement in IRR 4 within 2 percentage points.

▸ Exact and weighted agreement reduced for the AN-ACC classification from IRR 3 to IRR 4 by up to 5 percentage points.

▸ Weighted agreement statistics were very consistent (within 0.04) across the underlying assessment instruments. However, exact agreement generally decreased by up to 0.12 for most instruments except the DEMMI-modified and Braden.

EY | Building a better working world

EY is building a better working world by creating new value for clients, people, society and the planet, while building trust in capital markets.

Enabled by data, AI and advanced technology, EY teams help clients shape the future with confidence and develop answers for the most pressing issues of today and tomorrow.

EY teams work across a full spectrum of services in assurance, consulting, tax, strategy and transactions. Fueled by sector insights, a globally connected, multidisciplinary network and diverse ecosystem partners, EY teams can provide services in more than 150 countries and territories.

All in to shape the future with confidence.

ey.com