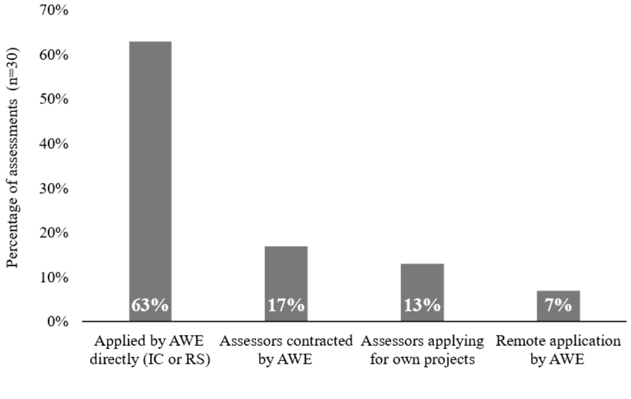
Supplementary Information 1

Collection of the C-Well data by assessors

In this paper’s dataset, the C-Well assessment was applied at 11 facilities a total of 30 times, where the data was collected by seven assessors. All assessors underwent several virtual training sessions regarding the methods of applying the C-Well measures, and five assessors conducted a pilot assessment by accompanying a trained assessor on-site, where they were able to practice data collection and analysis techniques. 19 out of the 30 assessments (63%) were conducted directly by AWE, i.e. either one of the authors of this paper, and initially as a response to the COVID-19 pandemic, a further two (7%) were conducted remotely by the authors, liaising with a trained assessor on-site (Figure 3). The remaining nine assessments (30%) were conducted by trained assessors contracted by AWE, where the results were then reviewed and reported on by the authors of this paper or by trained assessors who hold agreements with AWE to conduct the C-Well for their own projects.

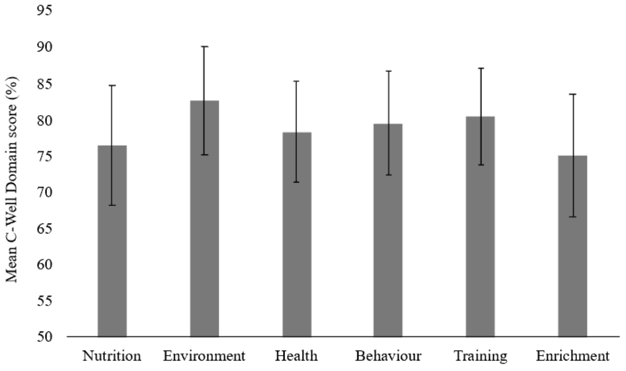


Supplementary Figure 1. Context regarding methods of data collection for the 30 C-Well assessment applications included in this paper

Supplementary Information 2

Variance across welfare Domain scores

The C-Well assessment measures are categorised into six Domains (FigureS2), and we hypothesised that from the 246 applications of the assessment, the average scores generated for each Domain would be statistically different. This hypothesis was largely correct: we first tested the normality of the Domain scores using the Shapiro-Wilk test, where all Domains returned p-values of less than 0.05, indicating non-normal distribution. Results from the subsequent Kruskal-Wallis test showed a significant difference between the Domains, with pairwise Wilcoxon Signed-Rank tests conducted to compare Domain scores (Bonferroni correction applied to control for multiple comparisons). The pairwise results (Table S2) indicated significant differences between total scores in all Domain pairs apart from the Environment-Training pair (P=1.00), suggesting that in general across the 11 facilities included in the study, welfare practices implicated within each Domain followed industry-wide trends, despite differences in species, group sizes, and habitat type.



Supplementary Figure 2. Mean C-Well Assessment Domain scores (as percentages) generated from 246 applications of the assessment on 111 individuals, with black lines showing standard deviation

Supplementary Table 1. Pairwise Wilcoxon Signed-Rank tests (following a Kruskal-Wallis test) were conducted to compare Domain scores (Bonferroni correction applied to control for multiple comparisons). The pairwise results indicated significant differences between all Domain pairs (as denoted by * and bold type) apart from the Environment-Training pair (P=1.00).

Domain Pair	Wilcoxon Statistic	Corrected P value
Nutrition vs Environment	12301	<0.001*
Nutrition vs Health	9193	<0.001*
Nutrition vs Behaviour	8664	<0.001*
Nutrition vs Training	10919	<0.001*
Nutrition vs Enrichment	12334.5	<0.001*
Environment vs Health	2964.5	<0.001*
Environment vs Behaviour	2132	<0.001*
Environment vs Training	6973	1.00
Environment vs Enrichment	11547.5	<0.001*
Health vs Behaviour	4763.5	<0.001*
Health vs Training	9429	<0.001*
Health vs Enrichment	10484	<0.001*
Behaviour vs Training	10546	<0.001*
Behaviour vs Enrichment	13172.5	<0.001*
Training vs Enrichment	10029	<0.001*

Supplementary Table 2. A full record of the changes made to the C-Well assessment measures from 2015 to 2024, with rationale and references.

Version	Release date	Type of change	Measures involved	Rationale and references
1.0	Dec 2013	N/A		Original version (Clegg, Borger-Turner and Eskelinen 2015)
1.0	Oct 2018	Adapted to belugas	5.2.2 Aerial behaviour -> Diving behaviour	Belugas do not show aerial behaviour frequently, ability to dive may be more ecologically relevant to welfare (Mann et al. 2000).
		Adapted to belugas	7.4.1 Blood values -> CRC Handbook beluga values	Gulland, Dierauf and Whitman 2018)
1.1	May 2019	Added	12.1 Positive Reinforcement training used	Positive reinforcement training can be a tool to promote good welfare in cetaceans (Brando 2010 2012; Clegg et al. 2018).
			12.2 Willingness to participate in training sessions	Research indicating willingness to participate is a sensitive welfare indicator, and correlated to overall health (Clegg et al. 2019).
			12.3 Anticipatory behaviour outside session schedule	First application to Indo-pacific bottlenose dolphins, no measures changed (no different blood value reference intervals available at this time, see later updates).
		Modified	9.1 Presence of Social behaviours	1.0 version detailed that no agonistic behaviours being observed in the observation period would receive the 'sub-optimal welfare' score, but since agonistic behaviour is not always shown frequently in different group compositions, this threshold was changed to 3 bouts of agonistic behaviour per hour indicating sub-optimal welfare (Samuels and Gifford, 1997; Scott et al. 2005).
		Modified	3.1 Time budget	The time budget measure, which captures time spent each day that the animals are trained versus 'free-time', was adapted to include a criteria in the sub-optimal welfare score for facilities who offer less than 1 hour of training per day, as this is likely too little stimulation for these animals (Brando 2012; Melfi 2013).
		Modified	5.1.2 Complexity of enclosure	Variation in pool topography more similar to variation in wild environment and can be used to facilitate exploratory behaviour (Clark 2013).
		Modified	10.2 Response to trainer while not under stimulus control	In version 1.0, non-food tactile interactions were required to be observed to achieve the good welfare score for this measure, as an indicator of good human-animal relationships. However, the word 'tactile' was removed after it was noted during applications that relationships could be positive without involving tactile interactions.
		Modified	5.4.1 Application of enrichment	In version 1.0, the original score for good welfare could be achieved for the enrichment measure by applying only 3 times per week. In line with updated recognition of the importance of enrichment, this was updated to 7 times per week.
1.1	May 2021	Adapted to Indo-Pacific bottlenose dolphins	7.4.1 Blood values	New research released with blood reference values for Indo-Pacific bottlenose dolphins (Lauderdale et al. 2021).
1.2	Jul 2021	Added	4.3.2 UV Avoidance policy	Incorporation of research highlighting importance of UV protection for eye health (Colitz, Walsh and McCulloch 2016; Colitz et al. 2019).
			5.1.3 Pool volume	No measure of pool volume included in original assessment, so measure added here (Rose et al. 2017; European Association of Aquatic Mammals 2019).

Supplementary Table 2. Continued.

Version	Release date	Type of change	Measures involved	Rationale and references
2.0	Jan 2022	Added	3.1 Resting behaviour 5.1.3 Interconnecting pools 9.4 Social group size 9.5 Social group management 10.2 Pattern swimming 13.1 Facility behavioural observation policy 14.3 Variability of training sessions 14.4 Variability of guest-facing sessions 16.1 Enrichment variability 16.2 Enrichment frequency 16.3 Enrichment novelty 16.4 Enrichment engagement records 16.5 Enrichment safety protocol 16.6 Average enrichment engagement time	In addition to updating in line with new literature and practical experience, key themes of the major update from version 1.2 to 2.0 were: • Increase of measures focussing on provision of positive welfare opportunities, particularly those evaluating choice, control and agency e.g. 14.3 Variability of training sessions • Balance of measures covering the functional domains (nutrition, environment and health) with the behavioural interaction domains (behaviour, training, enrichment) • Translation of qualitative to quantitative measures wherever possible to increase objectivity e.g. 16.6 Average enrichment engagement time • Refinement of methods and scoring criteria wherever possible e.g. 3.1 Resting behaviour
2.0	Jan 2022	Removed	2.1 Capillary Refill Time (CRT)	No evidence over years or application for CRT or respiration duration varying meaningfully with overall welfare, and no more literature published to support link to welfare.
			3.1 Time budget	Echolocation capabilities were present in all animals where measured, but time-consuming to apply and simple presence/absence may not indicate much about welfare.
			4.1 Frequency of water temperature testing (combined with another measure)	
			5.1.1 Echolocation	
			5.2 Ability to exhibit complex movements	Measure 5.2 Ability to exhibit complex movements more effectively replaced by pool volume.
			7.1.2 Respiration duration	
			8.4 Emergency Containment Training	Measure 8.4 Emergency Containment Training important for human safety but not very relevant to cetacean welfare.
2.0	Feb 2022	Adapted to killer whales and Pacific white-sided dolphins	5.1.3 Pool volume 5.2.2 Water temperature 7.4.1 Blood parameters	These measures were adapted to killer whales and pacific white-sided dolphins using the following references: (Rose et al. 2017; Gulland, Dierauf and Whitman 2018; European Association of Aquatic Mammals 2019; AMMPA 2020).