BIODIVERSITY CHALLENGE FUNDS

DARWIN INITIATIVE | ILLEGAL WILDLIFE TRADE CHALLENGE FUND | DARWIN PLUS

Application scores vs project performance

Biodiversity Challenge Funds: Building and Applying Evidence

Department for Environment, Food and Rural Affairs (Defra)

Date: 12 December 2023

NIRAS, Pentlands Science Park
Bush Loan, Penicuik, EH26 0PL
United Kingdom

UK International Development
Partnership | Progress | Prosperity

Department for Environment Food & Rural Affairs

Contents

Disclaimer

NIRAS is the fund administrator for the Biodiversity Challenge Funds and commissioned this work on behalf of the Department for Environment, Food and Rural Affairs (Defra) under Workstream 5 of the Biodiversity Challenge Funds.

NIRAS works with a range of specialists and consultants to carry out studies and reviews on the Biodiversity Challenge Funds. The views expressed in the report are entirely those of the author and do not necessarily represent the views or policies of Defra, NIRAS or the Biodiversity Challenge Funds. Defra and NIRAS, in consultation with wider stakeholders as relevant, are considering all findings and recommendations emerging from this study in how they manage the Biodiversity Challenge Funds.

Your feedback helps us ensure the quality and utility of our knowledge products. Please email BCF-Comms@niras.com and let us know whether or not you have found this material useful, in what ways it has helped build your knowledge base and informed your work, or how it could be improved.

Cover photograph: Blue Cranes – South Africa – Chris van Rooyen

1. Introduction

1.1. Background

The Biodiversity Challenge Funds (BCFs) are the UK Government's flagship biodiversity grant scheme, helping to protect biodiversity and the natural environment through providing innovative solutions for biodiversity conservation and poverty reduction in lower- and middle-income countries and UK Overseas Territories (UKOTs). The funds initially commenced in 1992 with the launch of the Darwin Initiative at the Rio de Janeiro Earth Summit. Since then, the BCFs have expanded and are now made up of three different funds: The Darwin Initiative, The Illegal Wildlife Trade Challenge Fund, and Darwin Plus. Since inception, these funds have awarded over £239m to more than 1,441 projects globally.

NIRAS – previously LTS International – has acted as the Fund Administrator for over 20 years. Part of NIRAS's role as Fund Administrators is to manage the assessment and evaluation of all projects funded under the BCFs. As part of this process, all projects are assessed and scored at multiple points of assessment throughout their lifetime. This offers grantees regular points of feedback, with an aim of building the capability and capacity of Project Leaders and organisations, and supporting projects to improve their overall performance. This is implemented through a:

1. Two-stage application process which sees Expert Group members score projects against set criteria before grants are awarded;
2. Request for annual progress reports, submitted by the project at the end of each financial year, and that are then reviewed and scored by independent reviewers; and
3. Final Project Report which is submitted by the project upon its conclusion and is then reviewed and scored by an Independent Reviewer.

NIRAS, Defra and the Fund Expert Groups seek to understand further the relationship between Application scores and project performance as reflected in Report Review scores. This work forms part of the BCF programme's "Workstream 5 - Building and Applying Evidence" which aims to collect and synthesise evidence and lessons from projects and processes across the three BCFs. The evidence will be used to evaluate the effectiveness of the application scoring and project performance scoring process, and identify any areas for improvement.

1.2. Objectives

The overall objective of this Deep Dive is to explore the relationship between application quality and project performance across all three funds, by identifying any trends in how projects score from the application stage through to project closure. To achieve this, this Deep Dive sets out to produce an overview of how project scores progress throughout the project-cycle and provide evidence and insights into the effectiveness of scoring and feedback processes from application stage to annual and final reporting stages, including any early indicators of project success or poor performance. This report seeks to describe:

- The extent to which lower scoring project applications can improve from Stage 1 to Stage 2;
- The extent to which application scores are correlated with project performance, as scored by independent reviewers of Annual Reports and Final Reports;
- The extent to which project performance can improve from the first Annual Report through to the Final Report; and
- The key similarities and differences between project scores across the funds and schemes.

Note: the Terms of Reference specified that the study should include an analysis of the role of geographies, and type of project Lead Organisation (i.e. University, International NGO, Local NGO). This was done in an exploratory way but given the relatively small dataset, no trends were observed and it was decided to focus on the fund and scheme-level analysis instead, leaving an analysis of geographies and partners to a future date when more data are available.

The following section outlines the methods used in this study. This is followed by a presentation of results and a brief conclusion.

2. Methods

The study's objectives were achieved through a two-stage process that included the formation of a dataset using various BCFs reporting data, and the analysis of this data to generate evidence to assist in answering the study questions.

2.1. Formation of the Dataset

The dataset for this study includes records on active and recently closed projects which have submitted an Annual Report Review or Final Report Review during the 2022/2023 reporting period (i.e. between April 2022 to March 2023). Data pertaining to Main Projects across the three BCFs – Darwin Initiative, IWT Challenge Fund, and Darwin Plus – are more comprehensive and lend themselves to more in-depth analysis. However, data from all schemes and funds have been included the dataset. The dataset includes projects reporting from the following Rounds:

- Darwin Initiative: Rounds 21, 22, 23, 24, 25, 26, 27 and 28
- IWT Challenge Fund: Rounds 1[1], 2, 3, 4, 5, 6, 7 and 8
- Darwin Plus: Rounds 3, 4, 5, 6, 7, 8, 9 and 10

The original Terms of Reference detailed assessing just five of the most recent completed Rounds. However, the Rounds assessed were expanded given that projects completed under the most recent three Rounds – i.e. Darwin Initiative Rounds 26, 27 and 28; IWT Challenge Fund Rounds 6, 7 and 8 and Darwin Plus Rounds 8, 9 and 10 – are yet to complete their Final Report. We therefore used older rounds as well as newer rounds to generate a larger dataset.

This following specific steps were taken to construct the database. The level of detail is intended to facilitate replication in future.

- Project data has been collated in a single Excel spreadsheet named 'App vs Report Score Template' attached with this report. The spreadsheet contains a 'Master' worksheet which holds a record of all projects and their respective scores at both the application and reporting stage, and then additional supporting spreadsheets named 'Apps ST1', 'Apps ST2' and 'All Reports'.
- The application scores have been collated from the Sift Master spreadsheets for the respective funding Rounds, whilst report scores are taken from the annual Report Master spreadsheet produced each year by NIRAS to manage the administration of Annual and Final Reports. The information from these various spreadsheets was collated into the supporting spreadsheets within the App vs Report Score Template. These sheets act as the full record of application/report scores.

[1] For the purposes of this analysis, IWT001–IWT005 have been classified as Round 1. In strict terms, however, these projects can be considered as Round 0 since they were funded prior to Round 1, just as the Fund was being established.

- Scores entered into these spreadsheets were then cleaned, ready to be populated into the 'Master' worksheet. The 'cleaning' process involved removing blank spaces from the end or start of scores, and deleting any irrelevant text or content. The data recorded in these three spreadsheets was then used to populate the scores in the 'Master' worksheet.
- The 'Master' worksheet was also populated with key project information including the different references allocated to projects at various stages of the application process, the Fund and Round under which the project applied, and the project's start and end dates.
- The Excel workbook is set up so that once project data has been included, and the record of scores has been inputted into the supporting worksheets, then the application and report scores automatically input within the Master spreadsheet.

2.2. Data transformation

The BCFs employ a unique scoring process for project applications, Annual Reports, and Final Reports respectively. Analysis of scores between these reporting phases needs to take this into account. Nevertheless, it is instructive to consider the relationship between project scores across these phases and we have attempted to normalise scores into a common metric to facilitate the visual assessment of trends presented further on. This was achieved as follows.

Application scores are recorded as integer values, provided alongside the maximum possible score. These numbers were divided by their maximum score and therefore normalised in a transparent and statistically consistent way, which is important as the total score available can vary between funding Stages, funding schemes and funding Rounds.

Annual Report Review (ARR) scores were coded as shown in Table 1. The consistency in the ordinal ranking of the AR score is relatively straightforward and the approach to reflecting this ranking in a normalised way is somewhat intuitive. However, the approach to coding was basic and pragmatic. We do not have a strong theoretical underpinning for the assertion that a 2 is equivalent to 70%, for example, and coded AR scores should be interpreted with some caution.

Table 1 Coding Annual Report Review scores

Description	ARR Score	Coded Score
Likely to be completely achieved	1	1
Likely to be largely achieved	2	0.7
Likely to be partly achieved	3	0.5
Only likely to be achieved to a very limited extent	4	0.3
Unlikely to be achieved	5	0
Too early to judge	X	N/A[2]

Final Report Review (FRR) scores were more challenging to normalise, given that there is no clear justification for assigning an A+ a different score in % terms, relative to an A++ or any other score. For this reporting phase, the approach to coding is based on the observation that scoring is largely around whether or not projects have

[2] For projects scoring X in at least one year of the relevant analysis, they were excluded given there would be no meaningful way to represent this score on a scale of 0–1.

delivered outcomes that have met or exceeded expectations. All projects falling into these categories have been coded as 1. The others have been, somewhat arbitrarily, coded as 0.7 and 0.5 respectively (see

Table 2).

Table 2 Coding Final Report Review scores

Description	FRR Score	Coded Score
Outcome substantially exceeded	A++	1
Outcome moderately exceeded	A+	1
Outcome met expectation	A	1
Outcome moderately did not meet expectation	B	0.7
Outcome substantially did not meet expectation	C	0.5

The normalised scores achieved using these techniques were not analysed statistically across phases, so the limitations inherent in our approach to coding have not affected any of the statistical results reported herein.

2.3. Analysis

We employed a combination of descriptive observation and statistical analysis to answer each of the questions, described further in the following sub-section.

2.3.1. Analysis of Stage 1 vs Stage 2 Applications

To assess the degree to which projects improve their scoring between Stage 1 and Stage 2, we considered only Darwin Initiative Main and IWT Challenge Fund Main applications, given that these receive a relatively high number of applications and these are the only two schemes which have had a 2-stage application scoring process running for long enough to generate sufficient data for analysis[3]. We first plotted the projects on a chart showing their Stage 1 scores relative to their Stage 2 scores, allowing for visual analysis of the scores complemented by the use of the Ordinary Least Squares method to investigate the degree to which Stage 1 scores can be used to predict Stage 2 scores. We then divided projects into low scoring (those with scores under 0.75) and high-scoring projects and analysed their average scores between the stages to reveal trends.

2.3.2. Extent to which Application scores predict project performance

We first present a box-and-whisker plot to analyse the spread of project application scores achieved by projects grouped according to their Final Score received (N=227). We then develop a model to assess the degree to which project application scores can be used to predict Annual Report Review scores using Ordinal Logistic Regression (OLR), given our interest in an ordinal dependent variable. We tested the strength of the model and verified its appropriateness considering the distribution of the application scores variable, as shown in Figure 1. This distribution follows a largely normal pattern with some overdispersion between the 0–0.5 range.

[3] Darwin Plus Main, IWT Challenge Evidence, Darwin Initiative Extra and IWT Challenge Fund Extra all have a 2-stage application scoring process but due to limited number of years running and limited number of projects these were not amenable to quantitative analysis.

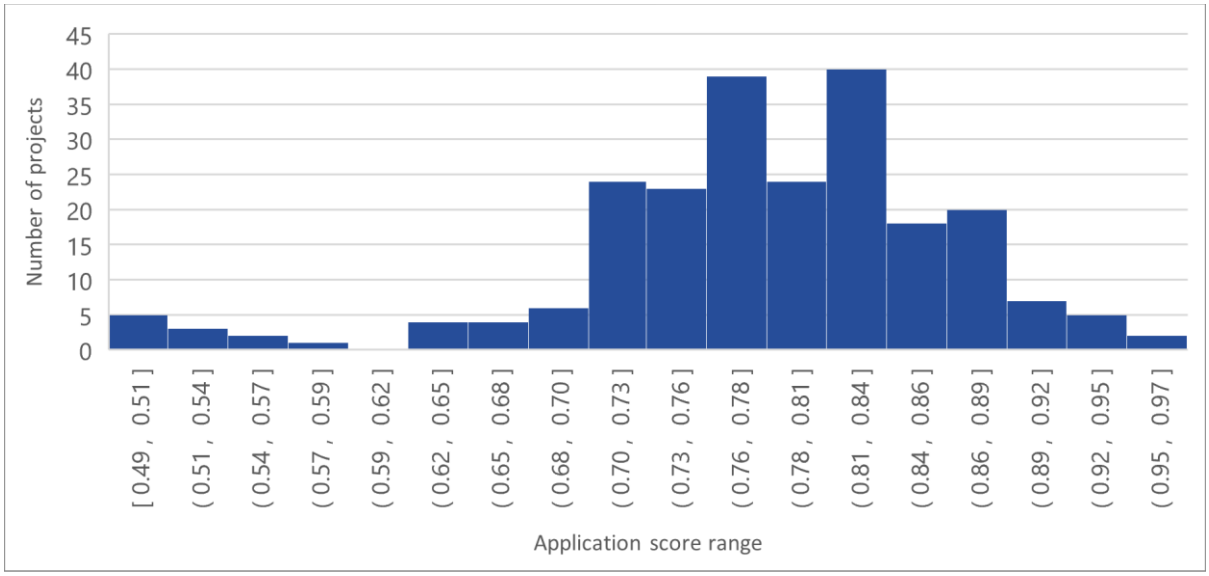


Figure 1 Distribution of application scores for projects that have received a Final Report score

2.3.3. Assessing project performance over time

We first plotted the normalised scores from all scoring phases onto a chart to visually assess the variation in scores over time across the different schemes. We then focused on Annual Report (AR1 and AR2) scores to more robustly assess the degree to which projects were experiencing mobility in scoring over time (N=256). We divided projects into low-scoring (0.5 and lower) and high-scoring (all scores higher than 0.5) groups and assessed the change in average scores between AR1 and AR2 to reveal trends.

2.4. Limitations

This study investigates the extent to which there is a relationship between the application scores and report review scores of successful grantees only. Given that the lowest scoring applicants do not qualify for funding, their scores are not represented in this dataset and we cannot know what their report review scores would have been, had they been funded. Furthermore, in some years, even some of the higher-scoring projects are not offered funding, given the amount of strong applications relative to the total amount of funding available in any particular year.

Annual and Final Report Review scores are determined by the extent to which grantees can demonstrate progress towards achieving their stated Outcome and Outputs. The scores are allocated based on a one-day desktop review of the Annual and Final Reports, as well as examination of documents provided as Means of Verification. The Report Review scores are therefore an incomplete picture of project performance to the extent that there are limited resources available to assess project performance comprehensively for every grantee (noting that the Fund Managers utilise several more comprehensive reviews for higher-risk schemes or projects).

The data transformation process had inherent limitations. While these have been made explicit above through a description of the coding process used, we are limited in the degree to which we can meaningfully compare scores from two different scoring processes, each designed with its own project phase-specific evaluation criteria(i.e. an application score compared to a final report review score).

3. Results

3.1. Lower scoring projects improve from Application Stage 1 to Stage 2

There is a broad correlation between the scores that projects receive at Stage 1 and the scores that they receive at Stage 2. This is shown in Figure 2, which plots each project's Stage 1 score against its Stage 2 score. To interpret the figure, consider that all projects which have achieved the same score at both stages fall precisely on the diagonal line. Projects displaying progress are displayed in the upper-left triangle, while projects which have experienced a drop in their overall score are shown in the bottom-right triangle.

The imperfect relationship between these scores is illustrated by the R^2 value associated with an Ordinary Least Squares regression fitted to these points – 0.48 for Darwin Initiative Main projects and 0.35 for IWT Challenge Fund Main projects. This reveals the extent to which there is potential for projects to progress, as well as to regress in their scoring during the project design phase.

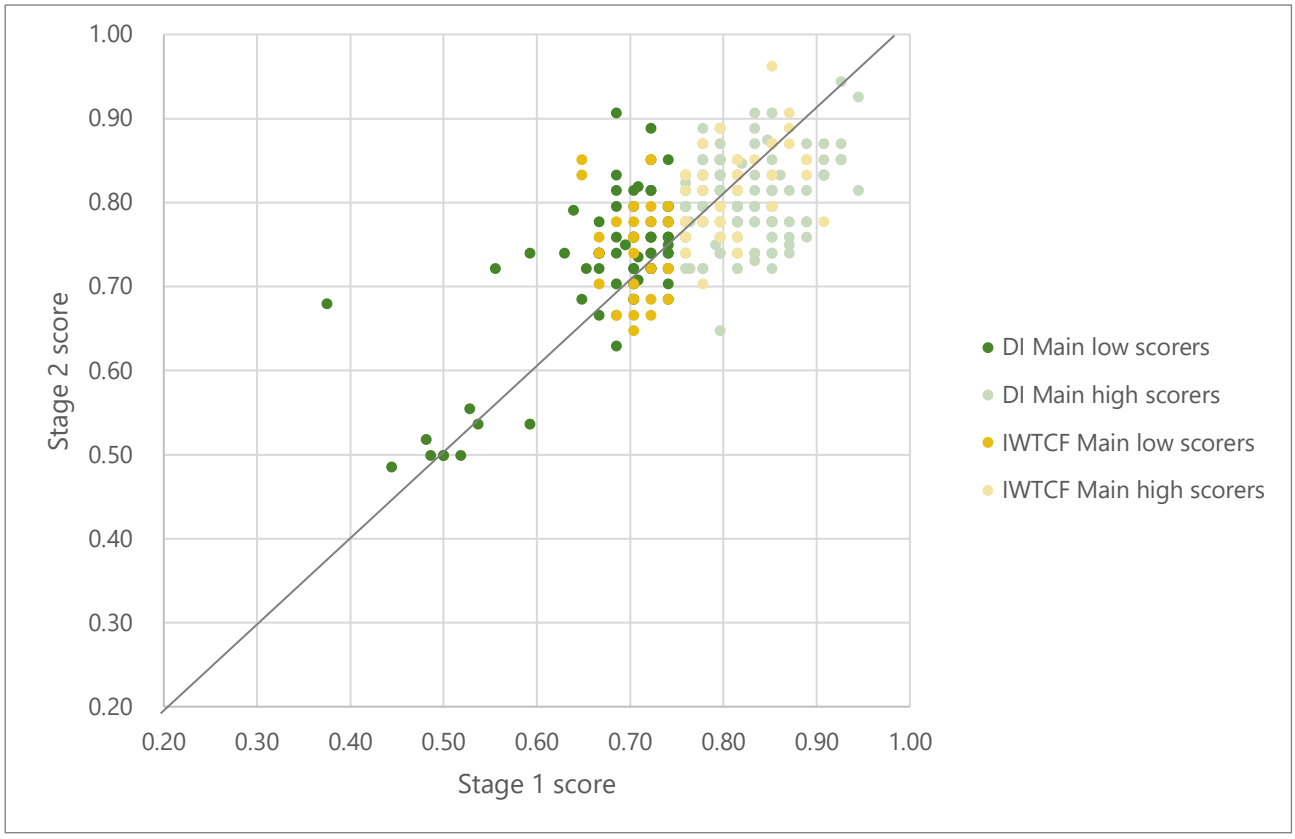


Figure 2 Relationship between Stage 1 and Stage 2 scores for Darwin Initiative Main and IWT Challenge Fund Main

Table 3 shows that low scoring projects funded under Darwin Initiative Main experienced an average increase of 10% in their scores between Stage 1 and Stage 2. Low scoring IWT Challenge Fund projects saw a lower increase of only 1% between stages. High-scoring projects experienced a slight decrease in average scores between the stages for both Darwin Initiative Main (2%) and IWT Challenge Fund Main (4%).

Table 3 Average scores for different DI Main and IWT Challenge Fund Main groups during the application process

Group	Average scores			
	Stage 1	Stage 2	Change	P-value[4]
DI Main low scorers	0.68	0.75	0.08	0.000
DI Main high scorers	0.82	0.80	- 0.02	0.001
IWTCF Main low scorers	0.71	0.71	0.01	0.007
IWTCF Main high scorers	0.81	0.77	- 0.04	0.561

3.2. Extent to which application scores predict project performance

Insofar as project performance can be assessed through the Final Report Review score, there does not appear to be a strong relationship between application scores and project performance. Figure 3 shows a box-and-whisker plot of application scores for projects grouped according to their Final Report Review score. While there does not appear to be a strong trend across the range of FRs, projects that scored Cs in their Final Reviews do appear to have lower-than-average application scores. There were only seven such projects, though, so this result should be interpreted with some caution.

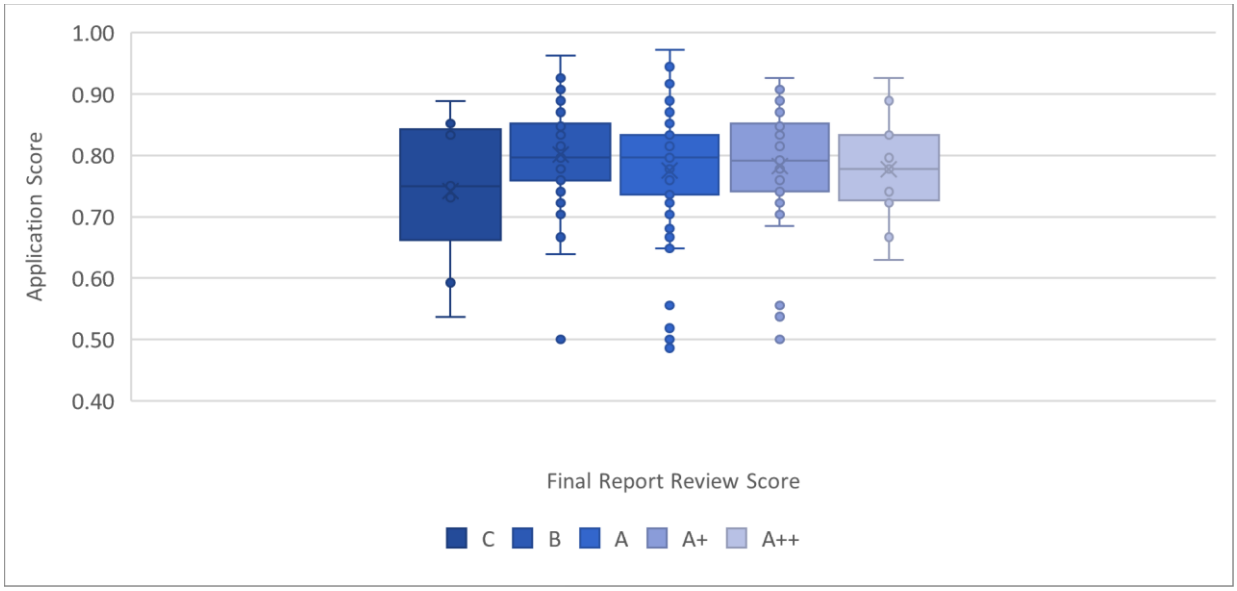


Figure 3 Application scores for projects grouped according to their Report Review Scores

To further understand the relationship between application scores and project performance, we tested the hypothesis that application scores are reliable predictors of Annual Report Review scores. Using Ordinal Logistic Regression, we failed to reject the null hypothesis, namely that there is no relationship between application scores and project review scores (a summary of the model is provided in Annex 1). We therefore observe that at this stage, given currently available and applicable data on 227 projects, we do not have sufficient evidence to suggest that project application scores are indicative of project performance.

[4] These p-values were obtained by running a paired, two-tailed t-test to ascertain whether the distribution of scores at Stage 1 is different from that at Stage 2 for each of the groups. Note that the results of t-tests become less reliable when they are applied to multiple clusters within the same dataset, as is the case here, so these results should be interpreted with some caution.

3.3. Extent to which project performance can improve

Overall, there does not seem to be a strong trend in the average scores that projects receive throughout their lifecycle. Figure 4 demonstrates this by showing that the average project score for each respective scheme remains within a relatively narrow band (between 60% and 85%). Within this margin, some of the schemes displayed show some variation in the average scores received by projects across review stages. Note that some of this variation, particularly where it is universal, is likely to be the result of the different scoring systems applied at the application, Annual Report, and Final Report stages. This figure should therefore be interpreted with caution.

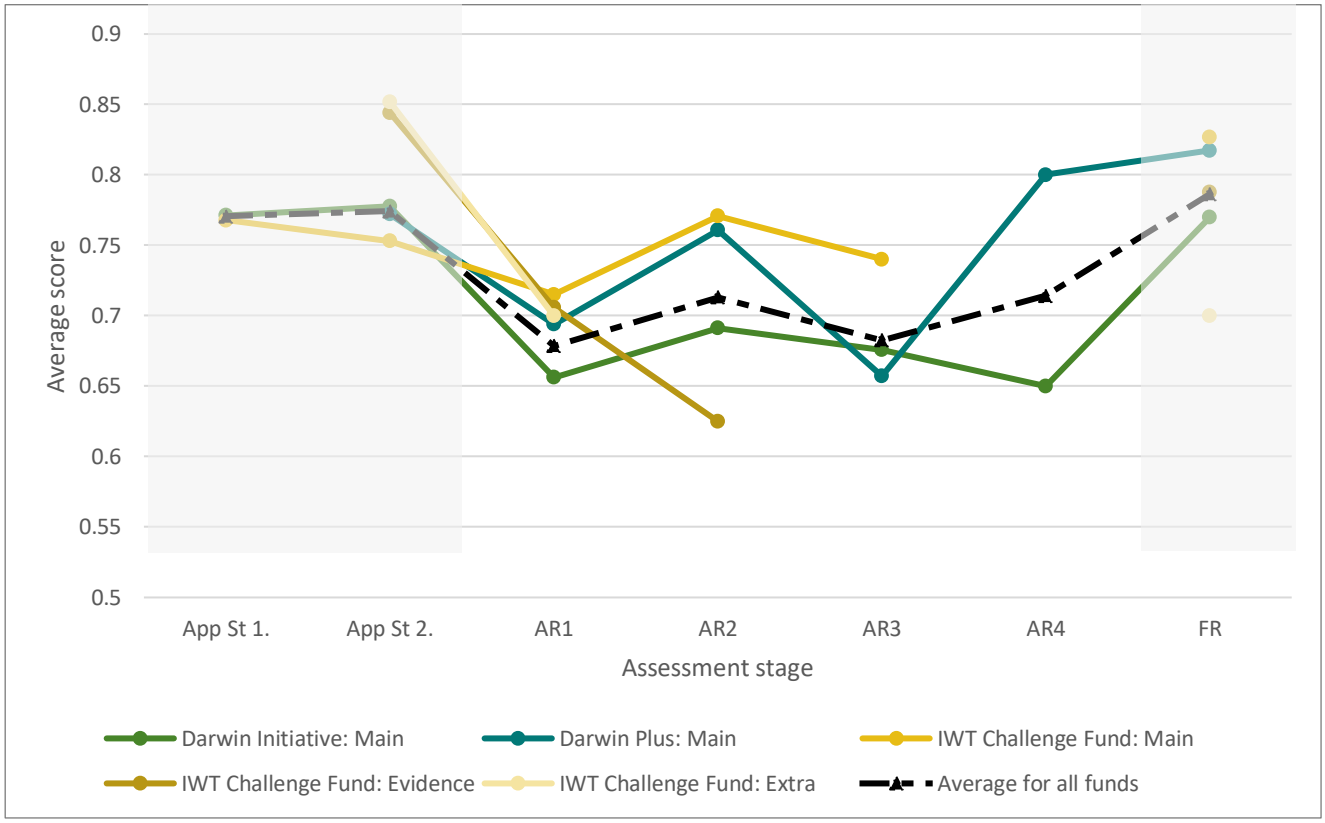


Figure 4 Average scores throughout project lifecycles

To assess the extent to which low-scoring projects can improve over time, we separated projects into two groups depending on how well they scored in their first Annual Report Review. We then looked at the average within-group score during the AR1 and AR2 processes. Table 4 shows that low-scoring groups were able to improve significantly, especially for Darwin Plus, IWT Challenge Fund, and Darwin Initiative Main projects. By contrast, projects that achieved high scores during their first annual review were likely to achieve similar scores in their second annual review.

Table 4 Average scores for projects of different groups between AR1 and AR2

Group	Average scores			
	AR1	AR2	Change	P-value[5]
DI Main low scorers	0.48	0.61	0.13	0.000
DI Main high scorers	0.77	0.73	- 0.03	0.133
IWTCF Main low scorers	0.39	0.65	0.25	0.018
IWTCF Main high scorers	0.78	0.81	0.03	small-n
D+ Main low scorers	0.48	0.75	0.27	small-n
D+ Main high scorers	0.75	0.79	0.03	small-n

4. Discussion and Conclusion

This study has revealed project scoring trends within the BCFs, to the extent that these trends exist and are apparent with currently available data. The results are discussed below, noting that these conclusions are subject to the limitations outlined in Section 2.4 and therefore expressed using cautionary language

Considering only schemes with a two-stage application process, we found that Darwin Initiative Main applicants which had achieved lower scores in their first stage were more likely to see improved second-stage scores relative to IWT Challenge Fund Main applicants. For both schemes, projects that scored highly in their first application were likely to score, on average, just slightly lower on their second application. This result suggests that feedback and guidance provided during the application stage is effective, and particularly effective for projects that scored lower initially.

When assessing the extent to which project application scores are reliable predictors of annual review scores, we did not find evidence to support this assertion. We found weak evidence to suggest that projects which scored a C at their final review stage achieved, on average, relatively low scores during their application stage. However, given that only seven projects in our database scored a C, this should be considered as a preliminary result to be investigated when further work is commissioned on this topic. One potential explanation for the lack of evidence around a trend is that the Expert Committees specifically intend to approve only projects which fall above a perceived threshold of quality and which can deliver against their objectives. This interpretation validates the existing strategy of Expert Committees approving only those projects which are perceived to be capable of delivering against their objectives, rather than approving all of the highest-scoring projects subject to a total budget constraint[6].Finally, when considering trends during implementation only, projects that perform poorly during their first Annual Report Review tend to see improved scores during their second annual review. This is true for Darwin Initiative Main, and is especially true for the IWT Challenge Fund and Darwin Plus grantees.

The dataset produced and attached with this report can be used for future analysis. This dataset includes the various project references and scores throughout their lifetime and can act as a future tool for understanding project performance. The dataset includes a guidance note, as a sheet in the Excel file, to help ensure that it is a useable tool.

[5] These p-values were obtained by running a paired, two-tailed t-test to ascertain whether the distribution of scores at Stage 1 is different from that at Stage 2 for each of the groups. Note that the results of t-tests become less reliable when they are applied to multiple clusters within the same dataset, as is the case here, so these results should be interpreted with some caution.

[6] A policy that is itself supported by flexibility in allocation across the full portfolio of funding schemes.

4.1. Recommendations

Completing this analysis has also led to identifying a number of recommendations to raise:

1. Projects have multiple different references – an initial Stage 1 Application reference, Stage 2 Application and then a final project reference. The BCFs team should maintain a central record to ensure that all application scores and review scores can be linked to their corresponding projects and to one another. The dataset produced through this Deep Dive should offer a point of reference for checking different references, but for future rounds it would help to have Stage 1, Stage 2 and Project references in a central database.
2. Relating to this, references are populated in different sheets in different formats, for example some include a shortened version of the full project reference, for ease of reference in large datasets. We recommend that the central record of project references suggested above include both short and long forms of the project references.
3. Applications, Annual Reports and Final Reports all have different scoring systems. Creating consistency where possible would simplify the process and allow for easier comparison of performance. However, we recognise that this is unlikely to be possible given the different purposes of the review phases. In addition, different schemes and Funding Rounds have different scoring systems. Therefore we suggest maintaining a clear record of the scores in relation to their maximum available scores, such that the scores can be coded simply and effectively to facilitate some degree of cross-phase comparison.
4. Future work could consider identifying projects which tend to cluster along a particular scoring pattern. As further data becomes available, such project clusters should become easier to identify and explore further. Of significance is the question of whether low-scoring projects tend always to score low from the start, or whether some projects score well initially and are subsequently impacted by external shocks that are reflected in lower later scores. Similarly, whether there are scoring trends according to geography of type of lead organisation. Exploring these questions further could yield insights for grantees and the fund management team.

Annex 1. Ordinal logistic regression summary

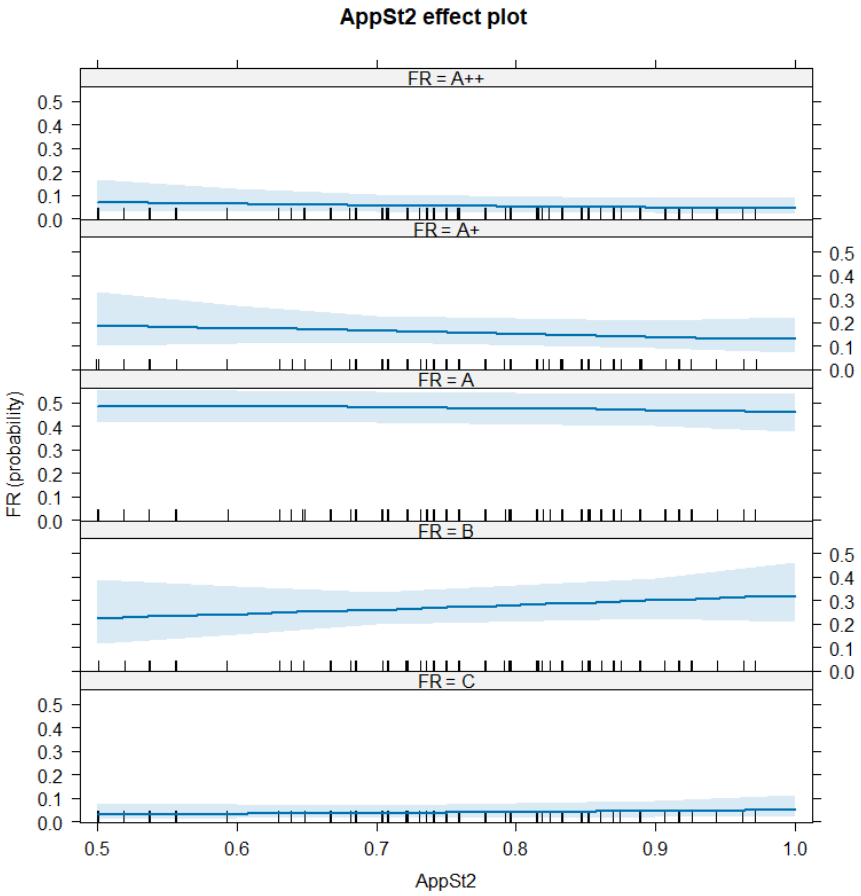
```
Model summary: Ordinal logistic regression (OLR)
FR = Final Report
AppSt2 = Application scores[7]

Call:
polr(formula = FR ~ AppSt2, data = Appdata, Hess = TRUE)

Coefficients:
        Value Std. Error t value
AppSt2 -1.088      1.422 -0.7656

Intercepts:
       Value   Std. Error t value
C|B    -4.0444  1.1723    -3.4499
B|A    -1.6198  1.1245    -1.4406
A|A+    0.4963  1.1163     0.4446
A+|A++  2.0392  1.1423     1.7852

Residual Deviance: 580.8858
AIC: 590.8858
```



[7] For schemes with a two-stage application process, Stage 2 scores were used