Norad
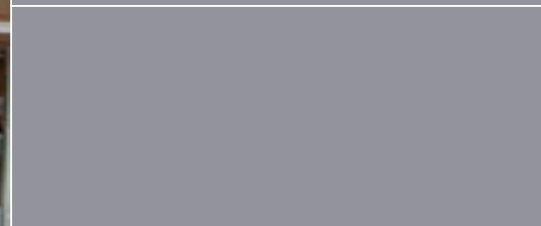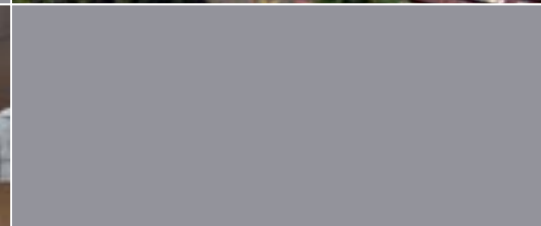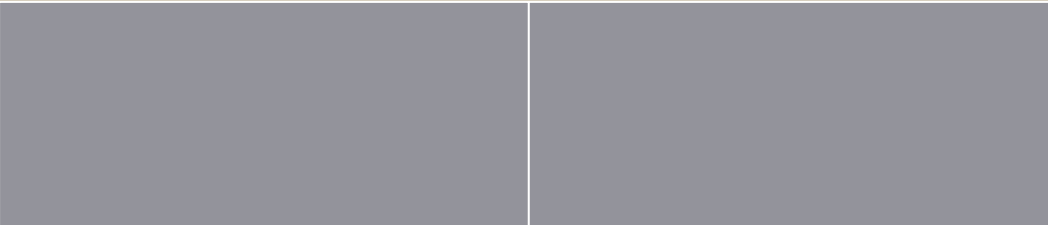
Evaluation Series of NORHED
Higher Education and Research for Development
Evaluation of the award mechanism

Report 6/2015

Evaluation Series of NORHED

Higher Education and Research for Development

Evaluation of the award mechanism Report

Development Portfolio Management Group

July 2015

USC University of Southern California

Acknowledgements

Table of Contents

Figures

Tables

Boxes

Acronyms and Abbreviations

CREAM	Clear, Relevant, Economic, Adequate and Monitorable
DPMG	Development Portfolio Management Group
FORSK	Section for Research, Innovation and Higher Education (Norad)
ICT	Information and Communication Technology
IDA	International Development Association
IRC	Independent Review Committee
LMIC	Lower-middle income country
M&E	Monitoring and Evaluation
NOK	Norwegian Kroner
NOKUT	Norwegian Agency for Quality Education
NOMA	Programme for Master Studies (Norad)
Norad	Norwegian Agency for Development Cooperation
NORHED	Norwegian Programme for Capacity Development in Higher Education and Research for Development
NUFU	Norwegian Programme for Development, Research and Education
PDO	Project Development Objective
RFF	Revised Results Framework
ToR	Terms of Reference

Executive Summary

A. Introduction

1. Norad's program to support higher education in developing countries, the Norwegian Program for Capacity Development in Higher Education and Research for Development (NORHED), is committed to strengthening higher education and research in low and middle income countries (LMICs). This objective is measured by whether the grantees produce more graduates, both men and women, and whether these graduates are better qualified in six identified areas/sub-programs.[1] It is also measured by the quantity and quality of research relevant to these same areas/sub-programs. In the long term, NORHED's investment in academic institutions should contribute to a more skilled workforce, increased knowledge, evidence-based policy and decision-making, and greater gender equality.

2. This document reports on the results for the second of three tasks specified in Norad's Terms of Reference for this call-off (See Annex A). This task was to describe and assess the awards mechanism that Norad used to award the NORHED grants.

3. The analysis was based on four data sources: a) meetings and subsequent communications with the staff for the NORHED program (Research, Innovation and Higher Education Section) and the Results Management Section to understand the details of the awards process; b) reviews of templates that structured the awards process and archival documents that recorded actions at each step; c) creation of databases from the archives and their analysis; and d) an independent evaluation of a proportional, stratified random selection of eligible applications.

4. The study is limited in certain ways, one being that it did not include interviews with either LMICs or Norwegian University partners who applied for a grant from the NORHED program. Obtaining feedback from the South/North partners could be especially useful for two aspects of the awards process: a) the application package; and b) the value of the feedback that they received from Norad on the results of the evaluation process. (The partners lacked information on other aspects of the process.) Many of these partners had also participated in the two predecessor programs to NORHED: Programme for Master Studies at Norad (NOMA) and Norwegian Programme for Development, Research and Education (NUFU). This experience could have given them a useful perspective on the fairness and quality of the NORHED awards process.

B. Findings

5. Between September 2012 and July 2013, the section for Research, Innovation and Higher Education at Norad (FORSK) designed and managed two calls for the NORHED program—a call for seed funding and the main call for applications. It then proceeded with the evaluation of both sets of applications and the final awards for both.

6. Receiving seed funding is statistically associated with stronger proposals, but in this case the call for seed funding had no random assignment, which would have allowed the Development Portfolio Management Group (DPMG) at the University of Southern California to test whether the benefits of seed funding outweigh the substantial costs. Without random assignment, the applicants that responded were self-selected, a process that introduces bias. There were also insufficient data to reliably differentiate lower capacity from higher capacity applicants. Therefore, **it was not possible to determine if lower-capacity applicants that received seed funding submitted better applications than lower-capacity applicants that did not receive seeding funding.**

---

[1] The six priority program areas are Education and Training; Health; Natural Resource Management, Climate Change, Environment; Democratic and Economic Governance; Humanities, Culture, Media, and Communication; and Capacity Development in South Sudan.

7.   The screens used in the evaluation process were generally well conceived and well managed, with only two having design problems: the Independent Review Committee (IRC) stage and the final selection stage. The IRC stage had four problems:

1) The workloads for the different IRC panels varied substantially. Panels for those thematic domains with large numbers of applications had an exhausting workload that could have affected the quality of their evaluations.

2) There was no training of the panels prior to assembling or when they assembled in Oslo in order to increase inter-panel reliability except for a two-hour discussion of the IRCs' deliverables for each application.

3) The IRCs were not asked to submit a final completed evaluation grid that recorded the points for each criterion and an explanation of the IRC's reasoning for the points allocated. Only the relatively short narrative could be used to determine if an evaluation properly reflected the quality of an application package. For this reason, there was no way to assess whether the IRCs' textual explanations and point allocations were logically consistent with each other.

4) There was no independent moderator for each IRC who could ensure that the IRC panel applied the criteria consistently across evaluations, that scores were justified, or (in coordination with other moderators) that the different IRC panels were consistent with each other.

8.   In the final selection stage, the director of Norad, informing the Ministry of Foreign Affairs about the process, made the final decision about those applications to be funded. Because this is bilateral aid funded by Norwegian taxpayers, it is important that the aid reflect the priorities of their democratically elected government. However, although in this case the final screen had a minimal effect on the selection of the final awards, the political considerations associated with this final stage could potentially distort a merit-based selection of applications.

9.   The main application form and NORHED evaluation grid have significant design flaws. As part of its independent evaluation of a random sample of 37 eligible applications, DPMG kept a log of design flaws that it found in the application package and the evaluation grid. Given these flaws, it is likely that Norad's award process did not select those projects from the eligible population of applications that responded best to NORHED's objectives and that were most likely to succeed.

10.   The main *application form* was poorly structured in four ways:

1) The organization of its requests for information did not let the applicant tell a coherent story about the project.

2) It did not elicit sufficient detail about the design of components and implementation arrangements to let IRCs and other reviewers judge whether the project could realistically be expected to implement.

3) It did not solicit some information that is important for judging the quality of the application.

4) It did not offer sufficient guidance to applicants to help them avoid mistakes.

11.   The NORHED *evaluation grid* also had a number of problems:

1) Although Norad's Independent Evaluation Committees and DPMG found that the definitions of the funding categories 1 ("fund") and 3 ("reject") were clear, they both found that the distinction between categories 1 ("fund") and 2 ("fund with reservations") was not clear. This lack of clarity introduced an undesirable degree of variability in the interpretations by the IRCs.

2) Some criteria (2A, 2B, 3A, and 3B) should have been treated as eligibility criteria, not as evaluation criteria.

3) One criterion (3B) penalized applications from upper middle income countries, even those in a priority region for NORHED funding. This reduced opportunities for South-South, not just South-North, capacity development arrangements.

4) Two sub-criteria (4B and 5B) were the only basis for judging an important dimension of the applications: the technical design of their educational and research components. However, these sub-criteria are stated so vaguely that they are vulnerable to multiple interpretations by reviewers.

5) The grid has evaluation items that seem to duplicate each other: risks and gender.

6) Most criteria have 2 to 4 sub-criteria, but the grid has no guidance on how FORSK wants evaluators to weight each of the sub-criteria in terms of points.

7) The grid's point allocations are puzzling: for example, it locks up (and virtually throws away) about 35 points in criteria 2 and 3, both of which are primarily eligibility criteria. Although the results framework (criterion 7), especially baseline data, is extremely important, it gets only 10 points. Criterion 9 that includes the financial plan, budget, and implementation plan, gets only 5 points.

8) The application form and evaluation grid are not completely aligned. For example, the grid asks for a judgment regarding the "financial sustainability strategy" of the project, but the application asks for an "assessment of sustainability" without specifying *financial* sustainability. It asks for an assessment of the cost-effectiveness of the "financial plan," but does not ask for comparative cost information for the various proposed expenditures. Unless average cost coefficients (or something similar) are provided to the reviewers, this sub-criterion cannot be judged.

12.   The categorization of the projects by DPMG and the IRCs was partially but not entirely consistent. DPMG evaluated a stratified and proportional random sample of the universe of eligible applications using the NORHED evaluation grid. Some differences between the two groups were expected, given the research on the variability between expert panels and the complexity of the material being evaluated.

▪ The IRCs and DPMG categorized almost half (46 percent) of the projects the same--both rated it a 1, a 2, or a 3. They differed by one category for more than a third (38%) of the sample--the IRC categorized the project as a 1 and DPMG as a 2 or vice versa (14%), or the IRC categorized the project as a 2 and DPMG as a 3 or vice versa (24%). They differed by two categories for about a sixth of the sample, with the IRC categorizing the project as a 1 and DPMG as a 3 or vice versa (16%).

▪ In 8 cases or 22 percent of the sample, the IRC would definitely not fund the project and DPMG would or might fund it. These cases represent potential missed opportunities. In 5 cases or 14 percent of the sample, the DPMG would definitely not fund the project and the IRC might fund it (the "2" cases). *However, none of these five cases were in fact funded* and thus do not represent potential mistakes. In 2 cases or 5% of the sample, the IRC would definitely fund the application and the DPMG would not fund it. These cases constitute potential mistakes.

**13.   It is important to examine the patterns of strength and weakness in the applications relative to the specific criteria of the evaluation grid. Such patterns can signal the need to redesign the application package and/or the evaluation grid or can alert FORSK to potential implementation problems downstream. In terms of the ratings of the specific grid criteria, the sampled projects performed the worst on the results framework criterion and on a criterion (8) that consisted of highly disparate sub-criteria. Projects also lost substantial points on criteria that assess the design of the educational and research components of the project.

C. Conclusions

Seed funding

**14.   Receiving seed funding and receiving a grant award are strongly related across all applications regardless of sub-program. To the extent that seed funding improves the chances of receiving a grant award, it occurs entirely through its impact on the quality of the proposal. It appears that applicants that received seed funding wrote significantly stronger proposals, which ultimately had much higher chances of

receiving funding. However, this finding may have nothing to do with seed funding and everything to do with selection bias.

15. It was not possible to determine if lower-capacity applicants benefitted from seed funding since there were inadequate data to differentiate lower-capacity from higher-capacity applicants.

Awards process

16. With some exceptions, the awards process itself was well conducted and in line with good international standards. Norad's Section for Research, Innovation and Higher Education that manages the NORHED program clearly tries to learn from experience, as evidenced by its two documents that report and reflect on the first call-off.[2]

17. The comparison of DPMG's independent evaluation of a random sample of applications and the results of the IRC panels showed a variation in scoring and categorization that was reasonable in most cases. However, it highlighted potential missed opportunities for about a fifth of the total sample and potential errors for about 5 percent of the sample.

18. The application package that evaluators were asked to assess and the evaluation grid that they were asked to use in their assessments were both poorly designed. These problems markedly increased the chances of evaluation mistakes—either by missing promising applicants or by selecting less promising ones.

D. Recommendations

1) Substantially revise the application package, especially the main application, as well as the evaluation grid, to eliminate their weaknesses.

2) To the extent possible, move political considerations in the selection process to the design of the main call as eligibility criteria.

3) Enlarge the number of IRC members, as needed, in their "reader" roles to ensure the panels have more balanced workloads.

4) Ensure that each IRC has a reasonable workload when meeting as a committee-of-the-whole, preferably by adding days in Oslo for the IRC.

5) Require IRCs to submit a completed evaluation grid for each application that reflects the IRC's discussion of the draft grid in order to conduct checks on inter-rater reliability and validity. FORSK could use the main readers to revise their draft evaluation grids to reflect considerations that were raised in discussions by the entire IRC.

6) Review the evaluation grid with all evaluators before they start their deliberations, clarifying the intent of each criterion and alerting them to possible misinterpretations.

7) In line with a proposal of the NORHED program staff, define the role of each IRC panel chair as the independent moderator that assures the quality and consistency of the IRC's work. Conduct modest training of the panel chairs for this role.

---

[2] *Innstilling og vurderinger av søknader,* 2013; *Erfaringsnotat for NORHED-programmet,* 2014a.

I. Introduction

A. Purpose

1.1. NORHED, Norad's program to support higher education in developing countries, is committed to strengthening higher education and research in low and middle income countries. This objective is measured by whether the grantees produce more graduates, both men and women, and whether these graduates are better qualified in six identified areas/sub-programs.[3] It is also measured by the quantity and quality of research relevant to these same areas/sub-programs. In the long term, NORHED's investment in academic institutions should contribute to a more skilled workforce, increased knowledge, evidence-based policy and decision-making, and greater gender equality.

1.2. This document reports on the results for the second of three tasks specified in Norad's Terms of Reference (ToR) for this call-off (See Annex A). This task was to describe and assess the awards mechanism that Norad used to award the NORHED grants.

1.3. NORHED's awards process is illustrated in Figure 1. NORHED is to be commended for managing the formidable logistics associated with evaluating 295 applications for seed funding and assessing 173 application packages for project funding—all in a short time frame.

Figure 1: NORHED's Awards Process—Steps, Volume, and Schedule

---
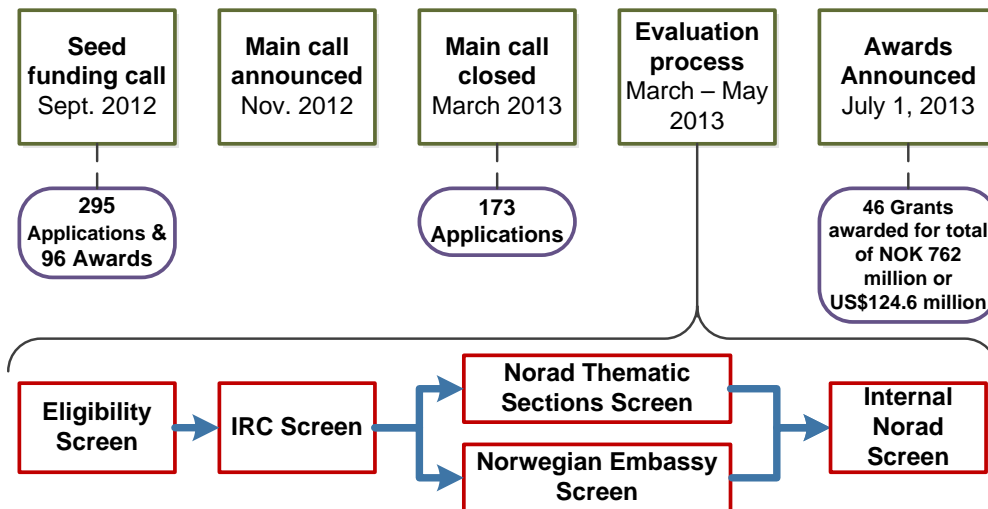
[3] The six priority program areas are Education and Training; Health; Natural Resource Management, Climate Change, Environment; Democratic and Economic Governance; Humanities, Culture, Media, and Communication; and Capacity Development in South Sudan.

B. Methodology

1.4. The analysis was based on four data sources:

- Meetings with the staff for the NORHED program (Research, Innovation and Higher Education, or FORSK) and the Results Management Section to understand the details of the awards process, as well as multiple clarification communications after these meetings took place;

- Reviews of templates and archived documents that recorded actions at each step;

- Creation of databases from the archives and analysis of the data, such as the database on seed funding used to assess if receiving seed funding enhanced the quality of proposals; and

- An independent evaluation of a proportional and stratified random selection of eligible applications.

1.5. The Development Portfolio Management Group (DPMG) at the University of Southern California had initially expected to review a proportional and stratified random sample of the evaluation grids used by the Independent Review Committees (IRCs) to rate the applications. The objectives of such a review would have been to assess: a) whether the IRCs had applied the criteria in the evaluation grid consistently within and between sub-program areas[4] (a check on inter-rater reliability); b) whether the evidence that the IRCs provided to justify their ratings of each evaluation criterion was persuasive (checks on validity[5]); and c) whether there were patterns of strengths and weaknesses in applications, as measured by the evaluation criteria, that could alert FORSK to potential problems with the application package and to potential downstream implementation issues.

1.6. DPMG had to jettison this plan because FORSK did not ask the IRCs to submit their final evaluation grids for the applications assessed. The main reader for each application completed the grid. S/he presented the grid to the full IRC panel for discussion, and presumably some ratings and supporting arguments were modified during these discussions. However, FORSK did not ask for either the draft evaluation grids or ask the IRCs to submit finalized grids that reflected their judgments. The only data were the IRCs' total score (sum of the points for the criteria), its categorization for each application as a "1", "2", or "3" (fund, fund with reservations, or reject), and a brief narrative that described the reasons for its judgments.

1.7. As a result, DPMG had no data for pursuing its three objectives. DPMG reasoned that the most important IRC decision for Norad was the IRCs' categorization decisions. Accordingly, DPMG felt that it was important to see if the categorization decisions by independent evaluators of a sample of projects approximately coincided with the IRCs' categorizations. Completing the evaluation grids for a sample of applications would also yield information on the relative strengths and weaknesses of the applications, as measured by the evaluation grid.

1.8. For this purpose it drew a 23 percent sample of the eligible applications that was random, proportional, and stratified. As described in detail in chapter IV, DPMG's evaluation duplicated some features of the NORHED IRC evaluation process. It did not duplicate others for two reasons discussed in detail in chapter IV: a) cost; and b) results of published evaluations of peer review processes (e.g., Marsh, Jayasinghe, and Bond, 2008; Bornmann, 2011; Abdoul et al., 2012). These studies show that variations can be expected between two different expert panels, *no matter how well their processes replicate each other.*

1.9. Thus, DPMG expected some differences between the judgments by the IRC and DPMG expert panels, some perhaps arising from differences in process and some from the natural variation associated

---

[4] The six priority program areas are: Education and Training; Health; Natural Resource Management, Climate Change, Environment; Democratic and Economic Governance; Humanities, Culture, Media, and Communication; and Capacity Development in South Sudan.

[5] The applications provided complex information susceptible to alternative interpretations. Thus, the validity issue was whether the IRC adduced reasonable arguments for its ratings, not whether a second evaluation might have emphasized different aspects of the same information referenced by the IRC to support its ratings.

with two different peer reviews. The applications being assessed are complex and susceptible to alternative judgments, depending on the backgrounds of the reviewers. The question was whether the DPMG and IRC categorizations were roughly similar.

1.10. Like NORHED, it used evaluators substantively expert in the sub-program areas in developing countries and in evaluation. Unlike NORHED, it trained the evaluators in order to increase inter-rater reliability. Each member of the DPMG team independently reviewed two applications, using the NORHED evaluation grid, before conducting its assessments. The entire team then assembled to discuss differences and causes of differences in their ratings of the two applications. Unlike NORHED, DPMG used one evaluator, not three, for each application. The same evaluator conducted the reviews of all applications sampled in his/her sub-program. DPMG did not replicate the IRC committee-of-the-whole that, depending on the sub-program, consisted of three to seven members. To simulate the effect of a discussion of each application by the full IRC panel, it did take two steps that are detailed in chapter IV.

C. Limitations of the Study

1.11. This study is limited in several ways. The study did not include interviews with either LMIC or Norwegian University partners who applied for a grant from the NORHED program. The Terms of Reference (ToR) for the scope of work for this task did not include interviews with applicants. However, as the work on the awards process proceeded, it was recognized that obtaining feedback from the South/North partners could be especially useful for two aspects of the awards process: a) the application package; and b) the value of the feedback that they received on the results of the evaluation process. (The partners lacked information on other aspects of the process.) Many of these partners had also participated in the two predecessor programs to NORHED: Programme for Master Studies at Norad (NOMA) and Norwegian Programme for Development, Research and Education (NUFU). Their experience with these programs could have given them a useful perspective on the NORHED awards process. Should Norad decide to contract for case studies of a sample of the grants, these studies can and should collect data from LMIC and Norwegian partner institutions on their views of aspects of the awards process.

1.12. A workshop for Norad and the stakeholders was held on the draft awards process report on June 3, 2015, at Norad in Oslo. Those attending were not representative of the population of applicants.[6] The workshop was not structured to elicit participants' views *systematically* on the application package and Norad's feedback on the awards decision. However, those stakeholders present confirmed the awkward design of the application package. They felt that Norad's not being able to share the finalized and complete evaluation grids with applicants compromised the transparency of the awards process and missed a learning/capacity development opportunity for applicants. One participant pointed out the problems with NORHED's instructions to Norwegian universities about setting priorities if they were submitting 15 or more applications, a process that FORSK had already concluded would not be used in future calls for proposals.

1.13. FORSK had used a seed funding call in the first round of grants and had found it demanding to manage. This study did not examine the cost/benefit tradeoffs between options that FORSK might use to support applicants' development of their proposals. The study also did not examine alternatives that FORSK might consider regarding its complaints resolution mechanism. Finally, it did not analyze the transparency and use of the often voluminous communications between FORSK/Norad and applicants during the award process.

---

[6] LMIC partners that had either won or not won an award and Norwegian partners who had not won an award were either absent or under-represented.

D. Organization of the Report

1.14. Chapter II assesses the relationships between receiving seed funding and receiving project funding. Chapter III describes and analyzes the overall awards process for the main call that is diagrammed in Figure 1.

1.15. Chapter IV describes DPMG's independent evaluation of a proportional and stratified random sample of the 164 eligible applications, using the NORHED evaluation grid. Chapter V analyzes the flaws in the application package and evaluation grid revealed by DPMG's independent evaluation. These flaws may have reduced the quality of projects selected. This chapter also evaluates the extent to which the application package and evaluation grid are aligned with the theory of change that underlies the NORHED program.

1.16. Chapter VI compares the outcomes of DPMG's assessment of its sample of eligible applications with the Independent Review Committees' results for these same applications, and Chapter VII presents the conclusions of the assessment and recommendations for future rounds of NORHED grants.

II.   Assessment of the Seed Funding Call

2.1.   FORSK started the NORHED award process by posting a call for seed funding proposals (see Figure 1). Awards ranged from NOK 150,000 to a maximum of NOK 250,000,[7] depending on the number of partners involved in the proposal. The intent was to defray the costs of developing complete applications, such as working out partnership arrangements or collecting baseline data. Receipt or non-receipt of seed funding did not affect institutions' eligibility to apply under the main call.

2.2.   Norad announced the call for seeding funding grants (and the main call) at www.norad.no. It informed potential university partners of the calls through different organizations and networks in Norway and informed potential partners abroad through the embassies and other organizations.

2.3.   Norad received 295 applications and made 96 seed funding awards. The thematic sections of the Ministry of Foreign Affairs/Norad and FORSK evaluated the applications, using the same assessment template that FORSK had developed. These sections then met to discuss their assessments, which were largely consistent. The resulting draft list of selected applicants to receive seed funding were next sent to the relevant embassies, which were asked to consider the applications' relevance to national priorities, other Norwegian efforts in that country, and the Embassy's experience with and assessment of the applicant institution. FORSK conducted a final review based on the overall assessment.

2.4.   FORSK found that the seed funding call was relatively costly to administer. Although the applications were small in scale, there were many of them and they had to be processed and evaluated by multiple sections in Norad. In planning future calls, the costs of a seed-funding phase have to be weighed against its benefits.

2.5.   DPMG conducted an analysis to shed some light on the benefits of seed funding. However, the analysis was limited because it could not properly control for selection bias since the award process of seed funding was not random. Those responding to the seed funding call were self-selected, and from that universe of applicants, FORSK selected the most promising concept papers for seed funding.

2.6.   DPMG coded Norad's master list of all eligible applications as having received or not having received seed funding. This number was 94 applications (two that received seed funding did not submit an application). DPMG then created a database of all eligible projects by seed funding (yes/ no), sub-program, IRC categorization (1, 2, or 3), and receiving a project award (yes/no). Annex B shows the distribution of applications by these categories.

2.7.   The statistical results and their interpretation are somewhat technical, and interested readers are referred to Annex C for the detailed results and discussion of the analysis. The main finding was that for all applicants, regardless of the sub-program, receiving seed funding and receiving a grant award were strongly related. This relationship was not affected by the nature of the sub-program to which proposals were submitted. It appears that applicants that received seed funding wrote significantly stronger proposals, which ultimately had much higher chances of receiving funding. However, this finding may have nothing to do with seed funding and everything to do with selection bias.[8] Applicants differed in their initial capacities, with some being notably more capable than others.

2.8.   It would have been desirable to test whether those applicants with lower capacities that got seed funding submitted better proposals than lower-capacity applicants that did not get seed funding. Unfortunately, NORHED did not collect much information for each applicant on system-level and

---

[7] At the NOK to dollar exchange rate in October, 2012, the seed funding grants ranged from about US$27,000 to about $45,000.

[8] FORSK is alert to this possibility, as indicated on p.5 of their *Erfaringsnotat for NORHED-programme.*

institutional-level enabling conditions that perhaps could have been used to differentiate among applicants with different capacities.

2.9.   Thus, the analysis was not able to resolve FORSK's dilemma about whether the benefits of seed funding outweighed the costs. There are many options for structuring a pre-proposal stage that FORSK might consider. These vary in their costs relative to benefits. [9]

---

[9] FORSK's idea of using an expression of interest stage rather than seed funding prior to the main call is worth considering. FORSK and Norad's thematic sections will still have to evaluate these expressions of interest, but FORSK will not have to manage the logistics of disbursing seed grants.

III.   Evaluation Process

3.1.   As Figure 1 shows, the core evaluation process for applications under the main call consisted of five screens.[10] Four of these five steps were good practice, and the fifth was reasonable.

1) Eligibility screen, which involved the whole population of applications and was conducted by FORSK;

2) Screen of all eligible applications by the Independent Review Committees, which assessed the quality of the applications against the criteria in NORHED's evaluation grid;

3) Screen by Norad's thematic sections. This screen involved a summary of each application and its IRC evaluation for all applications rated "1" and several rated "2" by the IRCs;

4) Screen by Norwegian embassies. The embassies assessed four questions for all applications recommended or being considered for funding. (They reviewed a set of those rated "2" that differed somewhat from those reviewed by Norad's thematic sections.) Like the thematic sections, their reviews were based on a summary of each application and its IRC evaluation;

5) Final selection from the proposed set of winners by Norad's management, including a check for balance in the portfolio relative to the Norwegian Government's priorities.

3.2.   Although there were issues with two of these screens, these screens were generally well conceived and well conducted.

A. Eligibility Screen

3.3.   FORSK screened applications against the basic eligibility criteria specified in *A Presentation of NORHED* (Norad, 2012)*.* This document was a combination of a tender and Terms of Reference for the main call and was available to all applicants.

▪ Applicants had to propose projects that had collaborative partnerships between higher education institutions in Norway and one or more higher education institutions in the South with the aim of building capacity in LMICs.

▪ Partners from LMICs had to be higher education institutions accredited or recognized by in-country national authorities in countries that are registered as OECD DAC official development assistance recipients, or as listed in the specific call for applications.

▪ Norwegian partners had to be higher education institutions accredited by the Norwegian Agency for Quality Education (NOKUT). They had to offer accredited degree programs and operate in accordance with Guidelines for Quality Provision in Cross-Border Higher Education (UNESCO/OECD 2005). Other Norwegian academic institutions or institutes that are not accredited by NOKUT can apply for NORHED projects in partnership with a Norwegian Higher Education institution accredited by NOKUT.

---

[10] In cases where higher education institutions submitted more than 15 applications, the *NORHED Guidelines for Applications* required that the concerned institutions prioritize the 15 projects, in no specific order, with a brief justification of the list. The *Guidelines* noted that applications not on the list would still be considered; that there was no need to rank the 15 projects; and that only 10 out of the 15 prioritized projects could be directed to the same NORHED Sub-programme. In its 2013 document on NORHED, *Innstilling og vurderinger av søknader*, FORSK reported on its initial reasons for asking for prioritization and the problems encountered by the universities and by Norad in using this process (p.17). It does not plan to use the prioritization process in future. DPMG saw no evidence that the results of universities' prioritization processes affected the awards process in the first call-off.

3.4. *A Presentation* specified other priorities: a) countries identified for long-term bilateral collaboration with Norway, b) projects germane to one or more of the six sub-program areas,[11] c) priority regions for each sub-program area, and d) a broad range of interventions eligible for support. However, these priorities were treated not as eligibility criteria but as evaluation criteria in the NORHED evaluation grid (see Annex D).

3.5. FORSK conducted the partner eligibility screen for the population of 173 applications. It excluded 7 projects on eligibility grounds. The IRCs discovered another 2 ineligible applications, bringing the total to 9 and leaving a population of 164 applications to be evaluated. There were no patterns among ineligibles. They were scattered across 5 of the 6 sub-program areas, and their reasons for being ruled ineligible varied—e.g., the LMIC applicant lacked a Norwegian partner, the Norwegian partner was not NOKUT-accredited, or the LMIC institution was not accredited. A tenth case could be considered ineligible. The IRC panel found that the two main partners had withdrawn after the application had been submitted. They rated the project a "3" because it was incomplete.

B. Independent Review Committee Screens

3.6. FORSK assembled six teams of experts to evaluate the applications, an evaluation strategy that sits squarely within the mainstream of grant award processes used in France (e.g., Agence Nationale de Recherche), the United Kingdom (e.g., Medical Research Council), Canada (e.g., Canadian Institutes of Health Research), the United States (e.g., National Science Foundation, National Institutes of Health), Australia (e.g., National Health and Medical Research Council), and the European Union (e.g., European Science Foundation).

3.7. Published studies of the expert review process reveal variability in reviews or ratings when two different expert groups are asked to evaluate the same grant proposals (e.g., Marsh, Jayasinghe, and Bond, 2008; Bornmann, 2011; Abdoul et al., 2012). However, grant review panels do considerably better than mere chance in choosing research proposals that result in research which eventually generates the most citations.[12] The consensus seems to be that peer review is imperfect, but that there is no better alternative as long as potential flaws, such as conflicts of interest, are avoided.

3.8. The Abdoul et al. study found that several factors helped reduce the scope for subjective judgments by reviewers. Carefully specified application forms, well-defined evaluation checklists, and instructions on how each evaluation criterion should be weighted increased inter-rater agreement by reducing the scope for subjective judgments by reviewers. FORSK followed several of these practices: it used a uniform application form and an evaluation grid or checklist, and it specified the maximum points for each criterion. However, as discussed below, the NORHED application package and evaluation grid need improvement.

3.9. The six IRCs assessed the quality of the 164 eligible applications against the criteria in NORHED's evaluation grid. There was one IRC for each sub-program. The members of the IRC conducted their work in two phases: prior to Oslo and as a committee-of-the-whole in Oslo. Prior to Oslo FORSK followed the good practice of having each application assessed by three members of the IRC for the sub-program to which the applicant applied. One of the three readers was designated as the "main reader" and this individual completed the evaluation grid. When the IRCs assembled in Oslo, the main reader for each application presented the sub-group's' evaluation to the full IRC for discussion by the full panel. The full IRC then decided how to categorize the application: fund, fund with reservations, not fund.

3.10. FORSK is to be commended for how they assembled the IRC panels. Anticipating that many qualified Norwegians would have to be excluded on grounds of conflicts of interest, they obtained recommendations on experts from a wide network of organizations in Norway and internationally, including

---

[11] These sub-program areas were: Education and Training; Health; Natural Resource Management, Climate Change and Environment; Democratic and Economic Governance: Humanities, Culture, Media and Communication; and Capacity Development in South Sudan.

[12] This citation measure does not control for the fact that the research generated by winning proposals was funded. Losing proposals may still generate research, perhaps funded by other sources, but not necessarily at the same levels as the winning proposals.

higher educational institutions domestically (through the Norwegian Association of Higher Education Institutions) and from abroad. They sought candidates that had expertise within the thematic area, knowledge of higher education and research, knowledge and experience with the countries and regions relevant for the program, and a balance in relation to gender and background (country). FORSK carefully screened all possible candidates for potential conflicts of interest, a step that, as expected, excluded many relevant Norwegian candidates. A review of a sample of the *curriculum vitae* of panel members indicates that panel members were technically qualified in the domains of concern to NORHED.

3.11. As Table 1 shows, the three larger thematic domains had somewhat larger panels than the three small domains. However, the good practice of assigning each application to three readers prior to assembling in Oslo meant that panel members in the large domains had heavy loads, with the health sub-program having the heaviest. Each panel member in the health IRC had to read 27 applications and act as the main reader on 9 applications.

Table 1: Distribution of Eligible Applications among Sub-Programs

Sub-program	Eligible applications	# of panel members per IRC	Ratio of panel size to # of eligible applications	Workload per panel member as one of 3 readers	Workload as main reader
Education	28	4	7	21	7
Health	36	4	9	27	9
Natural Resources	58	7	8	25	8
Governance	15	3	5	15	5
Humanities	17	3	6	17	6
South Sudan	10	3	3	10	3+
Total	**164**	**24**		**492**	

3.12. Panel members had a month (mid-April to mid-May 2013) between receiving their applications for review and assembling in Oslo with their evaluations completed. Each application package had six pieces to review: the main application, the budget, the implementation plan, a partner assessment, a letter of commitment, and CVs. For each application package, two readers were paid for two hours each to review the package, and the main reader was paid for four hours to review and complete the evaluation grid for that package. Obviously, hours paid do not indicate hours spent on a task. However, in its independent evaluation of applications (see section 3), DPMG found that two hours for the co-readers was the minimum needed and that the main reader task took between six and eight hours.

3.13. The bigger problem was the time available for the IRC panels when they met in Oslo as committees-of-the-whole to make funding recommendations for the entire portfolio of eligible applications in their respective sub-programs. Four IRCs had two days each (Education, Health, Governance, and Humanities). The Natural Resources IRC had 3 days; the South Sudan IRC, 1.5 days. At this juncture the workload for the larger sub-programs was stupefying. The Education IRC had to discuss and decide on 14 applications per day, the Health IRC on 18 applications per day, and the Natural Resources IRC on more than 19 applications per day. It has to be assumed that sheer mental exhaustion may have affected the quality of these panels' evaluations.

3.14. Especially for the Health and Natural Resource sub-programs, FORSK either needed additional IRC panels or additional days in Oslo for the panels. Although FORSK considered adding a second IRC panel for the large sub-programs, they decided against this option in order to eliminate the variability that is introduced by using two IRC panels. FORSK's decision to stay with one IRC per sub-program was certainly defensible--and, had these IRCs been able to work longer in Oslo, it would have had no downsides.

3.15. When the IRCs assembled in Oslo, they had a brief two-hour discussion of the IRCs' deliverables for each application. There were no training sessions to ensure that all panel members understood application packages and evaluation grid criteria appropriately and in the same way—a practice that

reduces inter-rater reliability problems. FORSK notes that the ToR and the Contracts for the IRC members, the information package on the NORHED program, Norad's results manual, and the evaluation grid guided the IRCs' work. Although these shared documents undoubtedly helped to create more uniform approaches to the evaluations of the applications, they do not substitute for structured training in actually evaluating application packages.

3.16. In fact, the training process probably should be done prior to Oslo, given that readers are assessing the applications prior to Oslo. FORSK could select 2 or 3 applications in each sub-program and ask all members of the IRC for that sub-program to evaluate them independently by a certain date. Once all draft evaluations have been received for all "trial" applications, FORSK can: a) tabulate the scores to show variations among team members for each application; b) send the completed draft evaluation grids to all members so that they can see whether different members were picking up different issues or just rating criteria more or less harshly; and c) hold a conference call with all IRC members to discuss each "trial" application in order to surface reasons for variations between raters for the same criterion and the same application. This process has the advantage of alerting both IRC members and FORSK staff to sources of undesirable variation. In addition, no matter how carefully FORSK revises the evaluation grid, reviewing trial evaluations will almost inevitably surface confusion about the intent and application of certain evaluation criteria.

3.17. The completed evaluation grids prepared by the main readers were the basis of discussions by the IRC as committee-of-the-whole. The deliverable for each application for the IRCs was to complete a sheet that stated:

- Their categorization of the application as "1" (fund), "2" (fund with reservations), or "3" (do not fund);
- Their total score for the application, arrived at by adding up the points allocated to each of the nine criteria in the evaluation grid;
- A brief narrative of at least 100 words on the main reasons for the IRC's recommended course of action on the application.

3.18. As previously noted, the IRCs were not expected to submit a final and completed evaluation grid that recorded the points allocated to each criterion and an explanation of the IRC's reasoning for the number of points allocated to that criterion. The draft evaluation grids prepared by the main readers were also not collected, let alone revised, to reflect discussions of the full IRC. Thus, there is no record of the IRCs' detailed reasoning. Only the relatively short narrative could be used to determine if an evaluation properly reflected the quality of an application package. For this reason, there was not enough information to assess whether the IRCs' textual explanations and point allocations were logically consistent with each other. Norad also could not share the completed and finalized grid with applicants to help them learn what they had done well and where they could improve for future calls.

3.19. As Table 2 shows, different IRCs used different ranges for each of the three categories of 1, 2, or 3. There were no fixed cut-off scores for classifying an application as a 1, 2, or 3, each IRC being allowed to set its own cut-off scores. This flexibility was reasonable.

Table 2: Variation in Cut-off Scores by Categorization and Sub-program Area

Sub-program area	Variation in scores by categorization		
	1	2	3
Education	70-85	55-65	23-50
Health	74-90	70-73	28-67
Natural resources	80-85	68-80	38-66
Governance	80-90	66-76	48-65
Humanities	75-94	60-73	0^1-48
South Sudan	85-87	No application categorized a"2"	54-64

[1] The zero here refers to the application where the two main partners withdrew after the application had been submitted. The IRC defined it as an incomplete application, gave it a zero score, and rejected it for funding.

3.20.   Each IRC had a chair and a member of FORSK joined each IRC in the role of observer and to answer questions. However, there was no independent moderator. Thus, no one filled the role of ensuring that: a) the IRC panel applied the criteria consistently across evaluations; b) scores were well-justified; or c) in coordination with other moderators, the different IRC panels were consistent with each other. FORSK has proposed that in future the IRC panel chairs could play this role. This is an excellent suggestion, assuming that the chairs' responsibilities are clearly specified and that the chairs receive some training together in discharging it.

C. Screens by Norad's Thematic Sections and by Norwegian Embassies

3.21.   Screens by the Norad's thematic sections and by the Norwegian embassies occurred simultaneously. Both parties reviewed only summaries of the applications and their IRC evaluations. Table 3 shows the sample of applications that each was asked to assess.

Table 3: Applications Evaluated by Embassies and Norad's Thematic Sections

Sub-program	Applications rated "1"	Applications rated "2"	
		Embassies	Thematic Sections[1]
Education	Both the embassies and thematic sections reviewed all applications rated "1".	All applications rated "2"	All applications rated "2"
Health		One application rated "2"	All applications rated "2"
Natural Resources		10 highest scored applications rated "2""	All applications rated "2"
Governance		No applications rated "2"	All applications rated "2"
Humanities		No applications rated "2"	No applications rated "2"
South Sudan		No applications rated "2"	No applications rated "2"

[1] Norad has no humanities thematic section. These projects were sent to the Section for Culture of the Ministry of Foreign Affairs for comment.

3.22.   NORHED used Norway's embassies to assess the relevance of and risks associated with applications being considered for funding. The embassies were asked to evaluate the following:

1) Assess project's relevance to national needs and priorities;

2) Assess project's relevance to Norwegian priorities and initiatives at country level;

3) Assess the Embassy's knowledge of and experiences with the institution applying, including the risk of corruption;

4) Assess other potential risk factors associated with the proposed project: political context, conflict/post-conflict situations, economic conditions, sensitive issues, etc.; and

5) Any comments and arguments on whether to support project proposals recommended in category 2.

3.23. Norad's thematic sections were asked to assess these questions:

1) Comprehensively assess the portfolio within the relevant sub-program to determine whether it reflects the priorities in the program and for Norwegian development cooperation within the specific field;

2) Identify and provide rationale for applications from category 2 judged especially relevant for support;

3) Assess individual applications where the IRCs disagreed.

3.24. FORSK summarized the comments by the embassies and Norad's thematic sections in a document that describes the awards process (*Felles Innstilling*). In general, these comments were germane to the awards process.

D. Final Selection of Winners

3.25. The director of Norad made the final decision about applications that would be funded, informing the Ministry of Foreign Affairs of the decision process prior to the announcement of the awards. These decisions were based on the recommendations from the selection process managed by Norad's FORSK. This final review looked for projects' relevance at the country level, relationships to other Norwegian priorities, risk factors, and thematic positioning within the portfolio.

3.26. At this stage there were no changes in the category 1 applications proposed for funding. One category 2 application was dropped due to the size of the project portfolio of the sub-program and budget limitations.

3.27. On the one hand, NORHED funding comes from the Norwegian taxpayers and should reflect the priorities of the elected Norwegian Government. As part of Government, Norad and the Ministry of Foreign Affairs are obligated to ensure that the final awards reflect Government policies and priorities.

3.28. On the other hand, this final stage could potentially distort a merit-based selection of applications. In this case this final screen had a minimal effect on the selection of the final awards. In addition, this concern is more theoretical than real for Norad than it might be for other organizations around the world because of the transparency culture of Norway's Government.

3.29. That said, theoretically, a better practice would be to protect the integrity of the procurement process by building political priorities into the eligibility criteria. In practice, building multiple priorities into eligibility criteria that are usually defined as yes/no might seriously restrict the universities that could apply. One possibility for addressing this problem would be to consider an approach similar to that used by the International Development Association (IDA). IDA uses a database, updated annually, that scores IDA-eligible countries on a range of factors. This database constitutes a performance-based allocation framework for allocating IDA resources among countries. The system assesses the quality of client countries' policies and institutions, and allocates IDA's resources on the basis of the IDA Performance Rating to ensure that good performers get a higher share of IDA's available resources.

3.30. Norad could create such a database that scores priority countries on issues of concern to Norway. The database could reflect both measures of countries' performance and their needs. It can measure countries' performances on country policies and risk factors important to Norway, such as a country's human rights record, policies and institutions for environmental sustainability, gender equity, equity of public resource use, or transparency, accountability and corruption in the public sector. It can also measure countries' needs, such as a fragile state status or higher education institutions that are low capacity. Norad could then allocate grant money among these countries on the basis of their scores on these criteria. Low-scoring countries would receive some funding, but not as much as higher-scoring countries. Only in exceptional circumstances would a country receive zero funding.

E. Final Distribution of Awards

3.31. Table 4 shows the final distribution of awards by sub-program, the IRC categorization, and the award rate relative to the number of applications in the sub-program. Although most sub-programs are tightly clustered between a 29 percent and 33 percent success rate, there is some variation in the success rate. Natural resources had the lowest success rate (22 percent), but the largest number of awards because it had the largest number of applications. Humanities had the highest success rate (36 percent) but also had a modest number of awards because it had a smaller number of applications.

Table 4: Final Distribution of Awards by Sub-program, IRC Categorization, and Award Rate

Sub-program	Number of Eligible Applications	Number of selected projects			Award rate (% successful)	Total NOK allocated (in millions)
		Total	**"1" score**	**"2" score**		
Education	28	8	6	2	29	123
Health	36	11	10	1	31	182
Natural Resources	58	13	7	6	22	217
Governance	15	5	5	0	33	87
Humanities	17	6	6	0	36	99
South Sudan	10	3	3	0	30	54
Total	**164**	**46**	**37**	**9**	**28**	**762**

IV. DPMG's Independent Evaluation of a Sample of Applications

4.1. As noted in Chapter I, DPMG had originally planned to assess a sample of the IRCs' evaluation grids for grant applications to determine: a) whether the IRCs had applied the criteria in the evaluation grid consistently within and between sub-program areas (a check on inter-rater reliability); b) whether the evidence that the IRC provided to justify its rating of each evaluation criterion was persuasive (checks on validity[13]); and c) whether there were patterns of strengths and weaknesses in applications that could alert FORSK to potential downstream implementation issues. However, since the IRCs did not submit a final and completed evaluation grid, there was no record of the points allocated to each criterion or an explanation of the IRC's reasoning behind the points allocated. The draft evaluation grids prepared by the main readers were also not collected.

4.2. Given the absence of these data, DPMG decided to conduct an independent evaluation of a proportional, stratified random sample of all eligible applications. Although we could not compare DPMG's detailed scores and reasoning with those of the IRCs, we could at least see if most of the IRCs and DPMG's categorizations came out approximately in the same place.

4.3. The sample was stratified by sub-program and by the IRCs' categorizations of the applications. The number drawn was proportional to the number of cases in each cell. Table 5 shows the structure of the final sample. It represents a 23 percent sample of the total eligible population of grant applications, a percent that meets standards for this type of exercise.

Table 5: Structure of stratified, proportional random sample for DPMG's independent evaluations

Thematic area	Categorization			Allocation of S. Sudan by thematic area	Total
	1	2	3		
Education	1	2	3		6
Health	2	1	5		9
Natural resources	2	4	7	2	15
Governance	1	1	1		3
Humanities	1	2	1	1	5
South Sudan[1]	1	0[2]	2	Reallocated to humanities and natural resources	
Total	8	10	19		37

[1] The South Sudan sub-program represents a fragile state focus, not a thematic area. Since applications for South Sudan had to select one of the five thematic areas as their priority, DPMG reallocated the sampled South Sudan applications to their thematic areas (natural resources and humanities). The evaluators for these projects had working experience in South Sudan or similar lower capacity, conflict-ridden countries in Africa.

[2] The Independent Review Committee for South Sudan categorized all applications either as "1" or "3". None was categorized as "2".

4.4. As noted earlier, the DPMG process duplicated the IRC process in some ways but not in others. Like NORHED, DPMG selected evaluators who approximated those assembled by FORSK for its IRCs: technical experts in the sub-program areas who had substantial development and evaluation experience. Unlike NORHED, DPMG used one reader to review each application and thus did not replicate FORSK's process of having three readers review each application. Each DPMG evaluator reviewed the entire sample for his/her sub-program area.

---

[13] The applications provided complex information susceptible to alternative interpretations. Thus, the validity issue was whether the IRC adduced reasonable arguments for its ratings, not whether a second evaluation might have emphasized different aspects of the same information referenced by the IRC to support its ratings.

4.5. Also unlike NORHED, to increase inter-rater reliability, each member of the DPMG team independently reviewed two applications, using the NORHED evaluation grid, before conducting its assessments. The entire team then assembled to discuss differences and causes of differences in their ratings of the two applications. At this juncture the team started a log that recorded problems identified with the application package and the evaluation grid itself. They added to this log continuously during the entire evaluation process.

4.6. Finally, DPMG did not have a full IRC panel that could function as a committee-of-the-whole. To simulate the effect of a discussion of each application by the full IRC panel, two steps were taken. First, the lead for the evaluation reviewed each of the 37 draft evaluations for plausibility, proper interpretation of the evaluation criteria, and sound justifications for the points awarded each criterion. Each draft evaluation was returned to the reader for revision. Second, when the lead's review was returned to the reader, the IRC's evaluation sheet for that application was also sent to the reader. Evaluators were instructed to review the IRC's sheet to determine if the reader had missed any positive or negative points identified by the IRC. Only in this case was the reader expected to revisit ratings of the criteria and possibly the categorization of the application. Readers were expressly told to alter their categorizations only if new information led them to revise their criteria scores and thus the project's standing.

4.7. DPMG considered replicating the NORHED IRC process in its entirety. It did not do so for cost/effectiveness reasons. In terms of cost, it would have been expensive to fully replicate the process. Replicating the feature of three readers would have doubled the cost of this feature alone, assuming that each of the two secondary readers was paid for 50 percent of the hours assumed for the main reader.

4.8. Replicating full-strength IRCs for the committee-of-the-whole discussions would have cost 3.5 times as much as using DPMG's simulated process. The Education and Health panels, which had four persons each, would each have required three additional people to bring them to the point of their committees-of-the-whole. Six persons would have had to be added to the Natural Resource panel, which had seven people. The Governance and Humanities panels, which had three persons each, would each have required two additional persons to create committees-of-the-whole. (Since DPMG had allocated the South Sudan applications to the thematic sub-programs, a full-strength IRC would not have been created here.) If the reader roles (main and two secondary) were divided among the number of individuals required for recreating full-strength IRC panels, the evaluation would also have become quite fragmented. Each person would have played the main or secondary reader role for a very small number of reviews. This situation would almost certainly have driven down inter-coder reliabilities within each panel.

4.9. In terms of effectiveness, DMPG was mindful of the consensus of published evaluations of peer review processes discussed earlier. DPMG judged that re-creating the NORHED IRC process faithfully would not necessarily have reduced the variability between the IRC and DPMG ratings. The applications being assessed are complex and susceptible to alternative judgments, depending on the backgrounds of the reviewers. For example, evaluators with extensive experience in a country would be much more aware of the contextual factors shaping the potential for success or failure of an application from a university in that country than those without their in-depth knowledge. In other words, variations between IRC and DPMG evaluations were expected *no matter how well the IRC and DPMG processes replicated each other.* DPMG judged that faithfully replicating the IRC process would not sufficiently reduce the natural variation expected from two different peer reviews to warrant the additional cost.

V.   Problems with the NORHED Application Package and Evaluation Grid

5.1.   Using the log created during the evaluation process, DPMG identified a number of design flaws for both the application package and the evaluation grid. Given these flaws, it is not clear that NORHED's award process selected the best projects from the eligible population of applications.

A. Application Package

5.2.   The main application was poorly structured in four ways.

- The organization of its requests for information did not let the applicant tell a coherent story about the project;

- The application did not elicit sufficient detail about the design of components and implementation arrangements to let IRCs and other reviewers judge whether the project could realistically be expected to implement;

- It did not solicit some information that is important for judging the quality of the application;

- It did not offer sufficient guidance to applicants that could help them avoid mistakes.

5.3.   Coherent and sufficiently detailed story. The structure of the application should "flow" in a way that lets the applicant clearly convey the information that reviewers need in order to be able to judge an application properly. Annex E has two tables. Table E.1 shows the structure of the main application that was used in the first call. Table E.2 shows a possible alternative structure for the main application. For example, the main application used in the first call starts by asking about the partnership. A possible alternative structure would first determine what the project hopes to achieve by project close (its development objectives), why these objectives matter and for whom, what activities or components will be implemented to achieve the objectives, and the detailed design of these components. It then asks how the applicant plans to implement the project. It is here that all of the partnership questions are asked—for example, how the selection of each partner in the partnership is expected to enhance the likelihood of achieving the project's objectives.

5.4.   Missing information. The application form does not ask about at least two dimensions that should affect reviewers' judgments of the application: enabling conditions and M&E arrangements. It does not ask for information on system-level and institution-level conditions found to enable capacity building at faculty or department levels. For example, is the partner university hampered by limitations in the national regulatory framework for higher education institutions that might affect project implementation?

5.5.   Although the application asks for the project's results framework, it does not ask for its M&E arrangements. For example, what data sources will be used to measure activities, outputs, and outcomes? How frequently will the values of indicators be updated?

5.6.   Guidance that would have helped applicants. FORSK created a very helpful example of a completed implementation plan for applicants. It also made available to applicants documents such as Norad's handbook on Results Management in Norwegian Development Cooperation and Norad's handbook on Assessment of Sustainability and key risk factors. However, the application package would have benefitted from guidance specifically tailored to the NORHED application, such as guidance on assembling the partnerships, a simple example of a results framework for a hypothetical NOHRED project, guidance on assessing the sustainability of capacity improvements that might be funded by NORHED, and the criteria against which the applications would be judged.

5.7.   There is substantial value, especially in higher education, in creating networks of universities, faculties, or research institutes, and FORSK used the concept of partnerships as a primary instrument for capacity building capacities in LMIC higher education institutions. However, more partners, other things being equal, will complicate implementation, especially when partners are from multiple countries. Such

complexity can increase transaction costs, confuse responsibilities and accountabilities, and slow (or stall) the implementation of a project.

5.8. Figure 2 shows that over a third of the 46 NORHED awards had only one LMIC partner, but almost a half had two or three partners, and 15 percent had four or five partners. The vast majority of awards had only one Norwegian partner, as shown in Figure 3. However, for the six cases with two or three Norwegian partners, five cases occurred in combination with two to four LMIC partners.

Figure 2: Awards by Number of LMIC Partners   **Figure 3**: Awards by Number of Norwegian Partners
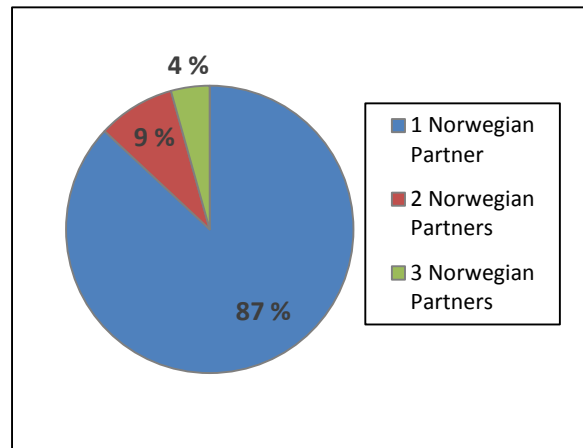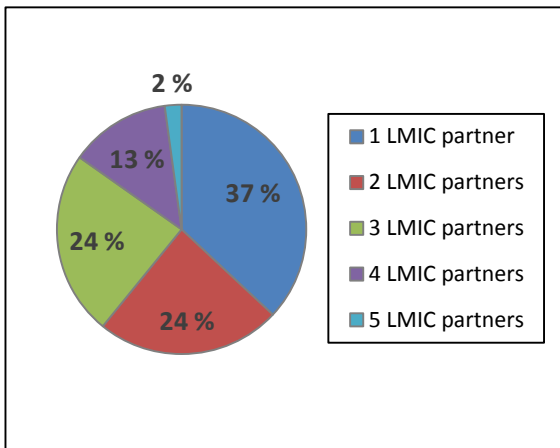
Source: DPMG proposal for call-off A-2.

5.9. FORSK now recognizes that applicants needed advice on assembling their partnerships. The main application form could have been structured to help applicants confront the transaction costs and accountability problems associated with multiple partners, especially when they are spread across different countries.

5.10. As explained in the next chapter, when DPMG evaluated a sample of applications, the lowest performance on the nine criteria of the evaluation grid was the quality of the applicants' results frameworks. Under a different task for Norad, DPMG assessed the revised results frameworks of grantees. (This task excluded losing applications.) To help grantees revise and strengthen their frameworks, the program staff for NORHED and the Results Management Section held a November 2013 workshop for grantees that conducted training in these frameworks. They also provided substantial technical assistance to grantees subsequent to the workshop.

5.11. Despite these efforts, DPMG still found that grantees had trouble writing an appropriate development objective. Some development objectives were too broad --"increase GDP per capita". Others just restated the immediate outputs, such as "more graduates with an MA or PhD". Applicants had trouble with key concepts. For example, they made the common mistake of confusing outputs (products, services or facilities that result from an organization's or project's activities) and outcomes (differences made by the outputs). Accordingly, they conflated or confused Sections 3.5 and 3.6 of the application, the first asking for baselines for outcomes and the second the baselines for outputs. Several applicants did not understand the concept of "target" in the results framework. They tended to give the final year of the grant (2018), not the targeted number to be trained, papers to be published, etc.

5.12. Clearly, applicants need guidance on how to prepare a results framework--*or something that serves its purpose but is simpler and intuitively easier to create.* Whatever device is used, applicants must think

through the causal path from activities to outputs to outcomes—de facto, the theory of change that underlies their project. They must figure out how to know whether the project is advancing or failing to advance along the path. This is hard work that demands analytic rigor. Unfortunately, there is no shortcut here. Projects based on muddy thinking will be at higher risk of failing--neither the grantee nor NORHED will know where the grantee is trying to go or whether what they are trying to do is even plausible. Although FORSK makes Norad's handbook on results frameworks available to applicants, a simple, concrete example based on a hypothetical NORHED project might help applicants.

5.13. Like any well-designed capacity development initiative, the NORHED program sought sustained capacity development. Section 4 of the main application asks applicants to assess sustainability after the project comes to an end. Although Norad has a guidance handbook on assessing sustainability, it probably would have helped applicants to see guidance specifically tailored to the types of sustainability challenges associated with NORHED projects—for example, retention of faculty trained under the project.

5.14. FORSK's plan to include the evaluation grid and the point allocations among the criteria in the application package for the second call is sound. Knowing the criteria and the point distribution among them provides additional guidance to applicants about where they should focus.

B. Evaluation Grid

5.15. Annex D shows the evaluation grid that the IRCs used to evaluate eligible applications for the first call. Although the results framework criterion (criterion 7) in the grid was well designed, the evaluation grid had a number of design flaws, as described below.

5.16. Unclear evaluation category. The IRC was asked to classify a project as a "1", "2", or "3". Category 1 was defined as "Proposal recommended for approval without changes (and no or only minor clarifications)." Category 2 was defined as "Proposal recommended for approval provided that clarifications or adjustments are met within a limited timeframe." Category 3 was defined as "Proposal not recommended for funding."

5.17. The definitions of categories 1 and 3, especially relative to each other, were clear. However, both the IRCs and DPMG found that the distinction between categories 1 and 2 was not clear. This lack of clarity introduces an undesirable potential for variability in the interpretations by reviewers, such as the IRCs.

5.18. Eligibility criteria treated as evaluation criteria. Four sub-criteria (2A, 2B, 3A, and 3B) should have been treated as eligibility criteria, not as evaluation criteria. (Box 1 shows the statements of these criteria.) Descriptions of NORHED's capacity development objectives and sub-program objectives in *A Presentation of NORHED* are sufficiently broad that almost all projects fit sub-criteria 2A and 2B. Sub-criterion 2C was properly treated as an evaluation criterion. In almost all cases sub-criteria 3A and 3B are de facto yes/no questions and thus eligibility questions. Thirty-five of the 100 total points that an application could earn were tied up in these two criteria. Since most applications met these criteria, the points that evaluators could in fact use to discriminate between applications were reduced from 100 points to 65 or 70 points. Evaluation criteria and points should meet the standard of helping Norad *discriminate* between applications.

Box 1: Definition of Criteria 2 and 3 in Evaluation Grid

---

2.   Project thematic relevance *(Section 2 of application form)*
A. Project relevance to NORHED's overall objectives of capacity development for higher education and research in developing countries (*refer to p.1-6 of document "A Presentation of NORHED"*)
B. Project relevance to the NORHED sub-programme(s) to which project is addressed (*refer to sub-programme descriptions on p. 18-26 in document "A Presentation of NORHED"*)
C. Project relevance to needs of institution(s), country and region where higher education institution capacity is to be developed *(section 2 of application form)*.

3.   Project geographic relevance *(section 2 of application form)*
A. Geographic relevance in terms of Norway's priority countries for bilateral cooperation *(refer to p. 9 of document "A Presentation of NORHED")*
B. Geographic relevance in terms of building capacity in NORHED sub-programme priority region. Note that Upper Middle Income Country-based projects will not score on this even if they are situated in a relevant region *(refer to p. 9-10 of document "A Presentation of NORHED" and the DAC list of ODA recipients)*.

5.19.   Docking points for upper-middle-income applicants. Sub-criterion 3B has a puzzling restriction. It states: "Note that Upper Middle Income Country-based projects will not score on this even if they are situated in a relevant region." DPMG appreciates that NORHED does not want its resources to help the "rich get richer." However, this criterion works against opportunities for South-South, not just South-North, capacity development arrangements. It seems preferable for Norad to require that an application with an upper-middle-income partner show in its budget and allocation of responsibilities how the upper-middle-income partner would be the source of capacity development for its LMIC partners.

5.20.   Unclear instructions for critical evaluation criteria. The whole technical design of the project components is judged in two sub-criteria that are vaguely stated. It is not clear, for example, what "quality" means.

- (Sub-criterion 4B) Quality of educational components of project in terms of strengthening capacity of LMIC(s) academic institution(s) to produce more and better candidates;

- (Sub-criterion 4B) Quality of research components of project in terms of strengthening the academic institution's capacity to conduct scientifically rigorous, nationally/regionally relevant research.

5.21.   Duplication. The grid has evaluation items that appear to duplicate each other: risks (sub-criterion 7C and sub-criterion 8A) and gender (criterion 6 and sub-criterion 8B). If FORSK sees these items as distinct, the wording in the evaluation grid needs to be clarified.

5.22.   Unclear weighting of sub-criteria. Most criteria have 2 to 4 sub-criteria, but the grid has no guidance on how FORSK wants evaluators to weight each of the sub-criteria in terms of points. Equally? At the evaluators' discretion?

5.23.   Questionable allocation of points. The grid's point allocations are puzzling. As already noted, it locks up—and virtually throws away—about 30-35 points in criteria 2 and 3, both of which are primarily eligibility criteria. Gender is important but receives only 5 points for criterion 6. (Additional points for gender can be earned under criterion 8, but as already noted, criteria 6 and 8 seem to duplicate each other.) Although the results framework (criterion 7), especially the baseline data, is extremely important, it gets only 10 points. Criterion 9 that includes the financial plan, budget, and implementation plan, gets only 5 points.

5.24.   Criterion 8 has four highly disparate and complex sub-criteria: a) relevance and quality of LMIC partner's institutional capacity analysis, including assessment of financial management, corruption and other potential risk factors, and relevant mitigation strategies; b) relevance and quality of gender equality

assessment in terms of project implementation and expected results; c) relevance and quality of conflict sensitivity assessment, human rights assessment, environmental/climate impact assessment, and other risk factors; and d) relevance and quality of project financial sustainability strategy, including sustainability and analysis of the LMIC lead partner's institutional capacity. However, this criterion is awarded only 10 points that have to be distributed among these complex components.

C. Alignment between the Evaluation Grid and Application Package

5.25. DPMG found three instances where the evaluation grid asks reviewers to judge something for which the application package does not ask.

- The evaluation grid, sub-criterion 8D, asks for a judgment regarding the "financial sustainability strategy" of the project. However, the application only asks for an "assessment of sustainability" without specifying *financial* sustainability. Although one would think that applicants' assessment of sustainability would include financial sustainability, evaluators should be asked to judge a factor that the applicant has been explicitly asked to address.

- Criterion 9 asks for both a budget and a financial plan. If "financial plan" is defined to specify how the costs of the project will be financed, the application only asks for a "detailed budget and breakdown of expenditures." It does not ask for a financial plan.

- Sub-criterion 9A asks for an assessment of the cost-effectiveness of the "financial plan." Aside from the definitional question about "financial plan," the application form does not ask for comparative cost information for the various proposed expenditures. Unless average cost coefficients, or something similar, are provided to the reviewers, this sub-criterion cannot be judged.

D. Alignment between the Theory of Change, Seed Funding Mechanism, Application Package, and Evaluation Grid

5.26. NORHED's purpose is to develop sustainable capacities of LMIC higher education institutions to perform their two basic functions of education and research. This purpose should be integrated into decisions and activities at all stages of the program, including the application and award processes.

5.27. In previous assessment work for NORHED, DPMG (Norad 2014b) developed **a theory of change**, based on international research, for increasing the capacities of higher education institutions to perform their educational and research functions. (See the diagram in Annex F.) The theory distinguishes three systems that are "nested": system-level enabling conditions (the tertiary education ecosystem within which universities develop),[14] institution-level enabling conditions,[15] and the causal links within universities that directly affect the number and quality of graduates and amount and quality of research.[16]

---

[14] At the system level, governance and financing are the most important factors: a) the degree of institutional autonomy and academic freedom enjoyed by higher education institutions; b) the accountability instruments (quality assurance and accreditation, assessment of learning outcomes, labor market results, research assessment) linked in one way or the other to performance measures of higher education institutions; c) student support and financial aid for students from disadvantaged groups; and d) the funding available for research and teaching, especially for the public universities that are heavily dependent on state resources.

[15] Institution-level factors that affect the university's capacity development are leadership, vision, and a development plan that includes a research strategy; internal quality assurance, which includes standards for student admission, standards for hiring and promoting faculty, and checks on the quality of learning and research; and resources and incentives for teaching, research, and student enrollment and retention (student aid).

[16] The third system consists of the factors that combine to produce the intermediate results of high quality and relevant instructional programs and dynamic research activities: appropriate qualifications, experience, and motivation of teaching and research staff; effective selection of incoming students, especially graduate students; favorable infrastructure for teaching and research; and linkages with industry and communities. The intermediate instructional and research results in turn directly affect NORHED's measures of capacity development (more and better graduates and more and higher quality research).

5.28.   In general, the reasoning behind the NORHED program was found to parallel this theory. For example, most activities eligible under the NORHED main call (Norad, 2012) target the third system (the causal links that directly affect the quality and number/amount of graduates and research).[17] Two were germane to the second system (institutional level).[18]

5.29.   FORSK's **seed funding mechanism** was entirely consistent with the NORHED program objective of capacity development. Seed funding was designed to help lower-capacity applicants prepare stronger proposals than they otherwise could have done. However, as noted earlier, whether this mechanism met its intent or not could not be determined. It would be necessary to discriminate convincingly between lower- and higher-capacity LMIC applicants for the main call in order see if lower-capacity applicants that got seed funding submitted better proposals than lower-capacity applicants that did not get seed funding.[19] Unfortunately, NORHED did not collect much information for each applicant on system and institution enabling conditions that might have been used to differentiate lower from higher capacity applicants.

5.30.   The application package and evaluation grid were not misaligned with the theory of change, but they were poorly designed to provide information required to assess applications relative to the theory of capacity development. For example, the application package did not require much information on the enabling conditions that the theory of change identifies as conducive to building capacity.[20] Knowing a grantee's capacity development status would have alerted FORSK and Norad to the capacity development challenges associated with each application. Funding very low capacity LMIC universities represents potentially high risk and high payoff. Presumably, Norad would want a portfolio that balances higher-risk with lower-risk grantees. Information about this status could also alert FORSK to grantees that should be supervised more closely.

5.31.   The **application package** does ask applicants for their results frameworks, and these frameworks constitute the theories of change underlying the projects. When these frameworks are reasonably well developed, they can be examined to determine if the projects proposed could plausibly produce enhanced educational and research capacities. As discussed earlier, the problem is that the results frameworks submitted were often weak.

5.32.   In terms of the **evaluation grid**, criterion 1 does ask for an assessment of the capacity-building aspects of the partnership. ("Partnership design including clarity of roles and responsibilities in terms of institutional capacity development objective, as well as relevance and competency of persons involved.") If the application package had asked how the applicant's implementation arrangements would allow the more capable partners to develop the capacities of the less capable partners, criterion 1 could have been more tightly tied to the capacity development purposes of the program.

5.33.   Criterion 8 asks about financial sustainability as one of four sub-criteria. However, as noted earlier, the application package does not ask applicants for information on financial sustainability, and the evaluation grid does not evaluate information on other factors that feed into sustainability, such as political support, likelihood of retaining key faculty, etc.

---

[17] These were in-country/ regional Masters and Bachelor education programs; PhD programs and post-doctoral doctoral fellowships; joint research projects in line with overall NORHED program aims and sub-program areas; and scientific equipment and small-scale infrastructure.

[18] These were strengthening of institution and systems (especially managerial and financial), and systems for knowledge management, information, and dissemination of results.

[19] Even then, the analysis would still be flawed because of the unmeasured variable problem.

[20] It does ask for information on three items related to enabling conditions: assessment of how project relates to the national and partners' institutions' strategic plans and priorities (in countries where capacity will be developed); assessment of partners' institutional capacity, including risk factors; and assessment of financial management and corruption.

VI.   Results of DPMG's Assessment of Applications Relative to IRC Results

6.1.   As previously noted, the literature shows that two different, but expert, panels will vary in their judgments of complex materials that are susceptible to different interpretations. Thus, we expected some differences between the IRC and DPMG panels in their categorization decisions. The question was whether the two sets of reviews were wildly different or varied within a tolerable range. In DPMG's judgment, the differences between the ratings by the IRCs and DPMG were within a tolerable range.

6.2.   Annex G records the data for DPMG's evaluation of a proportional, stratified, random sample of 37 eligible applications. The Annex includes not only DPMG's categorization and score for each application, but the scores awarded on each of the 9 criteria in the evaluation grid.

6.3.   Comparisons of scores. Table 6 compares the total scores awarded these 37 applications by sub-program. Except for the education sub-program, the average scores for each sub-program for the IRC and DPMG are reassuringly similar, although there was variation between them for individual applications within the sub-programs.

Table 6: Average Scores by Sub-program for DPMG and IRCs for the Same Applications

Sub-program	DPMG	IRC
Education	81.2	54.5
Health	69.9	62.3
Natural resources	68.8	67.5
Governance	63.7	65.0
Humanities	70.0	72.3
South Sudan	66.3	65.3
All sub-programs	70.0	64.2

6.4.   In the education case DPMG's expert consistently rated projects higher than the IRC--in one case, the difference was 48 points. In the absence of the completed evaluation grids for this IRC, specific sources of the differences could not be determined. For example, the DPMG expert and IRC for education might have been in relative substantive agreement, but the DPMG expert an "easier" grader, although this was not evident in the training session for DPMG's expert evaluators. The lead for this exercise and the DPMG expert for education had to rely on the IRC's narrative to identify possible sources of difference. Using this data source, the differences seemed to stem more from the value placed on the project concept and different weights accorded other aspects of the applications.[21] Interestingly, Table 7 shows that this expert's categorizations of the education applications were about as close to the categorizations by the IRC as were DPMG's categorizations for the other sub-programs.

6.5.   Comparisons of categorizations. More important than similar scores is whether the IRCs and DPMG categorized the applications in the same way. Table 7 addresses this question by sub-program. It shows that the IRCs and DPMG categorized almost half (46 percent) of the projects the same--both rated it a 1, a 2, or a 3.

6.6.   They differed by one category for more than a third (38%) of the sample--the IRC categorized the project as a 1 and DPMG as a 2 or vice versa (14%), or the IRC categorized the project as a 2 and DPMG

---

[21] For example, in one application the IRC seemed to downgrade the application because there was no evidence that the Norwegian partner had competence in Spanish, the native language of the LMIC partner. However, these two partners had been working together for several years. DPMG's expert took this as sufficient evidence that there was not a language problem.

as a 3 or vice versa (24%). They differed by two categories for about a sixth of the sample, with the IRC categorizing the project as a 1 and DPMG as a 3 or vice versa (16%).

Table 7: Consistency and Inconsistency between IRCs and DPMG in Categorization of Projects

Sub-program	IRC and DPMG agree on categorization	IRC and DPMG differ by 1 category	IRC and DPMG differ by 2 categories
Education	3 (50%)	2 (33%	1 (17%)
Health	5 (63%)	1 (13%)	2 (25%)
Natural resources	3 (23%)	8 (62%)	2 (15%)
Governance	1 (33%)	1 (33%)	1 (33%)
Humanities	2 (50%)	2 (50%)	0
South Sudan	3 (100%)	0	0
Total # of applications	17	14	6
% of sample	46%	38%	16%

6.7. At least at the margins, it was difficult for all reviewers—both the IRCs and DPMG—to discriminate between categories 1 and 2. Thus, the most meaningful differences occur in two situations.

1) The IRC or DPMG categorizes the project as a 2 and the other categorizes it as a 3.

2) The IRC or DPMG categorizes the project as a 1 and the other categorizes it as a 3.

6.8. In all of these cases either the DPMG or the IRC would not have funded the project and the other would definitely have funded it ("1" cases) or might have funded it ("2" cases). Cases where DPMG would definitely or might fund and the IRCs would definitely not fund are potential missed opportunities. Cases where the IRCs would definitely or might fund and DPMG would definitely not fund are potential mistakes. As Table 8 shows, about two-fifths of the total sample fell into these two situations.

Table 8: Differences between the IRC and DPMG in Cases Where One Does Not Fund the Project

Sub-program	IRC codes as 2; DPMG codes as 3	DPMG codes as 2; IRC codes as 3	IRC codes as 1; DPMG codes as 3	DPMG codes as 1; IRC codes as 3
Education	1	1		1
Health			1	1
Natural res.	1	3		2
Governance	1		1	
Humanities	2			
South Sudan				
Total # of applications	5	4	2	4
% of sample	14%	11%	5%	11%

6.9. In 8 cases or 22 percent of the sample, the IRC would definitely not fund the project and DPMG would or might fund it. These cases represent potential missed opportunities. In 5 cases or 14 percent of the sample, the DPMG would definitely not fund the project and the IRC might fund it (the "2" cases). *However, none of these five cases were in fact funded* and thus do not represent potential mistakes. In 2 cases or 5% of the sample, the IRC would definitely fund the application and the DPMG would not fund it. These cases constitute potential mistakes.

6.10. Comparisons between criteria. We cannot compare the IRCs' and DPMG's ratings by criterion because we lack the IRCs' complete evaluation grids. However, we can examine DPMG's ratings. Figure 4 and Table 9 show that, in DPMG's view, the sampled projects performed the worst on criterion 7 (the results framework) and on criterion 8 (the criterion with very disparate sub-criteria).

Figure 4: DPMG's Average Score on the 9 Criteria of the NORHED Evaluation Grid
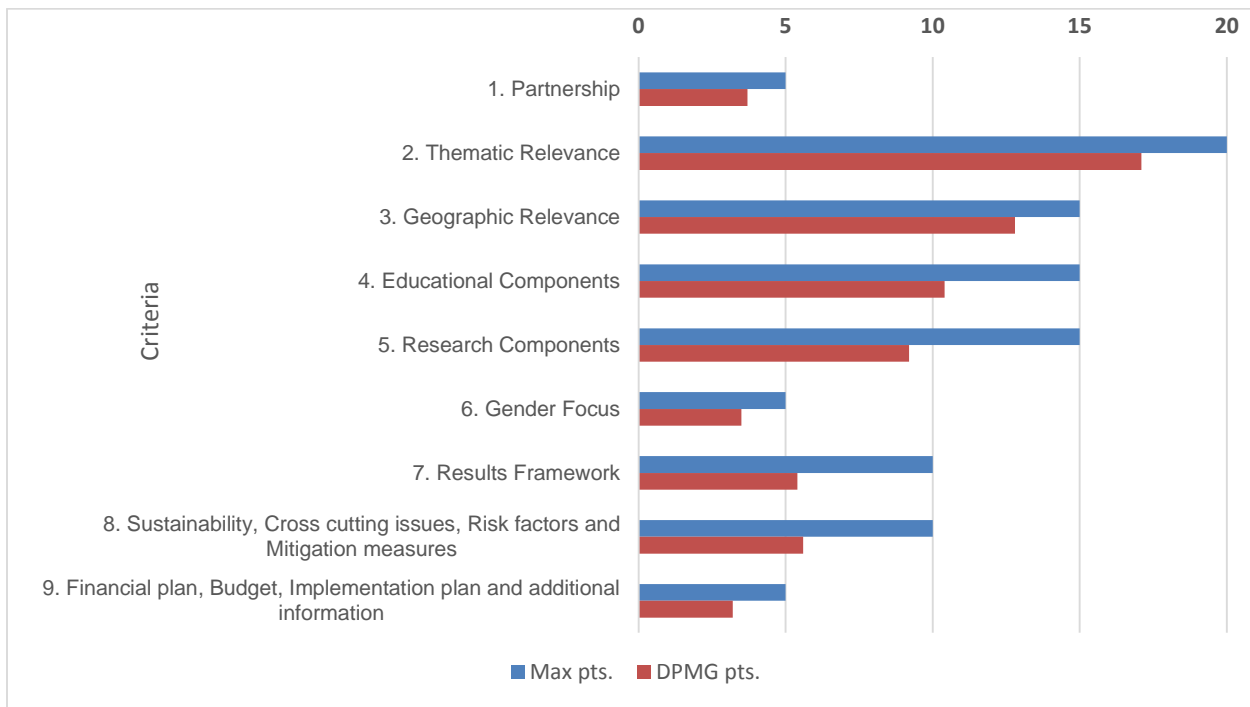
Table 9: DPMG's Average Score of the 9 Criteria of the NORHED Evaluation Grid by Sub-program

Sub-program	Criterion number								
	C.1	C.2	C.3	C.4	C.5	C.6	C.7.	C.8	C.9
Education	3.8	19.0	14.0	12.5	12.3	3.5	5.5	6.5	4.0
Health	4.3	17.8	12.9	11.4	6.9	4.1	3.8	5.1	3.8
Natural res.	3.8	17.3	10.9	9.2	10.3	3.6	6.6	5.2	2.3
Governance	4.0	10.7	15.0	6.7	8.3	3.7	4.3	7.3	3.7
Humanities	2.8	18.8	13.0	11.5	6.0	3.5	5.8	6.0	2.8
South Sudan	3.0	15.3	13.3	10.7	9.3	1.7	5.0	4.3	3.7
All sub-programs	3.7	17.1	13.2	10.4	9.2	3.5	5.4	5.6	3.2
Maximum points obtainable	5	20	15	15	15	5	10	10	5

Key: See figure 4 for the key for the criteria.

6.11.   It is noteworthy that applications also lost about a third of their points on two criteria (4 and 5) that assess the quality of the design of the project's educational and research components—factors that are critical to achieving the project's development objective. Each of these criteria gets 15 points, and each is split between two sub-criteria: relevance to the NORHED sub-program as specified in *A Presentation of NORHED* and the "quality" of components. Because most applications met the broadly defined relevance criteria, low scores reflected weaknesses in the design of the components.

VII. Conclusions and Recommendations

A. Conclusions

Seed funding

7.1. Receiving seed funding and receiving a grant award are strongly related across all applications regardless of sub-program. To the extent that seed funding improves the chances of receiving a grant award, it occurs entirely through its impact on the quality of the proposal. It appears that applicants that received seed funding wrote significantly stronger proposals, which ultimately had much higher chances of receiving funding. However, this finding may have nothing to do with seed funding and everything to do with selection bias.

7.2. It was not possible to determine if lower-capacity applicants benefitted from seed funding since there were inadequate data to differentiate lower-capacity from higher-capacity applicants.

Awards process

7.3. With some exceptions, the awards process itself was well conducted and in line with good international standards. Norad's unit responsible for the NORHED program (Section for Research, Innovation and Higher Education) shows that it learns from experience in order to improve future calls.

7.4. The comparison of DPMG's independent evaluation of a random sample of applications and the results of the IRC panels showed a variation in scoring and categorization that was reasonable in most cases. However, it highlighted potential missed opportunities and potential errors for about two-fifths of the total sample. In 8 cases or 22 percent of the sample, the IRC would definitely not fund the project and DPMG would or might fund it. These cases represent potential missed opportunities. In 5 cases or 14 percent of the sample, the DPMG would definitely not fund the project and the IRC might fund it (the "2" cases). *However, none of these five cases were in fact funded* and thus do not represent potential mistakes. In 2 cases or 5% of the sample, the IRC would definitely fund the application and the DPMG would not fund it. These cases constitute potential mistakes.

7.5. The application package that evaluators were asked to assess and the evaluation grid that they were asked to use in their assessments were both poorly designed. These problems markedly increased the chances of evaluation mistakes—either by missing promising applicants or by selecting less promising ones.

B. Recommendations

1) Substantially revise the application package, especially the main application, as well as the evaluation grid, to eliminate their weaknesses.
2) To the extent possible, move political considerations in the selection process to the design of the main call as eligibility criteria.
3) Enlarge the number of IRC members, as needed, in their "reader" roles to ensure the panels have more balanced workloads.
4) Ensure that each IRC has a reasonable workload when meeting as a committee-of-the-whole, preferably by adding days in Oslo for the IRC.
5) Require IRCs to submit a completed evaluation grid for each application that reflects the IRC's discussion of the draft grid in order to conduct checks on inter-rater reliability and validity. FORSK could use the main readers to revise their draft evaluation grids to reflect considerations that were raised in discussions by the entire IRC.

6) Review the evaluation grid with all evaluators before they start their deliberations, clarifying the intent of each criterion and alerting them to possible misinterpretations.

7) In line with a proposal of the NORHED program staff, define the role of each IRC panel chair as the independent moderator that assures the quality and consistency of the IRC's work. Conduct modest training of the panel chairs for this role.

References

Abdoul H, Perrey C, Amiel P, Tubach F, Gottot S, et al. 2012. Peer Review of Grant Applications: Criteria Used and Qualitative Study of Reviewer Practices. PLoS ONE 7(9): e46054.

Armstrong, D., A. Gosling, J. Weinman, and T. Martaeu. 1997. "The place of inter-rater reliability in qualitative research: an empirical study". *Sociology*, 31 (3), 597-606.

Bornmann, L. 2011. "Scientific peer review." *Annual Review of Information Science and Technology, 45*, 197-245.

Doran, George. 1981. "There's a S.M.A.R.T. way to write management's goals and objectives". *Management Review* (AMA FORUM) 70 (11): 35–36.

Gwet, K. 2014. *Handbook of Inter-Rater Reliability,* 4th Edition*: The Definitive Guide to Measuring the Extent of Agreement among Raters.*

Kapeller Jacob. 2010. "Citation Metrics: Serious Drawbacks, Perverse Incentives, and Strategic Options for Heterodox Economics." *American Journal of Economics and Sociology*, 69 (5).

Kusek, Jody Zall, and Ray C. Rist. 2004. *Ten steps to a results-based monitoring and evaluation system: a handbook for development practitioners.* Washington DC: World Bank.

Marsh, H.W., Jayasinghe, U.W., Bond, N.W., 2008. "Improving the peer-review process for grant applications: reliability, validity, bias, and generalizability." *American Psychologist* 63 (3), 160–168.

NORAD. 2012. *A Presentation of NORHED: The Norwegian Programme for Capacity Development in Higher Education and Research for Development.* Oslo.

NORAD. 2013. *NORHED 2013. Innstilling og vurderinger av søknader (translated as "Selection and Assessment of NORHED Applications*"). Oslo.

NORAD. 2014a. *Erfaringsnotat for NORHED-programme* (translated as "*An experience memo for the NORHED program"* or *"Lessons Learned about the NORHED program*"). Oslo.

NORAD. 2014b. *Evaluation Series of NORHED: Theory of Change and Evaluation Methods.* Oslo.

Schiavo-Campo, Salvatore. 1999. "'Performance' in the Public Sector." *Asian Journal of Political Science 7(*2): 75–87.

Baseline study of NORHED
Call-off A2

Terms of References (30th of September 2014)

1 Background

The Norwegian Programme for Capacity Development in Higher Education and Research for Development (NORHED) is operated by the Norwegian Agency for Development Cooperation (Norad) and has an annual budget of approximately 130 million NOK. The objective of NORHED is to contribute to development in the recipient countries by strengthening capacity in higher education institutions. In the programme document, this is operationalized as follows: *"The purpose of the programme is to strengthen capacity in higher education institutions in LMIC* [low and middle income countries] *to contribute to a) a more and better qualified workforce, b) increased knowledge, c) evidence-based policy and decision-making and d) enhanced gender equality. The future NORHED impact is sustainable economic, social and environmental development in low and middle-income countries. (…) Strengthening of higher education institutions refers to: 1. Producing more and better research relevant to the identified areas/sub-programmes [and] 2. Producing more and better qualified graduates, men and women, relevant to the identified areas/sub-programmes."* (Norad, 2013:4).

In the first call-off, NORHED received 173 applications[22] for project support, out of which 46 will receive funding. The projects are owned by 26 higher education institutions from 16 low- and middle income countries, and 10 higher education institutions from Norway (Norad, 2013). The higher education institutions have been awarded project funding for 750 million NOK over five years. The first disbursements were made in December 2013 and the first projects started in January 2014. Most projects are situated in Uganda, Ethiopia, Malawi and South Sudan. NORHED has developed an overall results framework and all project applicants have been asked to develop independent results frameworks in accordance with this overarching framework. To aid this process, all recipients of NORHED funding were invited to participate in a workshop in Addis Ababa in March 2014. The workshop resulted in a set of standard indicators (see reporting template for standard indicators (Norad, 2014a). The majority of institutions were required to submit a revised version of this framework by the 31st of March 2014, a deadline that was later extended to the 11th of April. After this, the NORHED administration has been in dialogue with the project owners to ensure that the results frameworks adhere to NORHED's overall results and monitoring framework. About half of the projects will hand in their first annual report in September 2014, and the first yearly project meetings will take place between October and November 2014. Many of these meetings are likely to take place at the South-led institutions.

Norad's Evaluation Department has commissioned a real-time evaluation of NORHED. The evaluation consists of a series of studies to be conducted over the period 2014-2017 by the University of Southern California's Development Portfolio Management Group. The purpose of the real-time evaluation is to enable learning within the field of higher education and development, so that future investments can become more effective in building capacity in higher education institutions as a contribution to California's Development Portfolio Management Group.[23] The purpose of the real-time evaluation is to enable learning within the field of higher education and development, so that future investments can

---

[22] As stated in the TOR for the framework agreement: all applicants should have accepted to be part of an evaluation, including those who did not receive funding.

[23] Norad's Evaluation Department has entered into a framework agreement with the University of Southern California's Development Portfolio Management Group. The agreement is for 2 years, and Norad's Evaluation Department has an option to renew the contract twice, for one year at a time.

become more effective in building capacity in higher education institutions as a contribution to development. The evaluation programme will increase knowledge about the extent to which NORHED builds capacity in higher education institutions, how capacity development of higher education institutions can be conducted most effectively and how higher education institutions affect development. This is the second call-off under the real-time evaluation.

2 Purpose and scope

The purpose of the assignment is dual: first to ensure that changes in institutional capacity (positive and negative) caused by the NORHED programme can be evaluated and plausibly attributed to the programme; and second to allow for learning from choices made early on in the programme period. This will be done through assessments of the existing monitoring and results framework and the selection mechanism for awarding funds, and by undertaking a baseline study of the current level of institutional capacity in institutions.

3 Objectives

3.1 PHASE 1: Inception

3.1.1 Assess the existing monitoring and results framework
In general, the evaluation team shall be specific about whether indicators and other data are at the level of input, output, outcomes and impact. The evaluation team shall respond to, but need not limit itself to, the following questions and topics:
a) Relevance of standard indicators: Using the general theory of change developed during call-off A1 as reference and taking into account costs associated with monitoring and evaluation; are the chosen standard indicators and the individual programme indicators developed by NORHED and the projects, appropriate for determining the effectiveness of the programme? Are any important indicators left out? If so, please be specific.
b) Operationalization of standard indicators: Is the operationalization of indicators sound? Is the definition of indicators specific enough to ensure that the same phenomena is measured across institutions? Are the operationalization of indicators likely to be biased?
c) Data quality: How will project managers go about to collect these data? Is it realistic that they can be collected as part of the routine monitoring/reporting procedures? Are they likely to be accurate?
d) Project-specific indicators: For project-specific indicators proposed by the individual NORHED projects, the usefulness of the indicators as such (not the operationalization and monitoring mechanisms) shall be assessed.
e) Need for additional data: Assess to what extent existing data (archives, registers, programme documents, etc.) can be used to undertake a descriptive analysis of the baseline situation and will allow for a future study of the effectiveness, or whether additional data need to be collected. The team shall clearly state if additional information is required and propose a method for data collection.
f) Update the evaluation team's database of NORHED-projects to take into account recent changes of the projects.

3.1.2 Describe the award mechanism
This part of the assignment shall include a description of the award mechanism[24] for the first NORHED call for proposals, including identification of award criteria, and mapping the process including relevant documents. The aim is to prepare for an assessment of the award mechanisms as such (3.2), and to make a recommendation regarding the use of non-grantees as a comparison group (3.1.3 b).

---

[24] The process of selecting 46 grantees out of 173 applicants.

3.1.3 Propose detailed methods to determine effects

a) The evaluation team shall propose a detailed case-study design, including specifying methods, and discuss whether the units of analysis should be institutions or projects. See recommendation in the report "Theory of Change and Capacity Development in Higher Education Institutions" (call-off A-1) (p.38), to evaluate "a sample of projects that have been implementing for a sufficient number of months to allow progress and problems to surface".

b) The evaluation team shall discuss which of the indicators that lend themselves to analysis using quasi-experimental design, and propose methods, including sample size strategy, data collection if required and costs. The sampling strategy should specify the control group, the number and level of cluster, units per cluster, the calculation of the intra-cluster correlation coefficient, and the minimum detectable effect of the outcome variables of interest for different powers. Please note that equations should be specified in an appendix.

c) Based on the assessment of available data, the evaluation team shall propose a detailed data collection plan for both qualitative and quantitative data, to ensure that effects can be documented and explained. In the report: *Theory of Change and Methods for Evaluating NORHED* (2014b), the evaluation team points out that the success of NORHED depends in part on system and university level enabling factors. The data collection strategy shall include a plan to detect whether such factors are likely to aid or impede NORHED funded projects from reaching their development objectives. In addition, the team shall discuss strategies for detecting adverse effects.

3.2 PHASE 2: Baseline

a) Collect additional data through open sources, interviews, surveys etc., as agreed during the inception phase.

b) Undertake a descriptive analysis of the baseline situation of the NORHED programme, taking the institutional context within which the institutions are operating into account. Unless otherwise is agreed during the inception phase, this analysis shall include the baseline situation for both institutions that applied for and were awarded funding and for institutions that applied for, but did not receive funding. The baseline analysis should include a discussion of factors that are likely to aid or impede NORHED funded projects from reaching their development objectives. The analysis may also highlight areas of particular interest for future call-offs.

c) Assess the award mechanism. This assessment shall include a discussion of the choices of projects and institutions selected to build capacity in light of the proposed general theory of change (Norad, 2014b) and other relevant literature on capacity building of higher education institutions. The assessment could include a comparison of types of projects and their contexts, available indicators and/or other factors for institutions that received and did not receive funding.

4 Fieldwork

During the inception phase the evaluation team should if possible, organize visit(s) to one or two of the annual meetings organised by the NORHED administration (likely to take place between mid-October and mid-November). Such visit(s) could combine observation of meetings, interviews, and on-site demonstration of procedures to collect indicator data in the institution hosting the meeting(s), to support the assessment of accuracy. Please note that any observation of closed meetings is subject to the consent of the organisers.

The inception phase will determine how later fieldwork should be conducted.

5 Team competencies

It is expected that methods such as interviews with the NORHED administration and other stakeholders, observation, literature and document reviews are the most important methods employed in this preparatory study. In addition, the team should have knowledge of quasi-experimental methods, as the preparatory study shall include a discussion of these methods, even though analysis, besides descriptive statistics, is not expected at this stage.

The team should at least cover the following competencies:
- Capacity building of higher education institutions
- Proficiency in qualitative methods, including observation, interviews and focus group discussions
- Proficiency in quasi-experimental methods
- Experience with conducting fieldwork in a low-income setting.

6 Deliverables

Deliverables will be organised in two phases corresponding to the organisation of objectives above.

6.1.1 PHASE 1: Inception

a) A technical proposal, a work plan specifying who should do what and when (responding to these terms of references) and two separate budgets: one for the inception phase and one for the baseline phase. The technical proposal, work plan and budgets should be submitted by e-mail, at the latest by the 10th of October 2014, and are subject to the approval by the Evaluation Department.

b) A draft inception report by the end of January 2015. The Evaluation Department will invite stakeholders to comment on this draft. The inception report shall include an assessment of the existing monitoring and results framework, a description of the award mechanism, and a detailed proposal for methods that can be utilised to assess the effects of NORHED, with respect to institutional strengthening.

c) A final inception report including a work plan and budget for phase two. The inception report is subject to the approval by the Evaluation Department.

6.1.2 PHASE 2: Baseline (dates are tentative and may be revised during the inception phase)

a) A draft report delivered no later than the end of May 2015. The draft report will be distributed to stakeholders for comments.

b) Based on comments from the Evaluation Department and stakeholders, the evaluation team shall revise the draft and submit a final report no later than the end of August 2015. The report will be made publically available [format TBA] after approval by the Evaluation Department.

c) Presentation of the baseline study at a seminar. Venue and date will be determined later. Two researchers who have been central in writing the baseline study shall attend.

Unless otherwise is agreed, the Evaluation Department's guidelines for report writing applies.

References

NORAD 2013. A presentation of NORHED: The Norwegian Programme for Capacity Development in Higher Education and Research for Development. *In:* DEPARTMENT FOR GLOBAL HEALTH, E. A. R. (ed.). Oslo Norad

NORAD. 2013 *46 projects on higher education and research to receive funding* [Online]. www.norad.no. Available: http://www.norad.no/en/support/norhed/46-projects-on-higher-education-and-research-to-receive-funding [Accessed 11.12.2013 2013].

NORAD 2014a. Narrative reporting template, P360 no. 1300798-25. Oslo Department for Global Health, Education and Research.

NORAD 2014b. Theory of Change and Methods for Evaluating NORHED. *In:* (DPMG), D. P. M. G. (ed.) *Real-time evaluation of NORHED* Oslo: Norad's Evaluation Department

Distributions of Applications by Category

Table B.1: Seed Funding by Sub-program Area

Sub-program	Number of eligible applications	Number and percent of total eligible applicants that received seed funding	Number of applications by categorization and whether received seed funding					
			Category 1 w/ seed funding	Category 1 w/o seed funding	Category 2 w/ seed funding	Category 2 w/o seed funding	Category 3 w/ seed funding	Category 3 w/o seed funding
Education	28	18 (64%)	4	2	6 (2 funded)	2 (0 funded)	8	6
Health	36	23 (64%)	6	4	3 (1 funded)	0	14	9
Natural Resources	58	28 (48%)	6	1	9 (2 funded)	11 (4 funded)	13	18
Governance	15	10 (67%)	4	1	4 (0 funded)	1 (0 funded)	2	3
Humanities	17	8 (47%)	4	2	3 (0 funded)	4 (0 funded)	1	3
South Sudan	10	7 (70%)	3	0	0	0	4	3
Total	**164**	**94 (57%)**	**27**	**10**	**25 (5 funded)**	**18 (4 funded)**	**42**	**42**

Research Method for Analyzing the Relationship between Seed Funding and Grants

Table C.1. Effects of Seed Funding on Receiving a Grant Award

	(1) Funded (1 if Yes, 0 if No)	(2) Funded (1 if Yes, 0 if No)	(3) Funded (1 if Yes, 0 if No)
Seed	0.14	0.136	-0.008
	(0.069)*	(0.071)	(0.033)
Education		0.198	0.007
		(0.095)*	(0.039)
Governance		0.243	-0.077
		(0.129)	(0.039)*
Health		0.219	0.008
		(0.092)*	(0.026)
Humanities		0.289	-0.097
		(0.117)*	(0.040)*
Natural Resources		0.158	0.027
		(0.062)*	(0.032)
South Sudan		0.205	-0.001
		(0.143)	(0.021)
IRC Score 1			1.024
			(0.015)**
IRC Score 2			0.224
			(0.066)**
Fraction of Funded Projects Among those that did not get seed funding	0.20	0.20	0.20
	164	164	164
Observations	**0.02**	**0.03**	**0.79**

R-squared
Robust standard errors in parentheses
* significant at 5% level; ** significant at
1% level

Discussion of Table C.1

The baseline regression equation model that we use to quantitatively examine the association between receiving seed funding and obtaining a NORHED project award is as follows:

$$Received\ Award_i = \alpha + \beta \cdot Seed_i + \varepsilon_i$$

Where $Received\ Award_i$ takes the value of 1 if project *i* obtained a NORHED award and 0 if it did not, $Seed_i$ takes the value of 1 if project *i* had received seed funding and 0 if not, ε_i is an error term, and α and β are the parameters to estimate. We estimate this regression equation using a linear probability model

(ordinary least squares) on the sample of 164 project applications with a valid IRC score. The estimate of α represents the probability that a project that did not receive seed funding ended up receiving a NORHED award. From column (1) this estimate is 0.2 or 20 percent. Because $Seed_i$ is a dichotomous variable, the estimate of β can be interpreted as the percentage point increase in the probability of receiving a NORHED grant for projects that received seed funding relative to projects that did not receive seed funding. From column (1) this estimate is 0.14 or 14 percentage points. In other words, in this baseline specification, receiving seed funding is associated with a 70 percent (0.14/0.20) increase in the probability of obtaining a NORHED award.

In columns (2) and (3) we modify the baseline specification as follows. In column (2) we include sub-program specific intercepts:

$$Received\ Award_{is} = \alpha_s + \beta \cdot Seed_{is} + \varepsilon_{is}$$

Where $Received\ Award_{is}$ takes the value of 1 if project *i* from subprogram *s* obtained a NORHED award and 0 if it did not, $Seed_{is}$ takes the value of 1 if project *i* from subprogram *s* received seed funding and 0 if not, ε_{is} is an error term, α_s are subprogram-specific intercepts and β is defined as above. As before, we estimate this regression equation using a linear probability model (ordinary least squares) on the sample of 164 project applications with a valid IRC score. Each subprogram-specific intercept estimate measures the subprogram-specific probability that a project that did not receive seed funding ultimately received a NORHED grant. For example, in Column (2) the probability that education sub-program projects that did not receive seed funding ultimately received a NORHED award is 0.198 or 19.8 percent (we reject the null hypothesis that award success rates for projects that did not receive seed funding is the same across sub-programs). The estimate on β in column (2) is 0.136 or 13.6 percent, very similar to that in column (1). This suggests that differences in subprogram specific award rates are not biasing the estimate of the association between obtaining seed funding and ultimately receiving a NORHED grant.

In column (3) we also control non-parametrically (i.e., using dummies) for the IRC score in addition to sub-program specific intercepts:

$$Received\ Award_{is} = \alpha_s + \beta \cdot Seed_{is} + \gamma \cdot 1(IRC = 1) + \delta \cdot 1(IRC = 2) + \varepsilon_{is}$$

Where $1(IRC = 1)$ is an indicator function that takes the value of 1 if a project's IRC score is 1 and 0 otherwise and $1(IRC = 2)$ is an indicator function that takes the value of 1 if a project's IRC score is 2 and 0 otherwise. The omitted category in this regression equation is projects that obtained an IRC score of 3 (the lowest possible). Since we also estimate this equation using a linear probability model and all of the right-hand-side variables of the model are dummy variables, they all have the same interpretation (percentage point increases relative to the omitted category) and the magnitudes are directly comparable. The results from column (3) indicate that once we control for IRC scores, there is no association between obtaining seed funding and ultimately receiving a NORHED award. In other words, the mechanism through which seed funding may be operating is through improving the quality of the application and its IRC score.

Evaluation Grid used for First Call

Recommendations and evaluation grid for NORHED IRCs

The IRCs shall review eligible NORHED proposals according to quality and relevance of project design, feasibility and potential for sustainability and impact, with reference to the overall priorities and geographic focus, sub-programme priorities and geographic focus, and results framework specified in the document "A Presentation of NORHED".

The evaluation grid follows the NORHED application form. It is divided into sections and subsections. Each subsection will be given a score based on the maximum score bracket for the particular subsection. Reviewers are encouraged to make use of the whole scale.

Directions: For each project to be reviewed, the Principal Assessor completes the detailed scoring grid on p. 2-3 and Section 1 below.

SECTION 1: FUNDING RECOMMENDATIONS, OVERALL ASSESSMENT AND TOTAL PROJECT SCORE		
Project title:		**Project number:**
Name of Principal assessor:		
Name of Second assessor 1:	**Name of Second assessor 2:**	
Recommendations to Norad *(please tick the relevant category)*		
Category 1		Proposal recommend ed for approval without changes (and no or only minor clarifications)
Category 2		Proposal recommended for approval provided that clarifications or adjustments are met within a limited timeframe
Category 3		Proposal not recommended for funding
Narrative, overall project assessment according to the NORHED objectives of capacity development for higher education and research including relevance of geographic focus (minimum 100 words) :		**Total score (of 100):** _____

Section 2	Comments & Score	Max. score	Score
1. Partnership *(Section 1 of application form)*		**Max 5**	
Partnership design including clarity of roles and responsibilities in terms of institutional capacity development objective, as well as relevance and competency of persons involved *(Section 1 of application form and separate CVs)*			
2. Project thematic relevance *(Section 2 of application form)*		**Max 20**	
Project relevance to NORHED's overall objectives of capacity development for higher education and research in developing countries (*refer to p.1-6 of document "A Presentation of NORHED"*)			
Project relevance to the NORHED sub-programme(s) to which project is addressed (*refer to sub-programme descriptions on p. 18-26 in document "A Presentation of NORHED"*)			
Project relevance to needs of institution(s), country and region where higher education institution capacity is to be developed *(section 2 of application form)*.			
3. Project geographic relevance *(section 2 of application form)*		**Max 15**	
Geographic relevance in terms of Norway's priority countries for bilateral cooperation *(refer to p. 9 of document "A Presentation of NORHED")*			
Geographic relevance in terms of building capacity in NORHED sub-programme priority region. Note that Upper Middle Income Country-based projects will not score on this even if they are situated in a relevant region *(refer to p. 9-10 of document "A Presentation of NORHED" and the DAC list of ODA recipients).*			
4. Project educational components *(section 2 of application form)*		**Max 15**	
Relevance of educational components of project to NORHED sub-programme *(refer to p. 18-26 of document "A Presentation of NORHED")*			
Quality of educational components of project in terms of strengthening capacity of LMIC(s) academic institution(s) to produce more and better candidates (*refer to p.4-6 of document "A Presentation of NORHED"*)			
5. Project research components *(section 2 of application form)*		**Max 15**	
Relevance of research components of project to NORHED sub-programme (*refer to p. 18-26 of document "A Presentation of NORHED"*)			

Quality of research components of project in terms of strengthening the academic institution's capacity to conduct scientifically rigorous, nationally/regionally relevant research *(refer to p.4-6 of document "A Presentation of NORHED")*			
6. Project gender focus *(section 2 of application form)*		**Max 5**	
Quality of gender perspectives in project, with gender dimensions reflected in both educational and research components of project *(refer p. 7-8 of document "A Presentation of NORHED")*			
7. Results framework *(Section 3 of application form)*		**Max 10**	
Relevance of result framework development impact (*application form section 3.1*) and project outcome (*application form section 3.2*) to the project description set out in section 2 of application form *(For guidance, please refer to p. 4-7 of document "A Presentation of NORHED")*			
Relevance and logical flow between project goals (*application form section 3.1*), outcomes (*application form section 3.2),* outputs (*application form section 3.3*) and activities (*application form section 3.4*), as well as relevance and measurability of indicators for outcomes (*application form section 3.*2) and outputs (*application form section 3.3. For guidance, please refer to p. 4-7 of document "A Presentation of NORHED")*			
Relevance and inclusiveness of assumptions and risks linked to outcomes (*application form section 3.2*), outputs (*application form section 3.3*) and activities (*application form section 3.4*), including relevant and realistic mitigation strategies			
Relevance and quality of baseline description (*application form section 3.5*) in relation to project outcomes			
8. Assessment of sustainability, cross cutting issues, risk factors and mitigation measures *(Section 4 of application form)*		**Max 10**	
Relevance and quality of LMIC partner's institutional capacity analysis, including assessment of financial management, corruption and other potential risk factors, and relevant mitigation strategies (*application form sections 4.2 and 4.3, see also separate Partner Assessment attachment. For guidance, please refer to p. 7-9 of document "A Presentation of NORHED")*			
Relevance and quality of gender equality assessment in terms of project implementation and expected results			

(*application form section 4.4. For guidance, please refer to p. 7-8 of document "A presentation of NORHED"*)		
Relevance and quality of conflict sensitivity assessment (*application form section 4.5*), human rights assessment (*application form section 4.6*), environmental/climate impact assessment (*application form section 4.7*) and other risk factors (*application form section 4.8*) including relevant mitigation strategies *(for guidance, refer to p. 8-9 of document "A presentation of NORHED").*		
Relevance and quality of project financial sustainability strategy (*application form section 4.9*)		
9. Financial plan, budget, implementation plan and additional information *(Sections 5 and 6 of application form, budget attachment and implementation plan attachment)*	**Max 5**	
Relevance, realism and cost effectiveness of financial plan and budget (*application form section 5 and separate budget attachment)*		
Relevance and quality of project implementation plan and procedures and routines to be followed in order to monitor partnership (*project implementation plan attachment and application form section 6.1 and 6.2*)		
10. Total Score	**Max 100**	

Current and Alternative Structures for Main Application[25]

As described in the main narrative, the main application should be organized in a way that lets the applicant clearly convey the information that reviewers need in order to be able to judge an application properly. DPMG's analysis identified a number of weaknesses in the structure of the application. This annex describes the sequence of tasks in the original application followed by a proposed alternative sequence. For example, the main application used in the first call starts by asking about the partnership. A possible alternative structure would first determine what the project hopes to achieve by project close (its development objectives), why these objectives matter and for whom, what activities or components will be implemented to achieve the objectives, and the detailed design of these components. It then asks how the applicant plans to implement the project. It is here that all of the partnership questions are asked—for example, how the selection of each partner in the partnership is expected to enhance the likelihood of achieving the project's objectives. This proposed structure would give applicants the opportunity to clarify the alignment among the projects' objectives, activities, outputs, and outcomes.

Structure of Original Main Application

Table E.1. Structure of current main application form

1.   **Administration and partnership**
1.1.   Partnership information
1.2.   LMIC partner institution
1.3.   Norwegian partner institution
1.5.   Additional partner institutions
• Background and history. Previous/ongoing cooperation projects
• Where main capacity development will occur
• Describe Norwegian partner (role in project; resources available, competence & qualifications, experience with geographic area concerned, other relevant experience)
• LMIC partner (same questions)
• Gender focal point
• Gender focal point
2.   **Description of project**
• Sub-program to which application applies
• Other sub-programs to which application is relevant
• Project title
• Project description and design of project
• Agreement partner institution. Describe current situation and need/justification for the project specified by institution/country
• Other partner institutions. (Same question)
3.   **Results framework**
3.1.   Development goal/intended impact on society
3.2.   Project goal
• Outcomes
• Indicators
• Assumptions/risk factors

---

[25] In addition to the main application, each application package also included a budget, the implementation plan, a partner assessment, a letter of commitment, and CVs.

3.3.	Expect results/services/products (outputs)
	• Outputs
	• Indicators
	• Assumptions/risk factors
3.4.	Planned activities
	• Activities
	• Inputs
	• Assumptions/risk factors
3.5	Present situation (baselines)
	• Indicators, baseline values, means of verification, targets for the final year
4.	**Assessment of sustainability, cross-cutting issues, risk factors, and mitigating measures**
4.1.	Assessment of how project relates to the national and partners' institutions' strategic plans and priorities (in countries where capacity will be developed)
4.2.	Assessment of partners' institutional capacity, including risk factors
4.3.	Assessment of financial management and corruption (refer to partner assessment form)
4.4.	Assessment of gender equality in relation to project implementation and expected results
4.5.	Assessment of conflict sensitivity
4.6.	Assessment of human rights
4.7.	Assessment of environmental and climate sustainability
4.8.	Identification of other risk factors and management of identified risks
4.9.	Assessment of sustainability after the project comes to an end
5.	**Financial plan (refers to budget and its allocation among the partners. In addition to separate detailed budget)**
6.	**Additional information**
6.1.	Explain procedures and routines for agreement partner to follow up cooperation with other partners (contracts, disbursement, procurement, monitoring and reporting)
6.2.	Any other information relevant to application

Proposed Structure for the Main Application

Table E.2. Possible alternative structure for the main application

Cover sheet
1. Project title and sub-program(s) to which applicant is applying
2. List of all departments and institutions that will collaborate in the delivery of the project and be accountable for its results
3. Information on Norad's eligibility criteria (e.g., Is the Norwegian partner accredited by NOKUT?)
4. Basic information on system-wide and institutional enabling conditions for each partner. (See Note at the end of this table for possible questions).

Main application
5. What does the project intend to achieve by the end of the project? This is the project's development objective, which should be clearly stated in a single sentence.
6. Why does this development objective matter and for whom—for the involved institutions, sub-groups, countries, or regions?
7. What does the project intend to do? Describe the nature and detailed design of each project component for which funding is sought, including how the design will further gender equality. Be sure that your components are aligned with your budget and implementation plan.
8. How does the selection and design of your components help to redress gender imbalances?

9. How do the partners plan to implement the project? Ensure that your responses here are aligned with your implementation plan.

9.1. Who are the partners?

9.2. Why were these particular partners selected? What capacities and resources does each bring to the partnership that will enhance the likelihood of achieving the project's objectives?

9.3. What specific responsibilities and accountabilities will each partner have relative to each of the other partners? Be sure that your budget properly reflects the responsibilities of each partner.

9.4. Describe why it is expected that the partners will be able to collaborate effectively with each other (past history, other factors). If any of the partners have collaborated in the past, identify the strengths and weaknesses of that past collaboration. Include in the application package any evaluations of their joint efforts.

9.5. Describe how the lead partner will monitor the performance of other members of the partnership in terms of accounting for disbursements, delivering on their implementation responsibilities, and monitoring and evaluation.

10. Describe what you expect the funded activities to produce in terms of outputs-—or example, a revised curriculum. State what difference you expect these outputs to make-; in other words, the outcomes of the project. For example, a better curriculum might result in better-trained graduates, as indicated by examination results or the ease with which they get jobs. This section states your theory of change: how your project activities will produce certain outputs that in turn will produce changes in outcomes.

10.1. Development goal: what are you trying to achieve with this grant?

10.2. Planned activities
- Activities
- Inputs (budget, people)

10.3. Expected outputs. These are the services or products that you "put out" as a result of completing your activities. Be sure to reflect gender in outputs and indicators.
- Outputs
- Indicators

10.4. Project outcomes. Outcomes are not the same as outputs. Outputs are the services or products that you "put out" as a result of completing your activities. Outcomes are differences that outputs make. Be sure to reflect gender in outcomes and their indicators or measures.
- Outcomes
- Indicators

10.5. Present situation (baselines). You will not know if you are progressing toward your output and outcome objectives if you do not know for each output and outcome where you started at the beginning of the project. If an output or outcome is entirely new and not a change in the number or quality of something that already exists, the baseline will be zero.
- For each indicator: state the baseline value, means of verification, and the target value. The target value is what change you plan to achieve by the final year—for example, how many more graduates, how large a percent increase in adoption of a new technology.

11. Describe the monitoring and evaluation procedures that will be in place. What data sources will be used to monitor each outcome, output, and activity: e.g., a survey, administrative data? How frequently do you plan to update the values for each outcome, output, and activity?

12. Risks.

12.1. Assess the main risks to the attainment of the project's outcomes. These can arise from the project design, the institution(s) in which the project is lodged, the higher education sector(s) in which the institutions are lodged, or political, economic, and social factors outside of the sector.

12.2. Realistically assess the likelihood that these risks will materialize.

12.3. Identify any measures that the project will take to mitigate any of these risks. If the project has no power to mitigate a risk and no ability to finesse the risk, candidly state this.

13. Sustainability. Discuss whether you expect the advances made under the project to be at least maintained after the project closes and why. Be sure to discuss financial, human resource, political, and other factors that could promote and/or undermine sustainability.

Note: The cover sheet should ask for information for each partner on the system-wide enabling conditions and institutional development dimensions. These can be Yes/No questions, or applicants can be asked to rate each condition on a five-point scale. Possible questions are these:

1. Does the partner country have an ongoing higher education development vision /plan?
2. Is the proposed project aligned with the current higher education developments in the country? State in 1-3 sentences how the project is aligned.
3. Does the partner university receive additional resources in the context of the present higher education development plan?
4. Is the partner university hampered by limitations in the national regulatory framework for higher education institutions that may affect project implementation? For example, do regulations restrict your university's autonomy and academic freedom? Please provide a few lines of explication.

These are possible questions on institutional development dimensions:

1. Does the partner university have an institutional development strategy?
2. If it has one, how does the proposed project fit into that strategy? State in a sentence how the project is aligned with the university's strategy.
3. Is the partner university facing financial difficulties?
4. Does the partner university have a favorable governance framework?
5. Does the partner university have an active Teaching and Learning Center?
6. Does the partner university have an effective graduate follow-up mechanism, such as tracking whether the graduates get jobs?

Theory of Change

Figure F.1.: General theory of change for producing more/better-qualified graduates and more/higher quality research

System-level enabling conditions

- Institutional autonomy & academic freedom

- Accountability mechanisms

- Student support and financial aid

- Funding for teaching and research

Institution-level enabling conditions

- Leadership, vision, & plan

- Internal quality assurance & institutional research

- Resources & incentives for teaching, research and student aid

Sequence of causal links within universities

Appropriate qualifications, experience and motivation of academic staff

Intermediary results
High quality and relevant programs/ courses; engaging pedagogical practices and more time spent on active learning activities

Dynamic collaborative and multi-disciplinary research

Increased number and quality of graduates; increased and better-quality research

Effective selection of incoming students, esp. graduate students

Favorable learning environment and research infrastructure

Linkages with industry & community; stakeholders' inputs

Source: Norad. 2014. *Evaluation Series of NORHED: Theory of Change and Evaluation Methods*. Oslo.

Data Used for DPMG Evaluation

Table G.1.: Data for DPMG's evaluation of 37 randomly selected eligible applications

Sub-program	DPMG category	IRC category	DPMG score	IRC score	Score by evaluation grid criterion and (maximum points)								
					1 (5)	2 (20)	3 (15)	4 (15)	5 (15)	6 (5)	7 (10)	8 (10)	9 (5)
Education applications													
1	2	2	81	61	5	20	15	12	11	4	3	6	5
2	3	2	80	61	3	20	15	10	10	3	7	7	5
3	3	3	60	40	2	14	15	10	8	3	4	3	1
4	1	1	93	80	5	20	15	15	15	4	5	9	5
5	2	3	85	45	4	20	12	13	15	4	7	6	4
6	1	3	88	40	4	20	12	15	15	3	7	8	4
Average score			81.2	54.5	3.8	19.0	14.0	12.5	12.3	3.5	5.5	6.5	4
% of max. points			81.2	54.5	76.7	95.0	93.3	83.3	82.2	70.0	55.0	65.0	80.0
Health applications													
1	3	3	46	55	2	15	10	9	5	3	0	0	2
2	1	2	72	72	5	16	15	12	12	2	3	4	3
3	1	1	80	74	5	20	15	10	10	3	8	5	4
4[26]	3	3	76	59	5	20	15	10	8	5	2	8	3

---

[26] This project (number 4) and number 6 are both categorized as "3's"--i.e., do not fund. However, their scores are higher than project #2 which is categorized as a "1"--fund. Both of these projects have the same story. In his draft evaluation, the evaluator rated both projects borderline 2's. After extensive discussions with the lead for the evaluation, he changed both to "3's". He found that both proposals scored well on the eligibility criteria of geographic and thematic relevance. However, they scored relatively poorly on the remaining criteria that captured much of the substance of the projects, and the evaluator lacked confidence that they

Sub-program	DPMG category	IRC category	DPMG score	IRC score	Score by evaluation grid criterion and (maximum points)								
					1 (5)	2 (20)	3 (15)	4 (15)	5 (15)	6 (5)	7 (10)	8 (10)	9 (5)
5	3	3	60	48	5	15	6	15	0	5	2	7	5
6	3	3	76	62	5	20	15	15	5	5	2	4	5
7	3	1	55	77	4	16	12	5	0	5	5	5	3
8	1	3	96	51	5	20	15	15	15	5	8	8	5
Average score			70.1	62.3	4.5	17.8	12.9	11.4	6.9	4.1	3.8	5.1	3.8
% of max. points			69.9	62.3	90	88.8	85.8	75.8	45.8	82.5	37.5	51.3	75
Natural resources applications													
1	1	3	62	50	5	20	0	15	8	2	5	4	3
2	1	3	85	63	5	20	15	9	15	5	4	8	4
3	1	2	82	78	5	20	15	10	8	5	9	7	3
4	2	3	74	64	5	20	0	15	10	2	10	9	3
5	2	2	78	79	5	20	10	10	8	5	10	8	2
6	2	1	75	80	3	20	15	10	15	4	6	5	2
7	2	3	75	65	3	20	12	10	15	5	7	2	1
8	3	2	42	75	2	10	15	2	5	2	2	2	2
9	2	3	72	63	2	15	15	10	10	3	10	4	3
10	2	1	73	81	5	15	15	7	7	5	9	7	3
11	3	3	37	47	2	8	8	2	10	2	2	2	1
12	3	3	53	54	3	17	8	10	8	2	2	2	1
13	1	2	86	79	4	20	13	10	15	5	10	7	2

would succeed if funded. Project 4 scored only 66% on the remaining 65 points; project 6, only 63% on the remaining 65 points. He felt strongly that in these two cases criteria 2 and 3, which applicants often satisfied, distorted the picture of the quality of these applications.

Sub-program	DPMG category	IRC category	DPMG score	IRC score	Score by evaluation grid criterion and (maximum points)								
					1 (5)	2 (20)	3 (15)	4 (15)	5 (15)	6 (5)	7 (10)	8 (10)	9 (5)
Average score			68.8	67.5	3.8	17.3	10.9	9.2	10.3	3.6	6.6	5.2	2.3
% of max. points			69.5	67.5	75.4	86.5	83.4	61.5	68.7	72.3	66.2	51.5	46.2
Governance applications													
1	3	2	51	67	2	2	15	5	12	4	3	5	3
2	3	3	68	48	5	10	15	9	7	5	5	9	3
3	3	1	72	80	5	20	15	6	6	2	5	8	5
Average score			63.7	65.0	4.0	10.7	15.0	6.7	8.3	3.7	4.3	7.3	3.7
% of max. points			63.7	65	80	53.3	100.0	44.4	55.6	73.3	43.3	73.3	73.3
Humanities applications													
1	3	3	73	0[27]	1	20	11	11	12	5	5	6	2
2	3	2	62	60	3	15	15	12	0	2	6	5	4
3	1	1	87	87	4	20	15	15	10	5	6	9	3
4	3	2	58	70	3	20	11	8	2	2	6	4	2
Average score			70	72.3	2.75	18.75	13	11.5	6	3.5	5.75	6	2.75
% of max. points			70	72.3	55	93.8	86.7	76.7	40	70	57.5	60	55
South Sudan applications													
1	3	3	53	59	2	15	10	9	7	0	3	5	2
2	1	1	89	83	4	20	15	14	15	2	8	6	5

---

[27] IRC learned during its evaluation of this application that two of the main partners had dropped out. They decided to treat the application as incomplete and gave it a score of 0. DPMG evaluated it on the assumption that the two partners were still in place.

Sub-program	DPMG category	IRC category	DPMG score	IRC score	Score by evaluation grid criterion and (maximum points)								
					1 (5)	2 (20)	3 (15)	4 (15)	5 (15)	6 (5)	7 (10)	8 (10)	9 (5)
3	3	3	57	54	3	11	15	9	6	3	4	2	4
Total			199	196	9	46	40	32	28	5	15	13	11
Average score			66.3	65.3	3.0	15.3	13.3	10.7	9.3	1.7	5.0	4.3	3.7
% of max. points			66.3	65.3	60	76.7	88.9	71.1	62.2	33.3	50.0	43.3	73.3
Total sample													
average score			70.8	64.2	3.7	17.1	12.8	10.4	9.2	3.5	5.4	5.6	3.2
% of max. points			70.8	64.2	74.6	85.7	85.6	69.2	61.3	70.8	54.1	55.7	63.2