Commonwealth Scholarship Commission in the UK

A study of research methodology used in evaluations of international scholarship schemes for higher education

COMMONWEALTH SCHOLARSHIPS

This report was written by Dr Matt Mawer, and published in June 2014.

For further information regarding the CSC Evaluation and Monitoring Programme, please contact:

Commonwealth Scholarship Commission in the UK
Woburn House
20-24 Tavistock Square
London WC1H 9HF
UK
evaluation@cscuk.org.uk
http://bit.ly/cscuk-evaluation

Contents

List of acronyms

ADA	Austrian Development Agency
AFGRAD	African Graduate Fellowship Program
ATLAS	Advanced Training for Leadership and Skills Program
AusAID	Australian Agency for International Development
CASS	Cooperative Association of States for Scholarships Program
CFSP	Canadian Francophonie Scholarship Programme
CGSP	Chinese Government Scholarship Program
CSC	Commonwealth Scholarship Commission in the United Kingdom
DAAD	German Academic Exchange Service
DFAT	Department for Foreign Affairs and Trade, Australia
DFID	Department for International Development, UK
EU	European Union
GMS	Gates Millennium Scholars Program
IFP	International Fellowship Program of the Ford Foundation
IPRS	International Postgraduate Research Scholarship, funded by Department of Innovation, Industry, Science and Research, Australia
JISPA	Japan-IMF Scholarship Program for Asia
LAC	Latin American-Caribbean Scholarships: funded by USAID under the CASS and SEED programmes
NFP	Netherlands Fellowship Programme
NGO	Non-Governmental Organisation
NICHE	Netherlands Initiative for Capacity development in Higher Education
NOMA	Norad's programme for Master's Studies
Norad	Norwegian Agency for Develop Cooperation
NPT	Netherlands Programme for Institutional Strengthening of Post-secondary Education and Training Capacity
OCEANS	Organisation for Cooperation, Exchange and Networking among Students
OECD	Organisation for Economic Cooperation and Development
OECD DAC	Development Assistance Committee of the Organisation for Economic Cooperation and Development (OECD)
RCT	Randomised Control Trial
SEED	Scholarships for Education and Economic Development Program: funded by USAID
UN	United Nations
USAID	United States Agency for International Development
VFM	Value for Money
VLIR-UOS	Flemish Interuniversity Council University Development Cooperation

Executive summary

International scholarship schemes for higher education are widely supported by governments, supranational bodies, and charitable organisations as part of both public diplomacy and developmental assistance commitments. Although the objectives of many scholarship schemes have evolved over the period of their administration, commitment to scholarships remains strong. Renewed investment by national governments and new investment by charitable foundations enable thousands worldwide to study outside of their home country each year.

Increasingly scholarship providers have invested time and resources into the evaluation of scheme outcomes, tracing alumni and examining how their experiences post-scholarship reflect progress toward the policy objectives of scholarship programmes. Despite the elevated importance of evaluation practices to both understanding outcomes and securing renewed funding, surprisingly little analysis has been conducted of evaluation practices currently employed across the scholarships 'sector'.

This scoping study was undertaken by the Commonwealth Scholarship Commission in the United Kingdom (CSC) in an attempt to identify the main trends and ambiguities in evaluation approaches and to analyse where different, or more detailed, methodology might lead to more robust evaluation. Reports and papers published by scholarship providers and evaluators in the governmental, non-governmental, and academic sectors were collated, supplemented by direct correspondence where required. Approximately 65 evaluation documents were reviewed. Four aspects of the documentation were analysed: methodology, methods, variables, and data analysis. Three thematic issues – counterfactuals, value for money, and aid harmonisation – were also considered.

In addition to descriptive findings, the scoping study offers more speculative conclusions regarding the ambiguities in the sector that might be addressed in order to strengthen evaluation methodology. Although some reference is made to research literature on evaluation methodology, this study has been designed to address current – not 'best' – practice. The results of the scoping study are thus intended to offer both an insight into the general trends in evaluation methodology within the sector and provide an opportunity for dialogue and cooperation going forward.

The main findings of the study are summarised below.

Research design

Methodology

- Methodology is not widely discussed within the sector, with most reporting of research design focused at the methods level.
- Almost all evaluations are ex-post: there are few examples of designed longitudinal evaluation having been undertaken. In many cases evaluations appear to be 'catching up' with previous years of scholarship administration during which concurrent evaluation was not conducted.
- Kirkpatrick's model for evaluation is the most frequently cited extant methodological framework, although it is used only within a small proportion of research reports. Contribution analysis has also received limited attention as a methodological approach, but as yet has not been widely reported within scholarship evaluation.

Methods

- Mixed-methods are prevalent: fewer than five evaluations reported solely quantitative or solely qualitative approaches.
- Surveys are the dominant tool for data collection, often administered as part of tracer studies and involving open-ended questions on post-scholarship employment and Likert-style questions on the benefits of the scholarship
- Interviews are also widely used, most usually with alumni in-country, but also not infrequently with other stakeholders, such as government officials, scholarship administrators, and employers.

Variables

- Almost all evaluation is concerned with similar topics: socio-demographics of candidates, scholarship process and satisfaction, return to home country, change in personal competencies, employment

trajectory post-scholarship, contribution post-scholarship to sector / profession / community / country, and links to donor countries.

- The OECD DAC's criteria for evaluating developmental assistance are used within several reports, but not widely across the sector. The DAC criteria focus at a more abstract / macro-level than much (particularly tracer) evaluation and thus usually require further operationalising before being a useful framework for structuring data collection.

Data analysis

- Both qualitative and quantitative data are analysed within the sector, but reporting of data analysis procedures is often vague. Approaches to analysing qualitative data, in particular, are rarely reported in sufficient detail, raising concerns that some elements of evaluations are based on anecdote rather than rigorous analysis.

- Most quantitative analysis is descriptive and this is often sufficient to give an overview of evaluation findings. Inferential statistics are less widely used, although there are examples of detailed statistical approaches to data analysis which have added value to evaluation findings.

- Lack of baseline data has been raised frequently as a concern with evaluation. Several studies have struggled to reconstruct baseline data effectively from retrospective measures and many studies have no baseline data at all. Robust monitoring and data management systems are required to facilitate long-term data collection and storage for the purposes of baseline to post-scholarship comparison, and in some cases these systems have been developed only relatively recently

Thematic issues

Counterfactuals

- Conditional counterfactuals have been addressed in a small subset of evaluation studies, examining alternative outcomes had a scholarship award not been received. Between-intervention type counterfactuals, examining scholarship outcomes in contrast to alternative programmes aimed at similar objectives, have not been widely discussed or investigated.

- There are several examples of comparative designs that have provided detailed counterfactual evidence, most usually contrasting scholarship recipients to a comparison group of non-selected finalists in the scholarship selection process.

- The cost and time investment in developing effective counterfactuals has been noted as a significant barrier, alongside the difficulty in reconstructing credible counterfactual data without a longitudinal comparison group.

Value for money

- Analysis of financial conduct within scholarship programmes is commonplace, most usually involving examination of budget administration efficiency. Some of these analyses have raised interesting issues for scholarships more broadly, such as the effect of disbursement schedules on accessing other sources of funding.

- Comparative value for money analysis, such as Cost-Benefit Analysis or Social Return on Investment, does not appear to have been widely conducted. Only one detailed value for money analysis was reported, based on a retrospective reconstruction of a baseline against which 'gains' (and thus value) could be established.

Harmonisation

- Discussion of aid harmonisation is very limited within evaluation reports. Commentaries that are available tend to reflect negatively on the harmonisation situation within countries, although there is evidence of greater harmonisation within the Pacific region.

- Synergy and interference effects between schemes that operate in the same geographical regions and / or target the same audiences are unexplored, but meta-analysis of these facets of scholarship impact would likely be very insightful.

1. Introduction

International scholarships schemes for higher education have an extensive history as tools of political engagement and overseas development assistance. Many countries have longstanding traditions of funding scholars to undertake protracted periods of academic study within their borders, and increasingly non-governmental organisations have funded large-scale scholarship schemes as part of their poverty alleviation endeavours. Recent events – such as the $500 million investment by MasterCard Foundation in a new scholarship programme and the trebling of the UK government investment in Chevening scholarships – indicate that confidence in scholarships as vehicles for developmental change and political influence is high.

Alongside the significant investment in scholarship schemes made by governments and NGOs has come an on-going engagement with evaluation. The face of evaluation has undoubtedly changed dramatically since the inception of older scholarship schemes (e.g. the Commonwealth Scholarship and Fellowship Plan in 1959); increasingly it is not sufficient to justify funding ex-ante, but necessary also to show the impacts of scholarship funding ex-post. An array of evaluation practices have evolved, administered both by scholarship providers and externally commissioned consultants.

Despite continued investment in scholarship schemes, analysis of the methodological approaches deployed to assess their impact has been relatively limited. The merits and demerits of particular research strategies used in scholarship evaluation have been examined primarily at the level of individual providers and rarely reviewed more comprehensively. The dialogue on evaluation issues has been stimulated by events such as the London International Development Centre's conference on 'Measuring impact of higher education for development' (19-20 May 2012), but the field yet lacks detailed international comparison of evaluation approaches.

It is into this space that the current scoping study is designed to enter, analysing trends in the practices of evaluators across a global community of international scholarship providers. In conducting the study the Commonwealth Scholarship Commission in the United Kingdom (CSC) has sought to interrogate evaluation practice with a view to understanding and learning from the experiences of colleagues. We now share our findings in the hope that colleagues across the sector may also discover useful insights and be moved to continue a vigorous international dialogue on how best to understand the impacts of international scholarships schemes for higher education.

Outline of the study

The scoping study has been designed with two main aims:

1. To identify trends in research practices and strategies used in the evaluation of international scholarships for higher education

2. To identify omissions, uncertainties, and ambiguities in current methodological approaches

To this end the study has involved locating and analysing evaluation reports and other documentation published by providers of (and commentators on) international scholarship schemes for higher education.

The study has examined evaluation research produced by scholarship schemes funded and administered through governmental bodies, funded through government but administered elsewhere (e.g. through NGOs), and funded and administered through NGOs, foundations, and supranational bodies (e.g. the European Union). Corporate and private sector scholarships have not been included in the study. The majority of evaluation documents gathered relate to schemes funded by OECD countries, although this is a reflection of the background of donors for whom evaluation documentation is available rather than a criterion applied to the study.

The focus of the scoping study has been on 'inbound' funding for applicants to study either in donor countries or other countries generally, the latter in cases where scholarships were funded by non-national entities (e.g. The Ford Foundation). 'Outbound' funding by national governments for their citizens to study abroad, such as Science Without Borders in Brazil[1], has not been considered as part of the scoping study[2]. Additionally, the study has prioritised evaluation documents relating to scholarship schemes that fully fund recipients. Although selection on this issue has been loose, the general rule has been that if the scholarship did not provide for full tuition fees, travel, and living costs then it would not be considered as part of the scoping study. Part-funding for study (for instance Scotland's Saltire Scholarships) was not included. Finally,

[1] See Ciência sem fronteiras: <http://www.cienciasemfronteiras.gov.br/web/csf-eng/home>
[2] An analysis of outbound scholarships has recently been provided by the British Council and DAAD (2014)

scholarships for professional development that did not yield academic qualifications (e.g. PhD: Master's degree) – such as military exchange programmes (e.g. Atkinson, 2010) - were not considered as part of the study, except where they were evaluated alongside other scholarship schemes that did yield academic qualifications.

International scholarship schemes for higher education are delivered with a variety of stated aims, but can usefully be considered within the two broad categories of international development scholarships and public diplomacy / soft power scholarships[3]. International development-oriented scholarships have humanitarian goals and are frequently funded through, and delivered by, government departments involved with overseas developmental assistance (e.g. USAID). Public diplomacy-oriented scholarships focus on political goals, such as creating persistent bilateral relations and garnering positive sentiment toward donor countries, and are frequently funded through foreign affairs or state departments. Both forms of scholarship are included in the scoping study, however, and following our interest at the CSC, international development-oriented scholarships are the main focus of the analysis. Scholarships often embody both development and public diplomacy goals to some extent and so it is expected that many of the observations and conclusions relevant to one will be relevant to both.
The search for evaluation documentation has not been exhaustive. There are, for instance, numerous other (particularly smaller) scholarship schemes that fit all of the general criteria used to structure the scoping study and which could have been included. However, as the aim of the study has been to identify trends, and given the difficulty in accessing evaluation research from smaller schemes (where such documentation exists), there has not been an effort to make the review an exhaustive analysis of all evaluation practice.

A comprehensive list of evaluation documents reviewed and personal correspondence held within the ambit of the scoping study can be found in appendices 1 and 2.

Structure of this report

Evaluation documents collected have been reviewed by the author and subsequently trends within research methodology identified and outlined. Omissions and ambiguities, where evident, have also been described.

The report is split into five chapters:

- Methodologies
- Methods
- Variables and indicators
- Data analysis
- Thematic issues

The initial four chapters each correspond to an area of research design, progressing from the conceptual framework underpinning evaluation to the strategies used to make sense of data collected.

In the thematic issues chapter, several additional topics are examined: the counterfactual, value for money, and aid harmonisation. The scoping study has been conducted with the intention to comment on these thematic issues as they are of particular interest to researchers and policymakers involved in international scholarship programmes.

Other parameters

The scoping study has been bounded in several additional ways.

Firstly, this study is primarily an analysis of published evaluation documents. Whilst some correspondence with researchers to discuss their experiences in evaluation has been a valuable part of the process, the corpus of evidence for the analysis presented has been evaluation reports, updates, journal articles, and, in some cases, books and strategy papers. The study is therefore mainly an analysis of what evaluators are *writing* about evaluation, not what they are *thinking* about evaluation.

[3] For a more comprehensive attempt at a typology of international scholarship schemes see Perna et al. (2014) or, on postgraduate scholarships specifically, Boeren et al. (2008)

Secondly, this study is neither a review of the merits and demerits of specific scholarship programmes or of scholarship programmes generally. It is not the purpose of this study to address the value of tertiary education scholarships or the impact of such scholarships on international development[4]. Rather, the purpose of the study is to examine how such issues have been evaluated, with a focus on research methodology rather than the substantive content of evaluation findings.

Similarly, it is important to make clear from the outset that in writing this paper no assumption has been made that more sophisticated research methodology, greater funding for evaluation, or any specific strategy for evaluation inevitably leads to better policymaking. The policy-evaluation interface within scholarship programmes is beyond the purview of the current work and is only noted in circumstances where frameworks have been specifically designed to facilitate dialogue between scholarship policy and evaluation evidence. The translation of evaluation results into policy change is not reviewed within the current study and, whilst the basic assumption is that rigorous and detailed evaluation data will benefit policymaking, it is not necessarily the case that lengthier, more costly, and more detailed evaluation automatically equals better programme policy.

Finally, although methodological texts produced by academic researchers, NGOs, and governmental institutes have been consulted as part of the study, they are used to contextualise and inform analysis of evaluation research rather than to establish what should be done. The scoping study is a review of the 'state of the actual', not the 'state of the art', and as such is concerned with trends and ambiguities more than with best practice.

Useful definitions

The term 'sector' is used in this paper as shorthand for the domain in which international scholarships for higher education, including their various stakeholders, operate. It is not used in an attempt to situate all scholarships within the same conceptual space (the public sector, the charity sector, etc.), given their varied interests and funding arrangements, but merely a way of allowing straightforward reference to the topic of the document.

It may also be helpful to distinguish between two types of reviews, both included in the scoping study but which are somewhat different in their aims and scope. Impact and outcomes evaluations examine the substantive results of the scholarship scheme, often at the level of an alumni cohort or specific country and with reference to the aims of the scholarship scheme. Policy and administrative reviews examine the management of scholarship schemes and their relative merit in national policy frameworks. The triennial review of the Marshall Aid Commemoration Commission (administrators of Marshall Scholarships) by the UK Foreign and Commonwealth office (2013) would fall into the latter category, whereas Nuffic's (2009) tracer study of the Netherlands Fellowship Programme would be better considered an impact and outcome evaluation. In practice most evaluation reports include some aspects of policy, management, and outcome evaluation, and so these are not categories as much as foci on which evaluators place more or less emphasis.

[4] On the latter topic the reader might consult Oktech, McCowan, and Schendel's (2014) recently published literature review.

2. Methodologies

In this study research 'methodology' is taken to be the framework informing evaluation design, from theoretical underpinnings to data analysis and reporting. The term methodology as used here is more in line with academic research literature and international commentaries on impact evaluation (e.g. Garcia, 2011) than is often the case in scholarship evaluation documents, where 'methodology' and 'method' are frequently used as synonyms. 'Design' has also been used to describe the encompassing logic of how research is conducted (e.g. Stern et al., 2012), but in this paper the term methodology is used throughout. Conflating methodology and method can be troublesome, particularly as many methodological approaches use the same basic methods (e.g. surveys, interviews) and are differentiated by their approach to the treatment of data or their conceptual stance.

Methodological frameworks for evaluation are discussed at a meta-level within the sector (e.g. Rotem, Zinovieff, and Goubarev, 2010), but in practice relatively few evaluations offer significant detail on methodology, with most listing data collection methods (e.g. interviews) under the heading of 'methodology' or 'method'. Nonetheless, several observations can be made of methodologies employed within the sector:

1. Almost all evaluations are ex-post

2. There are few examples of longitudinal analysis

3. Evaluations may focus on specific countries or a scholarship scheme generally

4. The favoured evaluation framework is the Kirkpatrick model, but it is not widely implemented if considered in proportion to the corpus of evaluation reports

Each trend is discussed briefly in the sections below, preceded by several initial observations on research methodology in the sector.

Amongst the evaluation reports studied there was very little engagement with experimental and quasi-experimental methodology, strategies recommended in other spheres of developmental intervention (e.g. Garcia, 2011). Randomised Control Trials (RCTs), regression discontinuity design, and difference-in-differences design, amongst other possible approaches, have not been reported within the sector. Elements of quasi-experimental methodology are planned for the evaluation of the MasterCard Foundation Scholars Program, including regression discontinuity design and, in one case, an RCT, but this evaluation has not yet been conducted and thus reports are unavailable[5]. There are doubtless a variety of reasons for this non-engagement, including the cost of methods such as RCTs, practical and ethical difficulty in applying experimental manipulation to complex social phenomena (Byrne, 2013), and limited engagement with comparison (counterfactual) cohorts within the sector (discussed in 'The counterfactual' on page 22). Whether quasi-experimental designs would be of significant value to evaluators in this sector is debatable, particularly given the complexity and diversity of developmental outcomes expected from investments in higher education scholarship schemes. The methodological challenges of, and subsequent solutions to, planned quasi-experimental and experimental evaluation of the MasterCard Foundation Scholars Program (Cosetino et al., 2013) will be informative in this regard.

A further absence from current methodologies is complexity thinking / theory. Complexity theory has gained popularity as an academic stance, originating in the natural sciences, toward social science and has seen application to evaluation of policy in health and other areas (see Sanderson, 2000: Callaghan, 2008: Byrne, 2013: Ling, 2013) . Complexity-informed approaches tend to challenge the legitimacy of causal attribution and (quasi-) experimental manipulation in favour of focusing on how whole social systems effect change (Byrne, 2013), albeit potentially catalysed by interventions. Whether complexity theory has anything to offer this sector specifically is, again, debatable, but there are some moves toward greater use of complexity-based research approaches. USAID, for instance, has recently published a note on 'complexity-aware monitoring' which draws upon complexity theory in consideration of evaluation (Britt, 2013), although in the context of developmental intervention generally rather than higher education scholarships specifically.

Ex-post evaluations

The majority of evaluations have been conducted ex-post, attempting to assess impact by retracing the post-scholarship experience of recipients. Tracer studies are the prototypical ex-post methodology for evaluating scholarship outcomes, the approach of which is well summarised by Nuffic:

[5] Barry Burciul, personal correspondence (February 19th, 2014)

> *'In general, tracer studies start with assembling time series data to measure the output of the fellowship programme in terms of degree attainment for all fellows. These data are further disaggregated by gender, home country/region, sector/discipline, etc.'* (2009: 8)

Almost all scholarship programmes have conducted some kind of alumni tracer study, either as a standalone ex-post programme evaluation (Nuffic, 2009) or as part of a longer term study using individual tracer reports to inform a longitudinal-style analysis (e.g. ICUag.net, 2013). The extent to which tracer studies assemble time series data, as the Nuffic definition indicates, depends greatly on the programme studied. Both the Graduate Impact Surveys for the Erasmus Mundus scheme (e.g. Säring, Spartakova, and Wegera, 2012) and DAAD's (2013) analysis of their postgraduate courses, for example, collate data from several surveys for cross-sectional and time series analysis. Conversely, there are several tracer studies of AusAID awards in Fiji (Bryant and Wrighton, 2008: AusAID, 2011) for whom the target populations overlap but in which the studies do not pool data for time series or comparative analysis. This partly reflects changing evaluation priorities over time (AusAID, 2011), and whilst some scholarship providers have conducted regular tracer studies over a period of 15+ years (e.g. DAAD) others appear to have conducted a relatively large number of tracer studies in a short period of time (e.g. AusAID).

Tracer studies can be more or less well defined methodologically. It is useful to delineate the use of the phrase 'tracing', the project of finding and contacting one's alumni, from 'tracer studies', which use the results of the tracing process as a participant group for assessing the outcomes of scholarships by examining the post-scholarship trajectories of alumni. The majority of tracer studies reported in the sector focus on the evaluation of programme outcomes (e.g. Webb, 2009), but there are a minority which are predominately about the tracing process itself (e.g. Chernikova, 2010). A limitation of all tracer studies is the efficacy of tracing: populations for study are derived from those successfully traced by evaluators and who have volunteered to be in contact with alumni programmes (either active participation or merely to have their contact details recorded). Tracing, and thus tracer studies, relies heavily on effective database systems and the decisions made by alumni officers in the tracing process (Creed, Perraton, and Waage, 2012). In some cases the lack of systematic tracing prior to the evaluation has been a difficulty for the evaluators (e.g. Bryant and Wrighton, 2008).

Whilst ex-post evaluations are dominant they are not universal in the sector. Several scholarship programmes have been designed and implemented with concurrent evaluation frameworks, including the Ford Foundation's IFP and MasterCard Foundation Scholars Program. In these cases monitoring and evaluation frameworks have been designed to collect data alongside scholarship schemes, compiling time series data on particular cohorts as well as investigating cross-sectional issues of interest to programme policy makers (see Enders and Kottmann, 2013: Cosetino et al., 2013). These frameworks extend regular monitoring and evaluation practice, conducted by most scholarship providers on selection and scholarship process (e.g. completion, immigration monitoring), but link concurrent programme evaluation closely with indicators pertinent to policy objectives. As those involved in the IFP (e.g. Clift, Dassin, and Zurbuchen, 2013: Enders and Kottmann, 2013) have observed, it is relatively uncommon to have concurrent evaluation programmes which interface directly with programme policy at a short time interval: ex-post evaluations with policy recommendations tend to be conducted at the conclusion of schemes or with lengthy time intervals between reviews.

Additionally, not all ex-post evaluations are tracer studies. Chesterfield and Dant's ex-post evaluation of USAID's LAC programmes, for instance, uses 'hybrid performance evaluation' (2013; 3) in which scholarship recipients are compared to a ''post-facto proxy control group' (2013; 29) of non-recipients. Similarly, Ramboll's (2012) ex-post evaluation of NPT and NICHE uses a 'post only non-equivalent comparison design' hybridised with Contribution Analysis (Mayne, 2011).

Nonetheless, if evaluation reports that are solely policy reviews and/or collect no original research data are excluded, ex-post evaluation through tracer studies is the prevalent analysis strategy in the sector.

Longitudinal data

Despite both the lengthy timeframe involved in scholarship awards and the extensive history of some scholarship schemes there has been relatively scarce investment in longitudinal analysis. Because most evaluation has been ex-post, the collection of longitudinal data with scholarship cohorts has been limited to rare instances in which both a concurrent evaluation strategy has been employed and this strategy has carefully tracked specific cohorts over the course of their scholarship and post-scholarship trajectory.

Of the evaluation reports examined, the clearest example of longitudinal design was employed by Amos et al. (2009) in the evaluation of the Gates Millennium Scholars (GMS) programme. In this instance data was collected from scholarship recipients and a comparison group of non-recipients at selection (baseline) and at several follow-up intervals. Yearly cohorts of scholarship recipients have been tracked uniquely, rather than

as a general pool of alumni as is the case in most tracer studies. Longitudinal evaluation of GMS allows a richer and more robust analysis of change than is available to many ex-post evaluations. Most notably, the collection of both baseline and multiple subsequent data points allows a more specific analysis of *when* change occurs, in addition to *whether* it occurs, and more comprehensive data on recipients' trajectories than may be available through ex-post reconstruction. Several of the advantages afforded by the GMS evaluation are, however, derived from both the longitudinal design and the use of a comparison group: the latter topic is addressed later in 'The counterfactual' on page 22.

Several other evaluations have employed designs that, whilst not truly longitudinal, approximate longitudinal data. The Erasmus Mundus Graduate Impact Surveys collect data on both alumni and current scholars, and, whilst the data does not form a panel study, approach respondents who may respond on multiple occasions as both current scholars and subsequently as alumni (see Säring, Spartakova, and Wegera, 2012: ICUag.net, 2013). The post-programme tracer study of the Ford Foundation's IFP (e.g. Tvaruzkova and Clift, 2013), scheduled to continue until 2023, will also generate a form of longitudinal analysis by collecting data from IFP alumni at multiple points within the next decade. Like the Erasmus Mundus Graduate Impact Surveys, the IFP tracer study is not designed as a baseline and follow-up cohort study, but it will study the experiences of alumni over time[6] and thus shares the aims of longitudinal research.

Country-level and scheme-level designs

Another potential distinction between the various approaches to evaluation is whether the methodological design focuses on countries / geographic regions or on segments or the entirety of a scheme.

There have been numerous country or region specific evaluation reports, including those relating to Vietnam, Cambodia, Fiji and Tuvalu, the Caribbean, Continental Europe, India, and Ethiopia. Conversely, there have been several scheme-wide evaluations that examine outcomes without focusing on a specific region, such as evaluations of NFP, NPT and NICHE, CFSP, and JISPA. Some evaluations which appear geographically bounded are also scheme-wide due to the specificity of the geographic origins of scholarship recipients, such as the evaluations of USAID's LAC (Chesterfield and Dant, 2013) and ATLAS/AFGRAD (Gilboy et al., 2004) programmes.

These are, of course, not mutually exclusive approaches. Danida's fellowship programme, for instance, uses both country case studies and a scheme-wide focus in its programme evaluation (Ministry of Foreign Affairs of Denmark, 2012). Only the most specifically targeted schemes can conduct empirical evaluations solely on a geographic basis, most conduct some cross-scheme analysis and use fieldwork in a subset of countries to examine particular cases and geographic issues (e.g. Nijathaworn et al., 2009). Evaluations of bilateral scholarship schemes (e.g. Marshall Scholarships) are, of course, inevitably focused at country-level.

The relative merits of scheme-wide and geographically-bounded evaluation strategies are complex and likely vary by scholarship and policy objective. For non-bilateral public diplomacy scholarships (e.g. Chevening) there is perhaps less need to conduct analysis at specific country-level, since the aims of such schemes are rarely linked to labour capacity in developing countries. Conversely, institutional capacity, skills shortages, and developmental impact are often more easily addressed at country- or region-level due to the variance between geographic areas. Some scholarship schemes are part of country-level initiatives and thus it logically follows that they be evaluated at country level: VLIR-UOS programmes in Ethiopia (Penny and Tefera, 2010) and Vietnam (Visser and Trinh, 2011) are examples of this practice.

There are, however, often good reasons to design evaluation at scheme-wide level. Some schemes, such as the Commonwealth Scholarship and Fellowship Plan, have a broad international base for scholarship awards and seek to impact developmental issues globally or within an extensive range of countries. Although analyses of impact in specific countries can be conducted, they are unlikely to well represent the impact of a scheme as a whole and so would need to be complemented by scheme-wide analysis. Additionally, the objectives of some non-developmental scholarship schemes focus on the individual scholars and their host countries and so it would be less useful to analyse outcomes geographically. Erasmus Mundus, for instance, provides awards for study in Europe under the auspices of raising the profile of European higher education, rather than influencing the developmental context of (or political relations with) a specific country sending recipients to European host institutions.

Although in broad terms the sector is evenly split between country-level and scheme-wide approaches, most large, externally commissioned evaluations of scholarship schemes tend to take a programme-wide approach with reference to countries as case studies (e.g. Ramboll, 2012). Country-level approaches, conversely, are more common for tracer studies and for scholarships which are designed as part of a bilateral developmental relationship (e.g. VLIR-UOS in Ethiopia).

[6] Mirka Tvaruzkova, personal correspondence (March 26th, 2014)

Kirkpatrick's model

Given the breadth of evaluation work on-going in the sector there are surprisingly few extant methodological frameworks referenced in published reports. Contribution analysis (Mayne, 2011) has received some limited attention (e.g. Rotem, Zinovieff, and Goubarev, 2010), but, with the exception of Ramboll's (2012) part-adoption of the approach, has yet to find extensive use within the sector. Kirkpatrick's (1994) four-level evaluation model has seen wider use, including in Denmark, Canada, the US, and prospectively within the Ford Foundation's 10-year post-programme study of IFP. Kirkpatrick's framework is thus the dominant evaluation model within those evaluations drawing on an extant approach, although in practice most evaluations do not draw upon an externally published methodological approach.

The Kirkpatrick evaluation model originally comprised four levels – reaction, learning, behaviour, and results – in which the effects of intervention (or training) were examined, from the immediate to the longer-term. The model has enjoyed enduring popularity in a variety of settings (see Alliger and Janak, 1989: Bates, 2004) – and some criticism (e.g. Holton, 1996) - and has been adopted by several evaluations of international scholarship schemes (e.g. Gilboy et al. 2004: Tvaruzkova and Clift, 2013). Because the Kirkpatrick model was designed for evaluating institutional / organisational outcomes, evaluations of international scholarship schemes have typically added an additional level to the model that examines sector, community, or national impact beyond the institution (e.g. Gilboy et al. 2004).

Whilst those analyses using Kirkpatrick's model are significantly shaped by the approach, there is little indication of the framework having shaped the landscape of research methodology in the sector more broadly. Kirkpatrick's model appears to have been taken up sporadically by programmes and is not used consistently by any donor or provider. The USAID funded ATLAS and AFGRAD schemes, for instance, are evaluated with the Kirkpatrick model at the heart of the process (Gilboy et al. 2004), whereas the USAID funded CASS / SEED scheme is evaluated without reference to Kirkpatrick (Chesterfield and Dant, 2013). Similarly, the Kirkpatrick model is not used within any of the current evaluation reports published on the Ford Foundation's IFP, but is used in the structuring of the 10-year post-programme study (Tvaruzkova and Clift, 2013). Nor can the use of the Kirkpatrick model be centred to a particular time period (used in 2004, 2005, 2012, and 2013) or particular group of external consultants.

The effect on evaluation practices of both Kirkpatrick's model and Contribution Analysis is thus somewhat unclear. There are indications that evaluators are considering scholarship outcomes in terms of the 'plausible contribution' of schemes to life trajectories (e.g. AusAID, 2011: Visser and Trinh, 2011), which would seem to partly reproduce the logic of contribution analysis, but this appears to emerge from a broader concern in the sector with the difficulty in attributing causality to scholarships[7], rather than an explicit influence from Contribution Analysis.

[7] Emily Hayter, Heath Thomson, and Beryl-Joan Bonsu, personal correspondence (February 20th, 2014). Joan Dassin, personal correspondence (February 21st, 2014).

3. Methods

As methodology has been differentiated above as the framework informing evaluation design, it follows that 'methods' are the practical tools applied to collect data within that framework.

Whilst there have been a number of variations in data collection techniques within the sector, the prevalent techniques are:

1. Self-report surveys,

2. Interviews with stakeholders

3. Documentary analysis

A brief outline of the use of these methods follows, followed by observations on the dominance of mixed-methods research within the sector and a commentary on the data collection instruments employed within evaluation research.

Self-report surveys are the dominant evaluation method within the sector, with almost all evaluation reports informed by one or more surveys. This is perhaps unsurprising given the flexibility of survey tools to address multiple topics and to be distributed at relatively low cost worldwide. Most surveys have tended to be administered ex-post, but there have been several examples of evaluations in which surveys have been administered both whilst scholarship recipients were undertaking study and after their scholarship had concluded (e.g. Säring, Spartakova, and Wegera, 2012: Enders and Kottmann, 2013). In the majority of cases surveys have been administered to scholarship alumni, but in several instances – notably scheme-wide, non-tracer study research – evaluators have also surveyed scholarship administrators, host institutions, project partners, and employers (e.g. Nijathaworn, Semblat, Takagi, and Tsumagari, 2009: van der Aa, Willemson, and Warmerdam, 2012). Surveys have also tended to be administered online where possible; however, in countries where tracing has been difficult or internet access uneven, surveys have also been conducted in face-to-face meetings (e.g. Bryant and Wrighton, 2008).

As with most elements of evaluation, providers appear to be at different stages in the strategic development of survey measures and as such the use of surveys has been uneven across the sector. Some providers have deployed systematic and regular survey research throughout the duration of the scholarship scheme: the concurrent evaluation of the Ford Foundation's IFP, for instance, involved 23 surveys conducted across the decade-long programme, including surveys of selected and non-selected applicants, current scholarship holders, programme partners, and at least 6 alumni surveys[8] (Enders and Kottmann, 2013). Other providers, such as AusAID, have invested more heavily in surveys as part of recent moves toward more regular and extensive evaluation (Gosling, 2008: AusAID, 2011). There are some difficulties evident in the latter approach, highlighted by Nugroho and Lietz's (2011) analysis that of 17 AusAID post-scholarship surveys only 5 could be considered 'high quality' and data comparability was a serious challenge because of the differing designs of the survey tools used. The CSC has had similar experiences with resolving historical differences in survey tools and it seems likely that, whilst not necessarily discussed widely in published evaluation reports, this concern has resonance across the sector.

Interviews have also been used extensively within the sector to collect evaluation data. Numerous evaluators have used semi-structured interviews as the primary mode for accessing qualitative data, typically interviewing alumni (e.g. Webb, 2009) but, and to a greater degree than surveys, also engaging with employers / managers and scholarship coordinators (e.g. Hansen et al., 2005: Bryant and Wrighton, 2008). Despite the global focus of scholarships, interview research has only rarely been conducted by computer-mediated-communication or by telephone (e.g. ECOTEC, 2009) and has usually been part of in-country field research (e.g. Norad, 2009: van der Aa, Willemson, and Warmerdam, 2012). Although not all reports specify the staff involved in each facet of evaluation, in-country research has often involved local researchers retained to conduct interviews (e.g. Chesterfield and Dant, 2013). When working in a large number of countries the involvement of local specialists ina broad research team seems inevitable (e.g. Ramboll, 2012: Tvaruzkova and Clift, 2013), but there have also been specific reasons for local consultants and research assistants to join evaluation teams: local language proficiency being an important example (e.g. Webb, 2009) [9].

[8] This tally excludes the global alumni surveys and global organisations surveys planned as part of the 10-year post-programme study: Tvaruzkova and Clift (2013)

[9] Awareness of cultural practices and interaction styles – particularly the best approaches to elicit research data in qualitative interviews– may be another advantage to involving local researchers.

Finally, almost all evaluation work in the sector involves an examination of project reports, financial accounts, briefing documents, and internal policy papers that can be broadly described as 'documentary analysis'. Particularly when external consultants are commissioned to evaluate programmes (e.g. Ramboll, 2012), it appears routine for researchers to begin their evaluation with a detailed reading and analysis of existing project documentation and reports; particularly with regard to the financial administration of programmes. Regular evaluation updates – such as tracer studies that form part of evaluation reports (e.g. World Bank Institute, 2010) – do not usually conduct these forms of documentary reviews, presumably because such research is unnecessary in the context of an on-going evaluation programme.

Dominance of mixed methods

Evaluation of international scholarships for higher education is predominately based on mixed-methods: using both qualitative and quantitative approaches to gather research data. There have been examples of (almost) entirely quantitative research (e.g. Amos et al., 2009) and qualitative research (e.g. Visser and Trinh, 2011), but these are exceptions to the general trend toward a mixed-methods design in which surveys collect quantitative data (and some free-text comments) alongside interviews that collect qualitative data.

There is very limited discussion in evaluation reports of the merits or demerits of mixed-method design; it is generally taken as assumed that mixed-methods research is appropriate for answering the research questions set out by the evaluation. Triangulation – or the testing of concordance between findings from different research methods – has been cited as a motivation for mixed-method approaches (e.g. Gilboy et al., 2004: Norad, 2009), although there have been no instances in which the actual process of triangulation has been detailed, given that there are multiple possible approaches (Madill, Jordan, and Shirley, 2000). Nor is it evident, for instance, what is the stance of any evaluators (or scholarship providers) toward philosophical issues – such as epistemology, ontology, and ethics - in research, but it is questionable whether evaluation reports would be the appropriate space to examine such academic issues and so it seems reasonable to assume that when discussions about these topics are conducted it is 'behind closed doors'.

More importantly, the detail given on some aspects of mixed-methods data collection is often quite limited. Interviews in particular are often merely stated to have been 'conducted', with evaluators rarely describing in what circumstances interviews took place, whether they involved closely defined or broad questions, were recorded and transcribed, and whether they involved only one or several informants. Similarly, language is not discussed by the majority of evaluations, yet may be relevant when dealing with countries in which English is not the main language and there is potential for informants' comments to be 'lost in translation' when reports are published in English (as is often the case). Lack of detail on these methods is certainly not universal - Visser and Trinh (2011), for instance, discuss the arrangements of data collection in significant detail – but it is sufficiently common as to be a sector-wide concern.

Data access has proven to be a challenge across the sector and this is undoubtedly one factor guiding evaluators toward mixed-method approaches. Response rates to surveys, for example, have varied, affected by both the circumstances of survey administration and the ways in which response rates are calculated. Large-scale post-programme surveys of alumni have tended to achieve lower response rates than concurrent evaluations involving current recipients (see, for instance, Enders and Kottmann, 2013). Although an average response rate could potentially be calculated from what published data is available, it would likely be uninformative since the circumstances of survey administration have differed so radically. One area in which response rates have been particularly poor is in surveys to employers, with Nuffic (2009) reporting only an 8% response rate to their survey of employers of scholarship alumni. Low response rates inevitably raise concerns about excessive sample variance (see Blair, Czaja, and Blair, 2014), but there is widespread recognition that a more pressing problem is nonresponse bias in which those who reply to sample surveys are likely to be engaged with alumni associations or tracing (e.g. Day, Stackhouse, and Geddes, 2009) and disproportionately represent the 'successful' outcomes of scholarship programmes.

In-country fieldwork and qualitative research methods appear to have had more success in garnering desired levels of participation, particularly in the context of reaching informants for whom current contact details were not available (Chesterfield and Dant, 2013). Yet some difficulties have also been reported with interview scheduling and securing participation (e.g. Visser and Trinh, 2011), and it is also important to consider that interviews do not have a 'response rate' per se (the number of persons invited to interview is rarely reported) and so it is difficult to compare levels of participation across interviews and surveys[10]. In this context, mixed-methods research may be advantageous to maximise opportunities for securing participation from the target population.

[10] Although it is interesting to note that the rate of participation refusal in Chesterfield and Dant's (2013) in-country interviews was very low, even amongst a cohort of non-selected scholarship applicants (not alumni)

Data collection instruments

The main data collection instruments within the sector are survey questionnaires and interview question schedules.

It is not uncommon for survey instruments to be included in a report annex (e.g. Ramboll, 2012), offering a useful insight into the types of questions being employed by evaluations. Survey questions of three main forms have been used: closed questions for discrete information (e.g. demographics), 'free-text' comments, and Likert-style questions. Free-text questions are not used as extensively as perhaps might be expected, although this might reflect the limited articulation of qualitative data analysis strategies in the sector (see 'Qualitative data analysis' on page 19); Nuffic (2009), for instance, used free-text comment questions in a survey but subsequently concluded that free-text data was too difficult to analyse within the scope of the research. Likert-style questions are the preferred format to collect data on perspectives and experiences, typically asking respondents to self-report their extent of agreement with statements about scholarship outcomes. None of the evaluation reports reviewed discussed their choices in the design of Likert-style questions - such as the number of scale points or labelling conventions[11] - but it appears that 5-point scales, with a neutral midpoint and each point labelled, are the preferred approach (e.g. World Bank Institute, 2010). As the use of multiple questions to indicate an underlying construct has been uncommon, ratings of reliability and validity have not been extensively reported[12].

More generally, external validation of instruments or the appropriation of instruments from previous research does not appear to have be prevalent, indicating that instruments are most often designed and used within a specific evaluation study. Some exceptions to this trend exist. Dong and Chapman (2008), in their study of the Chinese Government Scholarship Program, administered a survey substantially based on Pascarella and Terenzini's (1980) Institutional Integration Scales, examining the experiences of scholarship recipients in China and the predictors of their satisfaction. Dong and Chapman's work examines impact only through the lens of recipients' disposition toward China and Chinese culture ('soft power' outcomes) during the scholarship, however; it does not assess long-term impact. More usually, instruments have been created for the purposes of individual evaluation studies. As Nugroho and Lietz (2011) have observed of AusAID surveys, instruments have tended to be slightly different and this has caused difficulty in comparing findings, although there have recently been some strategic attempts to unify evaluation tools (e.g. DFAT, 2011). The variety of survey instruments used raises questions about how effective each instrument is in comparison to others and, especially where evaluations address similar variables (e.g. employment trajectory), there appears to be scope for 'best practice' to emerge: likely in the form of well-piloted, validated survey instruments.

Because survey research is the dominant method in the sector, evaluations rely heavily on self-report data from respondents and it is important to briefly recount some of the potential pitfalls with this form of data collection. Firstly, survey response rates are limited both to those who can be traced and, importantly, to those who choose to take part. The tendency for those at the poles of opinion – and particularly the highly positive - in alumni populations to respond to a survey is a representativeness concern and does not go unnoticed by some evaluators in the sector (e.g. Amos et al. 2009: Day, Stackhouse, and Geddes, 2009; van der Aa, Willemson, and Warmerdam, 2012). High response rates for some surveys (e.g. Enders and Kottmann, 2013) help to offset concerns about the representativeness of the sample, but, and particularly with long-running programmes that have accrued thousands of alumni, there remains a concern of positive response bias. The lack of negative case review (exploring the 'unsuccessful' outcomes) within the qualitative elements of many mixed-method approaches exacerbates the concern with survey positive response bias; the space for 'bad news' to find voice within evaluations often appears notably smaller than that for 'good news'.

More technical issues with self-report survey instruments have been seldom discussed within published evaluations. Perhaps the most important issue that has not been widely addressed is acquiescence bias: the tendency for respondents to rate more highly statements about successful outcomes, especially where this is seen as desirable to the evaluators. One common strategy for dealing with acquiescence is to balance positive and negative statements targeting the same issue – for example, including polar questions on having gained useful skills and on *not* having gained useful skills – and to reverse the subsequent statistical coding on negative statements. None of the evaluation reports indicated they had used this strategy, or indeed any strategy, for reducing possible bias in responses. There are, of course, strong incentives for evaluations to produce positive results (even when conducted by external contractors) that can help to guarantee the on-going funding of a scholarship programme and so evaluators must inevitably mediate a difficult political tension in ensuring rigorous research findings whilst maximising opportunities to show the programme in a positive light.

[11] See Dolnicar (2013) for an overview of research on Likert-scale design

[12] Amos et al. (2009) do use a compound variable to indicate a construct (leadership) and report Cronbach's alpha (internal consistency)

4. Variables and indicators

Various possible definitions are available for 'variables' and indicators, but this study takes a rather loose approach and considers variables to be the concepts and issues with which evaluations concern themselves. To direct data collection methods, evaluators must identify and operationalise variables on which they intend to collect data. Variables can be considered at a variety of levels, from the very broad (e.g. 'institutional capacity') to the very specific (e.g. 'number of alumni in public sector employment within 1 year of scholarship completion'). This chapter examines the abstract and specific variables regularly explored by evaluators, before considering issues arising.

Defining macro variables

The most common approach to identifying variables, both within this sector and elsewhere, is to work from a general aim or research question and operationalise increasingly specific levels of variables until evaluators have identified a set of indicators that can be measured. The process is demonstrated effectively by the working papers on Norad's programme for master's studies (NOMA) in which a research programme concurrent with NOMA attempted to define quantitative and qualitative indicators by which the core issue (higher education institutional strengthening in the South) could be assessed (Andersen and Tobiasen, 2007). Developing indicators in this way can be underpinned by theory from higher education and international development: Ramboll (2012) used existing theory on capacity development, for instance, to develop indicators for their evaluation of NPT and NICHE. The development of variables for analysis is not always fully explained in evaluation reports, but, as in the case of NOMA, may be occurring parallel to evaluation research.

In some cases the theory of change or underpinning conceptual logic of scholarship programmes has been sufficiently well detailed that both higher-level variables (e.g. 'institutional capacity building') and specific indicators (e.g. 'research publications in the five years post-scholarship) are set out in advance of evaluation studies being undertaken (e.g. Cosetino et al., 2013). Well-developed indicators that are coherent with core policy objectives can direct evaluation effectively, as long as evaluation research has sufficient latitude to report unexpected outcomes and findings that are not well captured by pre-designed indicators (Creed et al., 2012). Somewhat counter-intuitively, the analysis of unplanned outcomes can be carefully planned (e.g. Carpenter and de Vivanco, 2013), particularly when using qualitative methods that allow for more detailed elaboration on topics than closed survey questions.

An alternative approach to creating variables is to use an extant group of high-level criteria and tailor evaluation indicators to offer insight into these issues. The OECD's Developmental Assistance Committee (DAC), for instance, has produced criteria for evaluating developmental assistance, focusing on relevance, effectiveness, efficiency, impact, and sustainability (OECD, 1991: 2002). These criteria have been used to structure several scholarship programme evaluations in the sector, including programmes provided by AusAID (DFAT, 2010: Barber and Hel, 2012), Norad (Hansen et al., 2005), VLIR-UOS (Visser and Trinh, 2011), and the European Union (ECOTEC, 2009). The use of the OECD's DAC criteria is helpful insofar as it provides a common focus for evaluations conducted on different programmes and in different countries or time periods. As Creed et al. (2012) have observed, however, the DAC criteria are not in themselves operational variables: concepts such as 'impact' or 'relevance' clearly need further explication within the context of a particular programme to usefully structure evaluation (relevance to whom? In the context of what policy? Impact on what? And so forth). As such, the uses of DAC criteria have also required the further generation of variables to inform evaluations, except in the rare cases in which an evaluation has performed a secondary analysis of existing data through the lens of the DAC criteria (e.g. Barber and Hel, 2012).

Whether the use of macro-level variables is warranted in a particular evaluation likely depends on what purpose is determined for the research. Tracer studies are often conducted as part of a broader programme of evaluation (see, for instance, Nuffic, 2009) and so repeatedly referring to the 'bigger picture' may not be appropriate with only quite specific research data about a subset of alumni. Conversely, macro-level analyses which do not draw heavily on bodies of research evidence showing micro-level indicators (e.g. Feiler, Jager, and Reiter, 2007) can be somewhat abstract and leave questions about systemic impact unanswered. Methodological frameworks that specify levels of analysis – such as Kirkpatrick's model and Contribution Analysis – may assist in providing the balance between foci. It is notable that in the design of newer programmes (e.g. MasterCard Foundation Scholars Program) effort is being invested into determining core indicators (variables) for each desired policy objective prior to, or alongside, the programme being implemented (e.g. Cosetino et al., 2013).

A useful commentary, provided by both Penny and Tefera (2010) and Andersen and Tobiasen (2007), reflects on the need for indicators to reflect the perspectives of all stakeholders within programmes. In the case of the VLIR-UOS Ethiopia evaluation, for instance, Penny and Tefera observed that concepts such as

'relevance' and 'quality' may be defined quite differently by Northern and Southern partners. Evaluations taking only one perspective – most likely that of the Northern donor – are likely to present a limited analysis that may omit both successes and difficulties from the perspective of other stakeholders. This concern is perhaps more immediate for institutional capacity building programmes than for scholarship schemes, but there are significant parallels with schemes focusing on capacity development in employment sectors (e.g. JISPA: Nijathaworn et al., 2009).

More generally, and as Andersen and Tobiasen (2007) observe, indicators used need to be widely understood and accepted by all individuals and organisations involved. In pursuit of this end, we may benefit from considering carefully the indicators proposed by partner institutions or individuals rather than relying on measurement strategies solely developed by scholarship providers. The selection strategy employed by the Ford Foundation's IFP - involving a global patchwork of regional panels to define 'need' within particular countries or communities (see Dassin, Enders and Kottmann, 2012) - is an example of how some scholarship providers have already invested in the shared definition of policy objectives.

Value for money can also be considered a higher-level variable; however, discussions of value for money have been sufficiently important to the sector that they have been treated within a separate thematic section later in this report (see page 25).

Specific variables

At the level of specific variables examined in evaluation research, several key areas are focused upon by the majority of evaluations:

1. Socio-demographics of candidates
2. Scholarship process and satisfaction
3. Return to home country rate
4. Change in personal competencies
5. Post-scholarship employment trajectories
6. Post-scholarship contribution to sector, profession, community, or country
7. Links / networks to scholarship hosts

Foci differ slightly between reports, but the list above is broadly representative of the topics examined within almost all evaluation reports examined.

These variables are consistent with the policy objectives of most scholarship schemes, particularly those with an international development focus and which aim to catalyse labour market outcomes via capacity-building scholarships (e.g. Nijathaworn et al., 2009). Public diplomacy scholarships (e.g. Chevening) still tend to examine these topics, but with a greater emphasis on links / networks between alumni and their hosts and the perceived reputation of the host country or scholarship programme. Analysis of return rates can also differ somewhat for these schemes. Evaluation of Erasmus Mundus, for instance, examined recipients' disposition toward returning to work in the EU, rather than their rate of returning to work in their countries of origin (Säring, Spartakova, and Wegera, 2012). Similarly, the evaluation of Australian International Postgraduate Research Scholarships (IPRS) is not concerned with examining return home rates, given a policy objective of the scheme is to attract and retain skilled researchers in Australia (Department of Innovation, Industry, Science and Research, 2010).

> *'The evaluations are excellent at measuring what has been done (results, outputs, KRAs [Key Results Areas]) and at allocating scores (e.g. bad, good, excellent, better than planned etc.), giving a sense of objectivity and quantitative appreciation, but the real impact at the level of the individual, the department, the campus, the university, the local area and at regional and national levels, the society, is hard to materialize and quantify'*
>
> (Janssens de Bisthoven, 2009: 72)

Programme evaluations are concerned with immediate, medium-term, and long-term impacts, and so the variables examined span these timeframes. Analyses of socio-demographics and scholarship process primarily concern initial selection policy and experiences whilst on scholarships, both of which relate to immediate outcomes of providing scholarships. Change in personal competencies and links / networks to scholarship hosts

are short to medium term outcomes. Post-scholarship employment trajectories and contribution to development – or to bilateral relations – are medium to long term issues which have tended to be measured over a period of several years post-scholarship.

The majority of variables are assessed through alumni self-reporting, although data on some monitoring variables (e.g. gender of recipients) are collected during application processes. As noted in chapter 3, 'Methods', the prevalent approach to assessing issues of perspective (satisfaction, contribution etc.) is the Likert-style scale, and so it is very common to find alumni asked to rate their level of agreement with statements such as 'The skills and knowledge gained from my scholarship are relevant to my employment'. However, the data collected in addressing variables is not always quantitative. Indeed, quantitative data often fails to adequately capture the some of these issues: particularly notions of 'contribution' (Janssens de Bisthoven, 2009).

Several specific issues in variable choice warrant further examination: the use of compound variables, the types of socio-demographics examined, some of the less common (but potentially informative) variables explored, and the complexity of measuring concepts such as 'return'.

Compound variables

In a small minority of evaluation studies, compound variables have been used to develop more sensitive indicators for concepts such as 'leadership'.

Compound variables are created by collating results from other variables, sometimes weighting certain components more highly than others if they are deemed more influential on the construct being represented by the compound variable. Amos et al. (2009), for instance, used four survey items to create a compound score that they term a 'leadership index', which reported good internal consistency as an indicator of the leadership construct. Similarly, Chesterfield and Dant (2013) created 'aggregate' (compound) variables for leadership and management experience amongst scholarship recipients and non-recipients. Ramboll (2012) created both a 'cost-quality index' and a 'capacity development index' (CDI); the latter through a process of establishing the difference between expert-determined mean capability ratings prior to and after an intervention. These composite variables often allow for more sophisticated analysis than basic self-report variables alone. Using changes in CDI, Ramboll (2012) are able to offer not only a measure of baseline and post-project capacity in countries, but also to show the magnitude of these gains on a common scale. Perhaps most notably, the CDI allows Ramboll to offer a tentative figure for how much financial investment has been required in specific countries to increase this uniform index by a specific amount (e.g. 0.1).

Construct validity can be a concern with compound variables (i.e. do they measure what they claim to measure?), but valid uniform indices facilitate both systematic comparison across projects and countries and, often, more sophisticated and holistic analysis of individual post-scholarship trajectories. There are also potential pitfalls in quantifying the value of outcomes that are highly subjective. Capacity development is perhaps one of the variables more amenable to transformation into an index, but it is foreseeable that difficulties might occur if uniform, transnational indices were attempted in the cases of, for example, catalysing community change or social justice.

Demographics

Collection of basic demographic data on applicants is part of the selection process for many scholarship schemes and so this data has usually been available to evaluators, notwithstanding any problems working across the different databases of alumni and current scholarship recipients. Only one demographic variable – gender - receives sustained attention throughout the sector (within international development-oriented scholarships at least). Gender is frequently a policy issue for scholarship schemes (e.g. Feiler, Jager, and Reiter, 2007) and so many analyses include specific focus on the differential experiences or outcomes of male and female recipients. Few of these analyses go beyond descriptive reporting and analyse in detail the potential reasons for, and solutions to, differences between outcomes for male and female scholarship recipients. Conversely, some analyses of social disadvantage more generally include gender as a facet of possible disempowerment and have provided compelling evidence of the impact scholarship schemes can have on social change (e.g. Clift, Dassin and Zurbuchen, 2013: Mansukhani and Handa, 2013). Gender is often of paramount importance in understanding social disempowerment and, potentially, in differing long-term impact of scholarships, but it is also important to recognise, as some evaluators have (e.g. Enders and Kottmann, 2013), that gender is one of several demographic variables that might impact on social disempowerment and trajectories into and out of scholarship schemes.

A socio-demographic variable that has been reported in a minority of evaluations, but which perhaps deserves broader discussion, is the relative wealth and education level of scholarship recipients. Negin (2014), in his analysis of AusAID scholarships in Africa, has observed that a sizeable minority (34%) of

scholarship recipients had one or more parents who had undertaken tertiary education. Enders and Kottmann (2013) also examined familial wealth and education in profiling the socio-demographic background of IFP fellows. The relative wealth or social advantage of scholarship recipients is often salient to the aims of programmes and so may require further analysis. As Negin (2014) has observed, forming networks with elites is more likely to secure influence in local and national policy (either for development or public diplomacy purposes) than forming networks with disadvantaged or marginalised groups, but is also likely to perpetuate existing social inequalities, including in education and wealth.

Moreover, whilst programme such as the Ford Foundation's IFP have demonstrated that non-elite groups can be effective change agents (Clift, Dassin, and Zurbuchen, 2013), the kinds of impacts achieved by recipients with different social status and wealth can be disparate[13]. This raises the case for evaluation to also contextualise the long-term outcomes of awards within the socio-demographic circumstances of recipients, rather than solely in relation to the policy objectives of donors. Clearly there is a strategic balance to be achieved between understanding outcome variables (such as post-scholarship employment) through the lens of donor objectives and the lens of an individual alumnus' life circumstances.

Issues from understudied variables

Three issues infrequently examined, but which raise important questions for evaluation, are the comparison between expectations and outcomes, the impact on national workforces whilst scholars are absent in host countries, and the process of reintegration.

In the Erasmus Mundus graduate impact surveys (e.g. Säring, Spartakova, and Wegera, 2012) and ex-post evaluation (ECOTEC, 2009) a short analysis was conducted of the self-reported 'greatest impact' of the scholarship. The structure of the evaluations allowed comparison between responses by current scholars and programme alumni, and the results showed important differences between the greatest impact anticipated by current scholars and that experienced by alumni. Similarly, Gondwe and Schröder (2013) included unmet expectations as a variable for analysis in their evaluation of IFP in continental Europe. Comparison between expectations and outcomes is not widely addressed within current evaluation research, but in cases where scholarship outcomes are complex and differ from those intended by policymakers there is evidence that different expectations (between beneficiaries and policymakers, for instance) can be an influential factor (e.g. Webb, 2009).

A second variable rarely addressed is the impact of scholarship schemes on the national and organisational workforce *during* the recipients' absence. In their analysis of AusAID scholarships in Fiji and Tuvalu, Bryant and Wrighton (2008) included a short analysis of how employers cope with the absence of staff during the period of scholarships and whether scholarship policies may inadvertently contribute to labour shortages. Later tracer studies of scholarships in Fiji re-examined this issue and additionally noted the need for 'reintegration planning' in scholarship policy (AusAID, 2011). Unfortunately, no assessment of the net benefits – i.e. were the gains from the scholarship worth any impact from the absence of the scholar? – is forthcoming in currently published work. Although it may not be relevant to all providers to understand the impact of scholarships on recipient communities during the scholarship tenure, for those working at the level of specific countries (and arguably more so for small states) this form of programme impact could be highly pertinent.

Thirdly, and also touching upon 'reintegration planning', is the analysis of reintegration experiences. The majority of evaluation research analyses reintegration, but focus is usually on labour market trajectories (e.g. Ramboll, 2012). In a small minority of evaluations (e.g. Nuffic, 2009: Webb, 2009), re-integration back into home communities and professional or personal difficulties upon return from scholarships are addressed. As Clift, Dassin and Zurbuchen have observed, the 're-insertion challenges' (2013: 36) can be significant obstacles to realising the potential of scholarships and achieving policy objectives. As such, scrutiny of immediate experiences when scholars return home – in terms of labour market re-entry, community relationships, building networks to support activity, and so forth – could be useful to all scholarship schemes, notwithstanding that this must be at least partly distinct from impact analysis because socioeconomic impacts will almost certainly take much longer to realise.

Return and 'brain drain'

Return to home countries and the avoidance of brain drain has been much discussed throughout the sector. Certainly for developmentally-oriented programmes, the tendency of recipients to return to their home country or region is an important outcome of the scholarship. Different schemes have taken a variety of approaches to ensuring a suitable return rate, from selecting candidates with strong commitments to their local communities (e.g. Dassin, 2009) to enforcing a visa embargo to send recipients home at the end of

[13] Emily Hayter, Heath Thomson, and Beryl-Joan Bonsu, personal correspondence (20th February 2014)

their scholarship (AusAID, 2011). There are, however, a number of complexities in assessing return that belie simple representation as a percentage of scholars who returned to their home country post-scholarship, of which institutional brain drain, compound brain drain, and the contributions of non-returning recipients are perhaps most pressing.

Whilst most of the analysis concerning return and brain drain has focused at a national level – the avoidance of exodus from home countries as a result of the scholarship scheme – there have been occasional references to another potentially problematic form of labour movement: institutional brain drain. Scholarship evaluation seems to suggest programmes are generally successful in mitigating international brain drain with high return rates (e.g. 77%: World Bank Institute, 2010). There is much more limited analysis, however, of the effects within countries of post-scholarship employment trajectories on institutions, employers and sectors. Webb (2009), for instance, observed that within Cambodia although almost all recipients of AusAID scholarships returned, few considered their awards as part of a national development strategy for Cambodia, but rather as a route to professional advancement and out of the Cambodian public sector, where wages were very low. Similarly, the evaluation of the Netherlands Fellowship Programme indicated that whilst brain drain to other countries was not a significant problem, retaining scholarship alumni within institutions can be very challenging due to higher wages and opportunities elsewhere, leading to a form of institutional brain drain and gravitation toward certain parts of the labour market (van der Aa, Willemson, and Warmerdam, 2012).

As evidence from the DAAD's tracer studies have indicated, national return rates can be relatively high (66% within home country, 96% within region) whilst simultaneously occupational return rates are relatively low: 63% of DAAD alumni had changed employer after returning from their scholarship, most commonly motivated by financial incentives and opportunities for personal advancement (DAAD, 2013). When scholarship aims are conceived as focusing on individual empowerment, institutional brain drain may not be highly relevant. In the context of developing specific employment sectors, institutions, or communities, however, it may warrant extended analysis.

An ancillary strategic question for evaluation is how long is considered a reasonable 'return' on the scholarship, before which movement out of the country or into another sector might be considered brain drain. Whilst the time period involved might be arbitrary it bears consideration: does it matter if, for instance, alumni leave their home country 10 years post-scholarship[14]?

A final issue in analysing return is the 'compound drain rate' in countries where multiple scholarship schemes operate and each have a small non-returning cohort which, as an aggregate, is a relatively large population leaving the home country labour market. Differing aims from developmental-focused and non-development schemes can complicate analyses in this space, as many countries run schemes that attempt to mitigate brain drain alongside schemes that encourage inward migration (the AusAID scholarships and Australian IPRS are possible examples). Whether this interaction confuses the analysis of 'drains' and 'gains' at an international level is not entirely clear, but there does not appear to be any extensive analysis of scholarship interaction effects in this area. To conduct such an analysis would be under the purview of donor harmonisation (see 'Harmonisation' on page 27) and could only be achieved through coordination between scholarship providers.

[14] Jürgen Enders, personal correspondence (5th March, 2014)

5. Data analysis

In this study, data analysis is considered to be the procedure of interpreting collected research information (e.g. completed surveys, interview transcripts) and forming findings and conclusions based on this interpretation.

Data analysis strategies within the sector are not often elucidated in detail within evaluation reports. Report methodology sections have tended to offer a description of data collection, but data analysis has remained a silent partner, with only a minority of reports explaining statistical techniques applied and fewer still explaining qualitative strategies applied.

The two routes available to analyse data are statistical (quantitative) and qualitative; these analysis strategies are examined in separate sections below.

Statistical (quantitative) data analysis

Almost all evaluation reports present some or all research findings in quantitative terms, most usually through percentile case summaries and, to a lesser degree, cross-tabulation of variables. Evaluations tend to draw upon both monitoring data (e.g. demographics of applicants, completion rates) and 'impact' data (e.g. perceptions of success) and the reporting conventions for these data sources does not appear to differ greatly within or between evaluations.

Descriptive statistics

Descriptive statistics were included in all evaluation reports that collected any numeric data. Monitoring data and tracer surveys have tended to be predominately quantitative and so evaluators have often been able to report extensive basic information about cohorts of alumni.

The prevalent format for reporting quantitative data is through percentile case summaries, for example '54% of scholars were male' or '7% failed to complete the programme of study'. In the case of Likert-style scales, the majority of evaluations report relative proportions in the data - the percentage of the sample that fall into each category (strongly agree, agree, and so forth) - rather than a measure of the 'midpoint' of responses and variability of sample data around that midpoint. The use of proportions can help data display (see below), but can also make it difficult to ascertain a data midpoint 'by eye' and thus the overall level of agreement with a Likert statement can be somewhat opaque.

The trend for reporting descriptive data tends toward graphing response percentages, often as bar charts, or as stacked bars for Likert-style questions, such as in in this example from a CSC report:

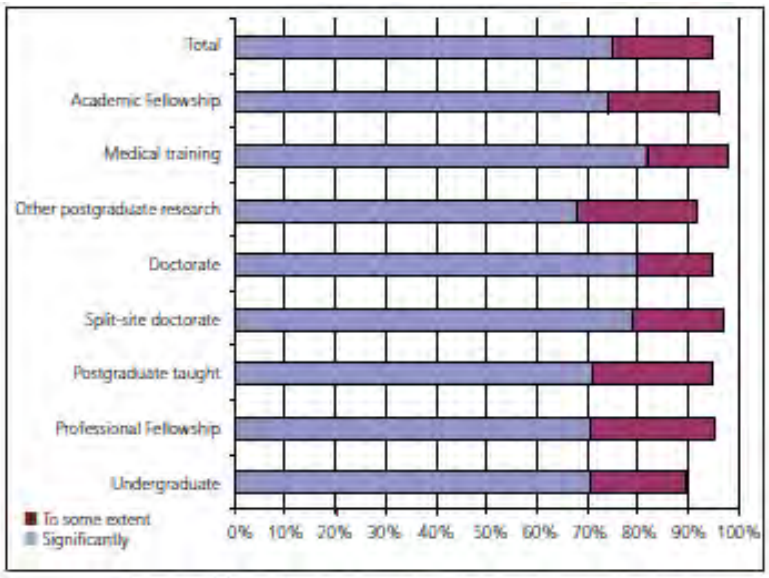

Figure 35: 'I use the specific skills and knowledge gained on award in my work', by level of study

(Day, Stackhouse, and Geddes, 2009: p51)

Whilst exact formats vary, this approach to quantitative data reporting is commonplace across the sector. Distribution graphs (histograms with normal curves) and statistics of the distribution shape – e.g. skewness and kurtosis – are not reported, although these graphs would only be of relevance to a minority of analyses (those that use inferential statistics; see below).

Percentile case summaries are frequently used to compare measures of specific variables between different groups within the survey population. A very common correlate in this kind of comparison is gender and, with a few exceptions (usually administrative or policy-level reviews), evaluations disaggregate findings by gender to examine differences between male and female scholarship recipients. Gender is notable as the only demographic variable to receive consistent attention across the sector in reporting descriptive statistics. Age and social status, for instance, are not regularly presented in analysis of either scholarship or post-scholarship experience[15].

Overall, the majority of evaluations that collect research data report quantitative descriptive analysis through percentile case summaries and a combination of bar, pie, and line graphs displaying this data. Exceptions exist - such as the VLIR-UOS country evaluation for Ethiopia (Penny and Teferra, 2010) in which statistical data informs a prosaic report and numerical data is restricted to the annexes - but are uncommon.

Inferential statistics

The relationship between scholarship evaluation and inferential statistical analysis is more ambiguous.

Inferential statistics aim to detect trends in data and make inferences from the characteristics of a sample to the characteristics of an entire population. Inferential statistics are thus tests of whether a difference or association is systematic in an entire scholarship population (e.g. all alumni) or merely due to random fluctuations in results ('chance') for a particular sample of that population (e.g. a single survey). In the case of evaluating scholarships, it may be important to know, for instance, if a gendered difference in levels of post-scholarship knowledge transfer is due to systematic differences between male and female recipients or merely random variation in our particular sample of male and female recipients. Descriptive statistics alone cannot offer this insight (although they can strongly hint at a particular conclusion); more sophisticated statistical tests of probability are required.

Use of inferential statistics has been somewhat inconsistent and has not always followed intuitive reporting conventions. A minority of evaluations have reported inferential analysis of data collected; most usually tests of association, such as the Chi-square test of contingency tables, and tests of correlation, such as Spearman's Rho (e.g. Webb, 2009). Other inferential strategies have been employed more rarely, including linear regression analysis (Amos et al., 2009: Ramboll, 2012), logistic regression analysis (Amos et al., 2009), and Student's t-test (ICUag.net, 2013). The overall trend emerging from evaluation practice in the sector, however, is that inferential statistics are used in a small minority of cases, although some of these minority cases present relatively sophisticated inferential analyses (e.g. Amos et al., 2009: Ramboll, 2012).

The approaches to reporting inferential statistics have also varied. Certain evaluations provide statistical test results according to reporting conventions accepted within the academic and evaluation research communities (e.g. ICUag.net, 2013; Webb, 2009), whilst in others only the test statistic or the p-value is reported (e.g. Heim et al, 2012), or a footnote is appended to note statistical significance without detailing the inferential tests used or further test results (e.g. Chesterfield and Dant, 2013). The reason for these discrepancies is not obvious, but in some cases may reflect the perceived audience of the report. Whilst evaluation experts, academics, and some of those involved in administrating scholarships may be interested to see reporting of statistical findings according to research conventions, it may not be meaningful to other stakeholders envisaged as the audience for a particular report[16]. It is also possible that, following the same logic, inferential statistics are used extensively but the results are irregularly (or non-transparently) reported, although there is no evidence to suggest this is the case. Nonetheless, if statistical reporting conventions are not followed within the body of a document it would be helpful if details were included in an annex (such as Webb, 2009) as many inferential tests are sensitive to factors (such as sample size and distribution shape) of which other researchers may wish to be appraised in order to have confidence in findings.

The extent to which limited use of inferential statistics in evaluation should concern evaluators and policymakers depends greatly on the nature of the evidence being presented. Describing the experiences of a particular group of scholarship alumni does not require statistical analysis more complicated than percentages and, perhaps, central tendency and dispersion. Thus if the aim of an evaluation study is to

[15] Social status has been addressed extensively in the Ford Foundation International Fellowship Programme (e.g. Volkman, Dassin, and Zurbuchen, 2009); particularly in the evaluation of scholarship recipients from India (Mansukhani and Handa, 2013).
[16] Jürgen Enders, personal correspondence (5th March, 2014)

examine a programme or alumni cohort as a case study, it is unlikely that inferential statistics are necessary. Alternatively, a 'rule of thumb' approach could be used in which repeated discovery of similar trends through descriptive analysis is assumed to demonstrate systematic trends. The rule of thumb technique potentially has an advantage over inferential testing when the conditions under which inferential statistics work most effectively are not met, although approximations must be managed carefully to avoid ignoring the problem (a low validity environment) in favour of producing desired results.

Another, more technical, group of issues relate to sample size and inferential analysis. One factor to consider in statistical data analysis is the sample size as a proportion of the total population; particularly in tracer studies for programmes that have ended. In some cases (e.g. Enders and Kottman, 2013) surveys have been conducted of entire alumni populations and response rates have been so high that the sample is the majority or entirety of the population. This renders statistical inference from the sample to the population largely irrelevant, unless the research is attempting to model predictive inferences for future scholarship programmes. In the more usual instances when evaluators are dealing with a relatively small sample, rather than the whole population, there are a variety of potential sampling errors to consider. In non-random, volunteer samples numerous sources of bias can limit the validity of inferences made from inferential statistics (Howell, 1999: Garcia, 2011), but such problems are not necessarily insurmountable. Statistical approaches can be used to help mitigate concerns and potential sources of error can be reduced by careful design of methods. Where the sample is of a specific sub-cohort of the total population - only traced alumni, for instance – it might be useful to include some articulation of how 'error' in results arising from this coverage bias (Blair, Czaja and Blair, 2014) has been mitigated.

There are, however, several ambiguities that arise from the limited reporting of sophisticated and robust inferential analysis and consequently areas that could benefit from further investment in these procedures.

With the exception of the few studies with extremely high response rates from the survey cohort (e.g. Enders and Kottman, 2013) most evaluations deal with samples of a population and conduct analyses deemed indicative of the whole population (i.e. of programme impact). In such circumstances it is appropriate to include at least some inferential analysis to determine the probabilistic strength of the relationship between the claimed characteristics of the population and the evaluation findings about characteristics of the sample. Several evaluation studies have already adopted this approach (e.g. Webb, 2009: Chesterfield and Dant, 2013) and it is evident from the data presented in many other evaluations that inferential techniques could be applied.

Another advantage to inferential analysis would be the opportunity to determine an effect size for observed correlations. Whilst it is relatively common for evaluations to at least comment on apparent connections between variables (e.g. gender and post-scholarship knowledge transfer), an effect size – measuring the magnitude or strength of the phenomenon – is rarely determined for these connections. Statistical effect sizes are useful for demonstrating the extent to which findings are important; large effect sizes represent influential relationships that might be important for programme policymakers to consider. At present, even the most statistically sophisticated analyses in the sector (e.g. Amos et al., 2009) do not routinely include effect sizes in research reports. Earlier it was noted that most descriptive statistics were percentile case summaries, not central tendency and dispersion. One approach to selecting effect sizes that inferential statistics are subsequently configured to detect is through assessment of past data on central tendency and dispersion (Howell, 1999), which may be difficult given the prevalence of standardised (percentage) descriptive statistics. This is an area of analysis that could benefit from a shift in reporting tendencies within the sector.

A more complex issue stemming from the discussion of effect sizes is the analysis of statistical power (see Cohen, 1992), which was not addressed in any of the evaluation studies reviewed. Power is essentially the discrimination of a statistical procedure to detect a relationship *if it exists*, as opposed to significance, which reports the probability that we have erroneously reported a relationship *when none exists*[17]. Statistical power, effect size, and ideal sample size are closely related; we can use mathematical formula to determine the sample size required to detect a particular effect size, given a desired level of power (error rate).

In practice, evaluation studies in the sector rely on non-random volunteer samples of alumni or current scholarship recipients, usually accessed through alumni tracing and without a sample size 'target' per se (an exception is Chesterfield and Dant, 2013). As such, specifying the sample size in advance might seem irrelevant since the answer to 'how many people should we survey from our population?' is usually 'as many as we can find'. However, this can be misleading as the statistical power of inferential tests will still be limited by sample size (amongst other variables) and as such small samples will be prone to miss significant relationships (see Cohen, 1992). Similarly, as a sample becomes larger inferential tests conducted on the sample increasingly tend toward significance (due to the mathematics involved) and so effect size becomes

[17] On significance and power, and type 1 and type 2 errors, see any good statistical handbook (e.g. Howell, 1999)

more important in identifying influential relationships. In either case, there is a strong argument to consider analysing and reporting effect size and statistical power in evaluations that use inferential statistics.

Finally, there is the potential of multivariate techniques to provide more complex statistical models of evaluation data. An important contribution of multivariate analyses is the ability to identify 'interaction effects' between influential variables and within subgroups in relation to a 'main' effect. To draw an example from Amos et al.'s (2009) analysis of Gates Millennium Scholars, the 'main' effect of the scholarship (i.e. recipients vs non-recipients) is broken down to show significant effects for subgroups, such as gender (i.e. male recipients vs female recipients vs male non-recipients vs female non-recipients). In the context of scholarships, particularly those funded within the ambit of international development, post-scholarship trajectories of alumni are likely to be complex and differences between outcomes likely influenced by a coalescence of factors (Ling, 2012: Byrne, 2013). Multivariate analyses offer the possibility of analysing how multiple variables interact to shape outcomes in differing ways. As with all statistical procedures, however, the data collected must be sufficiently robust and detailed to facilitate effective multivariate analysis.

Notwithstanding these comments, the World Bank Independent Evaluation Group, in its reflections on impact evaluation, sums up a sensible approach to designing quantitative analysis strategies: 'Better no numbers than silly numbers' (White and Barbu, 2006: 16). Given the difficulty in quantifying many of the outcomes of interest to scholarship providers, evaluators should proceed with caution when designing more complex statistical approaches to measurement.

Qualitative data analysis

The qualitative analysis procedures used to analyse data collected from interviews, focus groups, 'consultation', and site visits are reported to a very limited degree in evaluation reports. Despite the prevalence of interviewing in data collection, only a small minority of evaluations (Gilboy et al., 2004: Webb, 2009: Ramboll, 2012: Mansukhani and Handa, 2013) referred to qualitative data analysis; usually noting that interviews and/or free text survey comments were analysed through coding techniques. Of those completed or on-going studies that reported qualitative analysis strategies, forms of thematic analysis – clustering the data into prominent themes and elaborating on each cluster - appears to be the most common technique applied[18].

The majority of evaluation reports did not describe an approach to qualitative data analysis or refer to research methodology literature on the topic. Comments from interviews with alumni or other stakeholders were, in some instances, used to illustrate points made in evaluation reports without a prior description of a systematic data analysis strategy used to analyse qualitative data. Inevitably this raises concerns about the representativeness of comments included in the report (the 'cherry picking' criticism) and, without at least an overview of the data analysis strategy used, it is difficult to allay these concerns. It is also evident that some evaluation research has been better equipped to deal with quantitative data than qualitative data, despite collecting both. Nuffic (2009), for instance, have commented that in their NFP tracer study large amounts of free-text qualitative data collected as part of an alumni survey was subsequently used only for illustrative purposes because analysis of the free-text corpus was considered too difficult. The most detailed and well-articulated qualitative analysis strategies have been set out as part of evaluation research conducted by external consultants to scholarship administrators (e.g. Gilboy et al., 2004: Webb, 2009: Ramboll, 2012) and so might reflect additional time and expertise available to be invested into the evaluation.

Given the complexity and variation of possible approaches to qualitative data analysis (see, for instance, Saldana, 2012) it is of concern that the trend in reporting evaluation findings does not include sufficient detail to discern what strategies were employed. It would be helpful, at the very minimum, for evaluation work reporting 'interviews' or 'semi-structured interviews' to note whether these conversations were recorded and transcribed and, subsequently, whether (and how) they were coded and clustered to yield research results. The strategies used in qualitative data analysis are often not easily identifiable by their results (e.g. themes) and so leaving analysis methods unreported risks creating a gap in which readers of evaluation research remain unsure of how evaluators reached their conclusions and the credibility of the work is diminished. If it is undesirable to include such detail in the main body of an evaluation report then a methodology annex (e.g. Amos et al., 2009: Ramboll, 2012) or sub-report (e.g. Gilboy et al., 2004), offering detail on the process of transforming stakeholders' spoken words into a written evaluation report, would seem an effective alternative.

A prevalent use of qualitative data within the sector is for alumni profiles. Profiles tend to focus on a single alumnus and detail their experiences within a scholarship programme and their successes and (occasionally) difficulties post-scholarship. Such profiles tend to be used either as illustration within evaluation reports (e.g.

[18] For example, Negin (personal correspondence, April 16th 2014) noted that thematic analysis supported by the NVivo software was being used to analyse research interviews currently being conducted

DAAD, 2013: Clift, Dassin, and Zurbuchen, 2013), as annexes to reports (Mansukhani and Handa, 2013), as separate evaluation publications (e.g. CSC, 2014), or as part of the routine design of scholarship websites ('our alumni' sections and so forth). Detailed profiles of particular alumni provide both qualitative data for analysis and 'humanised' narratives for dissemination, but there is a delicate balance to establish between true case studies and journalistic-style alumni profiles that provide little content for rigorous evaluation.

One way to test the representativeness of case studies is to look for negative cases; instances where the alumni have failed to achieve either their objectives and/or the policy objectives of the programme have not been met, usually accompanied by an analysis of why this has happened. Negative case analysis in published reports across the sector is very limited. Almost all alumni profiles tend to show scholarship schemes in resolutely positive terms, as having transformed an individual's life and how, having overcome some barriers, that alumnus is now successful in their profession (e.g. Mansukhani and Handa, 2013). The depiction of any alumni as having not achieved their aims within a programme is rare and limited to specific issues - such as problems in monitoring and scholarship experience for a few IFP fellows in continental Europe (Gondwe and Schröder, 2013) – rather than as part of the programme's long-term impact analysis. This is not to say that case studies of a scholarship scheme's successes are unwarranted: they are important both in drawing attention to the (often extensive) value achieved by schemes and providing evaluation data on the ways in which schemes can have positive long-term impact on individuals and on policy objectives. It is not clear from individual alumni profiles, however, whether they represent typical or exceptional outcomes. Conducting broader thematic analysis of qualitative data in tandem with drawing out specific 'high impact' cases is a useful systematic qualitative strategy (e.g. Mansukhani and Handa, 2013), especially when it is unlikely, given the funding and political pressures on donors and evaluators, that negative case studies will be developed and publicised widely.

An earlier review of evaluation in higher education development interventions (Creed, Perraton, and Waage, 2012) found that it was *quantitative*, not *qualitative*, evidence within the sector that was lacking sophistication. Whilst there may be shortcomings to the use of quantitative techniques, it has been evident from this study that qualitative data analysis strategies are frequently lacking detail in reports. The use of qualitative evidence is widespread, but the analysis and treatment of this data – given the range of qualitative analytic strategies available – appears less well developed than for quantitative data.

Baseline data and comparative analysis issues

One persistent concern for data analysis has been the lack of baseline data to underpin comparisons. Although not all evaluations are concerned with comparative measures (either between groups or 'before and after'), those that do attempt comparative analysis frequently note the difficulty caused by lack of information about recipients prior to their scholarship experience against which to compare post-scholarship impact measures (e.g. Penny and Teferra, 2010; Chesterfield and Dant, 2013).

All scholarship programmes collect basic data on recipients as part of their application process and so typically demographic information, and potentially previous employment and academic histories, for recipients is available to evaluation studies. However, these data are often insufficiently detailed to provide insight into recipients' development and leadership activities or employment competencies prior to the scholarship. As such, monitoring data has not provided the basis for analyses of change in competence or involvement between pre- and post-scholarship activities. This leaves evaluations in the position to either (precariously) assume that post-scholarship competencies and development activities are directly attributable to the scholarship and not, for instance, a continuation of pre-scholarship trajectory, or to make a much more limited analytical case that there has been a contribution of some kind to those activities but without baseline data it is difficult to assess the *extent* of that contribution.

The concern of missing baseline data has not been widely addressed. Limited attempts to resolve the problem have involved retrospective analyses and the tailoring of survey questions to collect data on experiences or activities pre- and post-scholarship. Ramboll (2012), for instance, attempted to reconstruct baseline data by reviewing planning documents and asking survey respondents to comment on the situation before and after the programmes were implemented. Obtaining baseline data post-hoc is difficult and, whilst Ramboll have argued that their baseline reconstruction was successful, it is important to consider that self-report data on situations prior to, during, and after a project or scholarship are unlikely to be as reliable as comparative measurements taken contemporaneously (Garcia, 2011). In light of this difficulty, it is perhaps surprising that few evaluations have presented rigorous qualitative

> *'Little could be read that allowed outcome and impact of projects to be gauged against measurable aims defined and appreciated in the context of known conditions at the start...It is often 'forgotten' to document the beginnings and establish a baseline of data for future reference'*
>
> (Visser and Trinh, 2011: 16)

analysis of change in recipient's lives as a way to demonstrate the contribution of scholarships.

The most straightforward approach to collecting rich and reasonably reliable data in this domain is through research either at pre-scholarship or early tenure stages. Collecting pre-intervention baseline data is common in interventions within other fields, such as health (e.g. Antle et al., 2011), and facilitates comparison with impact measures used in later evaluations. One notable example of baseline data collection in scholarship evaluations has been in the use of propensity score matching in the Gates Millennium Scholars longitudinal evaluation to match recipients and a comparison group (non-scholar) along salient baseline characteristics (Amos et al., 2009). Of those evaluations reviewed as part of this study Amos et al.'s approach represents the most complete attempt to develop a baseline and post-test comparative framework for evaluating scholarship impact, using both initial data collection at pre-scholarship (or early in scholars' tenure) and a comparison group.

There are, however, other considerations that impact the effectiveness of baseline data collection. It is, for instance, a prerequisite of establishing baseline data that the aims of scholarships are clear and indicators of success (to be measured pre- and post-scholarship) sufficiently well-defined to allow data collection instruments to be developed. The conceptual underpinning and operationalised variables of scholarship programmes are not always established in sufficient detail to allow straightforward development of data collection instruments. In the case of the Netherlands Fellowship Programme (van der Aa, Willemson, and Warmerdam, 2012) for example, evaluation work included a phase of reconstructing a detailed 'intervention logic' that could subsequently be evaluated. Often aims and indicators of success change over time, meaning that scholarship programmes (particularly long-running programmes) may find that initial aims were detailed and operationalised in ways that do not necessarily cohere with current policy objectives. Similarly, evaluations often evolve over the period of scholarship programmes and so concurrent evaluation (e.g. Amos et al., 2009: Burciul and Sloan, 2013: Enders and Kottmann, 2013) collecting time series data may be better placed to cope with changes in policy and evaluation foci than entirely retrospective research.

> *'Those who are to collect baseline data must have an overall idea about how the project will be evaluated later on. Otherwise, the baseline data collected may be insufficient or no longer relevant at the time of evaluation.*
>
> (Garcia, 2011: 43)

Similarly, any form of pre-scholarship or early tenure baseline data collection requires a stable donor policy and administrative environment in order to be most effective. The variables assessed in pre-scholarship measures must remain consistently of interest to evaluators and policymakers from inception of a scholarship to an evaluation follow-up several years later. Some higher education scholarships, particularly PhDs with pre-programme training in language or other skills (e.g. Clift, Dassin, and Zurbuchen, 2013: DAAD, 2013), are planned to take 4-5 years from selection to completion. Additionally, time must elapse post-scholarship before impact can be measured and, depending on whether the evaluation focuses on contributions to an individual alumnus' career trajectory or diffusion of wider societal impact[19], this may yield a total period of 10-15 years between pre- and (final) post-scholarship data collection. If baseline measures collected in year 1 of a PhD scholarship are subsequently meaningless after 10 years have elapsed then the utility of those baseline measures is greatly reduced. Although the concern with long follow-up periods may be less acute with shorter scholarships (e.g. Master's programmes), the policy objectives of scholarships and foci of evaluations must maintain reasonable coherence across the research period in order for baseline measures to be effective analytic tools.

Finally, a related concern is the lack of reliable baseline and contemporary data on current labour markets, skills demand/shortage, and human resource planning, to provide a context within which the outcomes of scholarships can be interpreted (e.g. AusAID, 2011; van der Aa, Willemson, and Warmerdam, 2012). Benchmarking against national trends in employment has been used as a strategy for contextualising employment performance for scholarship recipients (Chesterfield and Dant, 2014), but this depends on reliable, contemporaneous data being available and so may be more applicable to certain countries or sectors. Developments in 'big data' for development and social research may offer some future solutions in this domain (Bollier, 2010: Boyd and Crawford, 2012: Letouzé, 2012), but are not implemented at present. More generally, evaluators cannot reasonably be expected to conduct primary research in order to construct employment or country-level contextual data in addition to examining the outcomes of scholarships. Where data is unavailable it may be advisable to use detailed qualitative methods to explore the context of skills shortage or human resource planning within specific companies into which alumni enter, rather than focus on sectors or national situations.

[19] Jürgen Enders, personal correspondence (March 5th, 2014)

6. Thematic issues

The counterfactual

A topic that garners significant discussion within the sector, and evaluation research more generally, is the counterfactual: essentially, what would happen in the absence of the scholarship scheme?

When broken down, there are three main variants of this question:

1. Do scholarship recipients perform better than non-recipients on outcomes pertinent to policy objectives?

2. Does the scholarship scheme as a whole perform better or worse than other, similar schemes?

3. Does the scholarship scheme produce better results than spending available resources into another intervention aimed at the same outcomes?

Of these questions, the most commonly addressed is the first, through comparison between scholarship recipient and non-recipient cohorts. This is not to say that it has been addressed widely: a point examined below.

Under-analysed counterfactuals

The latter two counterfactual questions – between-programmes and between-intervention types – are very rarely discussed in evaluation reports. It is relatively common for reports to include a short section or annex setting out scholarship objectives, eligible recipients, administration processes, and even financial arrangements for both the scholarship being evaluated and a selection of other schemes identified as comparable. However, these exercises tend to function as a way of situating the evaluated scholarship within the universe of scholarship schemes; they do not compare outcomes between schemes on the basis of research data. This is partly because meta-analysis of results across scholarship schemes is difficult: evaluation data is frequently not available for secondary analysis and even data collected within the ambit of the same scholarship programme is often incomparable (see Nugroho and Lietz, 2011). Although there has been some indication that the quality of evaluation research has been continually improving (see, for instance, Hageboeck, Frumkin, and Monschein, 2013), there is little evidence that the current state of evaluation research would allow scholarship providers to assess whether their scholarship scheme is performing better or worse than equivalents offered by others.

The between-intervention types counterfactual has been discussed to an even lesser extent in the sector, although it is perhaps questionable whether such an abstract policy topic ('which intervention shall we fund?') would necessarily be addressed in the evaluation of a specific scholarship scheme. In the context of programmes within which scholarships were part of a broader development approach (e.g. Penny and Tefera, 2010: Visser and Trinh, 2011) - or where researchers could consult stakeholders with experience of both scholarships and other intervention strategies aimed at similar policy objectives[20] - it is more straightforward to examine whether scholarships are contributing to policy objectives to the same extent as other strategies to which funding could be diverted. However, the lack of published, detailed value for money analysis for these programmes (see 'Value for money' on page 25) makes it difficult to assess whether any attempt to evaluate the between-intervention counterfactual would be fruitful. At the level of standalone scholarship schemes, those authors that have discussed the between-intervention counterfactual have been critical of whether such comparisons are practical or desirable. Dassin, Volkman, and Zurbuchen (2009), for example, argued that whilst the Ford Foundation's IFP was successful at achieving its policy objectives it is impossible to answer whether the programme was better value or more effective than contrasting funding options available to the donor because no comparison projects or agreed measures on which to judge diverse outcomes are available.

Conditional counterfactuals

Whilst between-programmes and between-intervention type counterfactuals have been addressed to a very limited degree within evaluation reports, the issue of whether recipients perform better on policy outcomes having participated in the scholarship scheme (the 'conditional' counterfactual) has been addressed in two ways.

[20] Joel Negin, personal correspondence (April 16th, 2014)

The first approach to the counterfactual has been through comparative designs, using a comparison (or 'control') group. Comparison groups are not widespread in the sector, but several evaluations have attempted to address the counterfactual in this way.

USAID's evaluation of LAC scholarships (Chesterfield and Dant, 2013) collected data from a comparison group of 214 non-recipient shortlisted applicants for scholarships, alongside 238 recipients, examining the same demographic, employment, and community participation variables for each group. Comparative statistical testing of outcomes for the recipient and non-recipient groups allowed the evaluators to make an assessment of the counterfactual scenario of not having participated in the scholarship scheme and remained within the LAC countries from where recipients were drawn. In order to form their comparison group, Chesterfield and Dant's (2013) research team invested significant resource into tracing non-selected scholarship applicants and contacting them to participate.

Evaluation of the Gates Millennium Scholars Program (GMS) by Amos et al. (2009) also employed a comparison group. Unlike USAID's evaluation of LAC scholarships, the comparison group for GMS was defined at the inception of the scheme and longitudinal data was collected on both cohorts of recipients and non-recipients across the duration of the programme. Two cohorts of scholars were examined, 483 recipients awarded GMS awards in 2002, and 664 recipients from 2003. Propensity score matching (Rosenbaum and Rubin, 1983) was used to select an appropriately matched, same-sized group of non-recipients, drawn from the longitudinal comparison cohorts, to be a comparison group for the evaluation. Logistic and linear regression were used to examine outcomes for the two groups and thus to draw conclusions about whether recipients had derived relevant benefits from participation in the scholarship programme.

The CSC also has a small counterfactual pilot study underway with a comparison group of non-recipient applicants[21]. Although evaluation reports are not yet available, the Monitoring, Evaluation, and Learning framework for the MasterCard Foundation Scholars Program also includes data collection from comparison groups, both as part of RCTs and quasi-experimental methodology.

The second approach to addressing the counterfactual has been through growth, or 'before-and-after', analyses. These forms of comparative analysis have examined the difference between conditions prior to the scholarship and those after the scholarship, with whatever happens between those time points being the growth contributed to by the programme (which would thus be absent in the counterfactual scenario).

ECOTEC's (2009) evaluation of Erasmus Mundus, for instance, adopted a before-and-after comparative design following the authors' concern that finding a suitable comparison group for the participants in Erasmus Mundus scholarships would be infeasible. Ramboll's (2012) post-only non-equivalent comparison design also focused on a before-and-after comparison, reconstructing baseline data using retrospective self-report measures to give an assessment of the prior conditions which could be compared to subsequent conditions. Research using comparison groups often also uses a before-and-after (time series) analysis structure (e.g. Chesterfield and Dant, 2013).

More generally, evaluators frequently use the before-and-after comparison as part of their qualitative data collection and reporting of alumni profiles. Mansukhani and Handa (2013), for instance, presented a collection of alumni profiles that explored the change in professional and personal trajectories for individual beneficiaries of the Ford Foundation's IFP. Counterfactual questions can also be included in surveys in an attempt to retrospectively assess conditional counterfactuals: both the CSC's alumni surveys and Negin's (2014) analysis of AusAID scholarships in Africa have used this strategy. Analysis of individual trajectories, rather than quasi-experimental assessment of change in specific variables, may also have greater affinity with contribution-focused (rather than attribution-focused) methodological strategies, such as those evaluations employing elements of Contribution analysis (Rotem, Zinovieff, and Goubarev, 2010: Ramboll, 2012). As has been noted above, however, reconstructing baseline data retrospectively can be troublesome and the reliability of recipients' perceptions of what would have happened had they not received a scholarship may be low.

Concerns with the counterfactual

Although the use of comparative designs and assessment of the counterfactual has been widely promoted (White and Barbu, 2005: Garcia, 2011: Vardakoulias, 2012), in this sector at least there have been numerous accounts of why designing effective methodologies for the counterfactual has been challenging.

One concern raised repeatedly is the difficulty in forming an appropriate group for comparison. This problem has been raised both in the context of finding comparator organisations not involved in funded projects (e.g.

[21] Rachel Day, personal correspondence (April 3rd, 2014)

Ramboll, 2012) and comparator individuals for scholarship recipients (ECOTEC, 2009: Säring, Spartakova, and Wegera, 2012). Without adequate matching between the comparison group and scholarship cohort the counterfactual analysis is likely to be subject to selection bias and lack validity (Garcia, 2011). The challenge, therefore, is to identify cohorts that are sufficiently alike to be valuable comparators but are neither participating nor affected by the participation of others.

Random assignment of participants to the intervention and comparison conditions, the standard usually employed in clinical and psychological research, is not usually possible with scholarship schemes, where non-random selection of candidates (e.g. based on academic ability) is highly desirable. As such, a matched cohort needs to be identified *after* non-random selection has taken place. Of those evaluations which have used (or are using) comparison group designs, the preferred cohort for comparison has been non-selected finalists: i.e. those who applied and made it to the final stages of the scholarship selection process, but were eventually unsuccessful.

The ease of access to a non-selected finalist group has varied between evaluations. Amos et al. (2009) have benefited from the longitudinal data collection conducted alongside the Gates Millennium Scholars Program and thus contemporary comparison group data on non-selected finalists was readily available. Amos et al. further mitigated against bias by using propensity score matching to ensure a well-fitting match between recipient and non-recipient groups evaluated. Although the international context of the scheme makes data collection procedures somewhat more diverse, the MasterCard Foundation Scholars Program has planned a similar process of collecting data on non-selected finalists to subsequently be used as a comparison group in evaluations[22]. For Chesterfield and Dant (2013), access to the non-selected finalist cohort was somewhat more complex because, although records were available on *who* was a non-selected finalist, data had not been collected previously and thus non-selected finalists had to be traced and petitioned to participate in the evaluation. This likely reflects more closely the situation across other scholarship schemes, where accurate, contemporary data on non-selected finalists is unlikely to be available, particularly given the difficulties experienced tracing even scholarship recipients in some evaluations (e.g. Bryant and Wrighton, 2008).

Beyond issues with accessing a suitable comparison group, there have been other concerns with counterfactual research. It has been noted, for instance, that whilst counterfactual research may be desirable, it can be a significant resource drain that ultimately may not prove sufficiently informative to warrant the investment of time and money[23]. Some weight is added to this concern by a brief examination of the time-intensive process required by the research team evaluating USAID's LAC scholarships to trace and interview non-selected finalists (see Chesterfield and Dant, 2013). Additionally, if there is no probability of funding for the scheme being renewed, or the scheme being scaled up, then it may make more sense to simply track the trajectories of alumni than to attempt to prove they are better off than non-recipients[24]. This critique of counterfactual research may not hold for all instances. Comparison between intervention and counterfactual scenarios remains a viable approach to analysing impact even when a programme has concluded. What may differ, however, is the necessity to supply counterfactual data to a donor or oversight body that places significant emphasis on such evidence.

At a more conceptual level of research design, it is important to note that most comparison approaches hold that time-varying factors unrelated to the scholarship are a constant across recipients and non-recipients. That is, whilst comparative approaches take account of factors identified within the scholarship scheme as differing between recipients and non-recipients – immersion in another country, making international links, the academic award studied, the funding itself – they do not account for other time-varying factors that may affect one group only. For instance, domestic conditions in a scholarship recipient's home country may change quite considerably over the period of a PhD undertaken in another country (e.g. through political instability or changing government investment priorities). Similarly, recipients may find themselves immersed in political or social phenomena unrelated to the policy objectives of a scholarship scheme, and which have no influence in their home countries, simply by virtue of studying abroad. As such, comparison groups tend to diverge from the point of selection, and by the time of comparison may differ substantially on variables that are not considered outcomes of the scholarship programme (and may not necessarily all be positive outcomes). In some cases this is circumvented by recipient and non-recipient groups remaining in the same general social setting (e.g. Amos et al., 2009), but for other comparative designs (e.g. Chesterfield and Dant, 2013) the complexities of individual and social trajectories may need to be carefully assessed as part of the counterfactual.

[22] Barry Burciul, personal correspondence (February 19th, 2014)
[23] Joan Dassin, personal correspondence (February 21st, 2014). Jürgen Enders, personal correspondence (March 5th, 2014)
[24] Mirka Tvaruzkova, personal correspondence (March 26th, 2013)

Value for money

Value for money (VFM) has been a significant topic of interest for development interventions (Fleming, 2011), with extensive debate about the conceptual, practical, and political consequences of increasing focus on economic efficiency (e.g. King and Palmer, 2012). This study does not engage directly with the themes developed in VFM debates, but rather examines the state of VFM discussion within evaluation reports.

Almost all scheme-wide evaluation reports, and some specific country reports, include an analysis of the financial conduct of the programme. These analyses, however, are almost never VFM evaluations, if VFM is defined as an assessment of the relative costs and benefits incurred during a programme and, consequently, a judgement as to the value accrued. The emphasis of financial analyses within evaluation has primarily been on administration and efficiency in the deployment of resources (e.g. Gilboy et al., 2004: CIDA, 2005: van der Aa, Willemson, and Warmerdam, 2012: Carpenter and de Vivanco, 2013: Chesterfield and Dant, 2013). In some cases evaluators have derived unit costs for the 'production' of, for instance, a PhD student or Master's student through the scholarship programme (e.g. Norad, 2009), but have also tended to caution readers that such unit costs are often difficult to establish and can be misleading given the differing long-term impacts of programmes.

Although it is not VFM analysis per se, the scrutiny of financial administration can be interesting beyond the audiences who are responsible for the financial auditing of scholarship programmes. Amos et al. (2009), for instance, discussed the phenomenon of the 'GMS tax', referring to the way in which scholarship funding issued to recipients actually diminished access to other funding distributed on a need-basis, even causing some financial hardship. Similarly, access to funds or equipment which may be either transferable or pooled between development interventions - rather than locked into a single programme - has been noted as a concern for institutional capacity building projects (Penny and Tefera, 2010). The GMS tax and desire for fluid resources across multiple interventions or contexts demonstrate how financial administration issues can be of relevance to project outcomes beyond typical concerns around inefficiency, waste, and lack of resource.

To return to VFM analysis in terms of input cost versus output value, Fleming (2011: 5) lists six main approaches to evaluating value for money:

1. Cost Effectiveness Analysis
2. Cost Utility Analysis
3. Cost Benefit Analysis
4. Social Return on Investment
5. Rank Correlation of Cost versus Impact
6. Basic Efficiency Resource Analysis

None of these approaches have been reported widely in evaluation reports. The reasons for this omission appear to vary between evaluations. Some evaluation programmes, such as the IFP 10-year study (Tvaruzkova and Clift, 2013), have considered and rejected VFM analysis as not aiding the evaluation of the programme; particularly given there is no prospect of the scholarship scheme being re-funded or scaled up[25]. Others, such as the MasterCard Foundation Scholars program, have placed VFM analysis on hold until the programme has run for a sufficient time period to reasonably assess outcomes[26]. Still other evaluators have either concluded that VFM is too difficult to assess given data and resources available (e.g. Penny and Tefera, 2012) or are sceptical about the ways in which VFM analysis will (or can) inform policy decisions[27].

Certainly one concern with VFM analysis has been the difficulty in quantifying, and particularly monetising, scholarship outcomes[28]. A useful example can be drawn from Lange's (2005) argument that facilitating students to study in partner programmes in the South allows far more students to benefit for the same financial investment than bringing students to donor countries for long periods of study. Whether this is or is not accurate of all (or any) programmes is beyond the scope of this study, but the argument highlights how some of the intangible aspects of value that have historically underpinned scholarship programmes – such as being immersed in the research culture of developed nations, or the 'soft power' of international relations

[25] Mirka Tvaruzkova, personal correspondence (March 26th, 2014)
[26] Barry Burciul, personal correspondence (February 19th, 2014)
[27] Joan Dassin, personal correspondence (February 21st, 2014)
[28] Emily Hayter, Heath Thomson, and Beryl-Joan Bonsu, personal correspondence (February 20th, 2014)

– can be difficult to establish in financial terms and thus their contribution to input-output based VFM analysis can be troublesome to calculate.

This is not to say that there has been no VFM analysis in the sector. It is entirely possible that VFM analysis may take place and remain unpublished. Whereas accounting for the use of government funds is usually a public activity in donor countries with a commitment to financial transparency, this process does not necessarily include sophisticated VFM analyses that examines inputs in the context of impacts. However, this may change with increased emphasis on VFM at donor agencies (e.g. DFID: Girdwood, 2012).

There has been at least one attempt at a detailed quantification of value and production of a VFM analysis: Ramboll's (2012) evaluation of NPT and NICHE. Ramboll's approach assessed three facets of VFM: cost efficiency in deployment of project resources, cost effectiveness in terms of overall capacity developed given the input costs, and the quality of specific outputs given resource inputs. The methods deployed to achieve these analyses were intricate and are well detailed in an annex of Ramboll's (2012) report, but in sum were a 'SMART matrix' for creating a cost-quality ratio, Responsible, Accountable, Consulted and Informed (RACI) analysis, and intra-programme benchmarking for input-output cost-efficiency (Ramboll, 2012). As noted in chapter 4, 'Variables and 'indicators', Ramboll's VFM analysis allowed them to offer an input cost to yield an increase in their compound Capacity Development Index (CDI) variable, statistically analysable between countries in which the programmes examined operated. Whilst the measures involved – and particularly the use of retrospective self-report data in compiling CDI – might be open to criticism, the evaluation of NPT and NICHE is certainly the most comprehensive published attempt at assessing VFM of the evaluations studied.

The lack of a comprehensive portfolio of VFM analyses akin to Ramboll's (2012) work on NPT and NICHE limits understanding of VFM within the sector. Whilst there are difficulties in producing rigorous cost-benefit analyses, there is also only a limited amount of useful information that can be extracted from data on financial administration. As Palenberg (2011) and Friedriksen (2012) have observed, measures of standalone efficiency are not very useful beyond the confines of a programme's administration and comparative efficiency is likely the subject of greater interest to evaluators and donors. Even Ramboll's (2012) VFM analysis cannot be fully appreciated without a corpus of other such analyses to provide comparators.

Notwithstanding the potential utility of VFM measures, however, a significant concern (more political than methodological) remains as to what benchmark VFM measures would be compared. Scepticism about the value of VFM analyses is not uncommon in development circles (e.g. Ellerman, 2012) and evaluators and administrators may hold legitimate concerns about both what VFM analysis would involve and how the results would be used (Barber, 2012). The methodological difficulties in credible VFM analysis certainly underpin these concerns to an extent, and so greater focus on VFM, if desirable, likely requires simultaneously political and methodological action to implement approaches in ways that are meaningful to donors, providers, and other stakeholders.

Harmonisation

Strategic coordination between the development activities of donor countries has been an issue of interest in all international development activity, including international scholarships for higher education.

Moves toward 'donor harmonisation' (OECD, 2005) have, for instance, yielded a strategic working group of (mostly European) scholarship providers aimed at bringing scholarship provision and evaluation into closer accord and avoid ing overlaps. The meetings of the Donor Harmonisation Group involve many of the providers who have commissioned the evaluation reports analysed in this study. Discussion of a different angle on harmonisation has also taken place within the ambit of 'soft power' and public diplomacy, with a committee of the UK House of Lords, for instance, recently asserting the need for UK scholarship programmes to 'offer a coherent package of engagement with the UK and its Embassies during the period of the scholarship and afterwards' (2014; 209). In a general sense, effective coordination of efforts either within-country or between-donors is a matter of interest in all scholarship policy.

Within evaluation reports, harmonisation has been raised, but not extensively discussed. Barber and Hel (2012), in their analysis of the impact of AusAID scholarships in Cambodia through the lens of the OECD DAC's criteria, commented on the state of donor harmonisation within the country. Their conclusions regarding harmonisation were not optimistic:

> *'There appears to be very little sharing of experiences or donor meetings on postgraduate scholarship programs; and little prospect of harmonisation occurring'* (Barber and Hel, 2012: 19)

This resonates with the analysis of Austrian Development Agency officials' experiences with donor coordination:

> *'ADA [Austrian Development Agency] Coordinators have no official role in harmonisation activities… they do not attend education donor coordination meetings and have no contact with other donors in HE. Occasionally (according to one coordinator), they are able to provide information about other relevant development partner funding or programmes if they happen to know about them.'* (Carpenter and de Vivanco, 2013: 46)

Whilst harmonisation is not widely examined in the evaluation literature, the discussion that does occur has tended to portray donor coordination as nascent and unsystematic. Penny and Tefera capture the broader theme on the evaluation of harmonisation when they remark that it '...occurs more by chance than design' (2010: 6).

Scholarship coordination within the Pacific region, and particularly the Pacific Island states, appears to have progressed more extensively than has been reported elsewhere. Gosling (2008) has noted that in Vanuatu and the Cook Islands, amongst other locations, there has been significant cooperation between the New Zealand Aid Programme and AusAID to deliver scholarships more effectively. More recently a joint monitoring and evaluation plan has been released for AusAID and the New Zealand Aid Programme to coordinate evaluation of their scholarships in Papua New Guinea (DFAT, 2011). Similarly, in Fiji there have been moves toward coordination between the New Zealand Aid Programme and AusAID, with recent evaluation research calling for comparative analysis of monitoring and evaluation data between scholarship providers (AusAID, 2011). In the case of Fiji particularly there is significant cause for greater cross-scholarship analysis of impacts and harmonisation of provision: AusAID have indicated that over 3000 scholarship places are open to Fijians each year, provided by some 57 scholarship schemes and funded through at least 8 donors. There may be similar cases in other locations, particularly Africa: where numerous donors have focused attention (e.g. Gilboy et al., 2004: Negin, 2014).

The concern with harmonisation within specific countries reflects one of the tensions for scholarship schemes (and their evaluation) generally. Whilst focusing at scheme-level may make sense from a donor perspective, particularly for those schemes that act across many countries, the individual requirements of countries differ and the best strategic fit with development (or even public diplomacy) objectives may require scholarships to act in different ways in different spaces. Penny and Tefera (2010), for instance, have observed that Southern partners in VLIR-UOS projects have tended to see harmonisation more as national or local coordination to ensure the availability of resources, from whatever donor source, in their particular areas of work.

It is perhaps plausible that scholarships are better placed than institutional cooperation initiatives with regard to availability, if not necessarily avoiding interference and promoting synergy. Since most scholarship competitions are open to a broad range of eligible candidates, and given the overlapping scholarship opportunities in countries such as Fiji, it seems likely that resources (in the form of scholarships) *will* be available in the areas relevant to Southern partners. There are, however, additional evaluation issues raised that would require multi-donor research to address, such as the possibility of oversupply, synergistic impacts of different scholarships, and the impacts of multiple concurrent scholarship programmes on the national labour force during recipients' study abroad period.

Evaluators considering harmonisation have promoted much closer discussion between donors simultaneously acting at country-level to both avoid interference and promote synergy (Visser and Trinh, 2011). High internal coherence between developmental activities is both desirable and often considered achieved (e.g. Penny and Tefera, 2010), but multi-donor coordination is clearly regarded as far less developed. Evaluation reports suggest this is an area that both requires development and is currently undergoing positive change, particularly in the Pacific. Despite this, the lack of emphasis on country- or sector-level synergy effects from schemes - including both negative interference and positive compound impact - is a major concern with current approaches to impact evaluation. In its most simple terms, and using the Fiji example, it is not clear how useful it will be to evaluate the impact of only one or two of the 57 scholarship schemes currently exerting developmental influence on the islands.

7. Conclusions

The purpose of this scoping study has been twofold:

1. To identify trends in research practices and strategies used in the evaluation of international scholarships for higher education

2. To identify omissions, uncertainties, and ambiguities in current methodological approaches

The methodology, methods, foci, and data analysis strategies reported in evaluation documents pertaining to numerous scholarship schemes worldwide have been scrutinised. This has been supplemented by a series of personal dialogues with evaluators and scheme administrators to garner additional detail, catalyse analysis, and achieve clarity.

It has not been the aim of the study to comment on either the utility of evaluation to policymaking (beyond commenting on the sophistication of the evaluation itself) or the most appropriate ways to evaluate schemes. The review has been concerned with the 'state of the actual' and the analysis of current evaluation practice.

The conclusions following are centred on themes emerging from the analysis of evaluation methodology and the critical reflections provoked by practices, ambiguities, and omissions.

1. The majority of evaluation is ex-post: it traces alumni sometime after their scholarships.

Using tracer studies, the sector invests heavily in evaluation following alumni after the conclusion of their scholarship programme. It is unsurprising and not inherently negative that ex-post evaluation has dominated the sector to date. Many scholarship schemes have been on-going for decades and rigorous evaluation may be relatively new to the administrative agencies or have been commissioned externally either part-way through or at the conclusion of a scholarship programme. In this sense, the sector has been catching up. Increasingly evaluation planning is at the heart of delivery and the decision to conduct ex-post tracer studies is out of design, rather than necessity. Yet there is some distance left to travel in this regard, particularly as external consultancy - widely used in the sector - engaged late in a scholarship programme can only ever be both ex-post and limited by the constraints of the monitoring data collected at baseline and during scholarship tenure.

As a corollary of ex-post evaluation dominating the field, designed comparisons such as longitudinal studies have been little used. Nonetheless, where comparative designs have been used – and particularly where counterfactual analysis has been included – some of the most detailed evaluation results have been produced. The lack of longitudinal data on participants is thus clearly a concern for the sector. Ex-post analysis is rarely conducted as an ex-post panel study (i.e. the same participants being surveyed over multiple time intervals) and thus neither pre- to post-scholarship data, nor post-scholarship panel data is readily available. Problems caused by missing baseline data would also seem to be primarily as a consequence of the retrospective approach to evaluation.

It would be misleading, however, to suggest that methodological difficulties with counterfactual analysis and baseline data would be eliminated solely by shifting to longitudinal designs and planned comparisons (although it would likely help). Counterfactuals may be easier to design and conduct with a longitudinal comparison group, for instance, but the resource cost of conducting such an analysis may be prohibitive: particularly on evaluation issues where rich qualitative data is required to facilitate a detailed analysis (re-integration, for instance). Additionally, data management systems would need to be sufficiently sophisticated to support the kinds of analysis demanded: in some cases baseline data on thousands of recipients and non-recipients would need to be stored for 10 or more years to inform evaluation.

2. Methodology – separate from methods – is not discussed very much within the sector.

Whilst methods of data collection are reported to some extent by all evaluation reports, methodology – the conceptual framework within which these methods sit – is not as broadly discussed.

The Kirkpatrick evaluation model has seen widest use within the sector, both historically and within on-going evaluation projects. Other methodological frameworks, such as Contribution Analysis, RCTs, longitudinal comparison designs, and post only non-equivalent comparison design, have been used in specific instances, but this does not appear to have been in more than one or two evaluation studies for each approach. As such, Kirkpatrick's framework is the dominant methodological logic for designing evaluation, although this inaccurately suggests a broader influence than is actually evident in the sector.

Lack of methodological discussion may indicate either a lack of engagement with methodology or a strategic decision to omit methodological detail from evaluation reports. Whilst the latter issue is somewhat concerning, the former is more so, particularly as conflation of methods and methodology in reports may reflect confusion in the evaluation process itself. Theories of change or detailed policy aims and indicators, for instance, have repeatedly been reconstructed ex-post by evaluations, which suggests an initial lack of cohesion between the elements of policy, administration, success indicators, methodology, and methods.

Externally commissioned, consultant-led evaluations have often been effective at developing detailed methodologically strategies, but, due to their frequent late involvement in scholarship evaluation, are limited to research designs that can accommodate the data available.

3. Surveys are the dominant tools for data collection. Interviewing in-person is also common.

Given the global audience of scholarship recipients, it is reasonable that survey methods have been widely implemented within the field. Surveys have the potential to reach participants quickly and relatively easily across a geographically diverse recipient group; they have commanded significant response rates for many evaluators in the sector. Delivery of surveys can also be tailored to use media accessible to the participants, such as online platforms, paper surveys, or surveys on mobile telephones (although there is no evidence of the latter having yet been used in scholarship evaluation).

Simultaneously there is a commitment to qualitative field work, with interviews being only marginally less common than surveys and most frequently involving face-to-face dialogue with scholarship stakeholders. Technology offers alternatives (telephones and internet telephony, for instance) and these are used within the sector, but visiting countries and speaking in person to alumni, employers, or government stakeholders appears to remain the preferred option.

The range of methods used by evaluators has been relatively small, consisting almost entirely of surveys, interviews, focus groups, and (to a very limited degree) participant observation. Although it should be noted that this list covers the dominant methods in many fields, it is interesting to observe that more participatory methods – such as journaling or techniques from visual anthropology - have not found popularity in the sector. Similarly, no evaluations deploying new, 'innovative' techniques in monitoring and evaluation[29] have yet been reported.

The narrow range of methods deployed may reflect a perceived lack of credibility for other methods (particularly participatory qualitative methods) within policy audiences. It is also indicative, however, of the way methods primarily appear designed to investigate specific research questions set by donors and providers. This may seem an obvious statement, but it is certainly not the only way to evaluate. Approaches such as Outcome Harvesting (Wilson-Grau and Britt, 2012), for instance, collect data on outcomes and work backwards, rather than overtly evaluating progress toward predefined impacts, helping to counter the notorious difficulty that survey designers encounter in exploring unintended outcomes. In doing so they take a recipient-centred approach that has been more common within wholly qualitative research in the scholarships sector than in, for instance, tracer studies.

It is also apparent that many of the innovative techniques currently capturing the imagination of evaluators and their sponsors focus on the immediacy of data gathering and the potential for technology to provide broader reach and depth at greater speeds: 'big data' focused methods (e.g. Letouzé, 2011) illustrate this effectively. Scholarships, conversely, are often considered in timespans of years and decades, with the time from funding to 'impact' unpredictable, but anticipated to be lengthy. In that context, it would seem that investment in methods for longitudinal data collection is likely to yield greater utility than more immediate data collection. A topic for consideration within the ambit of harmonisation, perhaps, is whether innovative research methods have anything to offer in examining the aggregate effect of scholarships on communities, countries, or regions.

4. Almost all evaluation is concerned with similar issues, such as completion rates, gains in knowledge and skills, return to home country, and employment trajectory post-scholarship.

Because the majority of scholarship schemes flow from similar policy objectives, the majority of evaluations address similar topics. At the level of specific variables, two prominent foci of evaluation have been the gains in knowledge and skills from study and employment trajectories post-scholarship. These foci follow from the objectives of most scholarship programmes being closely tied to human resource capacity and / or socioeconomic development.

[29] The UNDP Knowledge, Innovation and Capacity Group (2013) present a useful overview of many such innovations and, in particularly, the technologies that might aid monitoring and evaluation.

Analysis of recipients' tendency to return to their home countries – or at least leave the scholarship host country – has also been widely addressed. At a policy level, return is generally seen as essential to achieving scholarship aims (although there are some complexities in this). Analyses of return rates have tended to be relatively simple - percentages of recipients self-reporting having either left the host country or returned to the home country – but there is clearly concern within the sector over how return should be defined and how the contributions to policy objectives of those who fall outside the definition should be assessed.

Several evaluation reports have noted the difficulty in making judgements of impact based on historical data which lacked baseline information, focused only at an activity monitoring level, or both. Whilst the detail of impact analyses available has differed substantially across the sector, it is evident that any robust analysis will require data management and alumni tracking systems that are tailored to facilitate both accurate long-term record keeping and access to impact-level data longitudinally.

At a macro level, the OECD DAC's criteria for evaluating developmental assistance are used by a variety of evaluations, but there remains no single overarching framework within the sector. Widespread application of the DAC criteria could potentially facilitate a level of meta-analysis both within and between schemes that is not currently available to evaluators, assuming that any differences in foci and scale could be accommodated. Meta-analysis of scholarship outcomes, however tentative, would likely be a useful addition to understanding in the sector: particularly given the overlap of scholarship objectives and geographical areas of implementation.

5. The practices and standards of data analysis are not always clear in evaluation, leaving ambiguities about how data is treated within the sector

Although both qualitative and quantitative data has demonstrably been widely analysed, lack of clarity in data analysis procedures makes it difficult to assess the rigour of the research.

Most notably, the standard of qualitative analysis is unclear in all but a select few evaluation reports. The process of analysing interview data - through textual content analysis, coding, thematic analysis, or otherwise - is rarely detailed. More generally it seems that qualitative data is often treated at a descriptive level, with little analytic interpretation going beyond 'what' is being said to look at 'how' and 'why' it is being said. Because access to multiple types of data to form a complete picture on scholarship outcomes is often very difficult (the 'triangulation' problem) there is a danger of taking stakeholders' views at face-value and failing to interrogate the way they are influenced and constructed. That tendency should certainly be resisted, partly through documenting qualitative strategies more effectively (to demonstrate rigour) and partly through seeking ways to collect data that addresses the same topics from multiple perspectives. The qualitative data presented are often compelling, but it is never entirely clear without rigorous procedure documented whether such data are offering the whole story.

The situation with quantitative analysis is stronger and a body of well-documented research has been conducted on scholarship outcomes. This is, of course, not universal and some quantitative analyses are notably more robust than others. Given that almost all scholarship evaluations deal with samples, not populations, it would be advisable for the emphasis in quantitative analysis to be on inferential testing of whether patterns of results in samples were reflective of patterns of results in the scholarship population. When participant samples are very small it would be inadvisable to attempt inferential analysis. However, many scholarship evaluations collect data from a sample of many hundreds (sometimes thousands) of alumni, and so extending data analysis to go beyond describing patterns in the sample would appear to be a logical route to maximising the utility of evaluation data already being collected.

6. Discussion of harmonisation, and perhaps harmonisation itself, is very limited and this in turn limits useful analyses of synergy and interference

Donor harmonisation features only to a minor extent within evaluation reports. Harmonisation discussions have been limited to only a few evaluation studies and even in these cases the commentary could best be described as a reflection on harmonisation issues, rather than a detailed analysis of where harmonisation does and does not occur.

It is troubling that the tone of reflections on harmonisation are primarily negative, although it has not been within the purview of this study to explore harmonisation extensively. It is clear, however, that scholarship harmonisation, and particularly evaluation harmonisation, is nascent at best. There is some indication of coordination amongst scholarship programmes - particularly in the Pacific region – in designing complementary evaluation systems, but collaboration between providers to analyse scholarship outcomes is not evident more broadly.

Coordination between scholarship providers could provide major avenues for advancement in evaluation scope and detail. Exploring synergy and interference effects between schemes that act in the same geographical regions and / or with the same target audiences could offer an insightful analysis of developmental impact (however defined). It would also help to offset the current problem of multiple concurrent, unrelated evaluations examining the impact of scholarships in countries or regions without reference to the ways in which other scholarships (and development programmes more generally) feed into the milieu.

For 'soft power' scholarships, of course, this form of coordination could be self-defeating, unless the aims were for aggregate reputational gains (e.g. the image of European Higher education), in which case several involved countries might choose act in concert to evaluate outcomes (although they are unlikely to garner cooperation from other, 'rival' scholarships acting in the same space).

7. The corpus of evaluation data predominately relates to established OECD governmental scholarships

Finally, there is a lack of published evaluation data on scholarship schemes funded by emerging donors and non-OECD donors, such as China and India.

As a recent British Council and DAAD (2014) study has demonstrated, scholarship schemes are not only being funded through the overseas development and foreign affairs budgets of high income nations, but are increasingly a focus for a variety of middle-income countries. The absence of evaluation data on non-OECD scholarship schemes is notable, with only a small number of sources identified (e.g. Dong and Chapman, 2008) as part of this scoping study. There are several possible explanations for this absence. Amongst the most plausible are that concern with impact evaluation may not be as acute for the non-OECD donors, that evaluation is published internally only, or that evaluation is published externally but in a language other than English.

Whether the reason is a lack of published work or a limitation of the current study, it is a significant omission in the broader analysis of scholarships as either a developmental intervention or tool for public diplomacy that data is unavailable on several large schemes. The Chinese Government Scholarship Scheme, for instance, dwarfs many of its European and North American counterparts in scale (see Dong and Chapman, 2008) and so evaluation data on this scheme alone may add significantly to the collective understanding of scholarship impacts globally.

Appendix 1: Documents analysed

Amos, L.B., Windham, A, de los Reyes, I.B., Jones, W., Baran, V. (2009). Delivering on the promise: An impact evaluation of the Gates Millennium Scholars Program, Final report. Washington, DC: American Institutes For Research

Andersen, C., & Tobiasen, A. (2007). Effect measurement – Norad's programme for master's studies (NOMA): SNF-Working paper No. 13/07. Bergen: SNF

Atkinson, C. (2010). Does soft power matter? A comparative analysis of student exchange programs 1980-2006. *Foreign Policy Analysis*, 6(1), 1-22

AusAID (2011). AusAID Fiji ADA/ARDSImpact study 2011. Canberra: AusAID

Barber, D., & Hel, S. (2012). Cambodia: Review of the Australia awards program. Canberra: DFAT

Bryant, C., & Wrighton, N. (2008). Fiji and Tuvalu tracer study, 2008. Canberra: AusAID

Burciul, B., & Sloan, M. (2013). Measuring multi-dimensional change: The challenges and opportunities of the MasterCard Foundation Scholars Program. Paper presented at the 27th annual conference of the American Evaluation Association, Oct 16 - 19, Washington, DC

CIDA (2005). Evaluation of the Canadian Francophonie Scholarship Program (CFSP), 1987-2005: A need for reorientation. Gatineau, Quebec: CIDA

Carpenter, J., & de Vivanco, W. (2013). Evaluation: Austrian Partnership Programme in Higher Education and Research for Development (APPEAR): Mid-term evaluation - Final report. Vienna: Austrian Development Cooperation

Chernikova, E. (2010). Tracer study of IDRC-supported award-holders: Internship awards, professional development awards, young Canadian researchers awards, doctoral research awards, and Canadian windows on international development awards. Ottawa, Ca: IDRC

Chesterfield, R., & Dant, W. (2013). Evaluation of LAC Higher Education scholarships program: Final report. North Bethesda, MD: JBS International

Clift, R., Dassin, J., & Zurbuchen, M. (2013). Linking higher education and social change: Ford Foundation International Fellowships Program. New York: Ford Foundation International Fellowships Program

Cosentino, C., Dumitrescu, A., Moortby, A., Rangarajan, A., Shaw, A., Sloan, M., Thomas, C., & Burciul, B. (2013). Learning across borders: The collaborative creation of a monitoring, evaluation, and learning framework for the MasterCard Foundation Scholars Program. Paper presented at UKFIET international conference on education and development - Education and Development post-2015: Reflecting, Reviewing, Re-Visioning, Sept 10 - 12, Oxford

CSC (2014). Evaluation profile: James Achanyi Fontem. Retrieved 25th March, 2014, from: < http://cscuk.dfid.gov.uk/wp-content/uploads/2014/02/case-study-achanyi-fontem.pdf>

Dassin, J. (2009). "Return" and "Returns": Brain Drain and the path back home. In Volkman, T., Dassin, J., & Zurbuchen, M. (eds). Origins, journeys and returns: social justice in international higher education. New York: Social Science Research Council

Dassin, J., Enders, J., & Kottmann, A. (2012a). Social inclusiveness, development and student mobility in international higher education: the case of the Ford Foundation International Fellowships Program. In B. Streitwieser (Ed.), Internationalization of higher education and global mobility (pp. 1–13). Oxford: Symposium Books

Dassin, J., Enders, J., & Kottmann, A. (2012b). Social inclusion in international higher education and leadership for justice: the approach and achievements of the Ford Foundation International Fellowships Program (IFP). Paper presented at the IFP Symposium, 19 November, Hawaii,

Dassin, J., Volkman, T., & Zurbuchen, M. (2009). Beyond measure: Fellowships and social justice. In Volkman, T., Dassin, J., & Zurbuchen, M. (eds). Origins, journeys and returns: social justice in international higher education. New York: Social Science Research Council

Day, R., & Geddes, N. (2008). Evaluating the impact of Commonwealth Scholarships in the United Kingdom: Results of the alumni survey. London: Commonwealth Scholarship Commission in the UK

Day, R., Stackhouse, J., & Geddes, N. (2009). Evaluating Commonwealth Scholarships in the United Kingdom: Assessing impact in key priority areas. London: Commonwealth Scholarship Commission in the UK

Department of Foreign Affairs and Trade (2011). Scholarships PNG: Monitoring and evaluation plan, version 2. Canberra: DFAT

Department of Foreign Affairs and Trade (2012). 2011 Vietnam tracer study of Australian scholarships and alumni. Canberra: DFAT

Department of Innovation, Industry, Science and Research (2010). International Postgraduate Research Scholarships (IPRS) program evaluation. Canberra: DIISR

Deutscher Akademischer Austauschdienst [DAAD] (2013). Knowledge - Action - Change, Three alumni surveys in review: 25 years of DAAD postgraduate courses. Bonn: DAAD

Dong, L., & Chapman, D. (2008). The Chinese Government Scholarship Program: An effective form of foreign assistance? *International review of education*, 54(2), 155-173

ECOTEC (2009). Ex-post evaluation of Erasmus Mundus: A final report to DG Education and Culture. Brussels: ECOTEC-ECORYS Group

Enders, J. (2007). Mobilizing marginalized talent: The International Fellowships Program. International Higher Education, 46, 7-8

Enders, J., & Kottmann, A. (2013). The International Fellowships Program: Experiences and outcomes. Enschede, NL: University of Twente

Evans, G., Stackhouse, J., & Geddes, N. (2009). Evaluating the impact of Commonwealth Scholarships in the United Kingdom: Assessing impact in the Caribbean. London: Commonwealth Scholarship Commission in the UK

Feiler, L., Jäger, M., & Reiter, W. (2007). Evaluation of the education sector of Austrian development cooperation and cooperation with South-East Europe. Vienna: ÖSB Consulting GmbH

Foreign and Commonwealth Office (2013). Triennial review report: Marshall Aid Commemoration Commission. Available online: <https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/215692/MACC_review.pdf>

Gilboy, A., Carr, H., Kane, T., & Torene, R. (2004). Generations of quiet progress: The development impact of U.S. long-term university training on Africa from 1963 to 2003, Volumes I, II and III. Washington, DC: USAID

Gondwe, M., & Schröder, B. (2013). IFP in continental Europe: 2001 - 2012. The Hague: Nuffic

Gosling, M. (2008). Scholarships effectiveness review (parts 1, 2, & 3). Unpublished report: AusAID

Graham, M. (2007). Tracer study of awards programs supported by IDRC: Internships, professional development awards, young Canadian researchers awards, doctoral research awards, and Canadian window on development awards. Ottawa, Ca: IDRC

Hansen, S., Boeren, A., Lexow, J., Wirak, A., Sigvaldsen, E., Fergus, M., Mwaipopo, E., Vusia, S., Hossain, I. (2005). Evaluation of the Norad Fellowship Programme. Oslo: Norad

Heim, E., Engelage, S., Zimmerman, A., Herweg, K., Michel, C., & Breu, T. (2012). Tracking alumni career paths: Third NCCR North-South report on effectiveness. NCCR North-South Dialogue 42. Bern, Switzerland: NCCR North-South

ICUnet.ag (2013). Erasmus Mundus: Graduate impact survey, September 2013 Passau, De: ICUnet.ag

International Development Research Centre (2010). Internal review of IDRC's fellowships and awards program. Ottawa, Ca: IDRC

International Development Research Centre (2013). Evaluation at IDRC. Ottawa, Ca: IDRC

IREX (2010). Edmund S. Muskie Graduate Fellowship Program: Selected Results. Washington DC: IREX

Janssens de Bisthoven, L. (2009). Supporting development relevant research. Paper presented at the International Symposium on Evaluation of Development Research, 5 June, Brussels

Kiernan, M., Ssentengo, P., & Naggayi, R. (2012). Evaluation of Danida's fellowship programme: Uganda country case study report. Copenhagen: Danida

Kottmann, A., & Enders, J. (2011). Alumni of the International Fellowships Program: Main findings from the survey 2011. Enschede, NL: University of Twente

Lee-Woolf, N. (2014). Chevening Scholarships 2012/2013 arrival survey analysis. Unpublished report, The Association of Commonwealth Universities.

Lee-Woolf, N. (2014). Chevening Scholarships 2012/2013 departure survey analysis. Unpublished report, The Association of Commonwealth Universities.

Mansukhani, V., Handa, N.L. (eds) (2013). Opening Doors: Ten years of the Ford Foundation International Fellowships Program in India. New Delhi: Ford Foundation International Fellowships Program

Ministry of Foreign Affairs of Denmark (2012). Evaluation of Danida's fellowship programme. Copenhagen: Danida.

Negin, J. (2014). Australian aid programme scholarships: An effective use of Australian aid? Paper presented at the 2013 Australasian aid and international development policy workshop, Feb 13, Canberra, Aus

Nijathaworn, B., Semblat, R., Takagi, S., & Tsumagari, T. (2009). Final report of the committee to review the Japan-IMF scholarship program for Asia (JISPA). Tokyo: IMF-OAP

Norad (2009). Evaluation of the Norwegian Programme for Development, Research and Education (NUFU) and of Norad's Programme for Master studies (NOMA). Oslo: NORAD

Nuffic (2009). NFP tracer 2009: 2nd draft report. The Hague: Nuffic

Nugroho, D., & Lietz, P. (2011). Meta-analysis of AusAID surveys of current and former scholarship awardees (draft). Unpublished report: AusAID

Penny, A., & Tefera, D. (2013). Country evaluation Ethiopia. Brussels: VLIR-UOS

Ramboll Management Consulting (2012). Final report: Evaluation of NPT and NICHE. Berlin: Ramboll Management Consulting

Rathgeber, E. (2010). Evaluation of IDRC's fellowships and awards program: A forward looking analysis. Ottawa, Ca: IDRC

River Path Associates (2003). The FCO Scholarships Review. Wimborne: River Path Associates

Rotem, A., Zinovieff, M., & Goubarev, A. (2010). A framework for evaluating the impact of the United Nations fellowship programmes. Human Resources for Health, 8(7), available at: < http://www.human-resources-health.com/content/8/1/7>

Säring, J., Spartakova, N., & Wegera, K. (2012). Erasmus Mundus: Graduate impact survey, September, 2012. Passau, Germany: ICUnet.ag

SRI International (2006). Executive summary: Outcome assessment of the Visiting Fulbright Student Program. Arlington, VA: SRI International

Tvaruzkova, M., & Clift, C. (2013). IFP alumni tracking study: A 10-year journey. Paper presented at the 27th annual conference of the American Evaluation Association, 16 - 19 Oct, Washington, DC

van der Aa, R., Willemson, A., & Warmerdam, S. (2012). Evaluation of the Netherlands Fellowship Programme (NFP) 2002-2010. Rotterdam: ECORYS Nederland BV

Visser, J., & Trinh, QL. (2011). Country evaluation Vietnam. Brussels: VLIR-UOS

Volkman, T., Dassin, J., & Zurbuchen, M. (eds) (2009). Origins, journeys and returns: social justice in international higher education. New York: Social Science Research Council

Webb, S. (2009). Australian scholarships in Cambodia: Tracer study and evaluation. Canberra: AusAID

World Bank Institute (2010). Joint Japan/World Bank Graduate Scholarship Program: Tracer study VIII. Washington, DC: World Bank Institute

Wyatt, A., & Andah, S. (2012). Evaluation of Danida's fellowship programme: Ghana case study report. Copenhagen: Danida

Appendix 2: Personal correspondence

The analysis presented in this report would have been greatly impoverished without the contributions of colleagues worldwide who responded to our enquiries about evaluation procedure. In some cases this involved sharing evaluation documents and in other cases clarifying our understanding of evaluation work through informal conversations. In particular, our gratitude is extended to the following correspondents:

Ad Boeren (April 16th, 2014)
Annika Sundbäck-Lindroos (April 16th, 2014)
Barry Burciul (February 19th, 2014)
Colin Tannam (April 25th, 2014)
Emily Hayter, Heath Thomson, and Beryl-Joan Bonsu (20th February 2014)
Jannie de Graaf (13th May, 2014)
Joan Dassin (February 21st, 2014)
Joel Negin (April 16th, 2014)
Jürgen Enders (March 5th, 2014)
Luca Lo Re (January 16th, 2014)
Mary Denyer (January 21st, 2014)
Michael Scott-Kline (January 29th, 2014)
Mirka Tvaruzkova and Rajika Bhandari (March 26th, 2014)
Rachel Day (April 3rd, 2014)
Sebastian Sabogal (February 6th, 2014)
Stefan Wellens (April 18th, 2014)
Stephen Copek, Jean Sutor, and Ingrid Schwab (February 17th, 2014)

In addition, we offer our thanks to the delegates of the 5th Annual Forum of the Donor Harmonisation Group (11-13 June, 2014: Helsinki) for their comments and reflections on the first presentation of findings from this study.

Appendix 3: Other sources cited

Alliger, G., & Janak, E. (1989), Kirkpatrick's levels of training criteria: Thirty years later. Personnel Psychology, 42(2). 331–342

Antle, B., Sullivan, D., Dryden, A., Karam, E., & Barbee, A. (2011). Healthy relationship education for dating violence prevention among high-risk youth. Children and Youth Services Review, 33(1), 173–179.

Barber, O. (2012). Evidence-based, politically savvy ways of doing development aid. **In:** King, K. (ed) Norrag News, 47: Value for money international education: A new world of results, impacts, and outcomes? Geneva: Norrag

Bates, R. (2004). A critical analysis of evaluation practice: the Kirkpatrick model and the principle of beneficence. *Evaluation and program planning*, 27(3). 341-347

Blair, J., Czaja, R., & Blair, E. (2014). Designing surveys: A guide to decisions and procedures. London: Sage

Bollier, D. (2010). The promise and peril of big data. Washington, DC: The Aspen Institute

Boeren, A., Bakhuisen, K., Christian-Mak, A., Musch, V., & Pettersen, K. (2008). Donor policies and implementation modalities with regard to international postgraduate programmes targeting scholars from developing countries. Brussels: VLIR-UOS

Boyd, d., & Crawford, K. (2012). Critical questions for big data. *Information, Communication & Society*, 15(5). 662-679

British Council and DAAD (2014). The rationale for sponsoring students to undertake international study: an assessment of national student mobility scholarship programmes. Manchester: British Council

Britt, H. (2013). Complexity-aware monitoring. Retrieved January 31st, 2014, from: < http://usaidlearninglab.org/library/discussion-note-complexity-aware-monitoring>

Byrne, D. (2013). Evaluating complex social interventions in a complex world. *Evaluation*, 19(3). 217-228

Callaghan, G. (2008). Evaluation and negotiated order: Developing the application of complexity theory. *Evaluation*, 14(4). 399-411

Cohen, J. (1992). A power primer. *Psychological bulletin*, 112(1). 155-159

Creed, C., Perraton, H., & Waage, J. (2012). Examining development evaluation in higher education interventions: A preliminary study. Paper presented at the LIDC & ACU conference on Measuring the Impact of Higher Education Interventions on Development, 19 - 20 March, London

Dolnicar, S. (2013). Asking good survey questions. *Journal of travel research*, 52(5). 551-574

Ellerman, D. (2012). Do we need an institute for the study of development fads? **In:** King, K. (ed) Norrag News, 47: Value for money international education: A new world of results, impacts, and outcomes? Geneva: Norrag

Fleming, F. (2011). Evaluation methods for assessing Value for Money. Retrieved January 20th, 2014, from: < http://betterevaluation.org/resource/assessing-value-for-money>

Friedriksen, B. (2012). The trend for donors to treasure (and fund) what can be measured. **In:** King, K. (ed) Norrag News, 47: Value for money international education: A new world of results, impacts, and outcomes? Geneva: Norrag

Garcia, M. (2011). Micro-methods in evaluating governance interventions. *Evaluation Working Papers*. Bonn: Bundesministerium für wirtschaftliche Zu-sammenarbeit und Entwicklung.

Girdwood, A. (2012). DFID's use of evidence: Achieving value for money through knowing 'what works'. **In:** King, K. (ed) Norrag News, 47: Value for money international education: A new world of results, impacts, and outcomes? Geneva: Norrag

Hageboeck, M., Frumkin, M., & Monschein, S. (2013). Meta-evaluation of quality and coverage of USAID evaluations, 2009-2012. Retrieved February 5th, 2014, from: < http://usaidlearninglab.org/library/meta-evaluation-quality-and-coverage-usaid-evaluations-2009-2012>

Holden, J., & Tryhorn, C. (2013). Influence and attraction: Culture and the race for soft power in the 21st century. Manchester: British Council

Holton, E. (1996). The flawed four-level evaluation model. *Human Resource Development Quarterly*, 7(1). 5-21.

House of Lords select committee on soft power and the UK's influence (2014). First Report: Persuasion and Power in the Modern World. Retrieved March 28th, 2014, from: <http://www.parliament.uk/soft-power-and-uks-influence>

Howell, D. (1999). Fundamental statistics for the behavioral sciences (4th ed). London: Duxbury Press

King, K., & Palmer, R. (2012). Value for money in international education: A new world of results, impacts and outcomes? *Norrag News 47*. Geneva: Norrag

Kirkpatrick, D. (1994). Evaluating training programs: The four levels. San Francisco: Berrett-Koehler

Lange, S. (2005). The Norad programme in arts and cultural education: A review of the first phase, 2002-2004. Bergen: Chr. Michelsen Institute

Letouzé, E. (2012). Big data for development: Challenges & opportunities. New York: UN Global Pulse

Ling, T. (2012). Evaluating complex and unfolding interventions in real time. *Evaluation,* 18(1). 79-91.

Madill, A., Jordan, A., & Shirley, C. (2000). Objectivity and reliability in qualitative analysis: Realist, contextualist and radical constructionist epistemologies. British Journal of Psychology, 91(1), 1–20.

Mayne, J. (2011). Addressing cause and effect in simple and complex settings through contribution analysis. **In:** Forss, K., Mara, M., & Schwartz, R. (eds) Evaluating the complex: Attribution, contribution and beyond. New Jersey: Transaction Publishers

OECD (1991). Principles for evaluation of development assistance. Retrieved 24th March, 2014, from: < http://www.oecd.org/dac/evaluation/50584880.pdf>

OECD (2002). Glossary of key terms in evaluation and results based management. Paris: OECD Publications

Oktech, M., McCowan, T., & Schendel, R. (2013). The impact of tertiary education on development: a rigorous literature review. London: Institute of Education

Palenberg, M. (2011). Tools and Methods for Evaluating the Efficiency of Development Interventions. *Evaluation Working Papers*. Bonn: Bundesministe-rium für wirtschaftliche Zusammenarbeit und Entwicklung.

Pascarella, E, & Terenzini, P. (1980). Predicting Freshman Persistence and Voluntary Dropout Decisions from a Theoretical Model. *The Journal of Higher Education,* 51(1): 60–75.

Perna, L., Orosz, K., Gopaul, B., Jumakulov, Z., Ashirbekov, A., & Kishkentayaeva, M. (2014). Promoting human capital development: A typology of international scholarship programs in higher education. *Educational Researc*her, 43(2). 66-73.

Rosenbaum, P., & Rubin, D. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1). 41-55.

Saldaña, S. (2012) The coding manual for qualitative researchers. London: Sage

Sanderson, I. (2000). Evaluation in complex policy systems. *Evaluation*, 6(4). 433-454

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R. & Befani, B. (2012). Broadening the range of designs and methods for impact evaluations. Working paper 38. London: DFID

Terano, M. (2011). Initial study on the range and impact of international scholarships. Unpublished report, The Association of Commonwealth Universities.

United Nations Development Programme Knowledge, Innovation and Capacity Group (2013). Innovations in monitoring & evaluation results. Retrieved 9th April, 2014, from: < http://www.undp.org/content/undp/en/home/librarypage/capacity-building/discussion-paper--innovations-in-monitoring---evaluating-results/>

Vardakoulias, O. (2012). Measuring the economic impact of Commonwealth Scholarships: Identifying Methodologies for Cost Benefit Analysis and Value for Money. London: Nef Consulting

White, H., & Barbu, A. (2006). Impact evaluation – The experience of the Independent Evaluation Group of The World Bank. Washington, DC: The World Bank

Wilson-Grau, R., & Britt, H. (2012). Outcome harvesting. Cairo: Ford Foundation

Commonwealth Scholarship Commission in the UK

Woburn House
20-24 Tavistock Square
London WC1H 9HF
United Kingdom
T +44 (0) 207 380 6700
F +44 (0) 207 387 2655
evaluation@cscuk.org.uk
www.dfid.gov.uk/cscuk