Robert H. Chenhall, Matthew Hall and David Smith

Performance measurement, modes of evaluation and the development of compromising accounts

Article (Accepted version) (Refereed)

Original citation:
Chenhall, Robert H., Hall, Matthew and Smith, David (2013) *Performance measurement, modes of evaluation and the development of compromising accounts*. Accounting, Organizations and Society, 38 (4). pp. 268-287. ISSN 0361-3682 DOI: 10.1016/j.aos.2013.06.002

Full text copy of:

Chenhall, R., Hall, M., and Smith, D. 2013. Performance measurement, modes of evaluation and the development of compromising accounts. *Accounting, Organizations and Society*, 38, 268-287.

Abstract

In this paper we develop the concept of compromising accounts as a distinctive approach to the analysis of whether and how accounting can facilitate compromise amongst organizational actors. We take the existence of conflicting logics and values as the starting point for our analysis, and directly examine the ways in which the design and operation of accounts can be implicated in compromises between different modes of evaluation and when and how such compromises can be productive or unproductive. In doing so, we draw on Stark's (2009: 27) concept of 'organizing dissonance', where the coming together of multiple evaluative principles has the potential to produce a 'productive friction' that can help the organization to recombine ideas and perspectives in creative and constructive ways. In a field study of a non-government organization, we examine how debates and struggles over the design and operation of a performance measurement system affected the potential for productive debate and compromise between different modes of evaluation. Our study shows that there is much scope for future research to examine how accounts can create sites that bring together (or indeed push apart) organizational actors with different evaluative principles, and the ways in which this 'coming together' can be potentially productive and/or destructive.

Introduction

> *"There's still a big debate in VSO about whether the purpose is to make sure volunteers have a good experience overseas and then return back happy or do we have some sort of coherent development programmes which use volunteers as a key input. I think there are those two different views of the organization. I mean there's a whole lot of views between those two but those are the two extremes... it probably divides down the middle" (Regional Director 2, Voluntary Service Overseas).*

The role of accounting practices in situations of different and potentially competing interests has been a prominent feature in studies of accounting and organizations. Some studies have shown how accounting practices can be mobilized by organizational actors to introduce a new order and model of organizational rationality, typically one focused on market concerns (e.g., Dent, 1991; Oakes, Townley & Cooper, 1998; Ezzamel, Willmott & Worthington, 2008). Other research has emphasized the role of accounting in situations of multiple and potentially conflicting interests, logics and regimes of accountability (e.g., Cooper, Hinings, Greenwood & Brown, 1996; Ahrens & Chapman, 2002; Lounsbury, 2008). In these settings, organizational subgroups can hold differing views of organizational reality that are not displaced, but can become layered (c.f., Cooper et al., 1996) such that they persist over time, or, as the quote above suggests, remain "divided down the middle." Here, accounts such as costing, resource allocation, and performance measurement systems are involved in on-going contests and struggles as various groups advance particular interests and values (e.g., Nahapiet, 1988; Briers & Chua, 2001; Andon, Baxter & Chua, 2007). Research on the use of financial and non-financial measures in performance measurement systems (e.g., Kaplan & Norton, 1992; Sundin, Granlund & Brown, 2010), or the use of quantitative and qualitative information in financial reports (e.g., Nicholls, 2009; Chahed, 2010), can also be seen to relate to the ways in which accounting practices can give voice to different concerns and priorities. Often the outcome of struggles between groups is intractable conflict and confused effort with the eventual dominance of a singular perspective that limits opportunities for on-going contests and debate (e.g., Dent, 1990; Fischer & Ferlie, 2013). Alternatively, the processes taken by sub-groups to promote their preferred views can sometimes achieve a more workable compromise that generates constructive debate and on-going dialogue (e.g., Nahapiet, 1988; Sundin et al., 2010). Building on this literature, in this study we analyse directly the ways in which the design and operation of accounts can be implicated in compromises between different modes of evaluation and seek to illustrate when and how such compromises can be productive or unproductive.

As conflicting logics are probably unavoidable in any human organization (Gendron, 2002), our approach is to take the existence of, and the potential for, tension between different modes of evaluation as the starting point for our analysis. In doing so, we mobilize Stark's (2009: 27) concept of 'organizing dissonance', which posits that the coming together of multiple evaluative principles has the potential to produce a 'productive friction' that can help the organization to recombine ideas and perspectives in creative and constructive ways. The concept of organizing dissonance provides an analytical approach that views the co-existence of multiple evaluative principles as an opportunity for productive debate, rather than a site of domination or intractable conflict. As such, our approach extends prior research by privileging analysis of when and how the co-existence of multiple evaluative principles can be productive or unproductive. We summarize the focus of our study in the following research questions: How does the design and operation of accounting practices facilitate (or impede) compromise in situations of multiple evaluative principles? When (and how) is compromise between different evaluative principles productive or unproductive?

We argue that answers to these questions contribute to the literature by focusing directly on how accounting is implicated in compromising between different evaluative principles and the way in which such compromise can be productive or unproductive. Here the design and operation of accounting practices can help organizational actors to re-order priorities and integrate perspectives in situations of co-existing and potentially competing values (Stark, 2009). In particular, we show how accounts have the potential to provide a fertile arena for productive debate between individuals and groups who have differing values (Stark; 2009; Jay, 2013; Gehman, Trevino & Garud, 2013; Moor & Lury, 2011; Denis, Langley & Rouleau, 2007).

The findings from our field study of a non-government organization indicate that the potential for accounts to provide a fertile arena for productive debate is related to three important processes. First, designing accounts that productively manage tensions between different evaluative principles involves 'imperfection', that is, a process of 'give and take' that ensures that no single evaluative principle comes to dominate others. Here the design and operation of accounting practices represents a temporary settlement between different evaluative principles that will require on-going effort to maintain (c.f., Stark, 2009; Gehman et al., 2013). Second, the design and operation of accounts can facilitate productive friction by making visible the attributes of accounts that are important to organizational actors with different evaluative principles, a process that we term 'concurrent visibility.' This process is important because it serves to crystallize compromises between different modes of evaluation in a material form (Denis et al., 2007). Third, our study reveals an important distinction between the types of responses that can emerge in situations where compromises break down and accounting practices are viewed as 'not working.' In particular, we show how debates over the mechanics of accounting practices can be unproductive and lead to 'stuckness' (Jay, 2013) between different modes of evaluation, whereas debate focused on the principles underlying the account can help to integrate different evaluative principles in a productive way (Stark, 2009; Jay, 2013).

Overall, our approach improves understanding of how actors with different evaluative principles reach an acceptable compromise, the factors that promote and/or damage efforts to reach compromise, and the consequences for those individuals, groups, and organizations involved. Accounts are central to these processes because they are a site where multiple modes of evaluation potentially operate at once, with different modes of evaluation privileging particular metrics, measuring instruments and proofs of worth (Stark, 1996; 2009). Accounts of performance are critical because it is in discussions over the different metrics, images and words that can be used to represent performance that the actual worth of things is frequently debated and contested. An analysis of compromising accounts[1] provides a powerful analytical lens for examining whether and how compromise between different modes of evaluation is developed, established and destroyed. In particular, we show how the design and operation of accounts can create the potential for 'productive friction' to arise from the coming together of different evaluative principles (Stark, 2009).

Our study also makes a more specific contribution to research on performance measurement systems. There has been a wealth of prior management accounting

[1] We use the term 'compromising accounts' to refer to the role of accounts in facilitating (or not) compromise between actors with different evaluative principles. We develop this concept later in the paper.

studies focusing on the attributes of various performance metrics and their effects on individual and organizational performance (see, for example, research on subjectivity (Gibbs, Merchant, Van der Stede & Vargus, 2004; Moers, 2005), comprehensiveness (Hall, 2008) and financial/non-financial measures (e.g. Perera, Harrison & Poole, 1997; Baines & Langfield-Smith, 2003). However, most of these studies do not explicitly investigate how the metrics that comprise performance measurement systems are developed (see Wouters & Wilderom (2008) and Townley, Cooper & Oakes (2003) for exceptions). Thus, we extend this literature by examining explicitly the processes that take place in negotiating the scope, design and operation of the metrics included in performance measurement systems.

The remainder of the paper is structured as follows. In the next section we provide the theoretical framework for the study. The third section details the research method, with the fourth section presenting findings from our field study of a non-government organization, Voluntary Service Overseas. In the final section we discuss our findings and provide concluding comments.

Theoretical framework

Our focus is on whether and how accounting practices can aid compromises in situations of co-existing modes of evaluation. As such, in developing our theoretical framework, we draw on recent developments in the 'sociology of worth' to help conceptualize the co-existence of, and potential for agreement between, multiple evaluative systems (see for example, Stark, 2009; Boltanski & Thévenot, 1999, 2006; McInerney, 2008; Denis et al., 2007; Huault & Rainelli-Weiss, 2011). A focus of this perspective is to examine how competing values are taken into account when parties seek to reach agreement or resolve disputes. Boltanski & Thévenot (2006) conceptualize individuals as living in different 'worlds' or orders of worth, where each 'world' privileges particular modes of evaluation that entail discrete metrics, measuring instruments and proofs of worth (Stark, 2009).[2] Instead of enforcing a

[2] Boltanski and Thévenot (1999, 2006) specify six 'worlds' or orders of worth (the 'inspirational', 'domestic', 'opinion', 'civic', 'merchant' and 'industrial' worlds). The 'civic' world, for example, is based on solidarity, justice and the suppression of particular interests in pursuit of the common good, whereas the 'market' world is one with competing actors who play a commercial game to further their personal (rather than collective) goals. In this paper our key focus is on understanding why and how actors can reach compromises (or not) in situations that are characterised by the presence of multiple evaluative principles. In doing so, we follow the approach of Stark (2009, see p.13 in particular). That

single principle of evaluation as the only acceptable framework, it is recognized that it is legitimate for actors to articulate alternative conceptions of what is valuable, where multiple evaluative principles can potentially co-exist and compete in any given field (Stark, 2009; 1996; McInerney, 2008; Moor & Lury, 2011; Scott & Orlikowski, 2012; Kaplan & Murray, 2010).

As co-existing evaluative principles may not be compatible, a ‘clash’ or dispute may emerge between parties, who at a given point in time, and in relation to a given situation, emphasize different modes of evaluation (Jagd, 2011; Kaplan & Murray, 2010). Following Stark (2009), who extends the framework of Boltanski and Thévenot (2006), our focus is directed not at the presence of particular logics or orders of worth, but on exploring the ways in which the co-existence of different logics can be productive or destructive. In doing so, we draw on Stark’s (2009) notion of organizing dissonance. Stark (2009) characterizes organizing dissonance as being a possible outcome of a clash between proponents of differing conceptions of value, that is, in situations when multiple performance criteria overlap. The dissonance that results from such a clash requires the organization to consider new ways of using resources in a manner that accommodates these different evaluative principles. Here, rather than something to be avoided, struggles between different evaluative criteria can prompt those involved to engage in deliberate consideration about the merits of existing practices (Gehman et al., 2013). In this way, keeping multiple performance criteria in play can produce a resourceful dissonance that can enable organisations to benefit from the ‘productive friction’ that can result (Stark, 2009). However, as Stark (2009: 27) notes, not all forms of friction will be productive, as there is a danger that “where multiple evaluative principles collide...arguments displace action and nothing is accomplished.” This points to the critical nature of compromises when there are disputes involving different evaluative principles. In practice, such compromises can be facilitated by the use of conventions, as detailed in the following section.

Disputes, conventions and accounting practices

---

is, we do not confine our analysis to the six orders of worth as outlined by Boltanski and Thévenot (1999, 2006) but specify the different evaluative principles as is appropriate to the particular empirical setting. Given our approach, we do not elaborate further on the six orders of worth of Boltanski and Thévenot (1999, 2006) here. For further insight on the six orders of worth, see Boltanski and Thévenot (1999, 2006), and for their implications for accounting research, see Annisette and Richardson (2011) and Annisette and Trivedi (2013).

The negotiation and development of conventions is seen as a critical tool to aid compromise in situations of co-existing evaluative principles (Denis et al., 2007). A convention is “an artefact or object that crystallises the compromise between various logics in a specific context” (Denis et al., 2007: 192). Conventions can help to bridge different perspectives by providing an acceptable compromise between competing value frameworks (Biggard & Beamish, 2003; Denis et al., 2007).

Accounting practices as a convention can help to resolve disputes in two inter-related ways. One, the development and operation of accounts can provide a fertile arena for debate between individuals and groups with differing evaluative principles. The production of accounts is important to this process because different evaluative principles do not necessary conflict or compete continuously, but resurface at particular moments in time (Jay, 2013), such as during the design and operation of accounting practices. Two, the production of accounts can serve to ‘crystallize’ the compromise in a material form (c.f., Denis et al., 2007), thus providing recognition of, and visibility to, different values and principles.

Tensions over accounts and accounting practices are likely because they can have very real consequences for the ordering of priorities in an organization and, consequently, for the interests of groups within the organization who hold different views. It is well understood that accounting can make certain factors more visible and more important than others, provide inputs that affect decision-making and the allocation of resources, and can also provide authoritative signals regarding the very purpose and direction of the organization. In addition, research has highlighted the persuasiveness of numbers in accounts and the role of quantification in advancing particular views and interests (e.g., Robson, 1992; Porter, 1995; Vollmer, 2007).

Nahapiet’s (1988) study of changes to a resource allocation formula in the United Kingdom’s National Health Service showed how the formula made existing values more visible and tangible and thus acted as a stimulus which forced explicit consideration of three fundamental organizational dilemmas. In this setting, actors contested strongly the formula’s design and operation, and its interpretation by other groups. Different interpretations of the formula, and of accounting more generally, were problematic because they played a key role in establishing what counts and thus what is worthy. This tension is exacerbated in organizational settings where limited resources (e.g., money, time, space) mean that not all interests can be accommodated. In particular, the processes of evaluation inherent to the production of accounts are central to problems of worth in organizations (c.f., Stark, 2009). For example, the process of developing, adjusting and reconfiguring accounts can require groups to make mutual concessions (i.e., compromise) in order to agree on the final (if only temporary) form and content of the account. In this way, producing accounts can provide an arena where different understandings of value may be articulated, tested, and partially resolved (Moor & Lury, 2011). However, while debate over accounts has the potential to facilitate productive friction, this depends on whether and how the convention comes to be (and continues to be) viewed as an 'acceptable' compromise. Importantly, although accounts as conventions may help enact compromises, they can also be subject to criticism and thus require on-going efforts to maintain and stabilize the compromise.

Responses to breakdowns in compromise

Designing accounting practices in the presence of co-existing modes of evaluation is likely to result in situations where the practice is viewed, at least by some actors in the organization, as 'not working.' Here there is a 'breakdown' such that issues and concerns that have arisen can no longer be absorbed into the usual way of operating (Sandberg & Tsoukas, 2011). Some breakdowns can be viewed as temporary and so the focus is on what is problematic about the current practice and how to fix it (Sandberg & Tsoukas, 2011). For example, doubts and criticisms can arise about the difficulties of implementing the practice, about whether it will result in the desired behaviours, and how it will influence other practices (Gehman et al., 2013).

This resonates with research in accounting that shows how the introduction of new accounting practices can result in criticisms that they have not been implemented correctly and revised procedures are required to improve the design and implementation process (e.g., Cavalluzzo & Ittner, 2004; Wouters & Roijmans, 2011). A criticism of existing practices is also evident, for example, in the context of performance measurement systems that are seen to require more non-financial measures (Kaplan & Norton, 1992) and financial reports are viewed as needing more narrative information (Chahed, 2010). Such criticisms can result in changes to the existing accounting practices. Stark (2009) notes, however, that disputes over the mechanics of existing practices may not lead to effective changes, but rather result in a situation where nothing is accomplished. Here, co-existing modes of evaluation may not lead to innovation, but rather oscillation and ‘stuckness’ between logics (Jay, 2013).

A breakdown in practice can also be more severe such that existing ways of doing things no longer work and reflection at a distance from the existing practice is required (Sandberg & Tsoukas, 2011). Here actors can debate the principles and values underlying the existing practice and the changes that are required to move beyond the breakdown (Gehman et al., 2013). This type of criticism and debate can arise where people feel that some fundamental principles with which they identify are not being respected (Denis et al., 2007). This can be particularly problematic in debates over incommensurables, that is, the process of denying “that the value of two things is comparable” (Espeland & Stevens, 1998: 326). Claims over incommensurables are important because they can be “vital expressions of core values, signalling to people how they should act toward those things” (Espeland & Stevens, 1998: 327). It can also arise where the values evident in the existing practice clash with deeply held values obtained through prior experience (Gehman et al., 2013).

Debates over the underlying principles of accounting practices, and conventions more broadly, can result in what Stark (2009: 27) labels “organizing dissonance”, that is, a process of productive friction arising from debate between actors over different and potentially diverse evaluative principles. To generate productive friction in the context of such debates, the rivalry between different groups must be principled, with advocates offering reasoned justifications for their positions (Stark, 2009). In this situation actors become reflexively aware of latent paradoxes and directly confront and accept ambiguities, helping new practices that integrate logics to emerge (Jay, 2013). The resolution of breakdowns also requires recognition that such a compromise represents a “temporary settlement” (Stark, 2009: 27) between competing value frameworks that is fragile (Kaplan & Murray, 2010) and only likely to be maintained via on-going effort and reworking (Gehman et al., 2013).

Summary

This discussion highlights the potential role for accounts in developing compromises in situations where the co-existence of different evaluative principles is a common feature of organizations. In particular, it reveals how accounts have the potential to act as a convention to help develop and crystallize compromises. It also highlighted the way in which compromises are temporary settlements that require on-going work to stabilize. In particular, the merits of an accounting practice may be called into question, resulting in efforts to ‘fix’ the way in which the practice currently operates and/or debate focused on resolving tensions between underlying principles and values. In the next section, we empirically examine the role of compromising accounts through a detailed analysis of the development of a performance measurement system that we observed during a longitudinal field study at Voluntary Service Overseas (VSO).

Method

VSO is a non-governmental international development organization that works by (mainly) linking volunteers with partner organizations in developing countries. Each year approximately 1500 volunteers are recruited and take up placements in one of the over forty developing countries in which VSO operates. Our interest in VSO was sparked due to an initiative to develop a new performance measurement system, subsequently referred to as the ‘Quality Framework’ (QF). This framework attempted to combine different metrics and narrative content into a single report that would provide a common measure of performance in each of VSO’s country programmes.

The field study was conducted between July 2008 and August 2010. During this time we conducted 32 interviews, attended meetings, observed day-to-day work practices, collected internal and publicly available documents, participated in lunches and after-work drinks with staff and volunteers, primarily in London, but also during a one-week visit to the Sri Lanka programme office in January 2009.

Most of the interviews were conducted by one of the authors, with two authors conducting the interviews with the country directors. Interviews lasted from 30 minutes to two hours. Almost all interviews were digitally recorded and transcribed, and, where this was not possible, extensive notes were taken during the interview and further notes then written-up on the same day. We interviewed staff across many levels of the organization as well as staff at different locations. Face-to-face interviews were conducted at VSO’s London headquarters, and in Sri Lanka. Due to the location of VSO staff around the world, some interviews (particularly those with country directors) were conducted via telephone. Table 1 provides an overview of the formal interviews and observations of meetings. We carried out observations of 17 meetings and workshops in both London and Sri Lanka, primarily concerned with the QF and other planning and evaluation practices.

Throughout the study, we were also involved in informal conversations (typically before and after meetings, and during coffee breaks, lunches and after-work drinks) where staff and volunteers expressed their thoughts about the meetings, as well as other goings-on at VSO and the non-government organization sector. We kept a detailed notebook of these informal conversations, which was then written up into an ‘expanded account’ (Spradley, 1980) that on completion of the field study totalled more than 200 pages of text. We also exchanged numerous emails (over 700 separate communications) and telephone conversations with VSO staff.

We were provided access to over 600 internal VSO documents, including performance measurement reports, supporting documents and analysis. These reports included the complete set of QF reports from each VSO programme office for 2008 and 2009, documents related to other monitoring and review processes, as well as more general documents concerning organizational policies, plans and strategies. Finally, we collected publicly available documents, such as annual reports and programme reviews, newspaper articles, as well as several books on VSO (e.g. Bird, 1998).

Consistent with the approach employed by Ahrens & Chapman (2004), Free (2008), and Chenhall, Hall & Smith (2010), we employed Eisenhardt’s (1989) methods. This involved arranging the data (transcripts, field notes, documents) chronologically and identifying common themes and emerging patterns. We focused in particular on iterations in the content and use of performance measurement systems at VSO over time and then sought to understand why they came about and the subsequent reactions from different people within the organization. We then re-organized this original data around key events (for example, the ‘league table’ debates) and significant issues (for example, ‘consistency’) that emerged as we sought to understand the performance measurement and review systems at VSO. We compared our emerging findings from the study with existing research to identify the extent of matching between our data and expectations based on prior theory. In particular, findings that did not appear to fit emerging patterns and/or existing research were highlighted for further investigation. This process was iterative throughout the research, and finished when we believed we had generated a plausible fit between our research questions, theory and data (Ahrens & Chapman, 2006).

<insert Table 1 here>

Case context

VSO was founded in 1958 in England as an organization to send school leavers to teach English in the "underdeveloped countries" of the Commonwealth (Bird, 1998: 15). Volunteers were initially recruited exclusively from England, and later from other countries, including the Netherlands, Canada, Kenya, the Philippines, and India. The initial focus on the 18-year-old high school graduate was replaced over time by a (typically) 30-year old-plus experienced professional. Volunteers operated under a capacity building approach, being involved in initiatives such as teacher training, curriculum development, and advocacy.[3]

In 2004 VSO signalled it would adopt a more 'programmatic' approach to its work, which shifted attention away from each volunteer placement to one that focused "all our efforts on achieving specific development priorities within the framework of six development goals" (VSO, 2004).[4] This move to a programmatic model was coupled with explicit recognition of VSO's purpose as primarily a 'development' rather than 'volunteer-sending' organization, and the development of evaluation systems to support this change. Notwithstanding this explicit shift in organizational priorities, the focus on volunteering was still strong, particularly as many VSO staff were formerly volunteers. As such, a mix of different world-views at VSO was the norm:

> "There are some different kind of ideological views between people who feel that the important thing about VSO, it's just about international cooperation and getting people from different countries mixing with each other and sharing ideas. It doesn't matter what the outcome is really, it's going to be a positive thing but you don't need to pin it down. Versus it's all about pinning down the impact and the outcomes of our work and being very focused and targeted and being able to work out what is your return on your investment and all these kind of things so I think it is partly historical and partly differences in just a mindset or world-view." (Interview, Regional Director 2, November 2008).

[3] VSO operated what it calls a 'capacity building' approach by partnering volunteers with local organizations that require assistance or expertise in a variety of capacities. VSO describes its partnership approach as follows: "We work with local partners in the communities we work with, placing volunteers with them to help increase their impact and effectiveness" (VSO website, http://www.vsointernational.org/vso-today/how-we-do-it/, accessed 7 April 2010). Volunteers typically take up a specific role or position, often working alongside a local staff member, where partner organizations range in size from very small, local businesses, community groups and NGOs, to large organizations and government departments and ministries.

[4] The six development goals were health, education, secure livelihoods, HIV/AIDS, disability, and participation and governance (Focus for Change, VSO, 2004).

The different views on VSO's overall purpose created considerable tension, focused in particular on debates about the value of VSO's work. Originating from VSO's founding principles, many staff and volunteers felt that volunteering was, in and of itself, a positive and productive activity and any drive to specify an 'outcome' of this was secondary. In contrast, the programmatic approach, coupled with the recruitment of many staff from other international development agencies, gave more attention to poverty reduction and demonstration of the 'impact' of VSO's work. This situation was increasingly common in the wider NGO sector, where founding principles of volunteerism, the development of personal relationships, and respect for each individual were coming into contact with more 'commercial' values favouring professionalism, competition and standardization (see, for example, Parsons & Broadbridge, 2004; Helmig, Jegers & Lapsley, 2004; Hopgood, 2006).

As an espoused international development organization, VSO also existed in an environment increasingly characterized by the use of indicators and targets (a prime example being the Millennium Development Goals, see United Nations, 2011) and a greater focus on the effectiveness of aid (particularly the Paris Declaration on Aid Effectiveness in 2005).[5] VSO's main funder, the United Kingdom's Department for International Development (DFID), had aligned its development programme around the Millennium Development Goals, and was also a signatory to the Paris Declaration.[6,7] This had implications for the way in which VSO was required to report to DFID, particularly during the course of our study when a change in DFID's reporting format required VSO to track progress against its four agreed strategic objectives using a set of 17 indicators.[8]

Collectively, the changing context of the NGO sector, the move in international development towards an increased focus on aid effectiveness and the use of indicators, along with VSO's own progression from a volunteering to a more

---

[5] See www.oecd.org/dataoecd/11/41/34428351.pdf for the Declaration.

[6] See DFID (2000), DFID (2006) and DFID (2009).

[7] In terms of the overall funding environment, VSO's total funding increased steadily during the 2000s. In 2000 total income was approximately £28m, with approximately £22m from DFID (77% of total funds). In 2005 total income was approximately £34m, with approximately £25m from DFID (74% total funds). In 2009 total income was approximately £47m, with approximately £29m from DFID (60% total funds) (source: VSO Annual Reports).

[8] See www.vsointernational.org/Images/ppa-self-assessment-review-2010-11_tcm76-32739.pdf for the 2010/2011 report to DFID (accessed 31 May 2012). The first report issued under this format was for the 2009/10 reporting year. Prior to this, there was an absence of indicators, with VSO reporting against various development outcomes using descriptive examples of progress from different countries ('VSO Narrative Summary and Learning Report for PPA 2005-6').

programmatic focus, meant that the co-existence of different evaluative principles characterized the situation at VSO. In particular, we identify two primary modes of evaluation.[9] The first mode of evaluation, which we label 'learning and uniqueness', was focused primarily on reflection, the use of contextual and local interpretations, and a preference for narrative content. Discourses within VSO regularly emphasized the importance of this mode of evaluation, with one of VSO's three stated values a "commitment to learning" whereby VSO seeks to "continue to develop effective monitoring and evaluation methods so that we can learn from our own and others' works" (VSO, 2004). The second mode of evaluation, which we label 'consistency and competition', was focused primarily on standardization, the use of consistent and universal interpretations, and a preference for indicators. We outline the different modes of evaluation in Table 2, which we return to throughout our empirical analysis.

<insert Table 2 here>

Attempts at compromise between these different modes of evaluation became evident in debates about how to measure the value of VSO's work in each country. Measuring performance became particularly important because the move to be more programmatic had placed increased pressure on the allocation of resources amongst programme offices, as it required more expenditure on staff to support volunteer placements and develop and manage programmes.[10] However, the situation was characterized by a lack of commonly agreed criteria for measuring the performance of country programmes (c.f., Garud, 2008), where over time three approaches had been instigated; the 'Strategic Resource Allocation' (SRA) tool, the 'Annual Country Report' (ACR) and the 'Quality Framework' (QF).[11]

---

[9] As noted above, our approach here is to follow Stark (2009) and specify the different modes of evaluation in accordance with our empirical setting.

[10] VSO operated a geographic structure, whereby several programme offices were grouped together to form a specific region, for example, Sri Lanka, India, Bangladesh, Nepal and Pakistan formed the 'South Asia' region. Each Country Director reported to a 'Regional Director', with the Regional Directors reporting to the Director of IPG, based in London. IPG also had staff responsible for providing support to programme offices in areas such as funding, advocacy, and programme learning and reporting. Each programme office was a budget holder, and received core funding from VSO headquarters via the annual budgeting process. Core funding related to costs such as staff salaries and benefits, office and vehicle rental, and volunteer costs (including allowances and training/support costs). Each programme office received a limited amount of funding for 'programme' costs, with programme offices expected to apply for grants from donors to support further programme work.

[11] Our field study (July 2008 to August 2010) corresponded to the first year of the QF's operation and thus was subsequent to the use of the SRA and ACR. As such, we briefly describe the SRA and ACR

The SRA was developed in 2002 as VSO's first attempt to measure the effectiveness of each programme office.[12] The SRA relied almost exclusively on using numerical data to measure performance, where each programme office was required to score itself on 16 criteria related to the extent to which its work was focused on disadvantage, achieved certain outputs related to volunteers, and adopted a strategic approach. Each criterion was given a precise percentage weighting, e.g., 2% or 4% or 17%. Scores on the 16 criteria were to be aggregated with each programme office awarded a percentage score out of 100, with recognition that "the higher the overall percentage a Programme Office receives in this tool, the more "effective" it will be perceived to be based on this measure."[13] There was a strong emphasis on review of scores by staff in London "to ensure consistency between regions…in order to ensure transparency and to allow comparison between countries."[14] The SRA's implementation was problematic, however, and the approach was abandoned as a country director later explained:

> "The SRA was dropped because it was becoming increasingly apparent that some programmes were rather self-critical while others were not - but that this did not necessarily relate very closely to programme quality - in fact it appeared that sometimes the opposite was true, the programmes that had the capacity to critically assess their own performance (and give themselves a low score) were of a better quality than those who from year to year claimed that things were going well - and this resulted in some good programmes being closed down." (Interview, Country Director 1, November 2008).

Subsequent to the SRA, the ACR was developed in 2005 and focused its reporting on the 'activity' that a country programme had engaged in, such as 'so many workshops, so many volunteer placements.'[15] The ACR itself did not contain any quantitative scoring or ranking of programme offices but was a narrative report that provided descriptions of progress towards 'Strategic objectives' and contained a section focused on 'Lessons' to be learned.[16] The ACR also included one or more 'Most Significant Change' (MSC) stories' which focused on telling stories as a way to reflect upon and learn from programme experiences (see Dart & Davies, 2003; Davies & Dart, 2005).

---

to provide context to the development of the QF but do not analyze the development of the SRA and ACR in detail.

[12] See Appendix 1 which provides the 'summary page' of the SRA.

[13] SRA document, 2002.

[14] SRA document, 2002.

[15] Interview, Country Director 2, November 2008.

[16] ACR document, 2005.

The third approach (and our empirical focus) developed subsequent to the SRA and ACR was the QF, which attempted to combine scoring and narrative elements into a single reporting framework. We show how the QF was subject to criticism that resulted in changes in the use of narrative and quantitative measures, which favoured a mode of evaluation focused on 'consistency' over that which respected the 'unique' circumstances of individual country programmes. A further dispute emerged over the relative focus on 'learning' and 'competition' that precipitated more fundamental changes in order to develop an account that helped to compromise between the different modes of evaluation.

Development of the quality framework

The initial development of the QF occurred during a meeting of all country directors in late 2007.[17] Prior to this meeting, in an email sent to all country directors in May 2007, the Director of IPG gave his support to the development of the QF and outlined his rationale for its implementation:

> "We are very good at measuring volunteer numbers, numbers of tools being used, early return rates, levels of programme funding – but what about the impact of our work? How do we know if we really are working with poor people to contribute to positive change in their lives?...I believe that it is absolutely essential that we have a shared vision of success - that we all know what a high quality, successful VSO country programme could look like - that we know how to measure this - and that we have a culture that encourages, supports and celebrates this. Of course all of our country programmes could, and should, look very different. Local circumstances and development priorities above all should ensure this…However, there must be some fundamental principles that drive VSO's programme work and enable us to determine whether we are successful or not."

In this statement, the imperative for compromise between VSO's different modes of evaluation in the development of the QF was revealed. The reference to a 'shared vision of success' and knowing 'how to measure this' indicates a concern with developing common and standardized ways of measuring success. There is also recognition of the uniqueness of country programmes in that they 'should look very different.' In our analysis below, we focus on two central debates that emerged in the development of the QF; the first concerning the tension between standardization and

[17] As noted above, the first report to DFID that used indicators to track progress against strategic objectives was for the 2009/10 reporting year. Within VSO, work to address the new reporting requirements began in the second half of 2008, more than one year after the initial development of the QF. We also note that the QF reports were not provided to DFID or to any other external funders, although the IPG Director commented that he did inform DFID about the QF process and that this was considered by him to be 'helpful' in showing DFID that VSO was addressing issues around the impact of its work.

uniqueness, and the second regarding the most appropriate approach to improve programme quality.

Debate 1: How to standardize and respect uniqueness?

A key difficulty in developing the QF was tension between the desire to standardize, i.e., have indicators that provide a consistent method for measuring success in each programme office, whilst respecting the uniqueness of programme offices and the need for indicators to be ‘inspirational.’ It was the need to make choices about the content of elements, indicators and narrative components in the QF that provided an arena for debates and discussions regarding different modes of evaluation at VSO. Country directors and other programme staff were central to these discussions, and provided suggestions for elements and indicators that were collected in London, and then followed in late 2007 by a meeting of all country directors and senior IPG staff in Cambridge, UK. A central platform of this meeting was sessions devoted to dialogue and debate about the elements and indicators that would comprise the QF. Centred on the question “What is quality?”, it was here that staff were able to advocate for the inclusion and exclusion of particular elements and indicators. This resulted in a set of 14 elements relating to various aspects of programme quality, such as inclusion, volunteer engagement, innovative programming and financial management. Importantly, the elements relating to the impact of VSO’s work on partners and beneficiaries were given highest priority: they were the first two elements in the QF and were assigned the labels ‘Element A’ and ‘Element B’ to distinguish them from the other elements that were labelled with numbers one through twelve (see Appendix 2).

Testing the QF at the country level was a priority, with a country director offering the following reflections on a pilot test:

> “We worked through the different indicators to see whether the results that the framework spat out were recognizable…some of the results didn't give the right answer basically. So we changed some of the indicators...The framework itself allows for a small narrative at the beginning of each element, which can at least explain context as to why it may have a low score or conversely why it might have a high score. They may be working in a country that has a very easy operational environment. It might have lots of external funding and that for me is reflected in that short narrative section at the beginning.” (Interview, Country Director 1, November 2008).

This comment reveals how local knowledge was considered critical in that indicators were required to produce results that were ‘recognizable’ to programme office staff, partners and volunteers. Providing space in the QF report for narrative discussion to reflect the different circumstances of countries was also important. In particular, the narratives in the QF reports were typically very extensive and contained statements calling attention to the unique situation of each country. They also sought to celebrate achievements as a way to inspire staff, volunteers and other stakeholders, for example, by stating that "the success of the education programme in demonstrating beneficiary level impact in [Country X] is extraordinary and it is motivating for staff and volunteers to be able to see the impact of their work."[18] Further recognition of country uniqueness was evident in the design of the performance ranges for the indicators:

> "Many of the KPIs have got ranges set against them to outline what 'high performance', 'satisfactory performance' and 'room for improvement' looks like. However, in some cases it will be more relevant for CDs and Regional Directors to decide what results say about the performance of the programme within the context of the programme itself…it is recognised that what can be considered high performance will differ considerably between Programme Offices."[19]

In this quote, there is explicit recognition of differences between countries that prevents the use of standardized performance ranges for each and every indicator. As such, eight of the indicators in the QF were scored with guidance that "performance [to be] determined by PO [programme office] and RD [regional director]". Finally, in contrast to the SRA, elements were not given explicit weights and there was no calculation of an overall score for each programme office.

There was also scope for constructive debate lasting beyond the QF's initial development. Country and regional directors were required to discuss the scoring for an individual programme office together, and to analyze and resolve differences in scores that emerged from this process. Furthermore, many staff from programme offices completed the QF together, providing a way to review overall results, and many regional directors used the QF to help set objectives and action plans for country directors in the coming year. Some programme offices embraced the QF even further, using it to determine whether an office move would improve programme quality, or further disaggregating the QF so it could be applied to different parts of the programme office.

[18] QF Country Report, 2008.
[19] QF Guidance document, 2008.

Collectively, the input of country directors and other programme staff, the importance of local knowledge in designing indicators, providing space for narrative so that programme offices could reflect local circumstances, and recognition that performance on some indicators was best determined using programme office and regional director judgement, provided explicit recognition of the uniqueness of programme offices. Critically, however, the need for comparability was also recognized. Each programme office was required to complete the QF using a common template, with common elements and indicators, thus providing a standardized method of measuring performance across countries.

After its first year of operation, praise for the QF was widespread, with this comment from a country director echoing that of many others:

> "Overall it was a good move away from the Annual Country Report because one of the main things was it gave much more direction on being clear on what to report on but also through the report it identified what is important, what's quality but it's also important to reflect on as a programme. Now you can always argue about elements of that, that's not the point. I think it's just helpful to say well these are overall important parts to reflect on and I thought that was quite useful." (Interview, Country Director 2, November 2008).

In this statement, the QF is seen as better than the ACR because it provides clarity around what makes a quality country programme, and, in this way, provided a 'collectively recognized' reference regarding the way in which programme offices would be evaluated (c.f., Jagd, 2007; Biggart & Beamish, 2003). Importantly, this quote also reveals that although there is recognition that the QF was and would be the site of disagreements (e.g., over elements), it was the overall approach of focusing on what made a quality programme that was most important. This corresponded to the view of the IPG Director, who also praised the QF:

> "I think it's been great. It's not a perfect tool but I don't think any tool in development ever is perfect… there wasn't a lot of discussion about quality or about success and the discussions were more about how many volunteers have we got or how much programme funding have we got and the quality framework has been a really useful tool over the last 18 months for just getting people to talk more about impacts on poverty. Quality, what is quality like?…[The QF has] given me stronger evidence when arguing at senior management team level for where things aren't working. So when you've got 35 country directors saying things like the finance systems aren't working it gives you a lot of evidence to be able to really argue for that…so from [that] basis, I think it's gone really well." (Interview, IPG Director, December 2008).

Here, praise is directed at how the QF helped move discussions more towards the impact of programmes on partners and beneficiaries and less on volunteer numbers or funding levels. The ability to aggregate data across programme offices was important in providing arguments for more resources at senior management forums.[20] The statements that the QF was not ‘a perfect tool’ but was ‘quite useful’ and ‘worked out pretty well’ reveal an awareness of the importance of ‘making do’ (c.f., Andon et al., 2007), which helped to enact a certain stability between a mode of evaluation that privileged country uniqueness and one favouring standardization and comparability. However, such stability was temporary, highlighting the fragility of the compromise. While there was initial praise for the QF from many sources, there were also critics. Shortly after the completion of the QF in its first year, it was subject to strong criticism, aimed in particular at the process used to score the indicators and elements.

But you’re not consistent!

The scoring process for the QF was based on self-assessment by the country director (with programme office staff input), with review by the relevant regional director. This raised concerns, particularly from some staff in London, that scoring was inconsistent across regions and countries:

> “(The) key anomaly is that the ratings seem to have been applied differently in each region…I believe this is an inaccurate reflection of the current strengths and weaknesses of programme funding across IPG…I suspect there are different interpretations as to what constitutes good programme funding performance…I think there is a need to clarify what justifies a 1, 2, 3 or 4 within each indicator.”[21]

This comment reveals that the scoring methodology was criticized for producing inaccurate results, with the problem being a dislike of the different interpretations of good performance made by different countries. The suggested solution was to instigate changes to the scoring procedure to clarify the meaning of each score.

A senior IPG manager was the most vocal critic of the scoring process, which he believed was “extremely dubious.” He lamented the SRA’s demise and concluded that shifting the balance in favour of self-assessment in the QF had created what he believed were questionable results. He expressed a preference for taking the scoring of indicators and elements out of the hands of country and regional directors

[20] The Director of IPG was a member of the six-person executive management team at VSO called the ‘Senior Management Team’ (other members were the Chief Executive Officer, Director of the VSO Federation, Director of the UK Federation, the Chief Financial Officer, and the Director of Human Resources). This group was responsible for major resource allocation decisions, particularly the amount of funds that were allocated to each of the major divisions within VSO, including IPG.

[21] QF Element Summary document, 2008.

altogether. He first floated the idea of using an external assessment process akin to an "OFSTED-type inspection unit."[22] Another option was the use of an internal assessment unit within VSO to carry out an "independent performance assessment." These preferences strongly emphasized the importance of having a 'consistent methodology' with 'independence', which, in effect, placed the values of standardization and comparability above those of local context and country uniqueness.

Although the use of an internal or external performance assessment unit did not materialize, the criticism resulted in several changes to the QF for its second year of operation. Each indicator now included a description of each of the 1-4 levels, where previously only levels 1 and 4 had a description. Revised guidance documentation was also issued:

> "It is important to score yourself precisely against the descriptors. There may be very good reasons why you achieve a low score on a particular indicator, but it is important to score precisely – the narrative can be used to give a brief explanation."[23]

This guidance highlights two important changes. First, there was the explicit emphasis on the need to score precisely, with the reasons that lay behind particular scores considered secondary. Second, the narrative was now viewed as the space where scores can be explained, indicating that its primary value was its connection to the scoring process, not in providing information that can arise from other sources. In further changes, the guidance "what can be considered high performance will differ considerably between Programme Offices" was removed from the QF documentation, the scoring of only one (rather than the previous eight) of the indicators was to be assessed using judgement,[24] and ownership of some scoring was taken away from programme offices, with the explanation that this would allow "data to be comparable across programmes by using universal interpretations of the data."[25]

Concerns over the scoring process itself were also addressed, particularly in relation to ensuring consistency in the way that regional directors used indicator

[22] OFSTED is the Office for Standards in Education, Children's Services and Skills in the UK, an independent body that inspects schools and provides ratings of performance. For example, schools are awarded an overall grade from 1 to 4, where 1 is outstanding, 2 is good, 3 is satisfactory and 4 is inadequate (OFSTED website, www.ofsted.gov.uk/, accessed 24 July 2010).

[23] QF Guidance document, 2009.

[24] As an example, indicator 8.1 on funding was changed whereby the ability to assess performance on a 'country by country' basis was replaced with explicit monetary ranges that would apply to each programme office regardless of its size of operations or different external funding environments.

[25] QF Reporting Template, 2009.

scores to determine overall element scores. A series of meetings was arranged to address this issue directly. One of these meetings, which lasted for over two hours, involved regional directors working through a recently completed QF report in order to agree on how to score each element. One by one, through each of the 14 elements, the process for using indicator scores to determine an overall element score was discussed. Looking fed-up, a regional director said:

> "Can I just ask, do we really care how accurately we score? [quizzical looks from other regional directors]. No, honestly, so we could spend a lot of time working out how we score it and use it for comparison but I mean you could roughly get a score on an average without spending too much time on the scoring but concentrate on what they're saying, and concentrate on quality discussion which presumably we also want to do." (Regional Director 5, QF meeting, May 2009).

The ensuing discussion did not focus on what programme offices were 'saying', or on how to 'concentrate on quality discussion.' Rather, debate focused on whether the 'average' was a legitimate way to determine element scores, and whether the most important indicator in each element should be designated as the 'lead' indicator for the purposes of scoring. This reveals how debate and disagreement was focused exclusively on the mechanics of the scoring process, rather than providing a forum for the consideration of the different substantive issues in play. Despite a protracted discussion, at the meeting's end there was no established process, except for general agreement that an overall element score "won't be based on an arithmetic average of KPI results for that element."[26] Notwithstanding this guidance, virtually all changes to the QF resulting from this criticism privileged consistency in scoring over that of country uniqueness. While more consistent scores were (arguably) likely, the initial compromise appeared tenuous and a counter-criticism developed around the lack of inspiration evident in the QF.

But you're not inspiring!

The focus on consistency and precision in the changes to the scoring process meant that the use of indicators to inspire was given minimal attention. Concerns were expressed in a QF meeting in March 2009:

> Regional Director 3: The messages that we're giving to programs at the moment about thinking creatively, being ambitious, being innovative and so on, are not necessarily captured in this element, in this thing here [points at the QF document]…I think this is really good for telling us where we're at and measuring

[26] QF Guidance document, 2009.

what we're looking at measuring but in terms of really looking to shift and change, I just wonder how we're going to do that, and where that's captured.
PLA[27] Staff 1: I think there's a bit that's still missing in the quality framework because it's become a set of indicators, so the bit that I think is missing is that we don't really have anything about culture.
Regional Director 3: Yeah, that's what it is, yeah [enthusiastically].
PLA Staff 1: If people fulfil all these indicators…that might not be enough to achieve what we're really looking for, you know…we've got fixed on the elements but there's something behind it all that we haven't quite nailed…

Here, a strong criticism of the QF is made by likening it to a 'set of indicators', with such a description generally considered to be a damning indictment of any evaluation practice at VSO. Furthermore, concern is expressed that indicators cannot capture what is most valued, that is, the desire to 'shift and change', 'achieve what we're really looking for' and 'culture' are 'missing' in the QF. This concern was coupled with critical feedback on other changes made to the QF, for example, the emphasis on standardization meant that the indicators no longer captured performance 'accurately.' Reflecting general concern with the changes, this statement appeared in one country programme's QF report:

"The QF report has grown exponentially this year, and the indicators have changed and were only issued a month before the report is due…Good practice in monitoring and evaluation is to collect evidence and learning on a day-to-day basis, which is difficult to do if the ground keeps shifting under one's feet."[28]

Increasingly, discussions of the QF were focused on criticisms and counter-criticisms over the specific details of the scoring process. A regional director summarized the state-of-play after completion of the QF for the second year:

"People can see that we've tried to make it a little bit more objective in the way that it's done, but I am getting quite a lot of critical feedback that the quality framework is so big, so many indicators, stuff being sent really late…the whole thing is just a quantitative scoring tool and it's not about learning in any way, shape or form…so I am getting quite critical feedback" (Regional Director 5, QF meeting, May 2009).

In this quote, the focus on indicators and scoring processes is viewed as stifling opportunities for learning. Thus, an initial compromise between standardization through scoring and recognition of country uniqueness had faltered. The initial praise for the QF had dissipated and was replaced by critical feedback, particularly from country directors who felt that the push for consistency had moved too far such that the QF was no longer about learning and instead was labelled as a 'quantitative

[27] The PLA (Programme Learning and Advocacy) unit was a team within VSO whose main role was to support programme offices in learning from their own work and sharing good practice with other programme offices.
[28] QF Country Report, 2009.

scoring tool', a severe condemnation at VSO. We also see that the debates about indicators, performance ranges, or scoring methodologies were increasingly focused on the QF itself, often in the context of long meetings with little productive output. The previously fruitful discussions about how to improve quality or make compromises between differing values were almost non-existent. These developments were also evident in initial debates about how to use the QF to improve programme office performance.

Debate 2: How to improve quality?

Given the considerable effort that had gone into the development of the QF, there were high hopes that it would lead to improved performance of programme offices. On the one hand, there was a strong desire to improve through 'learning', whereby the QF would help identify examples of innovative practice that could then be shared amongst programme offices. Concurrently, there was a belief that the QF could be used to generate a sense of 'competition' amongst programme offices, which would lead to increased motivation and thus improved performance. These different positions on the avenues to improved performance presented many obstacles to enacting an acceptable compromise; obstacles that, at first, proved difficult to overcome.

Although not stated explicitly in QF documentation, the engendering of a sense of competition emerged through the way in which the results were distributed to country directors via email:

> Country Director 9: "We were sent back a kind of world-wide kind of scoring sheet and obviously all that that had was a series of numbers. Sort of 1, 2, 3 filled with red and green and orange. Although to be honest we came second in the world…I feel quite sorry for the countries that have scored quite low because I really don't think it's a valid scoring system…but for me it was quite handy to be able to say this and say "maybe next year- first" and all the rest."
> "How do you know you are second in the world? Was there some kind of ranking?"
> Country Director 9: "Yeah, they sent us a ranking. They sent us a worldwide ranking thing afterwards."
> "And so countries were rank ordered from 1 to 34?"
> Country Director 9: "Yeah it was a summary of results with ranking. And you could look against different scores so you could see that…globally you came second on something and third on something else but then there was an overall sort of score."
> (Interview, Country Director 9, November 2008)

Despite reference to an 'overall sort of score', the spreadsheet did not contain a summary score for each country and countries were not ranked, but listed in alphabetical order. As such, Country Director 9's country only appeared 'second in the world' by virtue of its name beginning with a letter near the beginning of the alphabet. Other similar stories emerged of countries with names that started with letters towards the end of the alphabet believing that they had performed poorly. These examples were the source of much joking at IPG staff meetings in London, with suggestions that 'Albania' will be top of the league table next year, and that the solution was to put countries in reverse alphabetical order. This light-heartedness about rankings belied an appreciation of how aware country directors were of the competitive mantra that lay behind the spreadsheet's distribution, and how this was viewed as stifling learning opportunities, as one country director commented:

> "When this whole [QF] thing was being started, some of the conversations were framed around what are the indicators of quality in a programme office that is doing well. How do we assess whether Ghana is better than Zimbabwe or vice versa? So I think the framing of the conversations around that time kind of planted the seeds of a league table…as long as people continue to see it [the QF] as a league table then we might see each other as competitors and therefore everybody [will keep] what he or she is doing very close to their chest" (Interview, Country Director 6, November 2008).

Importantly, several features of the spreadsheet served to reinforce the 'league table mentality' noted by the country director. First, it 'was a series of numbers' and did not contain any of the narrative discussion. Second, only the overall element scores were displayed without the specific indicator scores. Third, each element score of 1 to 4 was assigned a colour to make differences between scores in the spreadsheet visually distinct, with 'Low' performers particularly prominent as scores of '1' were assigned the colour red. This led regional directors to question the use of scores to promote learning, suggesting that it was their role, rather than that of a spreadsheet, to direct country directors to examples of good practice. More fundamentally, however, not only was comparison of countries in a spreadsheet not considered helpful for sharing best practice, but the uniqueness of each country also made such comparisons 'unfair':

> Regional Director 1: "I don't see the value of knowing that, for example, on maybe even six of the twelve criteria, West Africa comes out worse than say South-East Asia because my interpretation instinctively would be what are the cultural, educational, historical background, you know, accumulation of circumstances in South-East Asia that means that they're in a completely different environment."
> Regional Director 4: "It [a league table] makes it [comparisons] into a competition essentially."

PLA Staff 1: "Yeah, but it's an unfair competition…It's like getting Marks & Spencers compared with Pete's Café across the road where you've got totally different contexts."[29]
(QF meeting, March 2009)

Claims of unfairness speak to tensions arising because the principle that country uniqueness is important was not being respected. The process of reducing performance on an element to a standardized metric was seen by country directors to have ensured that the contextual information required to understand these scores had been stripped away. Even when this information was present in the narrative section, it did not accompany the scores in the spreadsheet and was thus seemingly ignored (or considered too difficult to take into account). In this way, the ideals embodied in different modes of evaluation did not co-exist as the values of competition had in effect 'swamped' the values of learning and country uniqueness.

This situation was also problematic because improving quality through learning was an explicit ambition of the QF, with one of its stated purposes "to help identify strengths and areas for improvement across programmes so that steps can be taken **to grow programme quality** and foster learning" (emphasis in original).[30] Country directors, however, felt that there was no formal mechanism for sharing learning between programme offices. Consequently, the view amongst country directors was that learning from the QF had been given a relatively low priority, and its potential for sharing good practice went unfulfilled. Seeking to reorder these priorities was an important debate in preparing the QF for its second year.

Reordering priorities of learning and competition

Over a series of three meetings, IPG staff debated ranking programme offices using QF data. This excerpt is from the first meeting in March 2009:

Regional Director 3: "You don't have to send it [the scores] out in a table to everybody but you can go and look and say 'right, ok, this country over here does really good volunteer engagement, why don't we arrange some sort of visit or some sort of support from that', that would absolutely make sense but to send something out and say look for the [countries that have scored] fours and talk to them doesn't."
Regional Director 2: […] "I think we all agree there's a reason to link people up according to where there is good practice, or good performance and there are ways to do that that aren't about a published table. So, who's for a published table, who's against a published table, who's for a published table at this stage?"

[29] Marks and Spencers is a large UK department store whose annual revenue in the financial year 2009/10 was £9.5 billion. Pete's Café was a small café immediately opposite the VSO building in London.
[30] QF Guidance documents, 2008 and 2009.

[General laughter]
Regional Director 4: "Obviously we [regional directors] are discouraging".
Director, IPG: "I would reflect on it a bit further…I'm not sure how helpful it is when the table only had the element scoring and it's very subjective…so I think work out if anyone else found it useful…I'm thinking all of this is worth picking up again in that next meeting."

Here, regional directors were generally against the idea of a spreadsheet being the vehicle for identifying good practice, instead arguing that they should take an active role in linking up programme offices to help improve performance. In particular, there was a strong argument against using scores of four (the highest score possible on an element) to identify what is considered good practice. The IPG Director believed that using the QF to create competition was sound, but could see weaknesses in the mechanics of the league table ('only…element scoring', 'very subjective'). As such, he was not yet convinced of the league table's apparent inappropriateness and stalled any decision to the next meeting.

Convened in May 2009, the next QF meeting was focused on convincing the IPG Director (not in attendance) that a ranking was not appropriate:

Regional Director 3: "I'm just wondering what the reason for having a league table is."
PLA Staff 1: "I think it's the idea you publish information and then people will be shamed, people will feel they got a low performance, they will feel forced to have to make improvement because it's public."
PLA Staff 3: "There is a real danger of labelling them [programme offices], isn't there? That's what's really horrible about this because someone then gets labelled as being the office that's rubbish at volunteer engagement or the one that's great at such and such."
PLA Staff 1: "Yeah, yeah I agree, yeah. The reason I really don't like it, I don't see how an organization's [that's] about volunteering and is very personal how that…sort of…philosophy could really fit with this [league table], but the other thing is I think it will change the quality framework from being a learning tool… my real fear is if you publish the scores people get fixated on doing well on particular indicators, which we're now saying aren't good enough, rather than the spirit of trying to actually improve…so I think it's a combination of philosophy in terms of what VSO is about but also, you know, keeping the quality framework as something that is a learning tool."

In contrast to the arguments used in the first meeting, this criticism was more fundamental, in that it directly criticized the very principles upon which the spreadsheet and (apparent) ranking system was based. Here, the use of competition to 'label' and 'shame' programme offices into improvements was viewed as 'horrible' and the league table considered incompatible with the purpose of the QF as a learning tool. Finally, and perhaps most telling, is that ranking programme offices was viewed as being against the ideals of 'volunteering' and personal engagement that are considered critical to VSO's philosophy, as expressed above by PLA Staff 1. In June

2009, a third and final meeting to debate the league table issue was convened. The above arguments were used in this QF meeting with the IPG Director, and a compromise agreed:

> Director, IPG: "All right, let's do it a different way…let's ask each element leader to highlight confidentially where they think there are real concerns."
> PLA Staff 2: "So, ok, that's fine, then what? What happens to that information?"
> Director, IPG: "So basically the element leaders are informing discussions about where we might prioritize…so the top three is highlighting good practice, giving an indication to countries across the world where they might want to talk to in terms of good practice, and the bottom one is just confidential for management purposes."

Importantly, the earlier appeals by the IPG Director to improve the mechanics of the league table were no match for arguments undermining the very principle upon which it was based. As such, the compromise between competition and learning, between comparisons and sharing good practice, was resolved by abolishing the league table and replacing it with a new practice of differential disclosure. That is, the identity of good performers would be made public whereas the identity of poor performers would be kept confidential. Disclosure of good performers would allow the sharing of good practice between programme offices, and disclosure of poor performers to the IPG Director would allow management action to be taken but without 'naming and shaming' programme offices in the process. It is here that debates about the league table facilitated productive discussion between those who viewed 'competition' as the route to improvement versus those who saw learning as the way to increase quality. Unlike the debates over consistency in scoring, discussion was not focused on the QF *per se*, but was connected to broader principles, such as uniqueness and innovation, competition and a volunteering ethos. In this way, principled argument led to a compromise between different evaluative principles, despite strong enthusiasm for the spreadsheet and ranking system to remain.

Epilogue

Towards the end of the field study, a review of existing "Quality Initiatives at VSO" was conducted, including the QF. While analysis of the QF was generally favourable, numerous "areas of improvement" were suggested:

> Its holistic and coherent nature allows people to think more broadly and reflect on the progress of the whole programme…The process of doing the report makes people take stock, consider areas of improvement and make action plans accordingly….It seems that the QF is not referred to or used as often as people would like…The numbers are not useful because they are too mechanistic, yet subjective and inconsistent across POs… [Self-assessment] is a great way for the PO to take stock and think about their performance and how to make improvements. But

> many people feel that this needs some kind of external support and verification… although the design of the framework is quick and simple…It has become too long and its design means that indicators are 'set in stone' to a certain degree in order to make comparisons from one year to the next.[31]

This analysis reveals that the QF enabled broad thinking but was not used enough; self-assessment helped to 'take stock' but needed external verification; and the QF was simple, yet too long. We see that positive features of the QF that were closely aligned to one mode of evaluation inevitably gave rise to suggestions for improvement that sought to address the concerns of those with different evaluative principles. The evaluation highlights how the compromises being made in the design and operation of the QF were not 'resolved' but formed a series of temporary settlements (c.f., Stark, 2009; Kaplan & Murray, 2010; Gehman et al., 2013) between different evaluative principles. In this way, the process of establishing and maintaining compromises between different modes of evaluation can be seen as a dynamic and enduring feature of a compromising account.

Discussion

Taking tensions between different logics and values as the starting point for our analysis, this study has focused directly on how accounting is implicated in compromising between different evaluative principles and the way in which such compromise can be productive or unproductive. Accounts are particularly important in settings of conflicting values because they are sites where multiple modes of evaluation all potentially operate at once (Stark, 2009). Our field study shows how VSO's attempts to measure the performance of its programme offices brought together differing modes of evaluation, one based primarily on 'Learning and Uniqueness' and the other based primarily on 'Consistency and Competition', where each mode of evaluation was distinguished according to its purpose, the desirable attributes of a good evaluation and subsequently the desirable attributes of a good account (see Table 2). Making choices about indicators, types of scoring processes, the identification of good and poor performers, and different methods of data analysis, created sites for debate between individuals and groups who espoused these different evaluative principles (Stark; 2009; Jay, 2013; Gehman et al., 2013; Moor & Lury, 2011; Denis et al., 2007). In this way, our analysis reveals how an account itself can

[31] Review of Quality Initiatives at VSO document, 2010.

act as an agent in the process of compromise between different evaluative principles. A compromising account is thus both the process of, and at particular moments the specific outcome of, a temporary settlement between different modes of evaluation. Analogous to Chua's (2007) discussion of strategizing and accounting, this draws attention to a compromising account as both a noun, i.e., the account itself that is produced in some material form (e.g., a balanced scorecard, a financial report), and as a verb, i.e., the processes of compromise that lead to and follow on from the physical production of an account.

Our study shows that differences in the design and operation of accounting practices can affect the extent of compromise between different evaluative principles, and whether such compromise is productive or unproductive. In particular, our findings reveal that the potential for accounts to provide a fertile arena for productive debate is related to three important processes: (1) imperfection – the extent to which the design and operation of accounting practices represents a 'give and take' between different evaluative principles; (2) concurrent visibility – the way in which desirable attributes of accounts are made visible in the design and/or operation of the accounting practice; and (3) the extent to which the discussions concerning potential problems with the accounting practice are focused on underlying evaluative principles (vs. mechanics/technical considerations). In the discussion below we elaborate the characteristics of these processes, and then conclude the paper by outlining the implications for future research and highlighting the insights for practice.

'Imperfection' and the potential for 'productive friction'

In organizational settings with multiple and potentially competing evaluative principles, the development of compromises reflects a temporary agreement (Stark, 2009; Kaplan & Murray, 2010; Gehman et al., 2013). In this setting, rather than reaching closure, the development and operation of compromising accounts entails on-going adjustment (c.f., Gehman et al., 2013). This was clearly evident in VSO's QF, where it was subject to on-going criticism and refinement and was 'loved by no-one.' We suggest that it is the 'imperfect' nature of the QF that was pivotal to its continued existence as a compromising account. We see that the constant shifting and rebalancing in the QF's design and operation enabled the co-existence, albeit often temporary, of different modes of evaluation. Changes privileging one mode of evaluation, such as a focus on a more rigorous and consistent scoring process, were accompanied by changes that shifted the emphasis back to another mode of evaluation, such as ensuring the analysis of QF data included a pairing of numbers with narrative. It was this 'give and take' between different modes that helped to resist pressures for recourse to a single and therefore ultimately dominant mode of evaluation (c.f., Thévenot, 2001), and enabled productive friction to arise from the coming together of different evaluative principles. In this way, compromises involving multiple evaluative principles are inherently 'imperfect' when enacted in practice (c.f. Annisette & Richardson, 2011).

We see our findings in this regard as having parallels with recent literature on the 'imperfect' nature of performance measures (see, for example, Andon et al., 2007; Dambrin & Robson, 2011; Jordan & Messner, 2012). These studies often stress the importance of organizational actors 'making do' with the existing performance measurement system, despite its perceived imperfections. For example, Bürkland, Mouritsen & Loova (2010) show how actors compensate for 'imperfect' performance measures by using other information, while Jordan & Messner (2012) find that actors respond to incomplete performance measures in two ways; by trying to repair them or by distancing themselves from the measures. However, in our study we find that rather than organizational actors merely 'making do' with imperfect performance measures, it was these 'imperfections' that helped to provide a fertile arena for productive dialogue and discussion between individuals and groups with differing values (c.f., Stark, 2009; Jay, 2013; Gehman et al., 2013; Moor & Lury, 2011; Denis et al., 2007). In this way accounts can play a role in surfacing latent paradoxes and providing space to work out ways to combine different evaluative principles (Jay, 2013). The struggles between different evaluative criteria can prompt those involved to engage in deliberate consideration about the merits of existing practices (Gehman et al., 2013). Here we see the importance of the accommodation of different perspectives and recognition by actors that the proposed solution (in our case the QF), although not perfect, provides a fitting answer to a problem of common interest (c.f. Huault & Rainelli-Weiss, 2011; Samiolo, 2012).

'Imperfect' accounts, such as VSO's QF, are therefore not just about 'making do', but can create opportunities for bringing together competing value systems and, thus, the potential for what Stark (2009: 19) terms 'productive friction.' This was most evident in the league table debates, where discussions between actors with different evaluative principles led to changes in the use of spreadsheets and element summaries that recognized a reordering of the priorities between learning and competition. Here we see the role of compromising accounts as creating a form of organized dissonance, that is, the tension that can result from the combination of two (at least partially) inconsistent modes of evaluation. A compromising account can thus be a vehicle through which dialogue, debate and productive friction is produced, where it is the discussion that can result from having to compromise on the design and operation of an account that can be productive.

Concurrent visibility

But how does a compromising account enable organised dissonance? Our study indicates that an important feature of a compromising account is that of 'concurrent visibility.' To facilitate organized dissonance it was critical that the QF made visible the features of an account that were important to different groups. We use the term 'visible' in a broad sense to refer to how the design and operation of a compromising account reveals the attributes of accounts that are important to organizational actors with different evaluative principles. For example, in the physical format of the QF, indicators were accompanied by narrative boxes, which enabled compromise between the evaluative principles of standardization and country uniqueness. In addition, the differential disclosure of good and poor performing countries (post the league table) facilitated compromise between the evaluative principles of learning and competition. The concurrent use of these different features gave visibility to the importance of different modes of evaluation. This resonates with Nahapiet (1988), where the resource allocation formula helped to make values more visible and tangible and prompted explicit consideration of three fundamental organizational dilemmas. More generally, it resonates with the way in which instruments like accounting and performance measurement systems are well suited to rendering visible the multiplicity of criteria of evaluation (Lamont, 2012).

We suggest that where the co-existence of different evaluative principles is an on-going feature of organizations, organizational actors are likely to be particularly concerned that their fundamental principles may not be respected and thus come to be dominated by others (c.f., Denis et al., 2007). It is here that 'concurrent visibility' in a compromising account can provide confirmation and reassurance that a particular mode of evaluation is, indeed, recognized and respected, thus making productive debate more likely. The visibility of different evaluative principles in the account also serves to crystallize the compromise between them in a material form (c.f., Denis et al., 2007).

The importance of concurrent visibility is evident by contrasting the views of the QF at the end of the first and second years of operation. The features of the QF during its first year of operation (narrative, local knowledge, judgement, common elements and indicators) gave explicit recognition to different evaluative principles and thus helped to develop a compromise between values of standardization and country uniqueness. In contrast, changes to make the QF more consistent removed many of the features that recognized country uniqueness as an important evaluative principle. Subsequently, the initial praise for the QF had dissipated and was replaced by 'endless' disagreements and critical feedback, which resulted in a situation where actors were 'stuck' between different evaluative principles (Jay, 2013).

Our study also reveals, however, that there are limits to the way in which concurrent visibility can facilitate organized dissonance, particularly where the strategy is 'additive.' That is, to address the evaluative principles favoured by different organizational actors, the account can simply encompass more and more of those desired features. Over time, however, the account is likely to become cumbersome and unwieldy, as we saw with the QF when, at the end of its second year of operation, it was described as 'so big, so many indicators.' As such, without careful attention, concurrent visibility could potentially be directed towards the appeasement of different modes of evaluation rather than serving as a necessary entry point for productive discussion over the merits of different evaluative principles.

Criticisms of accounts and breakdowns in compromise

Our study also highlights an important distinction between the types of responses that can emerge in situations where compromises break down and accounting practices are viewed as 'not working.' One criticism of the QF concerned the presentation of scores in a spreadsheet and the subsequent illusion of a league table ranking of countries according to their overall performance. Such a practice was viewed as privileging the value of 'competition' above that of 'learning' and was thus primarily a debate about the principles and values underlying the use and operation of the league table (c.f., Gehman et al., 2013). Here, there was a passionate response from those actors who felt that a fundamental principle was not being respected (Denis et al., 2007), particularly that the league table ignored their belief that the performance and hence value of country programmes was ‘incommensurable’ (Espeland & Stevens, 1998). This debate was not about how to ‘fix’ the league table per se but focused on whether the league table itself was an appropriate practice – revealing a situation where actors reflect at a distance on the values underlying the existing practice (Sandberg & Tsoukas, 2011; Gehman et al., 2013). This helped the actors to confront the latent paradoxes (Jay, 2013) evident in the use of a league table and facilitated ‘productive friction’ between those who viewed ‘competition’ as the route to improvement versus those who saw learning as the way to increase quality. As a result, a new practice was developed (i.e., differential disclosure of good and bad performers) that helped to integrate different evaluative principles in a more substantive way (Stark, 2009; Jay, 2013).

Another criticism of the QF was directed at its lack of consistency and thus inability to enable meaningful comparisons of country performance. This was primarily a criticism of the implementation of the QF’s scoring process, where discussion focused on what was problematic about the current practice and how to fix it (c.f., Sandberg & Tsoukas, 2011; Gehman et al., 2013) and not on whether scoring itself was an issue of concern. As such, subsequent changes to the QF focused on removing features of the existing scoring process that were seen not to align with the value of consistency, and adding features viewed as promoting consistency. Such changes clearly shifted the scoring process of the QF in favour of those organizational actors who held consistency in scoring as an essential feature of an evaluation process. Rather than integrating different perspectives, however, this response can be characterised by oscillation and ‘stuckness’ (Jay, 2013) between the evaluative principles of consistency and country uniqueness. Furthermore, as these debates were primarily focused on technicalities, they took up valuable meeting time that effectively preventing meaningful engagement (i.e., ‘productive friction’) with the underlying principles. This resonates with Stark’s (2009) warning that disputes over the mechanics of existing practices can limit effective changes and result in endless disagreements where nothing is accomplished.

Conclusion

Our study has highlighted the importance of examining the role of accounting in facilitating (or not) compromises in situations of multiple evaluative principles. Our results indicate that much can be learned by focusing on how accounts can potentially bring together differing (and often competing) evaluative principles, where such encounters can generate productive friction, or lead to the refinement of accounting practices and 'endless' debate and discussion over technicalities and the mechanics of the account. We view accounts as central to processes of compromise in organizations because it is in discussions over the design and operation of accounts that the worth of things is frequently contested by organizational actors. Drawing on Stark's (2009) concept of organizing dissonance, our study shows that there is much scope for future research to examine how accounts can create sites that bring together (or indeed push apart) organizational actors with different evaluative principles, and the ways in which this 'coming together' can be potentially constructive and/or destructive.

Our analysis also has implications for the ways in which performance measures and other accounting information can be mobilized by managers and practitioners as a resource for action (c.f., Ahrens & Chapman, 2007; Hall, 2010). In particular, our results indicate that 'imperfect' performance measures can actually be helpful, that is, they can be used by practitioners to generate productive dialogue, despite, or, as our analysis shows, because of, their perceived imperfections. This resonates with Stark (2009), who argues that entrepreneurship is the ability to keep multiple evaluative principles in play and exploit the friction that results from their interplay. Here, the 'imperfect' nature of compromising accounts can enable skilful organizational actors to keep multiple evaluative principles in play. In contrast, a focus on the continual refinement of accounts and a quest for 'perfection' can lead to the domination of a single evaluative principle, 'distancing' organizational actors who hold different evaluative principles, and limiting opportunities for productive friction.

A further implication of our study is to promote further research on how performance measurement systems, and accounting practices more broadly, are actually developed in organizations (c.f., Wouters & Wilderom, 2008). In particular, we analyzed the different responses that can occur when compromises breakdown and how they relate to the potential for productive friction. More broadly, it is unclear how organizational actors negotiate the development of performance indicators and what types of responses and arguments prove (un)successful in these encounters. This could prove a fruitful area for future research.

We conclude by outlining the practical implications of our study, which centre on imperfection and concurrent visibility. Although practitioners are no doubt aware of the need to 'make do' with the perceived inadequacies of performance measures, our study indicates that the productive discourse arising from performance measurement is perhaps more important than ensuring that such measures (or accounts more generally) are 'complete.' Our analysis of concurrent visibility indicates that practitioners should ensure that features of accounts that are of fundamental importance to particular groups are explicitly recognized, whether in the material content of the account, the associated scoring and evaluation processes, or in its use in wider organizational practices.

References

Ahrens, T., & Chapman, C.S. (2002). The structuration of legitimate performance measures and management: day to day contests of accountability in a U.K. restaurant chain. *Management Accounting Research 13*, 151-171.

Ahrens, T., & Chapman, C.S. (2004). Accounting for flexibility and efficiency: A field study of management control systems in a restaurant chain. *Contemporary Accounting Research 21*, 271-301.

Ahrens, T., & Chapman, C.S. (2006). Doing qualitative field research in management accounting: Positioning data to contribute to theory. *Accounting, Organizations and Society 31(8)*, 819-841.

Ahrens, T., & Chapman, C.S. (2007). Management accounting as practice. *Accounting, Organizations and Society 32(1-2)*, 1-27.

Andon, P., Baxter, J., & Chua, W.F. (2007). Accounting change as relational drifting: A field study of experiments with performance measurement. *Management Accounting Research 18*, 273-308.

Annisette, M., & Richardson, A.J. (2011). Justification and accounting: applying sociology of worth to accounting research. *Accounting, Auditing and Accountability Journal 24*, 229-249.

Annisette, M., & Trivedi, V.U. (2013). Globalisation, paradox and the (un)making of identities: Immigrant Chartered Accountants of India in Canada. *Accounting, Organizations and Society 38(1)*, 1-29.

Baines, A., & Langfield-Smith, K. (2003). Antecedents to management accounting change: a structural equation approach. *Accounting, Organizations and Society 28(7-8)*, 675-698.

Biggard, N. W., & Beamish, T. D. (2003). The economic sociology of conventions: habit, custom, practice, and routine in market order. *Annual Review of Sociology 29*, 443-464.

Bird, D. (1998). *Never the same again: A history of VSO*. Cambridge: Lutterworth Press.

Boltanski, L., & Thévenot, L. (1999). The sociology of critical capacity. *European Journal of Social Theory 2(3)*, 359-377.

Boltanski, L., & Thévenot, L. (2006). *On justification. The economies of worth*. (transl. C. Miller). Princeton: Princeton University Press.

Briers, M., & Chua, W.F. (2001). The role of actor-networks and boundary objects in management accounting change: a field study of an implementation of activity-based costing. *Accounting, Organizations and Society 26(3)*, 237-269.

Bürkland, S., Mouritsen, J., & Loova, R. (2010). Difficulties of translation: making action at a distance work in ERP system implementation. *Working paper*.

Cavalluzzo, K.S. & Ittner, C.D. (2004). Implementing performance measurement innovations: evidence from government. *Accounting, Organizations and Society, 29(3–4)*, 243-267.

Chahed, Y. (2010). Reporting beyond the numbers: The reconfiguring of accounting as economic narrative in accounting policy reform in the UK. *Working paper*.

Chenhall, R.H., Hall, M., & Smith, D. (2010). Social capital and management control systems: a case study of a non-government organization. *Accounting, Organizations and Society 35(8)*, 737-756.

Chua, W.F. (2007). Accounting, measuring, reporting and strategizing – Re-using verbs: A review essay. *Accounting, Organizations and Society 32(4-5)*, 487-494.

Cooper, D.J., Hinings, B., Greenwood, R., & Brown, J.L. (1996). Sedimentation and transformation in organizational change: the case of Canadian law firms. *Organization Studies 17*, 623-647.

Dambrin, C., & Robson, K. (2011). Tracing performance in the pharmaceutical industry: ambivalence, opacity, and the performativity of flawed measures. *Accounting, Organizations and Society 36(7), 428-455.*

Dart, J., & Davies, R. (2003). A dialogical, story-based evaluation tool: the Most Significant Change technique. *American Journal of Evaluation 24*, 137-155.

Davies, R. & Dart, J. (2005). *The 'Most Significant Change' (MSC) Technique: A Guide to Its Use*. www.mande.co.uk/docs/MSCGuide.pdf, accessed 20 September 2011.

Denis, J-L., Langley, A., & Rouleau, L. (2007). Strategizing in pluralistic contexts: rethinking theoretical frames. *Human Relations 60(1)*, 179-215.

Dent, J.F. (1991). Accounting and organizational cultures: A field study of the emergence of a new organizational reality. *Accounting, Organizations and Society 16(8)*, 705-732.

Department for International Development. (2000). *Eliminating world poverty: Making globalisation work for the poor.*

Department for International Development. (2006). *Eliminating world poverty: Making governance work for the poor.*

Department for International Development. (2009). *Eliminating world poverty: Building our common future.*

Eisenhardt, K.M. (1989). Building theories from case study research. *Academy of Management Review 14(4)*, 532-550.

Espeland, W.N., & Stevens, M. (1998). Commensuration as a social process. *Annual Review of Sociology 24*, 312-343.

Ezzamel, M., Willmott, H., & Worthington, F. (2008). Manufacturing shareholder value: The role of accounting in organizational transformation. *Accounting, Organizations and Society 33(2-3)*, 107-140.

Fischer, M.D., & Ferlie, E. (2013). Resisting hybridisation between modes of clinical risk management: Contradiction, contest, and the production of intractable conflict. *Accounting, Organizations and Society 38(1)*, 30-49.

Free, C.W. (2008). Walking the talk? Supply chain accounting and trust among UK supermarkets and suppliers. *Accounting, Organizations and Society 33(6)*, 629-662.

Garud, R. (2008). Conference as venues for the configuration of emerging organizational fields: the case of cochlear implants. *Journal of Management Studies 45*, 1061-1088.

Gehman, J., Trevino, L., & Garud, R. (2013). Values work: a process study of the emergence and performance of organizational values. *Academy of Management Journal 56(1)*, 84-112.

Gendron, Y. (2002). On the role of the organization in auditors' client-acceptance decisions. *Accounting, Organizations and Society* 27, 659-684.

Gibbs, M., Merchant, K.A., Van der Stede, W.A., & Vargus, M.E. (2004). Determinants and effects of subjectivity in incentives. *The Accounting Review* (April), 409-436.

Hall, M.R. (2008). The effect of comprehensive performance measurement systems on role clarity, psychological empowerment and managerial performance. *Accounting, Organizations and Society 33(2-3)*, 141-163.

Hall, M.R. (2010). Accounting information and managerial work. *Accounting, Organizations and Society 35(3)*, 301-315.

Helmig, B., Jegers, M. & Lapsley, I. (2004). Challenges in managing nonprofit organizations: a research overview. *Voluntas: International Journal of Voluntary and Nonprofit Organizations 15*, 101-116.

Hopgood, S. (2006). *Keepers of the flame: Understanding Amnesty International*. Cornell University Press, Ithaca.

Huault, I., & Rainelli-Weiss, H. (2011). A market for weather risk? Conflicting metrics, attempts at compromise, and limits to commensuration. *Organization Studies 32(10)*, 1395-1419.

Jagd, S. (2007). Economics of convention and new economic sociology: Mutual inspiration and dialogue. *Current Sociology 55(1)*, 75-91.

Jagd, S. (2011). Pragmatic sociology and competing orders of worth in organizations. *European Journal of Social Theory 14(3)*, 343-359.

Jay, J. (2013). Navigating paradox as a mechanism of change and innovation in hybrid organizations. *Academy of Management Journal 56(1)*, 137-159.

Jordan, S., & Messner, M. (2012). Enabling control and the problem of incomplete performance indicators. *Accounting, Organizations and Society 37(8)*, 544-564.

Kaplan, S., & Murray, F. (2010). Entrepreneurship and the construction of value in biotechnology. In Phillips, N., Sewell, G., & Griffiths, D. *Technology and Organization: Essays in Honour of Joan Woodward (Research in the Sociology of Organizations, Volume 29)*, Emerald Group Publishing Limited, 107-147.

Kaplan, R. S. & Norton, D.P. (1992). The balanced scorecard - Measures that drive performance. *Harvard Business Review* (January-February), 71-79.

Lamont, M. (2012). Toward a comparative sociology of valuation and evaluation. *Annual Review of Sociology 38*, 201-221.

Lounsbury, M. (2008). Institutional rationality and practice variation: new directions in the institutional analysis of practice. *Accounting, Organizations and Society 33*, 349-361.

McInerney, P-B. (2008) Showdown at Kykuit: field-configuring events as loci for conventionalizing accounts. *Journal of Management Studies 45*, 1089–1116.

Moers, F. (2005). Discretion and bias in performance evaluation: The impact of diversity and subjectivity. *Accounting, Organizations and Society 30(1)*, 67-80.

Moor, L., & Lury, C. (2011). Making and measuring value. *Journal of Cultural Economy 4*, 439-454.

Nahapiet, J. (1988). The rhetoric and reality of an accounting change: a study of resource allocation. *Accounting, Organizations and Society 13*, 333-358.

Nicholls, A. (2009). We do good things, don't we? 'Blended value accounting' in social entrepreneurship. *Accounting, Organizations and Society 34*, 755-769.

Oakes, L.S., Townley, B. & Cooper, D.J. (1998). Business planning as pedagogy: language and control in a changing institutional field. *Administrative Science Quarterly 43*, 257-292.

Parsons, E. & Broadbridge, A. (2004). Managing change in nonprofit organizations: insights from the UK charity retail sector. *Voluntas: International Journal of Voluntary and Nonprofit Organizations 15*, 227-242.

Perera, S., Harrison, G., & Poole, M. (1997). Customer-focused manufacturing strategy and the use of operations-based non-financial performance measures: A research note. *Accounting, Organizations and Society 22(6),* 557-572.

Porter, T. (1995) *Trust in numbers: The pursuit of objectivity in science and public life*. Princeton, NJ: Princeton University Press.

Robson, K. (1992) Accounting numbers as "inscription": action at a distance and the development of accounting. *Accounting, Organizations & Society 17*, 685-708.

Samiolo, R. (2012). Commensuration and styles of reasoning: Venice, cost–benefit, and the defence of place. *Accounting, Organizations and Society 37(6),* 382-402..

Sandberg, J., & Tsoukas, H. (2011). Grasping the logic of practice: theorizing through practical rationality. *Academy of Management Review 36*, 338-360.

Scott, S. V., & Orlikowski, W. J. (2012). Reconfiguring relations of accountability: materialization of social media in the travel sector. *Accounting, Organizations and Society 37*, 26-40.

Spradley, J.P. (1980). *Participant observation*. New York: Holt, Rinehart and Winston.

Stark, D. (1996). Recombinant property in east European capitalism. *American Journal of Sociology 101*, 993-1027.

Stark, D. (2009). *The sense of dissonance: accounts of worth in economic life.* Princeton: Princeton University Press.

Sundin, H.J., Granlund, M. & Brown, D.A. (2010). Balancing multiple competing objectives with a balanced scorecard. *European Accounting Review 19(2),* 203-246.

Thévenot, L. (2001). Organized complexity: conventions of coordination and the composition of economic arrangements. *European Journal of Social Theory 4(4)*, 317-330.

Townley, B., Cooper, D., & Oakes, L. (2003). Performance measures and the rationalization of organizations. *Organization Studies 24(7)*, 1045-1071.

United Nations. (2011). The Millennium Development Goals Report 2011. www.un.org/millenniumgoals/pdf/(2011_E)%20MDG%20Report%202011_Book%20LR.pdf, accessed 2 August 2011.

Vollmer, H. (2007). How to do more with numbers: Elementary stakes, framing, keying, and the three-dimensional character of numerical signs. *Accounting, Organizations and Society 32*, 577-600.

Voluntary Services Overseas. (2004). *Focus for Change: VSO's Strategic Plan*. www.vsointernational.org/Images/Focus_for_Change_tcm76-20678.pdf, accessed 10 August 2010.

Wouters, M. & Roijmans, D. (2011). Using prototypes to induce experimentation and knowledge integration in the development of enabling accounting information. *Contemporary Accounting Research, 28(2),* 708-736.

Wouters, M. & Wilderom, C. (2008). Developing performance-measurement systems as enabling formalization: A longitudinal field study of a logistics department. *Accounting, Organizations and Society 33(4-5),* 488-516.

Table 1
Formal fieldwork activity

Interviews	Location of staff	Number of interviews
Director, International Programmes Group	London	2
Deputy-Director, International Programmes Group	London	1
Regional Director	London, Ghana	2
Country Director	Sri Lanka(x2), Guyana, Ghana, The Gambia, Uganda, Vietnam, Nepal, Namibia, Cambodia	10
Head-Programme Learning and Advocacy	London	1
Team Leader-Programme Development and Learning	London	2
Executive Assistant to Director, International Programmes Group	London	3
Programme Learning Advisor	Ottawa	1
Systems and Project Manager	London	1
Head-Strategy, Performance and Governance	London	1
Director-VSO Federation	London	1
Volunteer Placement Advisor	London	1
Finance Manager	Sri Lanka	1
Programme Manager	Sri Lanka	2
Facilities and Office Manager	Sri Lanka	1
Volunteer	Sri Lanka	2
		32

Observation and attendance at meetings	Location of meeting	Number of meetings
Quality Framework meetings	London	6
Various planning and review meetings	London	6
Programme planning and review workshop	Sri Lanka	3
Office planning and logistics meeting	Sri Lanka	2
		17

Table 2
Modes of evaluation at VSO

Dimensions	**Modes of evaluation** <u>“Learning and Uniqueness”</u>	 <u>“Consistency and Competition”</u>
Purpose of evaluation	▪ Reflection, learning, improvement	▪ Standardize, compare, compete
Attributes of ‘good’ evaluation	▪ Contextual, detailed, 'local' interpretations	▪ Consistent, precise, objective, ‘universal’ interpretations
Attributes of ‘good’ accounts	▪ Narrative descriptions, case studies, stories, images ▪ Indicators that provoke creativity, ambition and innovation ▪ Avoid reliance on numbers as they provide only a partial account and do not tell the ‘real’ story	▪ Numbers, indicators, and scales, particularly those that can be compared between units ▪ Indicators that capture current performance accurately ▪ Avoid reliance on narrative as it is ‘selective’ and cannot be compared between units

Appendix 1: Strategic Resource Allocation tool – summary sheet

Assessment of Programme Effectiveness
Country : ………………………………….

The higher the overall percentage a Programme Office receives in this tool, the more "effective" it will be perceived to be based on this measure.

Section A. Focus on disadvantage (48 % of total score)

	Measure	% of total score
1	HDI	17%
2	Percentage of more disadvantaged people being reached through implementation of CSP aim	10%
3	Scored analysis of how well strategies are working in addressing the causes of disadvantage	10%
4	Disadvantage Focus in Current and Planned Placements	11%

Section B. Outputs of Country Programme (27% of total score)

	Measure	% of total score
5	What % of placements in the last 2 planning years have fully or mostly met their objectives (not including early return reports)?	13%
6	What was the Early Return rate (excluding medical & compassionate) over the last two planning years?	4%
7	What percentage of the ACP target of fully documented requests (i.e. with Placement Descriptions) was submitted on time over the last 3 planning years?	5%
8	What percentage of the ACP target number of volunteers was in country on 31/3/01, 31/3/00 & 31/3/99?	5%

Section C strategic approach (25% of total)
Note that the statements attached to each score are for guidance and are not absolute statements: we recognise that with some programmes no one statement will accurately describe the programme. The RPM must have a clear idea of the rationale behind the scoring, in order to ensure transparency and to allow comparison between countries. All of your scores should be based on an analysis of the current situation – i.e. not future strategy or placements.

9. Strategic approach based on programme at the current time	Score	% of total score
a) Placements working at different levels (micro/macro) towards strategic aims + planned links between them.		4%
b) Critical appraisal of placements with clear rationale linking placement to strategic aim + planned exit strategy		4%
c) Strategic & linked implementation of cross cutting themes		2%
d) In-country advocacy by the programme office		2%
e) PO proactive in promoting increased development understanding amongst volunteers		2%
f) Openness and commitment to learning		5%
g) Genuine partnership relationship with employers and other development actors		4%
h) Types of placements used most appropriate to needs of disadvantaged groups and based on strategic reasoning.		2%

Appendix 2

QF summary sheet 2008

Name of Programme:	**Please enter name of country here**			
Element	**Indicator**		**Indicator result**	**Element result**
Programme outcomes at partner level	A.1	Annual Progress in PAP objectives is achieved	**4**	**4**
	A.2	Programmes and partners monitor and review progress in capacity development and/or service delivery	**4**	
Programme impact at beneficiary level	B.1	Positive changes for target groups of partners are achieved	**4**	**4**
	B.2	Programmes and partners are able to understand, monitor and review changes for target groups	**4**	
Relevant and ambitious strategic plans are evolved in response to the development needs of the countryÕs disadvantaged communities	1.1	PO is responsive to changes in social, economic and political context	**4**	**4**
	1.2	PO is consulted by peer agencies and / or government bodies a credible development agency within its field of operation	**4**	
	1.3	Programmes working at different levels (e.g. national, provincial, districts, grass-roots) towards strategic aims	**4**	
Appropriate and innovative use of development interventions to deliver programme outcomes and impact	2.1	The contribution of National Volunteering (NV) to programme delivery has been maximised	**4**	**4**
	2.2.	The contribution of a range of different interventions to programme delivery has been maximised	**4**	
	2.3	Development awareness amongst volunteers and the wider community has been maximised	**4**	
	2.4	Opportunities to develop international resource partnerships are fully explored and developed	**4**	
	2.5	The contribution of advocacy to programme delivery has been maximised	**4**	
Programme delivery against plans	3.1	LTV/YFD & STV arrivals during year against reforecast plans	**4**	**4**
	3.2	Firm and documented placement delivery against reforecast plans	**4**	
	3.3	Quality of placement documentation	**4**	
	3.4	PIP milestones successfully completed across all programmes	**4**	
Inclusion	4.1	The PO has an inclusion statement and standards that are shared among all staff and that new staff sign. This is for both programme work and how the programme office is run	**4**	**4**
	4.2	Partner organisations include excluded groups in their work and as part of their target group	**4**	
Planning and review	5.1	Number of PIPs updated and signed off annually as a result of PARs in line with guidance	**4**	**4**
	5.2	All programmes are reviewed annually in line with guidance	**4**	
Partnership development and maintanance	6.1	Portfolio of partners in place relevant to PAP and CSP objectives	**4**	**4**
	6.2	Long-term (3-5 years) Partnership Plans are in place which include partnership objectives that are linked to the PAP objectives	**4**	
	6.3	Partners are actively involved in programme development and review	**4**	
	6.4	Partnerships are reviewed annually to assess progress towards Partnership and PAP objectives and quality of the relationship with VSO	**4**	
Volunteer enagement and support	7.1	Volunteer support baselines are being met by the Programme Office and partners are supported to manage volunteers	**4**	**4**
	7.2	PO celebrates volunteer achievement, responds to volunteer problems effectively and encourages the development of effective and accountable volunteer groups	**4**	
	7.3	Volunteers are engaged in programme development	**4**	
Programme funding	8.1	Value of proposals signed off by the CD/RPM against agreed quality criteria in Stage 2 of PMPG.	**4**	**4**
	8.2	Restricted income as % of PO Total Budget (including vol. recruitment costs)	**4**	
	8.3	Donor conditions for existing funding have been met (including financial, narrative and audit reports submitted on time and to the standard required by the donor) throughout the year	**4**	
Cost effectiveness	9.1	Percentage of total approved PO managed budget (restricted and unrestricted) budgeted on PC and VC costs in 2008/09 (Global average = 40%Regional averages range from32% and 57%)	**4**	**4**
	9.2	Possible areas of saving against costs identified during budget setting process through innovation and creative thinking	**4**	
	9.3	Percentage of total revised PO managed budget (restricted and unrestricted) budgeted on staff costs in 2007/08	**4**	
Financial management	10.1	Annual programme office expenditure variance (restricted plus unrestricted) for 08/09 against budget adjusted for macro-forecast	**4**	**4**
	10.2	Volunteer Unit Cost based on 08/09 budget	**4**	
Staff management and support	11.1	Performance management system are being actively implemented	**4**	**4**
	11.2	Evidence of major HR policies and systems being adhered to	**4**	
Legal and policy compliance and risk management	12.1	Number of outstanding category A and B internal audit actions for PO action relating to legal compliance	**4**	**4**
	12.2	Security Risk management plans signed off and implemented and tested according to Country's main security risks (e.g. avian flu, security, natural disasters etc)	**4**	