Michaelowa, Katharina; Borrmann, Axel

Working Paper

What Determines Evaluation Outcomes? Evidence from Bi- and Multilateral Development Cooperation

HWWA Discussion Paper, No. 310

Provided in Cooperation with:
Hamburgisches Welt-Wirtschafts-Archiv (HWWA)

Suggested Citation: Michaelowa, Katharina; Borrmann, Axel (2004) : What Determines Evaluation Outcomes? Evidence from Bi- and Multilateral Development Cooperation, HWWA Discussion Paper, No. 310, Hamburg Institute of International Economics (HWWA), Hamburg

This Version is available at:
https://hdl.handle.net/10419/19282

HWWA HAMBURG

DISCUSSION PAPER

What Determines Evaluation Outcomes? - Evidence from Bi- and Multilateral Development Cooperation -

Katharina Michaelowa
Axel Borrmann

HWWA DISCUSSION PAPER

310

Hamburgisches Welt-Wirtschafts-Archiv (HWWA)
Hamburg Institute of International Economics
2005
ISSN 1616-4814

Hamburgisches Welt-Wirtschafts-Archiv (HWWA)
Hamburg Institute of International Economics
Neuer Jungfernstieg 21 – 20347 Hamburg, Germany
Telefon: +49-40-428 34 355
Telefax: +49-40-428 34 451
e-mail: hwwa@hwwa.de
Internet: http://www.hwwa.de

The HWWA is a member of:

- Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz (WGL)
- Arbeitsgemeinschaft deutscher wirtschaftswissenschaftlicher Forschungsinstitute (ARGE)
- Association d'Instituts Européens de Conjoncture Economique (AIECE)

HWWA Discussion Paper

What Determines Evaluation Outcomes? - Evidence from Bi- and Multilateral Development Cooperation -

Katharina Michaelowa
Axel Borrmann

HWWA Discussion Paper 310

http://www.hwwa.de

Hamburg Institute of International Economics (HWWA)
Neuer Jungfernstieg 21 – 20347 Hamburg, Germany
e-mail: hwwa@hwwa.de

We are indebted to the evaluation departments of the major German donor institutions BMZ, GTZ and KfW as well as to the World Bank's Operations Evaluation Department (OED), in particular Patrick Grasso, for providing us with the data necessary for this analysis as well as with valuable information for their interpretation. Moreover, we would like to thank Matthias Busse, Carsten Hefeker and Aliya Khawari for their comments and suggestions on earlier drafts of this paper.

Edited by the research program Trade & Development

HWWA DISCUSSION PAPER 310
February 2005

What Determines Evaluation Outcomes? - Evidence from Bi- and Multilateral Development Cooperation -

ABSTRACT

Donor agencies invest considerable financial and human resources to evaluate the outcome of their development activities. To derive institutional conditions conducive to an efficient use of these resources, we develop a multi-level principal-agent model focusing on the various interests of the different actors involved in the evaluation process. The model highlights two central problems: (i) the aid agencies' conflicting objectives of transparency and self-legitimization, and (ii) the potential collusion between the evaluator and the project manager. Empirical evidence for the World Bank and different German donor agencies reveals concrete institutional requirements for a reduced evaluation bias and increased transparency.

Key words: Development cooperation, evaluation, political economy
JEL Classifications: F35, H43, D73

Katharina Michaelowa
Hamburg Institute of International Economics (HWWA)
Neuer Jungfernstieg 21
D-20347 Hamburg, Germany
Phone: +49-40-42834-291
Fax: +49-40-42834-367
E-mail: katja.michaelowa@hwwa.de

Axel Borrmann
Hamburg Institute of International Economics (HWWA)
Neuer Jungfernstieg 21
D-20347 Hamburg, Germany
Phone: +49-40-42834-436
Fax: +49-40-42834-367
E-mail: axel.borrmann@hwwa.de

1 Introduction

While evaluations are carried out in many areas of public policy, they are particularly frequent in development cooperation. All major donor institutions have particular evaluation departments monitoring the progress and outcomes of their activities in partner countries. At the World Bank, the Operations Evaluation Department (OED) carries out reviews of all Implementation Completion Reports (ICRs) produced by the operational staff, a task that is complemented every year by between 70-100 Project Performance Assessment Reports (PPARs), i.e. in depth reports based on OED investigations on the ground, additional analytical ex-post impact evaluations, as well as sector, thematic and country evaluations. OED has a staff of about 100 persons and an annual budget of approximately 20 million US$. Moreover, as regular monitoring and ICRs for each and every project are ensured by the operational staff of other departments, this represents only a minor fraction of the total staff hours and financial resources spent on the Bank's overall monitoring and evaluation activities which amount to annually over 200 million US$ (*Thumm* 1998, pp. 156ff.; *World Bank* 2001; *OED* 2003, Annex C).

Bilateral aid agencies make similar efforts to enhance information about the outcomes of their development activities. On average during the 1990s, the German Ministry of Economic Cooperation and Development (BMZ) carried out about 60 evaluations of ongoing projects, the cost of which have been estimated to lie between 2.6 and 5.1 million DM (*Borrmann* et al. 1999, p. 365). In addition, the two major German aid agencies, the German Technical Cooperation (GTZ) and the German Bank for Reconstruction and Development (KfW) carried out another 200-300 evaluations each year, and various smaller agencies and Non-Governmental Organizations (NGOs) produced additional reports.

Given the considerable financial and human resources involved, the question arises whether these resources are used efficiently and / or how the efficiency of their use could be improved. The central problem is that evaluations often serve more than one objective. They are simultaneously used as an instrument of transparency and control, accountability, legitimization and institutional learning. With respect to the legitimization function, evaluation can be thought of as a marketing device to "prove" the aid organization's successful work to the general public.

Generally, transparency would be considered the prime objective of evaluation. However, at least for some of the actors involved, the legtimation function seems to be dominating. Transparency and legitimization are clearly conflicting objectives in all cases in which actual development outcomes are not fully satisfactory. While the optimal control of operational staff on the ground could be easily derived in a principal-agent framework if the principal

were merely interested in transparency, the situation is more difficult if he also values the legitimization function.

It is the objective of this paper to clarify the underlying decision making problems and to derive institutional settings conducive to realistic evaluation results. We therefore introduce a political economic model focusing on the various interests of the different actors involved in the evaluation process (Section 2). In the second step, the results of this model are used to derive expectations about evaluation outcomes in different institutional environments (Section 3). These expectations are then examined in the light of empirical evidence for evaluation systems in different donor agencies (Section 4). This finally allows us to draw conclusions with some suggestions for institutional "best practices" (Section 5).

So far, literature addressing these topics is rather scarce. The literature on the economic theory of bureaucracy treats evaluation merely as a means of control (*Townsend* 1979, *Wintrobe* 1997, *Moe* 1997, *Dixit* 2002). Development policy oriented empirical studies generally consider evaluation outcomes as an indicator of true project results and focus on the relationship between donor preparation, analysis, supervision and/or management and the success of development programs (*Deininger*, *Squire* and *Basu* 1998, *Dollar* and *Svensson* 2000, *Hemmer* and *Lorenz* 2003). *Kilby* (2000) follows this tradition although he already includes some discussion about incentive structures which might lead to biased evaluation results. *Mann* (2000) adopts a nuanced, political-economic perspective on evaluation, but does not refer to development cooperation. Finally, *Carlsson*, *Köhlin* and *Ekbom* (1994) as well as *Nitsch* (2003) specifically consider the advantages of a political-economic perspective on the evaluation of development aid, and *Easterly's* (2002) fundamental critique of current donor practices also suggests this type of approach. However, a recent study by *Martens* (2002) is probably the only formal political-economic analysis of aid evaluation available so far. In addition, some practical insights are provided by *Brüne* (1998) and *Kadura* (1995) who highlight the dilemma of the evaluator, i.e. the person engaged to carry out the evaluation and who may fear to get sanctioned for unfavourable evaluation reports. We draw from these studies, but try to extend the theoretical framework in order to formalize the interrelations and interactions of various actors involved in the evaluation process. Moreover, besides pointing at the conflict of interest at the level of the aid agency and the evaluator, we also consider the possibility of collusion between the evaluator and the project manager on the ground.

2 Evaluation in development cooperation: a political economic model

Development cooperation differs from other policy areas because taxpayers of industrialised countries, who give policy makers a mandate for the delivery of aid, do not benefit directly from the results achieved. Mostly, there is a considerable geographical, political, social and cultural divide between the direct beneficiaries of recipient and the citizens of donor countries. Various intermediate institutions are involved in supplying aid services. Taxpayers of donor countries normally have no possibility to get in touch with aid recipients. And foreign beneficiaries have no voting rights in donor countries and thus no political leverage on donor politicians. As a result, there is a *broken feedback loop* that induces a performance bias in aid programmes (*Martens* 2002, pp. 154-155). Graph 1 depicts the constellation of players in the case of bilateral development cooperation, where voters mandate politicians to provide official development assistance (with the objective to alleviate poverty, say), and this mandate is delegated to the government or the relevant ministry and further on to public and non-governmental aid service suppliers (covering non-profit organisations and private consultancies). In partner countries, the mandate is forwarded to representatives of donor aid agencies, local consultants and NGOs as well as to partner institutions. It would be possible to include additional layers and their actors within these levels and institutions. However, this would further increase the complexity of our model.

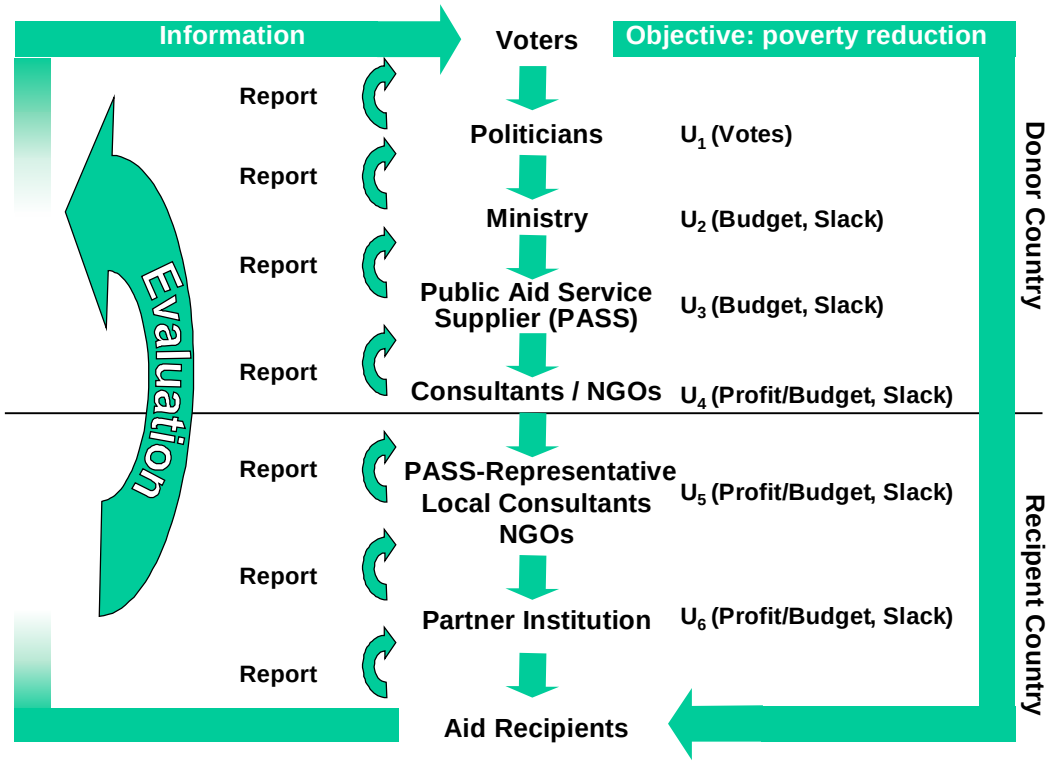
In the case of multilateral aid, the chain is slightly changed but at least equally complex. In particular, aid agencies then correspond to multilateral organizations that do not receive their mandate from any specific national ministry, but from the policy representatives of all of their member states, i.e. from multiple principals, whereby each is responsible for the tasks delegated by a different group of voters. At the same time, the chain of responsibilities within the recipient country may be shorter, as development cooperation of multilateral agencies mainly involves comprehensive, strategic projects or programs to be decided directly at the level of the recipient country government.

In any case, from this perspective, development cooperation appears as a hierarchical system of principal-agent relations. Members or institutions of each level aim at their specific objectives, which do not necessarily coincide with the objectives of their direct principal or the mandated objective of the voters, here: poverty reduction. Their objectives can be expressed by a set of stylised utility functions, assuming that the utility depends upon public approval (votes) in the case of politicians, upon their budget in the case of public aid service suppliers, upon profits in the case of private suppliers and, in the case of all suppliers upon pleasant working conditions (slack). Each report requested by the principal from his direct agent passes the filter of the agent's specific utility function. In such a framework, it appears hardly possible that citizens sitting at the end of the pipe are supplied with somewhat realistic

information on the effectiveness of aid. However, taxpayers depend on this filtered information because it would be far too costly for them to obtain first-hand insights on their own. Even taking into account the potential information transfer via NGOs and the media, it must be assumed that this information will remain very partial and incomplete. In order to bridge the broken feedback loop and compensate for the information cost disadvantage of taxpayers, evaluation seems to be an appropriate mechanism (*Martens* 2002, p 155).

Evaluation implies costs that the principal has to bear; however, it allows him to control the services of his agent(s) and actually provides the agent with the required incentive to perform in the principal's interest. Moreover, true reporting can be stimulated by sanctions to be imposed in case an evaluation reveals false declarations by the agent (see for example *Mookherjee* and *P'ng* 1989). This clearly explains that evaluations are of particular importance in the field of development cooperation or in any other policy field where direct feed-back channels - like direct information from programme users - are not available (*Martens* 2002, p. 155).

Figure 1: The broken feedback-loop – the case of bilateral development cooperation



However, as mentioned earlier, evaluations are not merely used as a control for the agent but also, among other things, to demonstrate the effectiveness of the development assistance as requested by the citizens. In fact as each principal up to the ministry is himself the agent of some other principal up to the voter, each of them has a genuine interest to let his work appear successful. If evaluation results are used for legitimization purposes, for each principal, there is a trade-off between requesting truthful evaluations which will imply the strongest working incentives for the agent, and positively biased evaluations which will make the principal himself appear very successful. Faced with the expectation of positive results, the evaluator engaged to carry out the project assessment – regardless of whether he is recruited externally or from the principal's own staff – gets into a similar dilemma. If he has to assume a strong preference of the principal for a positive image, he will be inclined to avoid a realistic evaluation in order to please the principal and – in the case of a private, external evaluator – to ensure follow-up contracts. However, in the long run, unduly biased evaluation results may spoil the evaluator's reputation so that he faces a trade-off just like the principal.

In addition, the incentives for truthful evaluation results also depend on the evaluator's relationship with the agent. Thinking of the agent at the lowest level of the hierarchical structure as the project manager or "expert" on the ground, it becomes clear that this person can greatly facilitate the evaluator's work by providing him with the relevant contacts and information. Obviously, these contacts may be selective and the information may be biased in order to hide potential problems in project management. But the evaluator may accept that and collude with the agent to the detriment of the principal.

Modelling the individual utility function of each actor involved in the delivery of aid can reveal a clearer picture of the problems resulting from this constellation. In order to decrease the complexity of interrelations to be analysed, the model is reduced to four main actors: the aid agency's project manager M working in the recipient country, the aid agency A as his principal, the evaluator E evaluating on behalf of the aid agency, and the politicians P of the donor countries acting as the aid agency's principal and being politically accountable for development assistance on the whole. Note that for reasons of simplification, we do not distinguish between several principals in the case of multilateral aid so that we can present a single model for both types of development assistance.

In case of an evaluation, let the project manager's utility U^M depend on the result of this evaluation $\hat{Y}_p$, and let it be determined by his own reporting about the success of his project $Y°_p$ otherwise. The benefits from reporting positive outcomes Π do not only include direct financial gains, but also increased reputation or job security via the continuation of the project under his responsibility. If he overstates the project success, however, and if this overstatement is revealed via an evaluation of his work, he will have to pay a fine of $T(Y°_p$-

$\hat{Y}_p$) depending on the extent of the divergence between Y°_p and $\hat{Y}_p$. Again, this fine does not need to be interpreted merely as a financial payment. It may imply that the project manager will no more be employed for interesting positions in the future, or that he will no more be offered any follow-up position by the aid agency at all. Moreover, in extreme cases of non-compliance with contractual obligations, the aid agency may hold back some of the payment agreed for his work. Let evaluations happen with a fixed probability of q.

A further relevant variable to be considered is the project manager's effort α which, on the one hand, positively influences the (potential) evaluation result, but, on the other hand, also induces utility losses via reduced slack S(α). The project manager's utility function can therefore be written as:

(1) $U^M = q[\,\Pi(\hat{Y}_p) - T(Y^\circ_p - \hat{Y}_p)] + (1\text{-}q)\,\Pi(Y^\circ_p) + S(\alpha)$,

with $\Pi'(\hat{Y}_p) > 0$, $\Pi'(Y^\circ_p) > 0$, $T'(Y^\circ_p - \hat{Y}_p) > 0$ und $S'(\alpha) < 0$.

Let the relationship between the true project outcome Y_p and the evaluation outcome $\hat{Y}_p$ be specified as follows:

(2) $\hat{Y}_p = Y_p + \lambda_p$, wherby $Y_p = Y_p(\alpha)$ with $Y_p'(\alpha) > 0$.

The parameter λ represents the evaluation error which is assumed to be exogeneous in typical principal-agent models, but which will be endogenized later in the context of our model.

It is a well known result from standard agency theory that given this specification of the project manager's utility function, an incentive compatible mechanism can be found which induces him to truthfully reveal the project outcomes in his own reporting and make all efforts which could be expected by the aid agency. This is true under the sole condition that the evaluation error typically remains small (*Mookherjee* und *P'ng* 1989, p. 414).

As discussed earlier, however, in our context, the project manager's principal, i.e. the aid agency, who could establish such an incentive compatible mechanism, is itself interested in a positive evaluation outcome. This evaluation outcome will influence the valuation of its own work by its own principal, the politicians, who will decide about the future budget for the agency on this basis.

Let the utility function of politicians be given by:

(3) $U^P = \text{votes}[I(\sum_i [\hat{Y}_i - g(\lambda_i)]) - B]$.

In this function, the politicians' utility from development cooperation U^P is defined as the difference between the satisfaction of the population about information on successful development aid I(·) on the one hand, and losses of votes due to taxation to finance the budget

B for development cooperation on the other hand. The information of voters by the means of which politicians try to attract voters depends positively on the evaluation outcomes of all individual projects i, including project p. At the same time, greatly upward biased evaluation results represent a certain risk for the politician: the greater the gap between real project outcomes and evaluation results, the higher the risk that these differences will be noticed. While the media generally do not show much interest in development issues, individual apparent failures of development aid have often become a topic of NGO campaigns discussed by newspapers and television (*Easterly* 2002, p. 29). This relationship is reflected by the function $g(\lambda_i)$ assuming $g'(\lambda_i)>0$ and $g''(\lambda_i)>0$, so that the risk of a revealed discrepancy between true results and official evaluation results rises over-proportionally with an increasing evaluation bias λ.

More generally, the information function $I(\cdot)$ reflects the assumption that evaluation results are not brought to people's knowledge in an unfiltered way.[1] Whether the population will be satisfied with glossy brochures describing selected projects and resuming aid outcomes in unspecific phrases like "80% of all projects have again been successful this year" depends upon the intensity of voters' interest which may differ from country to country and between multilateral and bilateral development cooperation. Clearly, the voters in each individual country can hold their policy makers responsible only partially for the outcomes of multilateral aid. At the same time they only finance parts of it with their own taxes. This should imply that whatever the outcome, it will have a lesser impact on votes than in the case of bilateral cooperation.

Let us now consider the utility function of the aid agency U^A. For reasons of simplification, we will assume that the aid agency is interested only in the volume of its budget B:

$$U^A = B \tag{4}$$

This implies that the distribution of funds within the agency is not of any interest here, and further, that the aid agency is indifferent between the use of these funds for operative purposes or for the evaluation of development outcomes.

The utility of the evaluator U^E depends on: (i) the acceptance $A(\lambda_p)$ that his report will find at the level of the aid agency who commissioned it and who may ensure follow-up contracts; (ii) the quality of his report $Q(\lambda_p)$ that guarantees his good name as an independent expert in the field, and (iii) pleasant working conditions (slack) $Z(\lambda_p)$. In this context, the parameter λ_p already defined above should now be understood as a variable directly depending on the choice of the evaluator. The evaluator can in fact purposefully obfuscate the true project

1 For the idea of the information function, see *Martens* (2002, p. 161).

outcomes and / or accept biased or unprecise evaluation procedures. This has a negative impact on the quality of his report ($Q'(\lambda_p)<0$), but a positive impact on his working conditions on the ground ($Z'(\lambda_p)>0$). The latter is due predominantly to the relationship between the evaluator and the project manager who can facilitate the evaluation task by answering the relevant questions, providing contacts for interviews, or accompanying the evaluator on his field trips. The evaluator's task will be easier if he accepts this assistance, but at the same time, this assistance must be expected to have a negative impact on the objectivity of the report. The influence of λ_p on A can be deduced only from the utility maximization process of the aid agency.

Finally, we introduce a diversification parameter θ which influences the evaluator's weighing process between the objectives to please the commissioning agency and to establish his name as a competent independent expert. If θ is small, the evaluator depends heavily on the aid agency and has only a few (if any) alternative employment options.

$$(5) \qquad U^E = (1-\theta)A(\lambda_p) + \theta Q(\lambda_p) + Z(\lambda_p), \qquad 0 \leq \theta \leq 1.$$

Assuming plausible concave functions and utility maximization of the different players, we can now derive unique equilibrium values for the success of the project Y_p as well as for the evaluation bias λ_p.

Let us consider politicians first. Their influence is limited to the decision about the budget for development cooperation B. Equating marginal cost and benefits in terms of additional votes yields their utility maximizing $B[I(\Sigma\hat{Y}_i - g(\lambda_i))]$.

We assume that this reaction function can be anticipated by the aid agency and will therefore be integrated into the agency's own utility maximization. To simplify notation, B will be written in the following as a direct function of $\hat{Y}_i$ and $g(\lambda_i)$. Taking into account the optimization procedure of the politicians the utility function of the aid agency can therefore be reformulated as:

$$(6) \qquad U^A = B\Big(\sum_i [\hat{Y}_i - g(\lambda_i)]\Big) = B\Big(\sum_{i \neq p} [\hat{Y}_i - g(\lambda_i)] + Y_p(\alpha) + \lambda_p - g(\lambda_p)\Big)$$

Moreover, as a well informed intermediate layer, the aid agency is not only aware of the optimization process of the politicians, but also about the optimization process of the project manager it engaged. The latter maximizes his utility over two parameters: his work intensity α and his reporting about project outcomes Y°_p. Consequently, the following first order conditions can be derived from his utility function (1):

$$(7) \qquad \frac{\partial U^M}{\partial \alpha} = qY_p'(\alpha)[\Pi'(\hat{Y}_p) + T'(Y^\circ_p - Y_p(\alpha) - \lambda_p)] + Z'(\alpha) = 0 \text{ and}$$

(8) $$\frac{\partial U^M}{\partial Y^\circ_p} = -q\, T'(Y^\circ_p - Y_p(\alpha) - \lambda_p) + (1-q)\Pi'(Y^\circ_p) = 0.$$

The intersection of these optimization condition yields two reaction functions: $\alpha(q,\lambda_p)$ and $Y^\circ_p(q,\lambda_p)$. The project manager's effort α increases with rising evaluation probability and decreases with rising evaluation bias, while overreporting of project success Y°_p is tempered with rising evaluation probability and encouraged with higher evaluation bias. Independently of the definition of individual curves, at q=0, utility maximization of the project manager always leads to $\alpha=0$ und $Y^\circ_p \to \infty$. At some positive probability of evaluation, the project manager's optimal effort becomes positive and then depends on the evaluation bias λ_p – just as the optimal exaggeration of project success. At q=1 finally, the project manager cannot draw any positive utility from overreporting. In this case, in order to avoid fines, he will report any project outcome between the true outcome $Y_p(\alpha)$ and the evaluation result $Y_p(\alpha) + \lambda_p$.

Let us now return to the aid agency. It knows the project manager's reaction functions. As his exaggeration of project outcomes is not directly related to the success of the project, it is only concerned about his effort. Inserting the project manager's reaction function $\alpha(q,\lambda_p)$ into the aid agency's own utility function (6) yields:

(9) $$U^A = B\Big(\sum_i \hat{Y}_i - g(\lambda_i)\Big) = B\Big(\sum_{i \neq p} \big[\hat{Y}_i - g(\lambda_i)\big] + Y_p\big[\alpha(q,\lambda_p)\big] + \lambda_p - g(\lambda_p)\Big)$$

Based on this function, the agency can now choose the relative frequency of evaluations q leading to the highest level of utility. As long as there are no budget constraints limiting the financial resources that can be used for evaluations, it will choose q=1, since B increases monotonously with q (indirectly via Y_p and α), i.e. increasing the probability of evaluation always raises the agency's utility. Nevertheless, even if the agency is theoretically indifferent between its operative business and evaluation activities, a certain budget constraint is provided by the fact that resources spent on evaluation must not exceed the overall budget and must also leave the financial resources allocated to the project which represents the basis of the evaluation. Therefore, q is not necessarily equal to 1 but always equal to the upper limit.

Once q is determined, B is defined as a function of λ_p. As we assumed concave functions, for small λ_p, the marginal benefits of a further increase of the evaluation bias via the improved evaluation outcome exceeds the marginal cost via lower project outcomes due to the reduced effort of the project manager $Y_p\big[\alpha(q,\lambda_p)\big]$ and the growing risk that failures may be discovered $g(\lambda_p)$. With growing λ_p the latter gain in importance until they fully outweigh marginal benefits. The agency reaches its utility maximum with the highest budget if λ_p is chosen in a way that marginal cost just set off marginal benefits.

While the agency cannot directly influence the evaluation bias λ_p it can make its preferences known to the evaluator to whom it delegates the project assessment and, in case of an external consultant, it can let him know about potential follow-up contracts in case his evaluation matches the agencies expectations. The approval of the evaluation $A(\lambda_p)$ is directly proportional to the agency's utility from λ_p, i.e.:

(10) $A(\lambda_p) = \gamma \left(Y_p[\alpha(\lambda_p)] + \lambda_p - g(\lambda_p)\right), \quad \text{with } 0 < \gamma \leq 1,$

where γ represents the relevance of project p within the agency's overall portfolio of development projects.

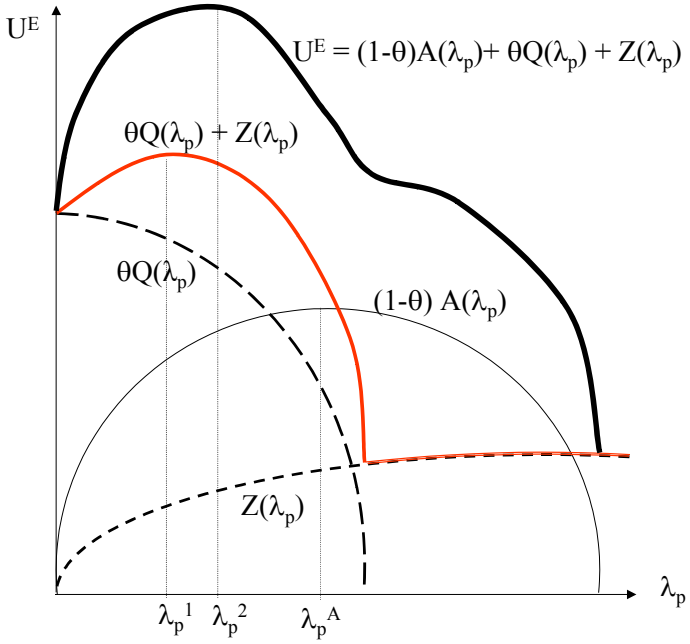
Inserting (10) in (5) leads to the evaluator's final utility function:

(11) $U^E = (1-\theta)\gamma \left(Y_p[\alpha(\lambda_p)] + \lambda_p - g(\lambda_p)\right) + \theta\, Q(\lambda_p) + Z(\lambda_p).$

The evaluator maximizes his utility via the determination of the degree of objectivity (or bias) of the evaluation result. This leads to the equilibrium value of the evaluation error λ_p which in turn influences the actual project result $Y_p[\alpha(\lambda_p)]$.

Figure 2 provides a graphical illustration of the evaluator's maximization problem. It shows that without any influence of the aid agency, the evaluation bias would be λ_p^1 because the evaluator would then only consider the trade-off between the loss in his professional image via reduced evaluation quality $Q(\lambda_p)$ weighted with the diversification parameter θ, and the utility gain resulting from improved working conditions on the ground $Z(\lambda_p)$. The sum of these functions is indicated by a grey line which has its maximum at λ_p^1. However, as the aid agency's valuation of the evaluator's results influences his future career perspectives, he will also take into account the agency's acceptance function $A(\lambda_p)$ weighted with $(1-\theta)$. The sum of all three functions is indicated by the bold black line representing the evaluator's overall utility. It has its maximum at an evaluation bias of λ_p^2, so that the optimal distortion is increased to this level. It follows that the optimal distortion always lies between the optimal bias without consideration of the aid agency λ_p^1 and the optimal bias for the aid agency itself λ_p^A.

Figure 2: The evaluator's utility maximization problem



λ_p^1: optimal bias of evaluation results without consideration of the aid agency

λ_p^2: optimal bias with consideration of the aid agency (equilibrium value)

λ_p^A: optimal bias from the aid agency's point of view

3 Project and evaluation outcomes in different institutional frameworks – some theoretical considerations and practical implications

The model derived above suggests a closer look at the reaction of the equilibrium values of λ_p and Y_p on changes of individual parameters, i.e. differences between the institutional frameworks in which the evaluations take place. As we will see, this framework varies considerably between different aid organizations so that the incentive problems mentioned above can be expected to be rather strong in some cases, and much less so in others.

3.1 Conditions ensuring the independence of the evaluator from the aid agency

Let us first stick to the analysis of the evaluator. Figure 2 provides a direct insight into the effect of the diversification parameter θ. If the evaluator is highly diversified, θ converges towards 1, and in this extreme case $A(\lambda_p)$ remains unconsidered in his utility function. Consequently, the equilibrium value of the bias will be relatively low (converging towards λ_p^1) which in turn implies a higher effort of the project manager and thus an improved project

outcome Y_p. This reflects a situation in which the aid agency cannot exert much pressure on the evaluator because the latter has a multitude of alternative future employment or consultancy opportunities. To the opposite, if θ converges towards 0, the optimization process of the aid agency plays a highly important role for the evaluator's future employment prospects. In this case, the evaluator is highly specialized on consultancy projects for this particular agency, or, equivalently, the agency holds a monopsony position in his area of specialization. Evaluation quality $Q(\lambda_p)$ which would have provided a crucial indicator of his professional knowledge to any new employer is not relevant here because no such alternative employment options exist.

In this case, whether the evaluation is carried out by a formally "independent" evaluator, i.e. a person external to the organization, or whether it is carried out by the agency's own staff, only plays a minor role. In fact, under certain institutional safeguard provisions, it may well be that the internal evaluator is actually more independent than a formally external evaluator elsewhere. As evaluations carried out by the agency's own staff play a predominant role in almost all aid organizations considered here, let us consider this situation in some more detail.

Theoretically an evaluator from within the agency can belong to: (i) the operational staff responsible for the project, (ii) the staff of a separate evaluation department, and (iii) the staff of other operational departments. At the World Bank, only in exceptional cases evaluations are commissioned to external consultants (*Thumm* 1998, p. 160). Continuous project supervision and monitoring as well as the Implementation Completion Reports (ICRs) belong to the responsibilities of the operational staff in charge of the project. However, a high number of projects undergo an additional detailed evaluation by the separate Operations Evaluation Department (OED). In the early 1970s when the overall portfolio of World Bank projects was much smaller, all completed projects were subject to these performance assessments (*Willoughby* 2003, p. 8). Today, Project Performance Assessment Reports (PPARs[2]) cover about 25% of all projects (*OED* 2003, p. 15). Institutional provisions ensure that OED is truly independent within the Bank. In particular, the Director-General of OED reports directly to the Executive Board, i.e. the different member country representatives, and not to the president of the World Bank. Moreover, he is engaged on a fixed term contract which can be renewed only once and may not be employed anywhere else in the Bank thereafter (*Stek* 2003, p. 493).

2 Note that in earlier World Bank publications PPARs are typically referred to as "Performance Audit Reports" (PARs). This terminology was changed to avoid the potential confusion with audits carried out by external courts of audit.

Within the German development organizations considered here, only the BMZ predominantly commissions evaluations to external consultants. The technical cooperation agency GTZ engages external consultants for some of its Project Progress Reviews, but has Final Reviews prepared by its own operational staff on the ground. Although a separate evaluation department does exist in the head office in Eschborn, this department does not yet carry out or commission any additional evaluations (except selected desk reviews of Final Reviews commissioned to an external auditing firm). The development bank KfW finally, also engages its own staff for most of the evaluations carried out. Until 2001, this staff typically came from other operational departments selected by the structurally not fully independent "Secretariat of international credit affairs". Since 2001, a newly founded independent evaluation department has been carrying out evaluations on its own, supplemented by operational staff and external consultants (*KfW* 2002, p. 3, *Borrmann et al.* 2001, pp. 91f.). In addition, it should be noted that as a bank whose major focus is on structural development within Europe and Germany, KfW's international development activities account for only 3% of its overall budget (*KfW* 2004, p. 7). The KfW can therefore be expected to be generally much more independent of the success of its development activities than other donor agencies.

In the framework of our model, self-reports of the operational staff like in the case of GTZ Final Reviews and World Bank ICRs must not actually be considered as evaluations, but as the project managers' reports. By definition, an evaluator is a person separate from the person to be evaluated.[3] However, assessments by staff of other operational departments or independent evaluation units can be considered as evaluations. In fact, these evaluators often have an incentive structure similar to the one of any external evaluator. Even extreme cases of $\theta = 1$ are possible if institutional provisions make sure that the "internal" evaluator has no career perspectives within the organization and only a fixed term contract like in the case of the head of OED. He can be expected to have a high incentive to excel via the quality and impartiality of his evaluations in order to pursue his career elsewhere thereafter. As this provision does not hold for other staff members of OED, they can be expected to weight their utility of high quality assessments valued highly by their Director-General against the utility of producing pleasant results for other parts of the Bank.

Let us now return to the discussion of evaluations commissioned to external consultants. As mentioned earlier, the evaluator's weights given to quality work increasing his reputation on the one hand and positively biased results to please the agency on the other hand strongly depend on his degree of diversification. Among the development organizations considered

3 This can be considered as a minimum requirement. The standard definition in the evaluation guidelines of the OECD Development Assistance Committee (DAC) includes a considerable number of additional criteria (*OECD* 1992).

here, only the evaluation department of the BMZ follows some kind of a tendering procedure for the evaluations it intends to commission. Therefore, personal contacts as well as "harmonic relations" currently dominate the recruitment of evaluators, whereas professional skills like sectoral know-how, language ability, and country and field specific experience play only a secondary role (*Kadura* 1995, p 14, *Brüne* 1998, p. 20). In the context of bilateral aid agencies, even where tendering takes place, it is often restricted to a narrow group of national consultants highly specialised on advisory services for development assistance. In the German case, evaluation reports are generally drafted in German so that language creates a natural entrance barrier to the market and protects German consultants from international competiton. At the same time, it increases their specialization on the German market and their dependence on the few German aid agencies requiring their services. The situation can be expected to be different for international organizations like the World Bank who recruit their external consultants from a larger international market which ensures a certain competition implying a higher valuation of professional reputation and reduced dependency on contracts with any individual aid agency. However, as the World Bank rarely employs external evaluators, this potential has remained largely untapped so far.

Except the features of the particular market, the model suggests that the dependency of evaluators on the aid agency may also be influenced by the hierarchical level at which the external evaluation report is commissioned. At higher levels, the units commissioning the evaluation have so many projects to deal with that even if they were held responsible for the overall outcome, any single project would only play a minor role, i.e. they become more independent. In the model, this is reflected in a lower value for the agency's acceptance function $A(\lambda_p)$ within the evaluators utility function and a lower marginal return to bias due to the parameter γ, i.e. an equally increased independence of the evaluator. The two extreme cases are evaluation units at the World Bank or the BMZ, dealing with several hundreds of projects annually on the one hand, and GTZ project managers dealing with a single Project Progress Review on the other hand. Indeed since 2000, the GTZ project manager decides himself about the person to engage for the Progress Review of the project he works on. This person may either be external to the organization or belong to other operational departments within GTZ which in turn may have some bearing on his valuation of his professional reputation. But in any case the evaluator can then be expected to be under extreme pressure for positive results. And in this particular case, as it is the project manager himself and not his principal who commissions the evaluation, a second problem not mentioned so far in this section becomes even more obvious: the potential collusion between the project manager and the evaluator to the detriment of the aid agency.

3.2 Conditions ensuring the independence of the evaluator from the project manager

Clearly collusion between the evaluator and the project manager must be expected to be extreme if the evaluation is not commissioned by any higher level institutional department but by the project manager himself. Theoretically, his type of collusion can lead to a higher bias than even the aid agency would approve. In this case, the functions in Figure 2 above must be drawn such that $\lambda_p^A < \lambda_p^1$. But even without assuming such extreme cases, it is obvious that the objectivity of the evaluation does not only depend on the evaluator's independence from the aid agency. As the evaluator values pleasant working conditions (slack) $Z(\lambda_p)$, he enjoys a positive working climate with the project manager who may take him for sight seeing or dinner invitations, and, even more importantly, who can greatly facilitate the evaluation work via prearranged meetings and field trips, a preselection of the relevant project documentation and readily prepared answers to any kind of question the evaluator may have. This support tends to be valuated very highly, in particular when time constraints and budgetary restrictions for the evaluation are very tight. Due to such support, the evaluator saves time, but he also accepts to adopt a less objective perspective on project outcomes. The project manager in turn obtains the possibility to reduce his effort without much risk that this attitude might find its reflection in negative evaluation results.

In the above mentioned case of GTZ Project Progress Reviews where the evaluation is commissioned by the project manager, common practice shows that the latter tends to choose a colleague who manages another project. At a different point in time, this colleague may then choose him for the same role. This institutional setting should be expected to lead to very positive evaluation results and roughly similar project outcomes as if evaluation was abolished altogether.[4]

Independent of the hierarchical level at which the evaluation is commissioned, various characteristics of the evaluation itself also determine the potential for collusion. First, the situation described above holds only for a traditional development project with a project manager responsible for the implementation on the ground. This is much more typical for technical than for financial cooperation. Second, collusion in the above described way requires the project manager to be still in place when the evaluation takes place. This is only the case for mid-term evaluations or project completion assessments before the project has actually stopped its operations on the ground. This is the typical context of evaluations commissioned by the German BMZ and GTZ. Ex-post evaluations, often several years after

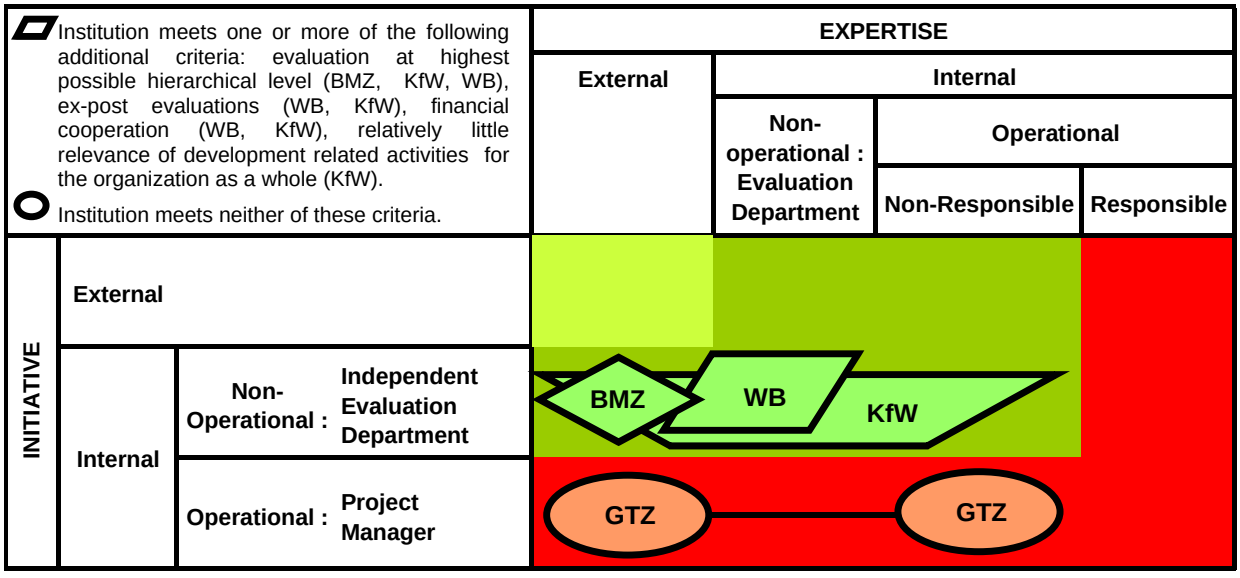
4 The GTZ has long been arguing that learning rather than control is the prime objective of its evaluations, and that the control function may be left to BMZ. However, it is questionable whether even the learning purpose can be met if evaluations do not reveal the existing problems and simply praise the work being done on the ground.

project completion clearly restrict the scope for collusion. Even if contacts between the former project manager and the evaluator are established before the latter leaves on his mission, the former cannot give him so much support any more. Moreover, as the former project manager is now involved in new projects, he can be expected to be less dependent on positive evaluation outcomes as he would have been if the evaluation result had decided about the continuation of his project (or his contract).

Ex-post evaluations are the rule at both the German KfW and at the World Bank (for OED's PPARs). In addition to these organizations' focus on financial rather than technical cooperation, this further adds to their limited risk of collusion as compared to other institutions like BMZ or GTZ.

Figure 3 provides a summary of the institutional settings of the different agencies based on the major arguments discussed so far. Each institution's evaluation system is ranked with respect to the institutional independence of the evaluator (horizontal line), and the institutional independence of the unit commissioning the evaluation reports (vertical line). Different shapes show additional aspects of evaluation characteristics more (round) or less (angular) prone to distorted results.

Figure 3: Institutional settings for evaluation*



*GTZ Final Reviews and World Bank Implementation Completion Reports (ICRs) are not considered as evaluations here because they are carried out by the project managers themselves.

3.3 Institutional conditions conducive to control from the general public

In the model, the aid agency's own interest in positive evaluation results is mitigated by the risk of a revelation of failures to the general public $g(\lambda_i)$. The corresponding trade-off is captured in the politicians' utility function and thus indirectly determines the agency's budget $B[I(\Sigma\hat{Y}_i - g(\lambda_i))]$. It depends upon the general interest of the population, and, in particular, the activity of critical NGOs and media to what extent $g(\lambda_i)$will be really relevant. This leads us to consider two questions:

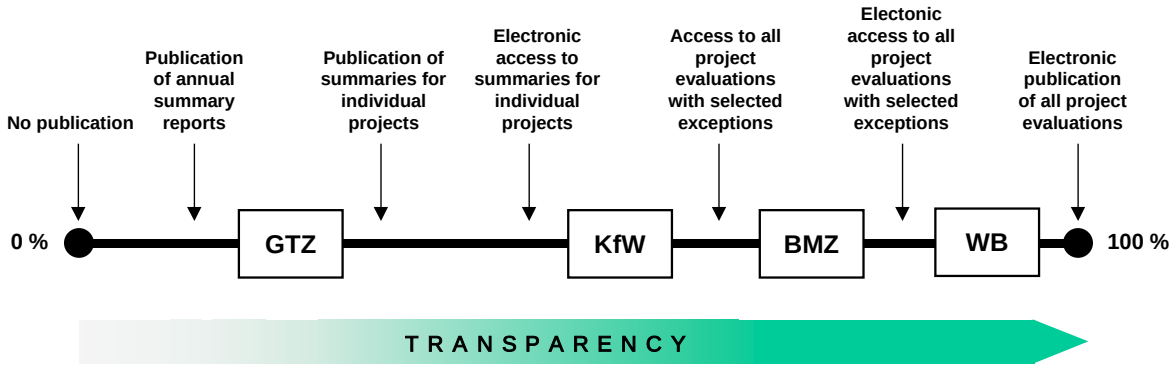
- Can the interest of NGOs, media and the general public be expected to be higher for multilateral or bilateral aid?
- Which general institutional provisions maximize the risk of revelation $g(\lambda_i)$ so that the agency will show a relatively low propensity to accept (or even pressurize for) overly positive evaluation results?

Let us start with the second question. NGOs and media interested in development cooperation have to shoulder the cost when gathering information. A simple measure to strengthen their role is to create transparency about evaluation results. Traditionally, German aid agencies have been very reluctant in this respect, and from the political side no attempt was made to formally request the publication of all reports. However, BMZ started a turnaround in 1999 by granting access to its evaluation reports with a few reservations only. Moreover, summaries and the BMZ's bi-annual evaluation programs are now available on the internet. Since the early nineties, KfW and GTZ have been publishing their regular cross-sectional evaluation reports. Recently, KfW started to publish a condensed version of each evaluation report including main findings and detailed success ratings (*KfW* 2002, p. 3).

At the World Bank, transparency has equally improved in recent years. While until the year 2000, with the exception of certain country, thematic or sector evaluations, evaluations were accessible only to internal stakeholders, since 2002, OED discloses virtually all evaluation reports to the general public (*World Bank* 2003, *Stek* 2003, p. 492). Transaction cost to access is low as reports are now available electronically via internet.

Overall, these developments point in the required direction. Nevertheless, important differences remain in the level of transparency provided by the agencies considered here. Figure 4 illustrates their relative position taking into account various institutional requirements for the transparency of evaluation results.

Figure 4: The transparency of evaluation results in different development agencies



Let us now consider the question of the information transfer to the voters in the light of general institutional differences between bi-and multilateral aid. It is often argued that decision making processes in international organizations are more difficult to follow by the general public than decision making processes and activities of national bureaucracies (see e.g. *Frey* 1997, pp. 116 and 119, *Vaubel* 1991a, p. 204, and *Vaubel* 2003, pp. 3ff.). As discussed earlier, one of the major reasons for voters' difficulties to receive unfiltered information from the beneficiaries of aid is geographical distance. Now if aid is channeled via multilateral organizations, the geographical divide is two-fold. This further reduces the chance to receive any direct first hand information and increases the voters' transaction cost. Moreover, national politicians can be only partially held responsible for undesirable outcomes of multilateral development cooperation. Each donor may send only one representative to the World Bank's Board of Governors and Board of Executive Directors. This further reduces any national voters' incentive to gather the relevant information. If the general public shows little interest, so will the media. This again raises information cost.

However, while the position of unorganized groups like the general public tends to be weakened when activities are carried out at the multilateral level the position of internationally organized lobbies is typically strengthened (see e.g. *Vaubel* 1991b, pp. 38f.). For them, following the activities of a single big international organization is much easier than observing each individual bilateral aid agency. As opposed to other policy fields where the interests of well-organized international lobby groups are antagonistic to the interests of the broader public, this does not appear to be generally true in the area of development cooperation. In fact, the predominant lobby groups to be considered in this context are international NGOs. Major international NGOs focus their lobbying activities on multilateral institutions like the World Bank and the IMF rather than on national development policies. In

fact, as these two international organizations often develop the policies adopted later on by numerous bilateral donor agencies, the NGOs' investment into information about their strategies has a long-term payoff both at the international and at the national level. As NGOs spread their knowledge to the media and across the general population in all member countries, voters' information cost decline and the above mentioned problem of the additional geographical divide is mitigated – if not overcompensated. In any case, for multilateral organizations, the threat of a big scandal set up by well organized NGOs on the international scene in case of any major failure of development policies certainly represents a major incentive for serious internal outcome assessments and a rejection of all-too positive evaluation results. From a theoretical point of view, it is therefore difficult to say, whether the risk of revelation $g(\lambda_i)$ plays a more important role at the bilateral or the multilateral level.

4 Some empirical evidence

So far, summing up the theoretical discussion, we obtain a relatively positive overall valuation of the institutional conditions for evaluations at the World Bank, the KfW and the BMZ. At the same time, the GTZ evaluation system shows major deficiencies in about all respects discussed above, especially since the year 2000 when GTZ project managers were entitled to select the evaluators. This leads us to expect highly upward biased evaluation results especially from this year onwards.

For all other agencies, we do expect greater objectivity, but some deficiencies remain. For BMZ evaluations which generally take place while the project is still ongoing, the weakest point might be the risk of collusion between the evaluator and the project manager on the ground. Concerning the ex-post evaluations of the KfW and the World Bank, some incentive remains to let the overall institution appear in a positive light. For KfW, this should be true in particular for the period before 2002, when the independent evaluation department started its operations. Unfortunately, it is difficult to carry out empirical tests differentiating clearly between these problems. Whether comparisons are made across different organizations or over time, generally, several parameters change simultaneously so that in many cases the most relevant of them cannot be unambiguously identified. Nevertheless, the available evidence can be assessed with respect to its general consistency with the results outlined above and thereby illustrate the relevance of the theoretical model.

The data used stem from OED's electronic database on all Project Performance Assessment Reports and OED checked ICRs since 1972, from the KfW electronic database with information for all evaluations between 1988 and 2001, and from GTZ's internal evaluation summary statistics for Project Progress Reviews and Final Reviews between 1993 and 2000.

As far as the BMZ is concerned, detailed project evaluation results of all projects covered by its annual cross-evaluation assessments from 1990 to 1996 are included in the analysis. From 1997 onwards, the BMZ shifted its focus on thematic, country and instrument evaluations and left individual project evaluations to its agencies.

Table 1 provides an initial overview of average project evaluation results at the GTZ, KfW and the World Bank over the period from 1994 to 2000, a period for which data are available for all three organizations.

Table 1: Share of successful projects according to the evaluation statistics of GTZ, KfW and the World Bank/OED, 1994-2000*

	1994-2000	2000	1998-99	1997	1996	1995	1994
GTZ Project Progress Reviews**	87%	97%	87%	87%	89%	88%	74%
KfW ex-post evaluations	70% (0.044)	73% (0.044)	66% (0.045)	69% (0.038)	85% (0.038)	72% (0.055)	64% (0.046)
World Bank/OED PPARs**	72% (0.018)	73% (0.048)	76% (0.036)	85% (0.043)	70% (0.046)	63% (0.047)	69% (0.042)

*Projects are considered "successful" according to the agencies' specific rating systems (GTZ: "Abschließende Berurteilung", KfW: "Projekterfolg KfW, OED: outcome). Generally, the category of successful projects includes the three subcategories "higly successful", "successful", and "successful with reservations" or "marginally successful". (Standard errors in parenthesis; not available for GTZ as only summary statistics were provided to the authors).

**GTZ project completion assessments and World Bank ICRs are not considered as evaluations here because they are generally carried out by the project managers themselves.

Sources: GTZ, KfW, World Bank/OED.

At first glance, we notice that while World Bank and KfW evaluation outcomes are roughly similar, GTZ results are clearly superior, not only on average, but even for each individual year. The gap is particularly prominent in the year 2000 where GTZ results show success rates of 97%, 24 percentage points above those of the other two institutions. What is the real explanatory power of these results? Has GTZ really been avoiding any mistakes in its current projects since the turn of the millennium? Are the World Bank and the KfW really so much less successful in development co-operation than the GTZ?

Given the institutional conditions for evaluations in the different aid agencies discussed above, it appears more probable to assume that KfW and OED have been simply applying tighter evaluation standards. In particular, the jump in GTZ success rates from 1998/99 to 2000 coincides perfectly with the change in GTZ regulations which entrusted the project managers to select the evaluators.

As in the German system, the BMZ carries out independent evaluations of GTZ and KfW projects, the question whether differences between the two agencies' success rates reflect differences in the evaluation framework or in real project outcomes can be counterchecked at a higher level. Table 2 reveals that, on average, results for both GTZ and KfW projects are less favourable when evaluations are not commissioned by the respective agency but by BMZ. However, the discrepancy is much more striking in the GTZ case. KfW ratings differ by 3 percentage points only, a difference that is statistically insignificant. The difference between BMZ ratings of GTZ and KfW projects is insignificant, and in fact, the share of successful projects is actually slightly higher for KfW than for GTZ.

Table 2: Share of successful projects by commissioning agency and type of project

Commissioning Agency	Type of Project	Share of Successful Projects*
BMZ (1990-1996)	All projects evaluated	71% (0.028)
BMZ (1990-1996)	KfW projects**	71% (0.095)
KfW (1988-2001)	KfW projects**	74% (0.080)
BMZ (1990-1996)	GTZ projects**	67% (0.038)
GTZ (1994-2000)	GTZ projects, Project Progress Reviews	87%

*As in Table 1, projects are considered "successful" according to the agencies' individual rating system. For BMZ this relates to outcome ("Zielerreichung") as in the case of the World Bank. (Standard errors in parentheses not available for GTZ as only summary statistics were provided to the authors).
Note that ratings by different agencies do not necessarily refer to identical projects as no joint project identification code was available. Comparisons are valid only under the assumption of random samples from identical populations.

**Joint projecst of GTZ and KfW excluded.

Source: BMZ, GTZ and KfW.

In a similar way, for the GTZ and the World Bank, we can compare the outcomes of assessments carried out by the project managers themselves with the outcome of evaluations by external consultants, members of other departments, or an external evaluation department within the agency. For GTZ, this implies comparing the Final Reviews drafted by project managers with the Project Progress Reviews discussed above. For the World Bank, results of ICRs drafted by the operational staff (or OED's reviewed version thereof) can be compared with the outcomes of PPARs. *OED* (2003, Annex C) provides an extended discussion of this comparison.

Similar to the difference between ratings at various hierarchical levels demonstrated in Table 2, the theoretical discussion leads us to expect a rather clear difference in evaluation results resulting from an upward biased result of the project managers' own assessment.

However, there are two distinct situations in which results should converge:

- The control system works extremely well, so that the project manager has no more incentive for upward biased reporting (perfect contract as derived in the principle-agent literature, i.e. ideal institutional conditions for evaluation)
- The control system works very badly and the collusion between the project manager and the evaluator becomes so strong that the latter turns into the mouthpiece of the former.

Table 3.1 shows that the case of the World Bank corresponds to neither of these extremes. As OED statistics provide individual project codes, the different valuations can be compared for identical projects. The last row of Table 3.1 indicates that for between 16 and 41% of these projects, PPAR outcomes have been somewhat less optimistic than ICR results. However, comparing the share of projects rated successful under the two assessment types indicates that the differences are not always clearly significant. In 2000, the share of projects rated successful in ICRs is even slightly higher than in PPARs. This can be interpreted as a positive sign of OED's success in inducing project managers to produce reasonably truthful assessment reports on their own. In order to come even closer to the optimal situation of truthful reporting, the frequency of PPARs and/or the negative consequences of upward biased ICRs once they have been discovered might have to be raised.

As opposed to the case of the World Bank, data for GTZ does not seem to be consistent with our hypothesis. Table 3.2 shows that evaluators tend to report even more optimistic results than project managers. Clearly, under these circumstances, evaluations cannot fulfill their role to induce project managers to be more realistic. But how can we explain this situation? Even if we assume that collusion between evaluators and project managers is extremely high, it cannot be higher than if the project manager himself writes the report. Moreover, even if the evaluator depends greatly on the agency's general approval for his work, it is difficult to see why this should be less relevant for the project manager. Potentially, one could imagine that a project manager with a long-term contract is more independent than an external consultant. But a more realistic explanation seems to lie in the fact that we consider evaluations only for mid-term reviews, while we consider the project manager's reporting upon completion. If the project manager himself has a short-term contract fixed to the duration of the project – a contractual situation which has become more and more frequent throughout the 1990s, it is vital for him that the project should not be stopped due to a negative result of the Project Progress Review. Once the project is completed, he has either found an alternative project already or at least he knows that he will in any case not be employed any further on the completed project. This puts him in a somewhat more independent position. At the time of the

mid-term evaluation, however, he has a strong incentive to have the evaluation outcome look positive. We can conclude that the Project Progress Reviews as currently carried out by GTZ have a high propability to come up with results that are close to meaningless.

Table 3: Average assessment results reported by evaluators and project managers*

Table 3.1: World Bank

	2002	2001	2000	1999	1998	1997	1996	1995	1994
Evaluation: PPARs	76% (0.051)	71% (0.054)	74% (0.048)	79% (0.050)	73% (0.052)	86% (0.042)	71% (0.046)	62% (0.048)	70% (0.042)
Project managers' reporting: ICRs (after OED desk control)	78% (0.049)	81% (0.047)	71% (0.049)	87% (0.042)	85% (0.042)	87% (0.043)	76% (0.044)	72% (0.046)	81% (0.036)
Share of projects with less positive ratings in PPARs	28% (0.053)	31% (0.055)	16% (0.040)	41% (0.060)	34% (0.055)	17% (0.045)	25% (0.044)	25% (0.043)	28% (0.041)

*Table 3.2: GTZ***

	1994-2000	2000	1998-99	1997	1996	1995	1994
Evaluation: Project Progress Reviews	87%	96%	87%	87%	89%	88%	74%
Project managers' reporting: Final Reports	84%	82%	88%	81%	85%	85%	78%

*Results are presented as the share of successful projects in terms of overall outcomes (as defined above).

**For GTZ, the projects subject to the different types of assessments are not necessarily identical because the individual projects could not be identified. Comparisons are valid only under the assumption of random samples from identical populations.

Sources: GTZ, World Bank/OED.

A complementary analysis of collusion between the evaluator and the operative staff can be carried out by exploiting the potential divergence of ratings for different evaluation criteria. We can distinguish between "nobody is responsible" or "third party is responsible" criteria on the one hand, and "project manager is responsible" or "aid agency is responsible" criteria on the other hand. A typical example for a "nobody is responsible" criterium is the development of the projects' external conditions, i.e. the project's economic and political environment. If a project really does not work so that it is hard to classify it as generally successful, the evaluator still has the possibility to put the blame on external conditions rather than on the bad performance of the project manager. Our theoretical results lead us to expect that this should happen most frequently under conditions of evaluations like those of GTZ and BMZ where evaluations are carried out during the lifetime of the project. As outlined above, in these cases, the evaluator and the project manager meet on the ground whereby the latter typically prepares the evaluation process, field visits etc. so that there is ample ground for collusion.

While the available GTZ data do not allow us to carry out such a detailed analysis, we can carry it out for BMZ. Data indeed show some evidence for the evaluators' tendency to attribute positive credits to the work of the operational staff while assigning shortcomings to the impact of the project environment. Figure 5.1 shows that the first and second best ratings (1 and 2 - corresponding to "excellent" and "very good") are attributed more frequently to the project manager's task of implementation while the unsatisfactory ratings (4 and 5 – corresponding to "unsatisfactory" and "strongly unsatisfactory") are more often used for external conditions. Table 4.1 provides a general overview over various BMZ evaluation criteria and confirms the relatively positive rating of implementation.

We can now try to replicate the same analysis for the World Bank or the KfW whose types of projects (financial rather than technical cooperation) and evaluation timing (ex post evaluations) have lead us to expect less collusion. We would therefore predict a more balanced rating of the different criteria. While no information on external conditions is available, OED data allow us to compare ratings of overall borrower and bank performance. If there were collusion between the World Bank's operational staff and evaluators, borrower performance should be, on average, rated more negatively than bank performance. However, as depicted in Figure 5.2, there is virtually no perceptible difference between the ratings of these criteria.

Figure 5: Rating of "Nobody responsible" versus "Project manager responsible" criteria

*Figure 5.1: BMZ evaluations, 1990-1996**

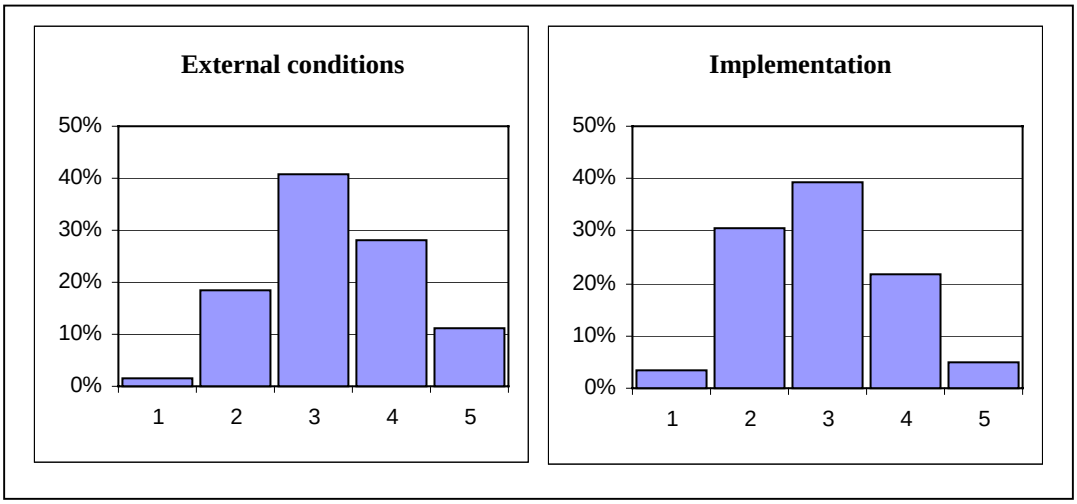
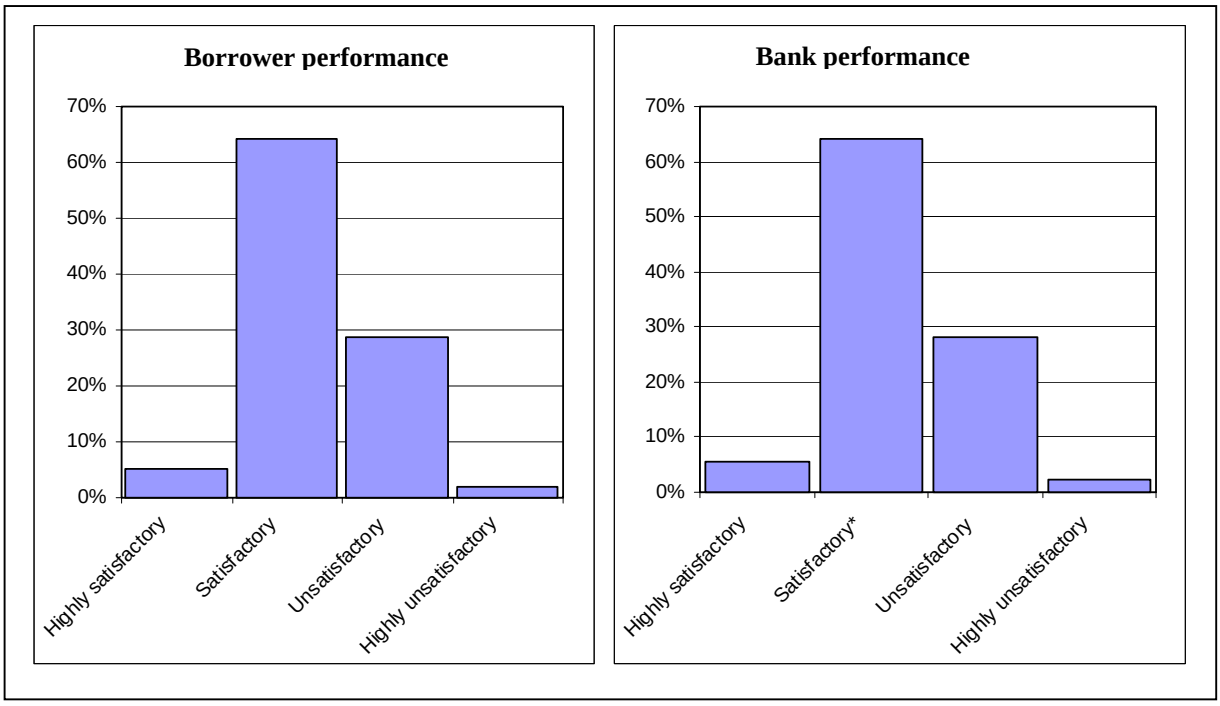


Figure 5.2: World Bank/OED PPARs, 1997 – June 2004



*BMZ ratings are from 1=excellent to 5=fully unsatisfactory. At the World Bank, for a single evaluation, bank performance was rated "marginally satisfactory". This is included in the category "satisfactory" here.

Sources: BMZ, World Bank/OED.

This clearly confirms our theoretical hypothesis and presents the World Bank evaluation system as almost immune against the collusion problem discussed above. The extremely positive picture is somewhat disturbed, however, when investigating further into the subcategories of "overall borrower performance" and "overall bank performance". The

positive rating of borrower performance is mainly driven by "borrower preparation" which obtains the best rating of all OED operative evaluation criteria (see Table 4.2). As opposed to "borrower compliance" and "borrower implementation", "borrower preparation" may well be a criterion for which the Bank's project manager may also be held responsible. In fact, if borrower preparation were bad, the project manager would probably have to face the question why he extended the loan in the first place. This implies that "borrower preparation" should probably not be included in the "nobody is responsible" or "third party is responsible" criteria. The above comparison must therefore be interpreted with some caution. If instead, we compare the ratings of "borrower implementation" or "borrower compliance" with "bank supervision" (see Annex 1), the charts look much more similar to the case of BMZ. However, a small advantage remains in favor of the Bank.

Table 4: Assessment of specific evaluation criteria, ratings and coverage

Table 4.1: BMZ evaluations, 1990-1996

Criterion	Share of projects considerd at least satisfactory*	Standard error	Number of projects	
Implementation	72.0%	0.028	261	Operational criteria
Definition of objectives	70.3%	0.027	283	
Project coordination	60.7%	0.030	262	
External conditions	60.6%	0.029	282	
Project planning	56.1%	0.030	278	
Efficiency	82.7%	0.025	225	Success criteria
Outcome	70.5%	0.029	255	
Sustainability	54.6%	0.032	249	

Table 4.2: World Bank/OED PPARs, 1993-2003

Criterion	Share of projects considerd at least satisfactory	Standard error	Number of projects	
Borrower preparation	78.1%	0.014	909	Operational criteria
Bank Supervision	74.8%	0.014	975	
Borrower compliance	67.5%	0.015	968	
Borrower implementation	65.7%	0.015	977	
Impact	83.7%	0.012	979	Success criteria
Efficiency**	72.6%	0.025	328	
Outcome	68.5%	0.015	988	
Sustainability	52.8%	0.016	967	

*Corresponds to a rating of 1-3 out of 5.

**As OED data contain only the rate of return but no indication above which threshold this rate of return is considered as satisfactory, we arbitrarily fix the threshold at 10%.

Source: BMZ, World Bank/OED.

As our theoretical model emphasizes the link between collusion and the evaluator's objective of slack maximization, it may be interesting to also search for further signs of such slack within the different institutional frameworks. One indicator to what extent the institutional system includes safeguard provisions against slack may be the number of entries for the different evaluation criteria. Theoretically, according to their terms of reference, the evaluators should report on all of them. Evidence for BMZ provided in Table 4.1 indicates, however, that there is a considerable number of missing values, in particular for criteria that are analytically more demanding like efficiency and sustainability.

An in-depth study of BMZ-commissioned evaluation reports reveals that evaluators often complain that either time and budget constraints or specific project characteristics do not permit a cost-benefit-analysis and/or an analysis of expected sustainability. Some evaluators circumvent the problem by assessing efficiency and sustainability in a brief, barely founded and intuitive way. Similar behaviour of evaluators can be observed for the development impact criterion which is not included in the BMZ-dataset. It has long been criticised that the ministry failed to introduce a more sophisticated set of empirical methods to evaluate such a fundamental issue (see for example *Stockmann* 1996). A comprehensive attempt has been made only recently (*BMZ* 2000).

At the World Bank, such impact assessments are part of the usual evaluation procedure and relatively well reported. Moreover, there is regular reporting on two other criteria of overall success, namely outcomes and sustainability. It seems as if the World Bank, which draws much of its reputation from being a think tank, had a higher incentive to comply with analytical and methodological requirements involved in the reporting of sustainability than the BMZ. However, at the World Bank, the efficiency criterion is missing even more often than for BMZ evaluations, namely in about two thirds of all cases. At the World Bank, information on efficiency implies the calculation of an explicit economic rate of return which is a much more complex endaveour than the rating of efficiency in BMZ evaluations. Moreover, circumventing analytical difficulties by providing loose general descriptions like it is frequently observed at the BMZ is impossible here. This may explain the difference in the share of missing values. We can thus conclude that we do find much evidence for reduced reporting on the most complex criteria in both organizations, but that at a similar level of complexity, the World Bank's OED provides a more exhaustive set of information.

A final check of whether the evaluation outcomes reflect superficial statements or well founded assessments can be obtained by an analysis of the internal consistency of the different evaluation criteria. It should be expected that serious evaluations show a strong and positive relationship between the rankings of the operational evaluation categories and the overall success categories. Running an ordered probit regression for the success variables

"outcomes" and "sustainability" respectively shows the expected relationship for the World Bank. In case of the BMZ, most operational criteria are equally significant, but there are exceptions. Project coordination is fully insignificant and the definition of objectives is significant only at a level of 10% for both outcomes and sustainability (see Annex 2).

Despite all apparent efforts for analytical excellence, even at the World Bank, however, overall evaluation outcomes can be shown to also follow other, non performance related trends. Based on the above mentioned ordered probit regression for "outcomes" the insertion of a simple time trend also produces a significant coefficient. This result is robust to the inclusion of additional control variables, i.e. regional dummies and sector specifications. Table 5 presents the results using East Asia & Pacific as a benchmark among the regions and keeping only "rural sector" among 15 possible sector specifications of which all others appear to be insignificant. Regression results are generally consistent with expectations. Sub-Saharan Africa, South Asia, Middle East & North Africa as well as East and Central Asia show significantly worse evaluation results than East Asia & Pacific, at given ratings for borrower and bank performance. Similarly, rural sector projects obtain significantly lower scores.

Table 5: Structural determinants of evaluation results, World Bank PPARs (1993-2003)

Ordered probit estimates					N =	882
					LR chi2(11) =	640.89
					Prob > chi2 =	0
Log likelihood =	-932.313				Pseudo R2 =	0.2558
Outcome*	Coef.	Std.Err.	z	P>|z|	[95% Conf.	Interval]
Bank supervision*	0.213	0.033	6.42	0.000	0.148	0.278
Borrower preparation*	0.237	0.035	6.73	0.000	0.168	0.306
Borrower implementation*	0.343	0.036	9.64	0.000	0.273	0.413
Borrower compliance*	0.164	0.030	5.38	0.000	0.104	0.223
Sub-Saharan Africa	0.484	0.123	3.94	0.000	0.243	0.724
East and Central Asia	0.354	0.153	2.32	0.020	0.055	0.653
Latin America & Caribbean	0.175	0.132	1.33	0.184	-0.083	0.434
Middle East & North Africa	0.426	0.163	2.62	0.009	0.107	0.745
South Asia	0.457	0.142	3.22	0.001	0.179	0.736
Rural Sector	0.477	0.098	4.87	0.000	0.285	0.670
Evaluation financial year	0.027	0.014	1.98	0.048	0.000	0.053

*Rated from 1=highly satisfactory to 6=highly unsatisfactory.

The actual variable of interest, however, is the trend variable "evaluation financial year" ranging here from 1993 to 2003. Its significantly positive coefficient shows that at given ratings for underlying operational evaluation criteria (and at given values for other control variables) overall evaluation outcomes have deteriorated over time. As the correction for other influences implies that there is no substance related reason for this deterioration, the most

obvious explanation appears to be that the rating of outcomes has become stricter over time.[5] This may be related to a gradual strengthening of OED's independent position within the World Bank leading to fewer qualms when it comes to ratings unpleasant to the operational staff.

Similar interpretations have been adopted by former OED officials to explain the general deterioration in OED ratings since the early 1970s (*Weiner* 2003, p 30). These positions sharply contrast other World Bank publications in which the decline in success rates is interpreted as a serious sign of declining effectiveness of operations over time (*World Bank* 2002, p. 73).

This leads us to a final – general – question of how to interprete evaluation results. As we have seen, comparisons across agencies are at least as problematic as comparisons over time because differences tend to reflect variations in the institutional setting for evaluation rather than real differences in outcomes. We have already noted this in the context of the upward jump in GTZ results for Project Progress Reviews from 1998/99 to 2000 when project managers were entrusted to select the evaluator. In a comparison of BMZ, KfW and GTZ, we have also observed the tendency of evaluations at higher hierarchical levels to show lower rates of success. And finally, we have observed a great number of structural differences between the evaluation systems at GTZ on the one hand, and KfW or the World Bank on the other hand, that lead to systematically higher reported success rates for GTZ (despite insignificant differences in outcomes as reported by BMZ, at least for GTZ and KfW).

A related problem is that there is no objective or exogenous rating scale for total success. In fact, success rates can be calibrated by aid agencies according to a desired range. While the GTZ may find it optimal to show success rates close to 100%, the KfW and the World Bank may prefer somewhat lower levels that increase their credibility. Partners or clients from the private sector would hardly find success rates close to 100% credible, particularly in the field of development co-operation. As overall success rates result from an aggregation of single ratings of a set of different evaluation criteria like effectiveness, efficiency, development impact and sustainability, the agency may include or exclude certain criteria or weigh them discretionarily. Data for BMZ and OED evaluations presented in Table 4 above revealed that average project success differs considerably depending on the criteria selected. It would also be conceivable to combine certain criteria such as outcome and sustainability and to consider

5 While annual monitoring results of ongoing projects follow a natural trend because outcomes tend to be updated rather than reassessed from scratch (*Kilby* 2000, p. 238), this is not the case for the final OED evaluations considered here. Therefore, this cannot explain the significant time trend observed in Table 5.

a project as successful only if the rating is satisfactory for both. Note that in this case, for both the BMZ and the World Bank the overall share of successful projects would drop below 50%.

To sum up, evidence presented here reveals that a meaningful interpretation of evaluation results requires considerable knowledge about the institutional setting in which the evaluation process takes place, as well as about the exact definitions of success used by the individual aid agency. This is where it becomes clear that transparency provided not only over general ratings, but also over the background of their computation is crucial for a true understanding of results. As transparency over results in turn reduces the transaction cost of control by national and international NGOs and the media, it is not surprising that the institutions providing the most comprehensive access to information also appear to be those with the lowest systematic bias of evaluation results.

5 Conclusions and recommendations

Using an extended principal-agent model, this paper reveals that the evaluator's dependency on the acceptance of the results by the principal (i.e. the aid organization), and the potential collusion between the evaluator and the agent (i.e. the project manager) strongly influence the credibility of evaluation results. Institutional conditions conducive to relatively unbiased results are shown to include: the institutional independence of the unit commissioning the evaluation, the institutional and professional independence of the evaluator, the predominance of ex-post evaluations, a high degree of transparency, and a specialization on financial rather than technical cooperation.

All except the last of these criteria are amenable to institutional reform which can greatly enhance the efficiency of the evaluation system. In particular, making sure that evaluations are commissioned by an evaluation unit placed at the highest possible level of the organization is a low-cost measure which can be quite effective. In addition, carrying out ex-post rather than mid-term reviews or evaluations directly upon completion can significantly reduce the risk of collusion. Among the donor agencies considered here, currently, GTZ's situation is most problematic with respect to all of these criteria, and the BMZ still shows some reluctancy to move towards ex-post evaluations. At the KfW and the World Bank, these requirements are fully satisfied.

The additional requirement of professional and institutional independence of the evaluator may entail considerable cost once the agency decides to move from evaluations carried out by their own staff to evaluations carried out by external consultants. So far, among the agencies considered here, only the BMZ systematically engages external evaluators. In addition, even the fact that they are external to the organization does not ensure that they are truly

independent. It appears that their actual independence could be enhanced via an opening up of the relevant market inducing stronger competition both on the demand and on the supply side of evaluation services. This would require transparent tendering procedures and a reduction of language barriers. While evaluation reports for German aid agencies are generally written in German, this could easily be opened up to at least English, but possibly also French and Spanish depending on the local language in the recipient country concerned. This would have the additional advantage of facilitating the integration of local consultants into the evaluation team.

Considering the trade-off between the cost and benefits of such external evaluations, it should be noted that far from all projects would need to be subject to this procedure. In fact external evaluations could be used as spot-checks. In order to decrease the bias inherent in self-evaluations, the general principal-agency literature suggests to carry out selected result-based re-evaluations with the probability of a re-evaluation increasing if the result of the self-evaluation is positive (*Mookherjee* and *P'ng* 1989, p. 408 f.). Possibly, project appraisals like those carried out in the framework of the DAC Peer Reviews could be extended to full-fledged evaluations of projects selected by the reviewers rather than by the country under review. In addition, periodic meta-evaluations of the evaluation systems and its methods should be forseen. This would improve and maintain their institutional evaluation quality.

Finally, transparency requirements call for an electronic availability of all individual evaluation reports as well as a detailed description of evaluation methods and structural procedures. Only the World Bank comes close to this ideal situation. BMZ so far holds up transaction cost by sending reports only on request. KfW publishes only summaries and ratings and GTZ only publishes aggregated results in its annual report.

Generally comparing multilateral and bilateral aid agencies might have initially lead to the expectation of stronger problems at the multilateral level due to the geographical distance to the voters in the member countries and their limited influence on decision making. However, at least at the World Bank, it seems that this effect is compensated by a stronger interest of international NGOs who follow the Bank's activities and inform the media and the national populations of member states. In any case, the empirical evidence presented in this paper suggests that the World Bank possesses one of the best functioning evaluation systems in terms of credibility, analytical excellence and transparency. In many respects, it could be seen as a model case for other agencies. It should be noted, however, that recent trends to increase the number of country, thematic and sector evaluations to the detriment of project evaluations, may at some point lead to a situation where they do no more represent a sufficient control mechanism. As sophisticated as the evaluation may be, if its probability becomes too low, it will no longer prevent project managers from significant overreporting. As long as projects

remain the predominant form of aid, project evaluations should also remain the predominant activity of evaluation units. The BMZ which has delegated individual project evaluations to its agencies should at least make sure that a credible evaluation system is in place in all major German aid agencies, including the GTZ.

References

BMZ (2000): Langfristige Wirkungen deutscher Entwicklungszusammenarbeit un dihre Erfolgsbedingungen, Eine Ex-post-Evaluierung von 32 abgeschlossenen Projekten, BMZ Spezial No. 19, Bonn (BMZ).

Borrmann, Axel, Karl *Fasbender*, Manfred *Holthus*, Albrecht *von Gleich*, Bettina *Reichl* and Rasul *Shams* (1999): Erfolgskontrolle in der deutschen Entwicklungszusammenarbeit, Analyse, Bewertung, Reformen, Baden-Baden (Nomos).

Borrmann, Axel, Albrecht *von Gleich*, Manfred *Holthus* and Rasul *Shams* (2001): Reform der Erfolgskontrolle in der deutschen Entwicklungszusammenarbeit, Eine Zwischenbilanz, 2001, Baden-Baden (Nomos).

Brüne, Stefan (1998): Evaluierung als öffentliche Kommunikation - Zu den politischen and entwicklungsbezogenen Rahmenbedingungen entwicklungsbezogener Wirkungsbeobachtung, in: Stefan Brüne (ed.): Erfolgskontrolle in der entwicklungspolitischen Zusammenarbeit, Hamburg (German Overseas Institute), pp. 9-26.

Carlsson, Jerker, Gunnar *Köhlin* and Anders *Ekbom* (1994): The Political Economy of Evaluation - International Aid Agencies and the Effectiveness of Aid, London (Palgrave Macmillan).

Deininger, Klaus, Lyn *Squire* and Swati *Basu* (1998): Does Economic Analysis Improve the Quality of Foreign Assistance?, in: The World Bank Economic Review Vol. 12, No. 3, pp. 385-418.

Dixit, Avinash (2002): Incentives and Organizations in the Public Sector: An Interpretive Review, in: Journal of Human Resources Vol. 37, No. 4, pp. 696-727.

Dollar, David and Jakob *Svensson* (2000): What Explains the Success or Failure of Structural Adjustment Programmes?, in: Economic Journal Vol. 110, pp. 894-917.

Easterly, William (2002): The Cartel of Good Intentions: The Problem of Bureaucracy in Foreign Aid, in: Journal of Policy Reform Vol. 5, No. 4, pp. 223-250.

Frey, Bruno (1997): International Organizations, in: Dennis Mueller (ed.): Perspectives on Public Choice, Cambridge, (Cambridge University Press), pp. 106-123

Hemmer, Hans-Rimbertand Andreas *Lorenz* (2003): What Determines the Success or Failure of German Bilateral Financial Aid?, in: Review of World Economics Vol. 139, No. 3, pp. 507-549.

Kadura, Bernd (1995): Wie frei ist ein freier Gutachter? Zwischen bürokratischer Gängelung, selbstauferlegter Befangenheit and lasziver Gestaltungsfreiheit, in: Stefan Brüne (ed.): Erfolgskontrolle in der entwicklungspolitischen Zusammenarbeit, Thesen and Materialien zu einer Tagung des Deutschen Übersee Instituts, Hamburg (German Overseas Institute), pp. 9-14.

KfW (2002): Cooperating with Perspective: Building on Opportunities, seventh evaluation report on projects and programmes in developing countries, Frankfurt (KfW).

KfW (2004): Annual Report 2003, Frankfurt (KfW)

Kilby, Christopher (2000): Supervision and Performance: The Case of World Bank Projects, in: Journal of Development Economics Vol. 62, pp. 233-259.

Mann, Stefan (2000): The Demand for Evaluation from a Public Choice Perspective, in: Vierteljahreshefte für Wirtschaftsforschung Vol. 69, No. 3, pp. 371-378.

Martens, Bertin (2002): The role of evaluation in foreign aid programmes, in: Bertin Martens, Uwe Mummert, Peter Murrell and Paul Seabright: The Institutional Economics of Foreign Aid, Cambridge (Cambridge University Press), pp. 154-177.

Moe, Terry (1997): The Positive Theory of Public Bureaucracy, in: Dennis Mueller: Perspectives on Public Choice: a Handbook. Cambridge (Cambridge University Press), pp. 455-480.

Mookherjee, Dilip and Ivan *P'ng* (1989): Optimal Auditing, Insurance, and Redistribution, in: The Quartely Journal of Economics, Vol. 14, No. 2, pp. 399-415.

Nitsch, Manfred (2003): Evaluationskriterien für Mikrofinanzinstitutionen: Finanzielle Nachhaltigkeit, Erreichung ärmerer Zielgruppen und Einkommenswirkungen – a comment, in: Heinz Ahrens (ed.): Neuere Ansätze der theoretischen und empirischen Entwicklungsforschung, Berlin (Duncker&Humblot).

OECD (1992): DAC Principles for Effective Aid, Development Assistance Manual, Paris (OECD)

OED (2003): Annual Report on Operations Evaluation, Washington (The World Bank).

Stek, Pieter (2003): Evaluation at the World Bank and Implications for Bilateral Donors, in: Evaluation Vol. 9, No. 4, pp. 491-497.

Stockmann, Reinhard (1996): Die Wirksamkeit der Entwicklungszusammenarbeit . Eine Evaluation der Nachhaltigkeit von Programmen and Projekten der Berufsbildung, Opladen, (Westdeutscher Verlag) .

Thumm, Ulrich (1998): Erfolgskontrolle bei der Weltbank: Von der Projektverabschiedung zur Durchführung und Ergebnissen vor Ort, in: Stefan Brüne (ed.): Erfolgskontrolle in der entwicklungspolitischen Zusammenarbeit, Hamburg (German Overseas Institute), pp. 152-167.

Townsend, Robert (1979): Optimal Contracts and Competitive Markets with Costly State Verification, in: Journal of Economic Theory Vol. 22, pp. 265-293.

Vaubel, Roland (1991a): The Political Economy of the International Monetary Fund: A Public Choice Analysis, in: Roland Vaubel and Thomas Willet (eds): The Polictical Economy of International Organizations: A Public Choice Approach, Boulder (Westview Press), pp. 204-244.

Vaubel, Roland (1991b): A Public Choice View of International Organization, in: Roland Vaubel and Thomas Willet (eds): The Polictical Economy of International Organizations: A Public Choice Approach, Boulder (Westview Press), pp. 27-45.

Vaubel, Roland (2003): Principal-Agent-Probleme in internationalen Organisationen, HWWA Discussion Paper No. 219, Hamburg (Hamburg Institute of International Economics).

Weiner, Mervyn (2003): Institutionalizing the Evaluation Function at the World Bank, 1975-84 in: Patrick Grasso, Sulaiman Wasty and Rachel Weaving (eds.): World Bank Operations Evaluation Department – The First 30 Years, Washington (World Bank), pp. 17-30.

Willoughby, Christopher (2003): First Experiments in Operations Evaluation: Roots, Hopes, and Gaps, in: Patrick Grasso, Sulaiman Wasty and Rachel Weaving (eds.): World Bank Operations Evaluation Department – The First 30 Years, Washington (World Bank), pp. 3-15.

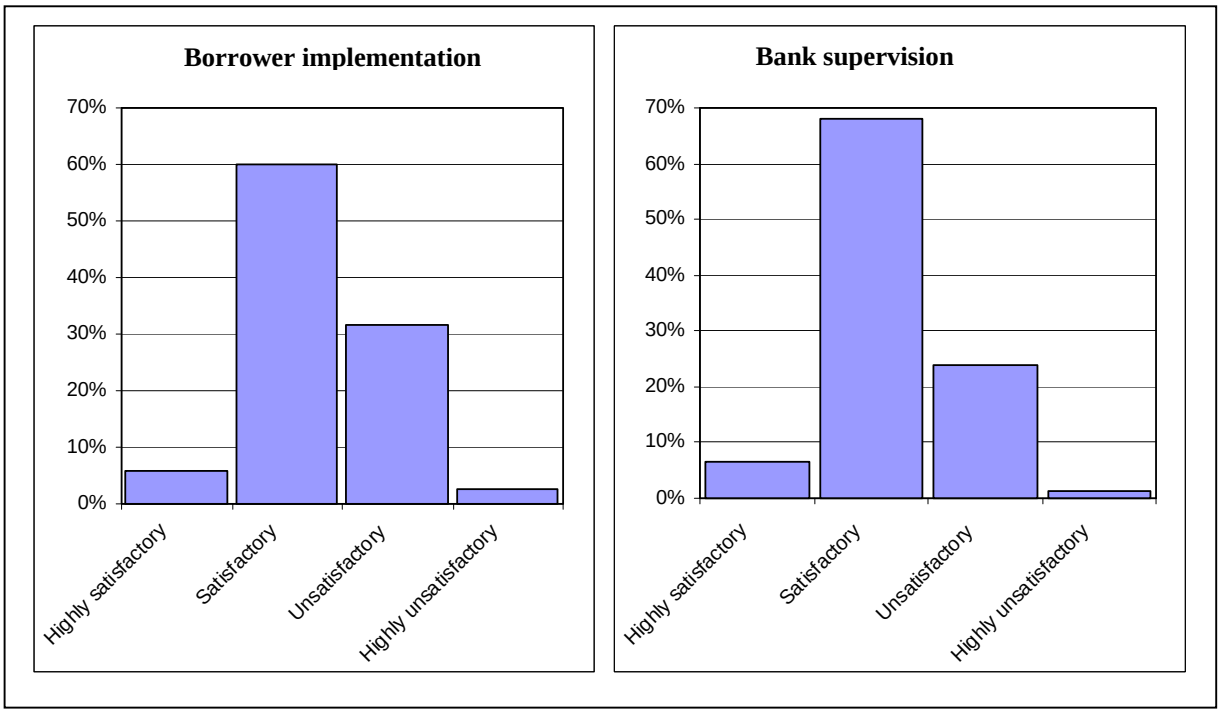
Wintrobe, Ronald (1997): Modern Bureaucratic Theory, in: Dennis Mueller: Perspectives on Public Choice, a Handbook, Cambridge (Cambridge University Press), pp. 429-454.

World Bank (2002): The Role and Effectiveness of Development Assistance, Lessons from World Bank Experience, Development Vice-Presidency Research Paper, Washington (World Bank).

World Bank (2001): OED's Work Program: Consultation and Budget Realignment Support New Program Priorities, in: OED Reach No. 30, http://lnweb18.worldbank.org/oed/oeddoclib.nsf/DocUNIDViewForJavaSearch/00C4C3E0E0D6F8C085256A9C005772B7/$file/OED_budget.pdf, May 10 (accessed on 6/12/04).

World Bank (2003): Independence of OED, in: OED Reach, http://lnweb18.worldbank.org/OED/OEDDocLib.nsf/DocUNIDViewForJavaSearch/50B9B24456B788BE85256CE0006EE9F8/$file/Independence_Reach.pdf, February 24 (accessed on 6/12/04).

Annex 1: Assessment of borrower implementation versus bank supervision, World Bank/OED (1993 – 2003)



*For a single evaluation bank performance was rated "marginally satisfactory". This is included in the category "satisfactory" here.

Source: World Bank/OED.

Annex 2: Operational performance as determinants of overall success*

Ordered probit estimates: BMZ, regression on outcome (1990-1996)

N = 277
LR chi²(5) = 127.72
Prob>chi² = 0
Log likelihood = -470.519
Pseudo R² = 0.1195

Outcome	Coef.	Std.Err.	z	P>|z|	[95% Conf.	Interval]
External conditions	0.166	0.074	2.25	0.024	0.021	0.311
Definition of objectives	0.138	0.074	1.85	0.064	-0.008	0.284
Project planning	0.186	0.086	2.16	0.031	0.017	0.355
Implementation	0.570	0.096	5.96	0.000	0.382	0.757
Project coordination	-0.007	0.088	-0.07	0.940	-0.178	0.165

Ordered probit estimates: BMZ, regression on sustainability (1990-1996)

N = 272
LR chi²(5) = 100.97
Prob>chi² = 0
Log likelihood = -435.479
Pseudo R² = 0.1039

Sustainability	Coef.	Std.Err.	z	P>|z|	[95% Conf.	Interval]
External conditions	0.151	0.075	2.02	0.043	0.004	0.298
Definition of objectives	0.135	0.076	1.78	0.075	-0.014	0.283
Project planning	0.385	0.088	4.36	0.000	0.212	0.557
Implementation	0.196	0.094	2.10	0.036	0.013	0.380
Project coordination	0.035	0.089	0.40	0.689	-0.138	0.209

Ordered probit estimates: World Bank, regression on outcome (PPARs, 1993-2003)

N = 882
LR chi²(4) = 588.76
Prob>chi² = 0
Log likelihood = -958.379
Pseudo R² = 0.235

Outcome	Coef.	Std.Err.	z	P>|z|	[95% Conf.	Interval]
Bank supervision	0.215	0.033	6.53	0.000	0.150	0.279
Borrower preparation	0.230	0.035	6.62	0.000	0.162	0.297
Borrower implementation	0.340	0.035	9.79	0.000	0.272	0.408
Borrower compliance	0.152	0.029	5.26	0.000	0.095	0.208

Ordered probit estimates: World Bank, regression on sustainability (PPARs 1993-2003)

N = 868
LR chi²(4) = 338.85
Prob>chi² = 0
Log likelihood = -754.096
Pseudo R² = 0.184

Sustainability	Coef.	Std.Err.	z	P>|z|	[95% Conf.	Interval]
Bank supervision	0.178	0.034	5.26	0.000	0.111	0.244
Borrower preparation	0.181	0.035	5.15	0.000	0.112	0.249
Borrower implementation	0.219	0.035	6.21	0.000	0.150	0.288
Borrower compliance	0.116	0.030	3.92	0.000	0.058	0.174

*Rated from 1=excellent to 5=fully unsatisfactory for BMZ, and 1=highly satisfactory to 6 highly unsatisfactory for World Bank/OED.