Autoren:
Hans-Dieter Daniel & Rüdiger Mutz

Titel:

What does impact mean for grantees? Cultural consensus in perceived personal, organizational and societal impacts of small-scale funding initiatives of the Volkswagen Foundation

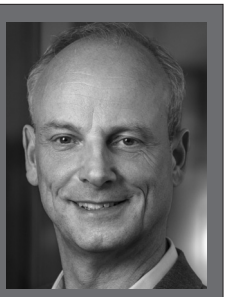
Ursprünglich erschienen in:

Zeitschrift:	Qualität in der Wissenschaft (QiW)
Erscheinungsjahr:	2023
Ausgabe:	1
Jahrgang:	17
Seiten in Druckversion:	24-32
ISSN:	1860-3041
Verlag:	UniversitätsVerlagWebler
Ort:	Bielefeld

Impressum/Verlagsanschrift: UniversitätsVerlagWebler, Bünder Straße 1-3 (Hofgebäude), 33613 Bielefeld

Hans-Dieter Daniel & Rüdiger Mutz

Hans-Dieter Daniel



Rüdiger Mutz

What does impact mean for grantees? Cultural consensus in perceived personal, organizational and societal impacts of small-scale funding initiatives of the Volkswagen Foundation*

Compared to large-scale funding programs, small-scale ones receive little attention in research. For instance, the funding initiatives of the Volkswagen Foundation (VWS) generates impetus for scientific developments. In small-scale funding programs the knowledge of fellows as informants of funding programs and their impacts can be better considered. With Cultural Consensus Theory and the notion of an "impact culture" a methodological framework will be suggested for evaluation purposes of small-scale funding programs. "Impact culture" is defined as the shared knowledge about perceived impacts of a funding program among fellows. In the study commissioned by the VWS two-stage surveys with different levels of impacts were conducted (individual, institutional, societal) for two funding initiatives of the VWS. One single impact culture of each initiative can be found with consistent response set across all levels.

1. Introduction

The central interest of evaluations of funding programs is often directed at their impacts, especially the individual or personal impacts. What impacts do funding programs have on their fellows (e.g., career development, international visibility, publication output)? These efforts are not only supported by the choice of study designs (e.g., control group), but supported by modern statistical methods of causal inference as well, which make a causal interpretation partially possible even in the case of missing control groups (e.g., Jaffe 2002; Mutz et al. 2017; Rubin 2007). Especially large funding programs with a large number of fellows/grantees allow more statistically sophisticated impact evaluations. For example, the FP 7 Marie Curie Actions Interim Evaluation was based on data from a population of 33,400 fellowships (2007-2011) (Public Policy and Management Institute (PPMI) 2013, p. 89), the "Evaluation of NSF's International Research (IRFP)" assumed a population of 1,660 applications (Martinez et al. 2012). The figures were similarly high for the Evaluation of the Swiss National Science Foundation`s Ambizione Funding Scheme with 1,347 applicants and 308 funded projects (Balthasar and Iselin 2014, p. 15).

Large-scale evaluations have not remained without criticism. McGarvey (1979) already pointed out the limitations of large-scale studies. "The literature on evaluation, typically arising out of formally commissioned studies built into curriculum projects at the initial design stage …, may not offer appropriate guidelines for small-scale studies on limited resource. The scope, style and ethics of major evaluations … have attracted disproportionate attention, whereas more modest evaluations must somehow carry their limitations in 'real situations'" (McGarvey 1979, p. 63). Funding programs that are only aimed at a small circle of scientists are left out or, if necessary, abandoned. But it is precisely smaller programs that allow for much faster adaptation to scientific developments.

Since 1962, the Volkswagen Foundation has covered the full spectrum of scientific fields "ranging from humanities and social sciences, through natural sciences and engineering, to life sciences and medicine" (Volkswagen Foundation 2010, p. 1). The Volkswagen Foundation's funding intents to give impetus for new and promising ideas in science and research, to support especially young researchers and to cross both national and disciplinary borders. International funding has developed since the early 1960s. Foreign scholars have been invited to Germany by supporting regional studies

* Funding: This work was supported by the Volkswagen Foundation.

on developing and transition countries. Throughout the years, the focus has shifted from cross-border mobility towards "building sustainable research capacities and infrastructures in developing and transition countries, and from a focus characterized by 'research on' a certain region ... towards 'research with' … " (Volkswagen Foundation 2010, p. 10) and thereby aspiring symmetric partnerships.

A second limitation of these large-scale evaluation studies is that they tend to focus on the personal effects of the programs on the fellow, while institutional or even societal effects are usually ignored. An exception is the "FP7 Marie Curie action interim evaluation" (Public Policy and Management Institute (PPMI) 2013) that distinguished between benefits on the personal, organizational (e.g., host, partner organizations) and system level (e.g., strategic priorities for EU). These evaluation studies, which focus strongly on personal effects, are in contrast to developments such as the open science movement, in which the societal benefits of research are also demanded. "… open science is defined as an inclusive construct that combines various movements and practices aiming to make multilingual scientific knowledge openly available, accessible and reusable for everyone, to increase scientific collaborations and sharing of information for the benefits of science and society, and to open the processes of scientific knowledge creation, evaluation and communication to societal actors beyond the traditional scientific community." (UNESCO 2021, p. 7). In this respect, smaller funding programs have an advantage, since goals and intervention logics can be readjusted more easily when conditions change compared to large-scale ones with a long-term perspective.

A third limitation results from the methodological approach. Usually, a so-called intervention logic is formulated for a research funding program that connect the input of funding programs with their possible short-term results, medium-term and long-term effects (Kellogg Foundation 2004; McLaughlin/Jordan 2015). These intervention logics are used to formulate top-down evaluation criteria. The perspective of the beneficiaries, who have directly experienced the funding and who as experts could best provide information about impacts on different levels, is left out. Ecker et al. (2022) examined a number of fellowship and research cooperation programs of the BMBWF (Austrian Federal Ministry of Education, Science and Research), including outgoings and incomings. Again, only personal expected and achieved goals were asked (Ecker et al. 2022), but these were given without a preceding exploratory phase (Ecker et al. 2022, p. 58f). Personal goals, teaching-related goals, study-related goals, and research-related goals were distinguished, operationalized with three to seven statements in a questionnaire.

From the variety of initiatives, our study focuses on the two international ones (Daniel et al. 2021, p. 20). The Volkswagen Foundation's funding initiative "Postdoctoral Fellowships in the Humanities at Universities and Research Institutes in the U.S. and Germany" (HUMAN) granted postdoctoral fellowships in the humanities at universities and research institutions in the U.S. and Canada to researchers from Germany (i.e., the outgoing dimension) (Volkswagen Foundation 2018). In 2012, the Andrew W. Mellon Foundation (New York) started financing the reciprocal dimension of the initiative i.e. postdoctoral fellows based at American universities and research institutions who intended to pursue a research stay in Germany (i.e. the incoming dimension). The initiative is now completed; the last grants were awarded for the academic year 2019/2020.

Under the funding initiative "Knowledge for Tomorrow – Cooperative Research Projects in sub-Saharan Africa" (CAPACITY), the Volkswagen Foundation supported researchers from Africa to conduct research projects in their home countries. The fellowships were aimed at postdoctoral fellows based at universities or research institutions in sub-Saharan Africa who intended to conduct a research project there (Daniel et al. 2021, p. 20). However, any postdoctoral research project funded under this scheme had to be carried out in collaboration with a German academic partner institution. Postdoctoral fellows had the possibility to get funding for a maximum of eight years, if a person benefitted from the junior (3 years), senior (3 years) as well as the extension (2 years) funding. Projects were awarded in the program lines Neglected Tropical Diseases, Natural Resources, Engineering Sciences, Social Sciences and Humanities. The initiative has been phasing out since 2018 when the last calls (restricted to already funded fellows only) were launched.

The aim of this paper is to use the concept of "impact culture" as a shared knowledge about impacts of a funding program to illustrate the possibilities of impact evaluation for small-scale funding programs, with surveys on two funding initiatives of the Volkswagen Foundation as an example. It is assumed that grantees are experts of a funding program and that their implicit knowledge about the impact of a program can be tapped through panel surveys. The theoretical background is the Cultural Consensus Theory (CCT), which has its roots in cultural anthropology (e.g., Romney et al. 1986; Weller 1984) and statistically strongly developed in the last years (e.g., Anders et al. 2014; Anders et al. 2018; Aßfalg/Klauer 2020). Initial results on CCT have already been published in a monograph (Daniel et al. 2021, p. 72; pp. 210-15).

In the following, the notion of an "impact culture" is derived from CCT.

2. Impact Culture and Cultural Consensus Theory

Funding programs are designed to have a variety of impacts on the fellows or grantees. It is assumed that despite different research foci or scientific disciplines, a funding program nevertheless triggers a uniform network of impacts in the grantees, about which only the fellows can make a statement, a kind of "tacit knowing" (Polyani 1966). This implicit shared knowledge about impact is defined here as "impact culture". The impact culture needs not be identical to the impact model formulated in an intervention logic of a program. For example, negative impacts are usually not mentioned in intervention logics.

The most natural way to tap this implicit knowledge as consensus about perceived effects is to interview the grantees themselves and to determine the consensual items from the survey results using statistical methods. This is done by the CCT as part of "Cognitive Psychometrics" described by William Batchelder (Batchelder 2016). The CCT has roots in cognitive anthropology (e.g., Romney et al. 1986). In anthropology, the term "culture" is used in different ways. What these different notations of "culture" have in common is "... that culture (as anthropologists use the term) is best understood as a heterogeneously distributed collective system of pragmatic knowledge." (Kronenfeld 2018, p. 2).

The CCT models the response behavior of the respondents, here the fellows, by means of a "test theory without answer key" (Batchelder/Romney 1988; Oravecz et al. 2015). In a normal knowledge test, the correct answers are known in advance. In CCT, the "true" answer key Z is reconstructed as a latent variable from the answers to the items, which in the simplest case are binary (e.g., "agree or disagree", "occur or not occur"). The "true" response key Z represents the "impact culture" or network of impacts about which there is consensus in the sample of fellows as informants. Four components are distinguished in the psychometric model (Batchelder et al. 2018, p. 3):

1. *Competency* (Θ): While there is consensus on perceived impacts of the program, informants differ in their knowledge about these shared impacts. It indicates the proportion of shared impacts each informant knows. For example, fellows whose funding dates back a few years might make more accurate statements than fellows whose funding dates back years.
2. *Bias* (g): When informants do not know the correct answer, i.e., whether an item is consensual or not, they guess the correct answer with probability g. "No bias ... in true-false a bias of 1/2 means that if the informant does not know the answer to the question that they will choose either alternative with equal probability." In this case there is no bias or response set (Romney et al. 1986, p. 317). If the bias deviates in the direction of 1.0, then "acquiescence" is present, informants are more likely to agree with the item, regardless of whether it is a consensual item or not.
3. *Item difficulty* (δ): The items vary in their difficulty as a measure of cultural salience. As a result, some items generate higher consensus than others. The lower the item difficulty, the easier it is for informants to identify an item belonging to the "impact culture".
4. *Subgroups:* There may be subgroups of informants, each with different consensus on a set of impacts, i.e., different "impact cultures" are present.

CCT is not identical with the concept of "wisdom of the crowd" (Sunstein/Hastie 2015, p. 143f), i.e., different response frequencies of impacts alone are not sufficient to identify the "impact culture". Responses from individuals with high cultural competency are weighted more heavily than responses from individuals with low cultural competency. Additionally, response biases are considered. Another advantage of CCT is that it does not require an explicit definition of impact, as is usually necessary in research evaluation (e.g., Donovan 2011; Penfield et al. 2014). The statistical model will be outlined in the section "Data and Methods".

3. Research questions

In detail, the focus of the paper is on the following four research questions:

- Are there one or more impact cultures that encompass the personal, institutional, and societal impacts of a funding initiative?
- How can the impact cultures be characterized?
- To what extent are fellows capable of making a statement about the different impact levels? Is this competency a trait, which does not differ much across the different impact levels (personal, institutional, societal)?
- To what extent do fellows show response sets, e.g., acquiescence? Is this response tendency a trait, which does not differ much across the different impact levels?

4. Data und Methods

4.1. Sample of fellows

The sample of individuals is taken from a comprehensive study of former fellows of Alexander von Humboldt Foundation funding programs and Volkswagen Foundation funding initiatives (Daniel et al. 2021). However, only the fellows of the two funding initiatives of the Volkswagen Foundation are included in the analysis which were aimed at a comparably small number of fellows compared to the funding programs of the Alexander von Humboldt Foundation and thus preclude the application of common multivariate statistical analyses that require large numbers of cases.

A total of 90 former fellows (36 "incomings", 54 "outgoings") for the funding initiative "Humanities at Universities and Research Institutes in the U.S. and in Germany (together with the Andrew W. Mellon Foundation)" were awarded in the year 2008 to 2018 (HUMAN). The two groups "incomings" and "outgoings" are combined, which is consistent with the intention of the funding initiative. The response rate for HUMAN-incomings amounts to 78%, for HUMAN-outgoings 89% (overall N = 76 respondents).

A total of 76 former fellows of the funding initiative "Knowledge for Tomorrow – Cooperative Research Projects in sub-Saharan Africa (neglected tropical diseases, humanities, social sciences, livelihood management, natural resources, and engineering)" were awarded (CAPACITY). The response rate for CAPACITY was 74% (N = 56 respondents). Overall, the response rates are high if the response rates of comparable studies are taken into account. For example, the response rate of the evaluation study of fellowship and research cooperation programs of the BMBWF (Austrian Federal Ministry of Education, Science and Research) was on average 33.6% (incoming) and 35.9% (outgoing) (Ecker et al. 2022, p. 28; Table 11).

4.2. Two-stage survey

A two-stage online survey of former fellows was conducted in the year 2019. In the first round open-ended questions were asked about impacts and benefits at the personal, institutional and at the societal level.

In the second round close-ended questions were asked. In order to stimulate respondents to provide as broad answers regarding the impacts and benefits they perceived and to clarify what was meant by "impact" or "benefit" two sets of three different examples, for instance, retrieved from academic literature were formulated. The sets were randomly assigned to one part or the other part of the respondents. Based on the analysis of answers to open-ended questions from the first round, lists with impacts/benefits/aspects of added value, shortly "impact items", were drawn up for the three different levels (personal, institutional, societal). The societal level was divided into "research system" and "other aspects of societal life" with respect to the home country (HUMAN: Germany, CAPACITY: sub-Saharan Africa). Up to 19 till 43 impact items in English were provided in the questionnaire for the second round. The list with impact items were given in random order in each questionnaire to avoid systematic position effects (e.g., fatigue). The respondents were asked to indicate whether each item had occurred in their case. At the end of the questionnaire, questions were asked with regard to career development and socio-demography. Only quantitative results of the second round are reported here.

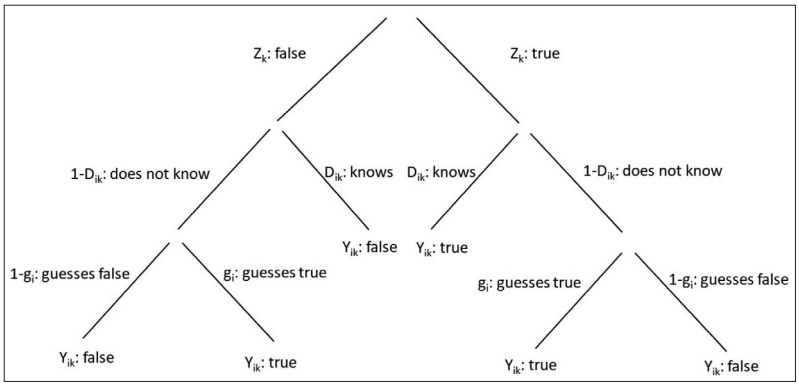
4.3. Cultural Consensus Theory (CCT) and Latent Class Analysis (LCA)

CCT is an application of the multinomial processing trees (Batchelder/Riefer 1999). From the answers to binary questions the set of items about which there is consensus in a sample of informants is inferred (Anders/Batchelder 2012; Anders et al. 2014; Anders et al. 2018; Aßfalg/Klauer 2020; Batchelder/Romney 1988; Batchelder/Anders 2012; Batchelder et al. 2018).

Comparable to a knowledge test, the key Z is identified from the answers with the correct answers ("test theory without answer key"). **Figure 1** shows the processing tree for the response of an informant *i* to a single impact item *k*. Each path is identical to a sequence of decisions leading to the final response Y_{ik}, where Y_{ik} = true ("occurred") or false ("not occurred") (see Batchelder/Anders/Oravecz 2018, p.8). D_{ik} represents the probability that informant *i* knows, whether an impact item *k* is true and belongs to the set of consensus items or false and does not belong to this set. D_{ik} is a function of the difficulty δ_k of an impact item and the overall competency Θ_i of an informant to identify the consensus items, which can be represented by a Rasch model, a psychometrical test model:

$$D_{ik} = \frac{\theta_i(1-\delta_k)}{\theta_i(1-\delta_k)+\delta_k(1-\theta_i)} \quad (1)$$

Fig. 1: Processing tree for the response of an informant *i* to impact item *k*.



Each path is identical to a sequence of decisions leading to the final response Y_{ik}: Y_{ik} = true ("occurred") or false ("not occurred") (see Batchelder/Anders/Oravecz 2018, p. 8).

For example the response probability for Y_{ik} given z_k = true (=1) and y_{ik} = true (=1) can be derived according to **Figure 1** as follows: $P(Y_{ik} = y_{ik}|Z_k=z_k, g_i, D_{ik}) = D_{ik} + (1-D_{ik})\, g_i$. Either the informant knows that the item *k* belongs to the set of consensus items (D_{ik}) or he or she does not know it ($1-D_{ik}$) and guesses it with the probability g_i. The probability D_{ik} is the higher, the lower the item difficulty δ_k of an impact item and/or the higher the competency Θ_i of an informant in comparison to the item difficulty δ_k. The model parameters can be estimated by a Bayesian statistical model, which is available in the R package CCTpack (Anders 2017).

Besides the model estimation, it is important to assess the adequacy of the chosen model. Bayesian posterior predictive model checking was carried out after random samples are drawn from the posterior distribution of an estimated parameter (Batchelder et al. 2018, p. 38f). In addition to the Deviance Information Criterion (DIC) scree plots (plots of eigenvalues) might help to find the optimal number of factors (i.e., cultures). "… the culture number check is satisfied when the black line (scree plot of the data) is overlaying or highly similar to the gray lines (many scree plots produced from multiple data sets predicted by the model fit)." (Batchelder et al. 2018, p. 54). The Variance Dispersion Index (VDI) helps to decide whether item heterogeneity (differences in item difficulties) should be included in the model (Anders/Batchelder 2012, p. 457). "Then the item difficulty check is satisfied for the GCM when the VDI statistic … for each culture lies within the 10th to 90th percentile of the distribution of VDI statistics, which is calculated from these many predicted data sets of the model fit." (Batchelder et al. 2018, p. 54).

5. Results

5.1. Model comparison CCT

The first step is to check whether the items at the different levels have the same item difficulties and item biases, and whether there are different latent classes reflecting different impact cultures. Unfortunately, the latter question can only be answered for the personal impacts due to estimation problems. The model comparison with BIC revealed that the items are strongly

Table 1: Model comparison for the two funding initiatives "*Postdoctoral Fellowship in the Humanities at Universities and Research Institutes in the U.S. and Germany incoming and outgoing fellows*" (HUMAN) and "*Knowledge for Tomorrow – Cooperative Research Projects in Sub-Saharan Africa*" (CAPACITY)

Impact level	MNo	Latent Classes	Item difficulties	Item biases	HUMAN pD	HUMAN DIC	CAPACITY pD	CAPACITY DIC
Personal	0	1	=	=	57.0	3,766.6	69.1	2,725.1
	1	1	≠	=	131.0	3,617.8	124.5	2,436.3
	2	1	=	≠	153.2	3,531.0	104.1	2,566.2
	3	1	≠	≠	222.7	3,351.3	175.2	2,321.3
	4	2	≠	≠	987.4	4,038.7	252.1	2,367.1
Institutional	0	1	=	=	89.6	1,377.4	62.7	1,345.6
	1	1	≠	=	132.9	1,242.8	100.2	1,274.1
	2	1	=	≠	693.1	1,871.1	108.2	1,198.8
	3	1	≠	≠	1,216.3	1,216.3	182.6	1,151.6
Societal	0	1	=	=	81.8	2,214.5	24.7	2,434.5
	1	1	≠	=	136.3	1,936.0	88.0	2,335.2
	2	1	=	≠	124.1	1,976.8	86.2	2,189.1
	3	1	≠	≠	186.9	1,907.4	182.2	2,093.8
All	0	1	=	=	88.4	7,365.4	79.3	6,462.4
	1	1	≠	=	180.4	6,736.1	183.8	5,911.9
	2	1	=	≠	154.6	6,824.2	116.9	5,878.9
	3	1	≠	≠	255.0	6,340.7	237.8	5,354.2

Note: "=" parameters are restricted to be equal, "≠" parameters are not restricted, "pD" effective number of parameters, "DIC"

heterogeneous in the item parameters (difficulty, bias) (**Table 1**, M_3). The perceived impact items show different cultural salience and biases. The strong assumption of CCT cannot be hold. "Each respondent has a fixed 'cultural competence' over all questions. Cultural competence is defined as the proportion of the cultural questions for which the correct answer is known by the respondent. This strong assumptions says that questions are all of the same level." (Romney 1999, p. 107). A weaker assumption is valid in our case "… that the respondents who do better on one subset of questions will do better on another subset of questions" (Romney 1999, p. 107).

The thesis of multiple cultures can be rejected at least at the personal level. To differentiate between two subgroups ("incomings", "outgoings") in the funding initiative HUMAN is, therefore, not necessary. The Bayesian predictive check revealed (Figure 2) that the VDI statistical "lies within the 10th to 90th percentile of the distribution of the VDI statistics" (Batchelder et al. 2018, p. 54). The plot of eigenvalues (**Figure 2**), which are ordered from largest to smallest, is used to determine the number of factors (scree test), where the eigenvalues lift off from the randomly distributed factors. The scree tests also do not justify more than one factor or culture.

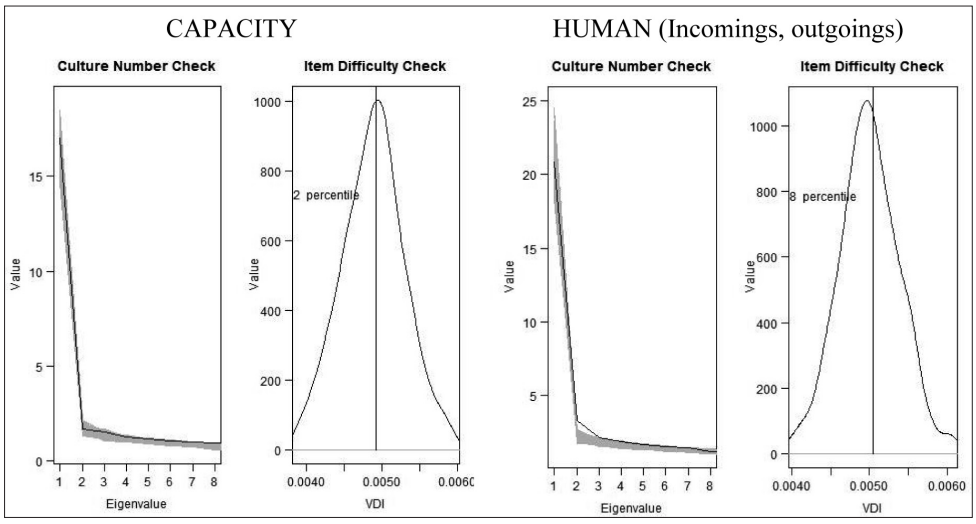
5.2. Impact cultures for the two funding initiatives

The impact cultures of the two funding initiatives are presented below, based on the analysis of all items at all levels. A Z score of 1.00 was chosen as the criterion for selecting items for the "impact culture", resulting in a slightly lower number of consensus items than in the original study of Daniel et al. (2021, p. 212) with a less strict criterion of Z > 0.80.

With regard to the funding initiative HUMAN (**Table 2**), there are 15 items on the personal level (out of a total of 43 items in the questionnaire), 0 items on the institutional level (out of a total of 19 items), and 1 item on the societal level (out of a total of 35 items), on which there is consensus among the fellows surveyed. This set of items constitutes the impact culture. The higher the level, the fewer impact items there is consensus on among the fellows.

The following two impacts are seen as particularly important with low item difficulty and thus high cultural salience at the *personal level*: "I had (more) time to concentrate on research" and "I increased my visibility in international research". Overall, fellows tend to emphasize aspects of their own research in personal impacts (e.g., concentrate on research, visibility, reputation, advance my career, academic confidence, personal development).

Fig. 2.: Bayesian posterior predictive checks for all levels, separately for the two funding initiatives and one culture



Shared items on the *institutional level* are completely missing, although they were indicated in the questionnaire, e.g., "I helped increase the institution's visibility", "Other projects at the institution benefited from my contribution", "I encouraged other researchers at the institution to apply for international fellowships".

On the *societal level*, there is consensus only with regard to the item "I conveyed my favorable impressions of my host country to friends, colleagues or family" Other items concerning the "research system" and "societal life" are missing, although

Table 2: Impact culture (consensual items) for the funding initiative HUMAN on personal (IND) and societal level (SOC), sorted in descending order by "Yes %" for each impact level (N = 76 fellows)

Impact level	Label	Item difficulty	Yes %
IND	I had (more) time to concentrate on research.	.20	89.47
IND	I increased my visibility in international research.	.23	88.16
IND	My reputation increased.	.44	77.63
IND	I advanced my career in research.	.42	77.63
IND	I increased my academic confidence.	.42	77.63
IND	The research stay meant a lot for my personal development.	.42	77.63
IND	I had access to expertise, human resources or intellectual community.	.53	72.37
IND	I improved my intercultural skills.	.54	71.05
IND	I broadened my network by new collaborative partners.	.60	69.74
IND	I increased my independence as a researcher.	.58	68.42
IND	I sharpened my research profile.	.62	65.79
IND	I improved my publication performance.	.67	64.47
IND	I increased my competitiveness on the job market.	.65	64.47
IND	I improved my language skills.	.65	64.47
IND	I broadened my research spectrum.	.67	63.16
SOC	I conveyed my favorable impressions of my host country to friends, colleagues or family.	.56	69.74

they were marked as "occurred", however, with low frequencies, e.g., "I informed German researchers about research systems of other countries", "The project increased the international visibility of research conducted in Germany", "The project strengthened international research networks of Germany", "My research stay had a positive impact on Germany's image abroad" and "I was involved in public outreach activities".

Interestingly, *negative impacts* are not part of the impact culture. These items are ticked as applicable in the questionnaire by only a few grantees, e.g., "My research network in my home country worsened because of my research stay abroad", "I faced competition rather than cooperation", "The re-integration in the research system in my home country was difficult after the stay abroad".

The results for the CAPABILITY funding initiative are different (**Table 3**). 29 of a total of 93 items specified in the questionnaire are assigned to the impact culture, i.e., 18 of 43 items at the personal level, 5 of 19 items at the institutional level, and 6 of 31 items at the societal level. The items on the societal level refer to the research system and societal life in sub-Saharan Africa and not to Germany as in the funding initiative HUMAN.

On the *personal level*, the impact culture is very broad with a focus on improvement of skills (e.g., mentoring, management), increase of capacity (e.g., high quality research, leadership) and visibility. As the funding initiative aims to strengthen the capacity of the fellows in the home country, the consensus impacts are in line with the intention of the funding initiative.

At the *institutional level*, the focus is on the impact on research of the institution at which the fellow works, with the consensus items "I encouraged others to apply for international fellowships", "I increase the institution's visibility", "I improve the institution's publication performance" or "I taught or advised (PhD) students at the institution". The impact "I conducted research on global issues (e.g. climate change)" or "I introduced new lines of enquiry, methods, or theories to research in sub-Saharan Africa" does not play an important role.

At the *societal level*, only impacts that affect the research system in the home country are mentioned. There is no sufficient consensus on impacts that affect "other aspects of societal life in sub-Saharan Africa", for example, "I conveyed my favorable impressions of Germany to friends, colleagues or family", "The research project helped form a network with different societal stakeholders" or "I reached a position in academia where I can influence society".

Negative impacts, for instance, "My research network in my home country worsened because of my fellowship" or "Finding a job after the end of the fellowship was more difficult than I expected." do not play a role or are hardly ever ticked off.

5.3. Correlations among bias and competency across levels

In the following, we will answer the questions to what extent fellows are comparably able to make valid statements about the different levels (personal, institutional, societal) and whether the response tendencies/biases remain constant across the levels. Pearson-Bravais correlations of the two parameters "competency" and "bias" were calculated across the different levels (**Table 4**, **Table 5**). While both funding initiatives show medium to high positive correlations for bias across the different levels, the correlations for competency are rather low to moderate. Unlike competency, the bias component can be understood as a trait, a response tendency that is consistently evident in the sets of items. The correlation between bias and competency is inconsistent and specific to the funding initiative.

In addition to the person parameters, the item parameters, which were composed of the CCT analyses of the three levels, were also correlated with the item parameters from the CCT analyses of all items at all levels (**Table 6**). Interestingly, there was a high agreement with regard to the parameters Z and item difficulty independent of the funding initiative. The item parameters are thus independent of whether single or overall analyses across all items are conducted.

6. Discussion

Small-scale funding programs have little chance of becoming the subject of scientific (evaluation) studies and, if studies are conducted, results are not published in scientific journals. Statistical data analyses are usual-

Table 3: Impact culture (consensual items) for the funding initiative CAPACITY on personal (IND), institutional (INST) and societal level (SOC), sorted in descending order by "Yes %" for each impact level (N = 56 fellows)

Impact level	Label	Item difficulty	Yes %
IND	I increased my capacity to conduct high quality research (methods, techniques, approaches, etc.)	.07	94.64
IND	I broadened my network by new collaborative partners.	.09	92.86
IND	I improved my research management skills.	.15	89.29
IND	I increased my visibility in international research.	.16	87.50
IND	I improved my leadership capacity.	.17	87.50
IND	I improved my mentoring skills.	.16	87.50
IND	I advanced my career in research.	.19	85.71
IND	I sharpened my research profile.	.22	83.93
IND	I increased my independence as a researcher.	.22	83.93
IND	I improved my ability to acquire further funding.	.22	83.93
IND	I improved my publication performance.	.26	82.14
IND	The fellowship meant a lot for my personal development.	.25	82.14
IND	My reputation increased.	.28	80.36
IND	I increased my academic confidence.	.34	78.57
IND	I conducted interdisciplinary research.	.38	76.79
IND	I increased my co-authorship network.	.36	75.00
IND	I broadened my research spectrum.	.42	73.21
IND	I moved into a more senior managerial or research role.	.54	66.07
INST	I encouraged other researchers at the institution to apply for international fellowships.	.10	92.86
INST	I helped increase the institution`s visibility.	.28	80.36
INST	I helped improve the institution`s publication performance.	.37	75.00
INST	I taught or advised (PhD) students at the institution.	.37	75.00
INST	The institution benefited from a continued collaboration with me.	.51	69.64
SOC	I conducted research relevant to the development of my home country.	.18	87.50
SOC	The project increased the international visibility of research conducted in sub-Saharan Africa.	.26	82.14
SOC	I conducted research on pertinent issues affecting local populations.	.27	80.36
SOC	I helped build research capacity in sub-Saharan Africa.	.30	78.57
SOC	I raised awareness of research opportunities available in Germany.	.38	76.79
SOC	The project strengthened international research networks in sub-Saharan Africa.	.48	71.43

ly not worthwhile due to the small number of cases, and the generalizability of the results is considered questionable. With its small-scale funding initiatives the Volkswagen Foundation successfully demonstrates how smaller programs can also provide effective support for fellows and respond very quickly to certain social developments, generate impetus for scientific developments.

With the Cultural Consensus Theory and the concept of an "impact culture", quantitative analyses of the impacts of funding programs can be made, especially for funding programs that are only aimed at a small group of fellows. The term "impact culture" refers to the shared knowledge about the perceived effects of a funding program. Not only the shared knowledge of impacts can be empirically identified, but also the competency and the tendency of bias of fellows. CCT was used here for the first time in the field of research evaluation, and it is the first time in the scientific literature that CCT parameters have been correlated across different sets of items to identify traits.

The results of the CCT themselves can be used for evaluation purposes or the consensus impacts can be further operationalized as possible evaluation criteria in subsequent evaluations.

The approach chosen here is participant-oriented in two senses. Former fellows were given the opportunity in a first round to name possible impacts of fellowship programs and additionally had the opportunity in a second round to indicate to what extent the impacts mentioned in the first round (or in the scientific literature) had actually occurred in their case.

With regard to the research questions, the following results can be formulated:

- *Number of "impact cultures":* In the two funding initiatives, only one impact culture was evident in each case (i.e. only one impact culture, for example, for incomings and outgoings in the funding initiative HUMAN).
- *Description of the "impact cultures":* For both funding initiatives, the higher the level (personal, institutional, societal), the fewer impact items are part of the impact culture, especially for the funding initiative HUMAN. There are similarities and differences between the two funding initiatives. For the funding initiative HUMAN unlike CAPACITY there are no consensus items on the level of institutions. With respect to personal impacts fellows tend to emphasize aspects of their own research (e.g., concentrate on research, visibility, reputation, advanced my career, academic confidence, personal development). Regarding CAPACITY the range of personal impact is very broad with a focus on improvement of skills (e.g., mentoring, management), increase of capacity (e.g., high quality research, leadership) and visibility. At the societal level, only impacts that affect the research system in the home country are part of the shared impact culture, not other aspects of societal life in sub-Saharan Africa. Negative impacts were not part of the impact cultures in either funding initiative.
- *Fellows' competency:* The competency to identify the consensus items is not comparable across the personal, institutional and societal level, if CCT analyses were done for the different levels separately. The correlations were heterogeneous. There is no overall competency of fellows in the sense of a trait.

Table 4: Correlation between competency and bias parameters, estimated on the personal (IND), institutional (INST), societal (SOC) level and all (ALL) levels (logit-scale) for the funding initiative HUMAN (N = 76, model M_3).

Area	Impact level	Mean (STD)	Competency (Θ) IND	INST	SOC	ALL	Bias (g) IND	INST	SOC	ALL
Competency	IND	-0.15 (0.41)	1.00	.04	-.01	.84*	.07	.07	.07	.13
	INST+	0.50 (0.33)	-	1.00	.41*	.36*	-.38*	-.48*	-.43*	-.52*
	SOC+	0.04 (0.50)	-	-	1.00	.33*	-.49*	-.43*	-.41*	-.58*
	ALL	-0.05 (0.46)	-	-	-	1.00	.00	-.20	-.20	-.09
Bias	IND	-0.03 (1.08)	-	-	-	-	1.00	.48*	.47*	.90*
	INST	0.02 (1.02)	-	-	-	-	-	1.00	.58*	.73*
	SOC	0.03 (0.78)	-	-	-	-	-	-	1.00	.76*
	ALL	-0.13 (1.03)	-	-	-	-	-	-	-	1.00

*p<.05 + one outlier was eliminated (N = 75)

Table 5: Correlation between competency and bias parameters, estimated on the personal (IND), institutional (INST), societal (SOC) level and all (ALL) levels (logit-scale) for the funding initiative CAPACITY (N = 56, model M_3).

Area	Impact level	Mean (STD)	Competency (Θ) IND	INST	SOC	ALL	Bias (g) IND	INST	SOC	ALL
Competency	IND	-0.16 (0.79)	1.00	.48*	.29*	.88*	.52*	.51*	.39*	.50*
	INST	-0.44 (0.52)	-	1.00	.23	.68*	.00	.02	.09	.03
	SOC	-0.68 (0.32)	-	-	1.00	.61*	.31*	.21	.10	.19
	ALL	-0.46 (0.60)	-	-	-	1.00	.41*	.38*	.23	.33*
Bias	IND	-0.12 (0.81)	-	-	-	-	1.00	.82*	.70*	.90*
	INST	0.01 (1.22)	-	-	-	-	-	1.00	.74*	.90*
	SOC	-0.28 (1.33)	-	-	-	-	-	-	1.00	.93*
	ALL	-0.07 (1.15)	-	-	-	-	-	-	-	1.00

*p<.05

Table 6: Correlation of item parameters (true answer vector Z, item difficulty on logit scale) estimated for each single level separately (personal, institutional, societal) and estimated across all levels for the two funding initiatives HUMAN and CAPACITY.

	Number of items	Separated analysis of each level	Overall analysis across all levels	r
HUMAN	89	Z	Z	.96*
		Item difficulty	Item difficulty	.97*
CAPACITY	97	Z	Z	.98*
		Item difficulty	Item difficulty	.97*

- *Fellows' bias:* Unlike competency, bias is moderately to strongly positively correlated across levels when CCT is conducted separately for the different levels and can also be interpreted in terms of a disposition. For example, there are fellows who tend to tick more items (acquiescence) than the average, not only at the personal level, but also at the institutional and societal levels.
- *Item parameter:* The item parameters characterize the impact items (Z, item difficulty). There are high correlations among item parameters, which were estimated and composed from single-level analyses, and item parameters estimated by the overall analyses across all levels. Item parameters are more or less independent of the modus of analysis (single-level, total).

Nevertheless, the approach cannot completely eliminate the usual limitations of small-scale studies, for example, the problem of the extent to which the results can be generalized to other fellows and programs. It could be shown, however, how more generalized statements are possible if more than one funding initiative is investigated. Furthermore, the statistical approach requires a certain methodological know-how and practical applications are lacking so far.

References

Anders, R. (2017): CCTpack: Cultural Consensus Theory applications to data. R package version 1.5.2.

Anders, R./Batchelder, W. H. (2012): Cultural consensus theory for multiple consensus truths. In: Journal of Mathematical Psychology, 56 (6), pp. 452-69.

Anders, R./Oravecz, Z./Batchelder, W. H. (2014): Cultural consensus theory for continuous responses: A latent appraisal model for information pooling. In: Journal of Mathematical Psychology, 61, pp. 1-13.

Anders, R./Alario, F. X./Batchelder, W. H. (2018): Consensus analysis for populations with latent subgroups: Applying multicultural consensus theory and model-based clustering with CCTpack. In: Cross-Cultural Research, 52 (3), pp. 274-308.

Aßfalg, A./Klauer, K. C. (2020): Consensus theory for multiple latent traits and consensus groups. In: Journal of Mathematical Psychology, 97, article nr. 102374.

Balthasar, A./Milena, I. (2014): Evaluation of the Swiss National Science Foundation's Ambizione funding scheme. Swiss National Science Foundation.

Batchelder, W. H. (2016): Cognitive psychometrics. In: Haupt, J. W./Blaha, L. M. (eds.): Mathematical Models of Perception and Cognition. New York: Psychology Press, pp. 245-166.

Batchelder, W. H./Romney, A. K. (1988): Test theory without an answer key. In: Psychometrika, 53 (1), pp. 71-92.

Batchelder, W. H./Riefer, D. M. (1999): Theoretical and empirical review of multinomial process tree modeling. In: Psychon Bull Rev, 6 (1), pp. 57-86.

Batchelder, W. H./Anders, R. (2012): Cultural Consensus Theory: Comparing different concepts of cultural truth. In: Journal of Mathematical Psychology, 56 (5),pp. 316-32.

Batchelder, W. H./Anders, R./Oravecz, Z. (2018): Cultural Consensus Theory. In: Wixted, J. T. (ed.): Stevens' Handbook of Experimental Psychology and Cognitive Neuroscience. New York: John Wiley & Sons, pp. 1-64.

Daniel, H.-D./Bobokova, J./Mutz, R. (2021): Funding Programmes and Initiatives for Internationally Mobile Postdocs: Perceived Impacts on Individuals, Institutions and Society. Bonn.

Donovan, C. (2011): State of the art in assessing research impact: Introduction to a special issue. In: Research Evaluation, 20 (3), pp. 175-79.
Ecker, B./Pfaffel, A./Régent, V. (2022): Mehrwert der Stipendien- und Forschungskooperationsprogramme des BMBWF [Added value of the BMBWF's fellowship and research cooperation programs]. Wien: WPZ Research GmbH.
Jaffe, Adam B. (2002): Building programme evaluation into the design of public research support programmes. In: Oxford Review of Economic Policy, 18 (1), pp. 22-34.
Kellogg Foundation (2004): WK Kellogg Foundation logic model development guide. https://wkkf.issuelab.org/resource/logic-model-development-guide.html
Kronenfeld, D. B. (2018): Culture as a System – How We Know the Meaning and Significance of What we Do and Say. London: Routledge.
Martinez, A. et al. (2012): Evaluation of NSF's international research fellowship program. National Science Foundation.
McGarvey, J. E. B. (1979): 1. Some problems in small-scale evaluation. In: British Educational Research Journal, 5 (1), pp. 63-68.
McLaughlin, J. A./Jordan, G. B. (2015): Using logic models. In: Newcomer, K. E./Hatry, H. P./Wholey, J. S. (eds.), Handbook of practical program evaluation. New Jersey: John Wiley & Sons, pp. 62-87.
Mutz, R./Wolbring, T./Daniel, H.-D. (2017): The effect of the "very important paper" (VIP) designation in Angewandte Chemie International Edition on citation impact: A propensity score matching analysis. In: Journal of the Association for Information Science and Technology, 68 (9), pp. 2139-53.
Oravecz, Z./Anders, R./Batchelder, W. H. (2015): Hierarchical Bayesian modeling for test theory without an answer key. In: Psychometrika, 80 (2), pp. 341-64.
Penfield, T., et al. (2014): Assessment, evaluations, and definitions of research impact: A review. In: Research Evaluation, 23 (1), pp. 21-32.
Polyani, M. (1966): The Tacit Dimension. Chicago: Doubleday & Company.
Public Policy and Management Institute (PPMI) (2013): FP7 Marie Curie actions interim evaluation. Final report. European Commission.
Romney, A. K. (1999): Culture consensus as a statistical model. In: Current Anthropology, 40, pp. 103-15.
Romney, A. K./Weller, S. C./Batchelder, W. H. (1986): Culture as consensus: A theory of culture and informant accuracy. In: American Anthropologist, 88 (2), pp. 313-38.
Rubin, Donald B. (2007): Statistical inference for causal effects, with emphasis on applications in psychometrics and education. In: Rao C. R./Sinharay S. (eds.): Handbook of Statistics, 26. New York: Elsevier, The Netherlands, pp. 769-800.
Sunstein, C. R./Hastie, R. (2015): Wiser – Getting Beyond Groupthink to Make Groups Smarter. Boston: Harvard Business Review Press.
UNESCO (2021): 'UNESCO recommendation on Open Science', in UNESCO (ed.). Unpublished.
Volkswagen Foundation (2018): Postdoctoral fellowships in the humanities at universities and research institutes in Germany. Information for applicants 97.
Weller, S. C. (1984): Consistency and consensus among informants: Disease concepts in a rural Mexican village. In: American Anthropologist, 86 (4), pp. 966-75.

■ **Hans-Dieter Daniel,** Prof. Dr., Assoziierte Professur für empirische Hochschulforschung, Psychologisches Institut, Universität Zürich, E-Mail: hans-dieter.daniel@uzh.ch
■ **Rüdiger Mutz,** Dr. rer. nat, Senior Researcher am Center for Higher Education and Science Studies, CHESS, Universität Zürich, E-Mail: ruediger.mutz@uzh.ch