Darryl M. Hunter, 2017

Volume 3 Issue 2, pp. 833 - 855

Date of Publication: 18th September, 2017

DOI-https://dx.doi.org/10.20319/pijss.2017.32.833855

This paper can be cited as: Hunter, D. (2017). Sunny Ways or Somber Weather? International Management Consultants and Appraisal of Policy Capacity. PEOPLE: International Journal of Social Sciences, 3(2), 833-855.

SUNNY WAYS OR SOMBER WEATHER? INTERNATIONAL MANAGEMENT CONSULTANTS AND APPRAISAL OF POLICY CAPACITY

Darryl M. Hunter
University of Alberta, Edmonton, Canada
dhunter2@ualberta.ca
darrylinvic@hotmail.com

Abstract

Nearly three hundred judgments of policy capacity within one of three Ministries of Education in a South Pacific country are closely scrutinized for their accuracy. A clearly discernable and measurable halo, or its horned converse, a less positive interpretation, was apparent when internal and external raters used a modified United Nations Development Programme scale during interviews. The scale was augmented with criteria recognizing some desirable traits for civil servants. Those criteria centered on recognizing the warranting nature of various types of evidence, professional knowledge of education policy/processes, and persuasive ability of the person. North American research in cognitive science, organizational behaviour, social psychology and pragmatic-legal theories of evidence offer four explanations. The latter two theories appear to best explain overly positive or negative views of an organization's capacity. Implications are drawn for management consultants and researchers alike when working on international development projects.

Keywords

Management, Consultants, Judgement, Halo effect

1. Introduction

Policy capacity is key when citizens seek changes in organizational learning (Honig & Coburn, 2008; Ramos, 2017), knowledge building (Singh & Kumar, 2017), and improvements in their governments' ability to deliver services—whether in education, health care or public safety. Building capacity is now a central objective in many international development projects as sponsored by the World Bank, the United Nations, and national agencies that work abroad. Policy capacity, however, is notoriously difficult to define and measure. There is a vigorous scholarly debate about alternate definitions for policy—whether a synonym for politics, an expression of public aspirations, a set of governmental processes, or a material representation of goals in bureaucratic documents. The concept of capacity can encompass civil servant competencies, collections of programs, pecuniary allocations, diffuse processes in policy formation, implementation and adaptation to local contexts. Thus, benchmark studies become crucial for looking at growth, sustainability, or deterioration in policy capacity over time. Accurate appraisal becomes a problem for management consultants who undertake such projects.

Indeed, independent management consultants have an ill-explored and potentially powerful impact in education systems within Canada and abroad. Many are veteran educational administrators contracted on an interim basis to provide advice in sometimes tempestuous circumstances. Others are deliberately engaged as free-lance, third-parties to evaluate programs or operations, to accomplish feasibility studies, to appraise capacities, and to make (in)formal recommendations for change (Lapsley & Oldfield, 2001). Some argue such consultancies are crucial so that leaders receive dispassionate and objective advice in the midst of organizational failure (Halstead, Morash & Ozment, 1996). Consultancies are essential to any organization in need of flexible and immediate responses to unexpected events. Yet others contend that consultants bring either sage advice or innovative ideas to educational organizations gone awry or grown stale. In the United Kingdom, critics argue (Ball, 2009; Saint-Martin, 2004) that the proliferation of such independent contractors is evidence of privatization forces at work in policy and practice. However, we do not have a scholarly understanding of the degree to which consultants' opinions are at variance with those working within the educational institution.

Regardless of circumstance and contractual scope, it seems incontrovertible that most management consultants are engaged because their judgement (the application of values to evidence) is trusted by the decision-maker and the contractor. The reputation of the consultant is essential for their own continued work prospects. Indeed, those permanent members of an inside educational management team may distrust the advice of an external managerial expert who has been parachuted near contested ground. "How can the outside expert possibly understand our workplace demands and actual resources without direct, hands-on experience in our work?" those inside the organizational fold might lament. On the other hand, those wanting the third-party consultant might desire a fresh view or experienced outside opinion precisely because they harbor doubts about that "disinterested" advice offered from within. A key issue is why the judgments offered by organizational members often exhibit a sunny halo, while more negative opinions from those outside may be offered and prevail in final reports.

Critics (Rosenzweig, 2007) contend that halos are a form of hubris and a serious "business delusion" when appraising performance, contaminating many management studies. Others (Czarniawska-Joerges, 1990), who disparage the management consultant as a 'merchant of meaning,' deem that more negative appraisals relate to a role discrepancy where consultants are asked to symbolically bring a new metaphor to an organization, introducing alternative world views, ideologies, ideas, rationalizations and interpretations. Because the consultant displaces more traditional forms of managerial control—such as force, incentives or interpersonal persuasion— internal managers will amplify the status of existing systems to support their structures and their personnel, becoming more lenient in their evaluations of performance than the outsider (Bol, 2011; Bol & Smith 2011; Grund & Przemeck, 2012).

In other words, the management consultant must balance on the horns of a dilemma, promoting change without denigrating organizational efforts. From the perspective of those inside the organization, the outsider's judgements may be seen as excessively negative and too technically-driven rather than immersed in practice. From the perspective of an external observer, inside organizational members may hold a substantively different picture of reality with little understanding of theory and technique. The role of the management consultant is to accurately view reality and enlighten the relevant ministry or development partners to what that reality is. At the nub of the problem, therefore, are the sources for halo and horned effects in judgement.

2. Literature Review

At least four, possibly overlapping, theories have been posited in the academic literature for explaining the overly-optimistic responses of participants. Each theory mirrors the primary concerns of various branches in organizational theory. The sources can be found, respectively, first in cognitive science's focus on perception; next, in organizational behaviour specialists' study of reward systems and performance; as well, in social psychologists' studies of interpersonal dynamics; and finally in legal scholars' interest in reasoning patterns with warranting evidence. Each theory is based on a different premise about causal mechanisms and adopts different assumptions about the interplay of perspective and structure in the construction of reality.

Table 1: *Theories for Over-Optimism*

(Original) Recent Researchers	**Disciplinary Orientation**	**Name for Phenomenon**	**Assumptions**	**Features**	**Causal Mechanism**
(Thorndike) Tractinsky, Katz, & Ikar (2000).	Cognitive psychology	Halo and Horns Effect	Theory of perception Difficulty distinguishing among relevant and detailed attributes	Extrapolation from general impression to specific attributes Transposing positive qualities onto others' less known qualities	Perceptual distortion by affect Ambiguous information
(Vroom) Isaac, Zerbe & Pitt, 2001 Savolainen, 2012	Organization Behaviour	Expectancy Theory	Theory of motivation Distinctions between valence, expectancy and instrumentality	Personnel evaluation	Organization incentive systems
(Thomas & Kilmann)	Social Psychology	Impression Management	Theory of regulating	Convincing others that	Interpersonal dynamics

Congo-Poottaren, 2017 Forgas, 2013 Rosenzweig, 2007		and Social Desirability	others' perceptions Many forms: intimidation, ingratiation, self-promotion, exemplification, and supplication	one is a model employee or organization Self-focused and other-focused behaviours	whereby people put best face or foot forward
(Peirce) Conley, O'Barr & Lind, 1995 Freidson, 2001 Kassin, Dror & Kukucka, 2013 Laudan, 2007 Lipton, 2006 Nickerson, 1998	Legal Pragmatism	Reasoning to the best explanation	Theory of practical reasoning Multiple types of evidence with inability to distinguish which has greater probative value: direct, indirect, hearsay	Abductive reasoning from consequent to antecedent Confirmation bias: select what we observe to confirm beliefs. Ordinary or practical knowledge of what works	Reasoning patterns with warranting evidence Hope that our hypotheses explain events Confusion of loveliest explanation with most likely explanation

Regardless of whether over-optimism stems from perceptual distortion in psychological mechanisms, from reward systems, from interpersonal relations, or from flaws in reasoning, researchers and third party evaluators frequently discern a corona hovering over judgments, overly optimistic or perhaps biased and bleak accounts of organizational behaviour, and inflated ideas about capacity. For management consultants, the issue of halo effects, or conversely, whether their commission yields overly-horned judgements, becomes important when establishing the credibility of the results by a variety of readers for any given report.

The purpose of this study, then, is to shed light on sources by looking at results from an international development project that was undertaken in 2016. My primary research questions are:

i. What was the magnitude of the halo or horn effect in the ratings of policy capacity in a United Nations Development Programme-framed project in one South Pacific country?

ii. Among the rival explanations offered to date, which best accounts for biases in appraising policy performance?

3. The Study

This study draws on the tenets of evidence-based judgement within a February-May 2016 project carried out by the author, a Canadian academic, in the South Pacific. To preserve the confidentiality of participating countries, pseudonyms are used in this study. The Pacific Community (SPC) contracted a management consultant to conduct a pilot study of educational policy capacity in the Gemini Islands, Andromeda Islands, and Sagittarius ministries of education (Dean, D. & Guild, D., 2016). The Pacific Community is a regional development organization that provides technical support and advice to ministries of education in participating countries.

The study was based on a conceptual framework drawn from the United Nations Development Programme (UNDP, 2006, 2008a, 2008b, 2010). The roles of evidence, professional knowledge, and persuasive argument in policy making were emphasized, using analytic scales to construct a global five-point criterion scale. A team from SPC gathered data through interviews and on-line surveys of ministry respondents in March 2016. Team members were assessment specialists within SPC's Educational Quality and Assessment Programme. In addition, one official from within each national ministry participated in data collection in each country. Consistent, accurate information was sought in four policy domains: assessment and evaluation; curriculum; school administration and governance; and teacher quality. The purpose of the project was to enable the three participating countries to establish baselines and constructively discuss the implications of those baselines as part of a South Pacific project which benchmarked for educational results among the countries involved. The evidence was to be used to direct and drive system changes in order to improve educational outcomes for children in the

Pacific. Fin this paper, data from Gemini and Sagittarius were excluded, for reasons of space. Nevertheless, the trends reported here were similar in these two countries.

For this project, United Nations Development Programme definitions were adopted. One such definition is capacity, defined as the ability of individuals, institutions, and societies to perform functions, solve problems, and set and achieve objectives in a sustainable manner (United Nations Development Programme, 2006, Slide 3; 2009, p. 54). Capacity development is the process through which individuals, organizations and societies obtain, strengthen and maintain their capabilities to set and achieve developmental objectives over time (United Nations Development Programme, 2009, p. 54). Within this UNDP study, policy was defined as "a set of principles which derive from conflicting interests to enable the formulation and operation of different programs."

Because the project drew on appraisals from both those policy spokespeople within a Ministry and those external raters working at an external coordinating agency, independent judgements of capacity in policy formation, formulation and adaptation were collected. Evidence-based appraisals from both insiders and outsiders were sought to identical questions that were rated with an identical criterion scale.

3.1 Instruments

A 75-question interview guide/questionnaire was created, using the United Nations Development Programme Capacity Assessment User's Handbook and the overall Institutional Capacity framework. The questions could be linked back to multiple dimensions of capacity, such as point of entry, core issues, enabling environment, capacity or competency, or domain. These questions were customized for looking at capacity in four educational policy domains: student assessment and evaluation; curriculum and materials; school governance and administration; and Teacher quality. Questions were consensually modified in their terminology during a training session by the external raters and by senior representatives of each Ministry to fit the South Pacific context.

The United Nations Development Programme sets out a 5-point rating scale ranging from "no evidence of policy capacity"/ "anecdotal evidence of capacity"/ "partially developed"/ "widespread but not comprehensive evidence of capacity"/fully developed evidence of capacity." No attributes are provided for each of these scale points in the UNDP scale. Nor are different kinds of evidence pre-specified in the rating scale. To support the criterion-referenced rather than norm-referenced project design, the management consultant drafted three rubrics which expanded upon and elaborated on this aforementioned scale. To render external judgements more explicit and detailed, participants in the training session created a global or holistic rubric that merged the traits from the three analytic rubrics. The analytic rubrics looked at three dimensions of a civil servant's competencies for their policy multiple capacities:

- their ability to identify, articulate and use various kinds of evidence when explaining policy (evidence rubric)
- their professional knowledge of an educational domain, its processes and its associated policies (professional knowledge rubric)
- their persuasive ability to verbally explain government policy, positions and processes (persuasive explanation rubric).

To construct these analytic rubrics, particular traits were drawn from the research literature on evidence-based decision-making (Sanderson, 2002), on policy making (Head, 2008; Marston & Watts, 2003; Nutley, Davies & Smith, 2000), on the law of evidence (Goode & Wellborn, 2007) and on Bloom's taxonomy of knowledge (Anderson, Krathwohl & Bloom, 2001). The fourth global scale, shown in Table 2 below, therefore became a key instrument for elaborating the criteria reference points and make comparisons among countries in the project.

Table 2: *Holistic Criterion Scale for Rating Policy Capacity, Pacific Community, March 2016.*

5	**Public evidence is** used to warrant **policy statements** so that principles are actually **communciated to external audiences or stakeholders**. These may take the form of laws, regulations or comprehensive statements of values and principles. Applies current policy to **multiple situations** (concepts, principles, approaches) and can **adapt/synthesize** to create new policy where gaps are evident. A sophisticated explanation for how policy principles **logically link to particular situations with a clear view of the reasons**.
4	**Documentary evidence** such as **actual documents**, briefing notes, or policy statements **that are not publicly communicated**, are used to illustrate principles and their incorporation in programs. **Analyzes policy strengths and weaknesses** (concepts, principles, approaches) based on experience and expertise, and can indicate how policy revisions will better enable the government's civic engagement. Can point out defects in current policy and why they are inapplicable, and **can point out new policy principles**, but **unable to explain how they might apply to unusual situations**.

3	**Direct testimonials or demonstrative evidence** is a **detailed description** of events which shows the **application of policy** in particular circumstances. This testimony involves "objective" descriptions, drawing on **more than one source of information** to provide a respondent's own perceptions of the "facts" in an elaborate manner. **Understands** organization's current policy (concepts, principles, approaches) and is able to **connect to their professional knowledge and expertise**. A mechanistic application of policy and principles to the current situation, but is **unable to explain how they apply in new situations**.
2	**Circumstantial evidence or anecdotes** are **short stories** that people tell to show they behave according to policy principles but with **fragmented evidence** that they actually do so. Evidence does **not relate directly to the policy statement**, and more so is **based on opinions rather than pertinent facts**. **Presumes** policy (concepts, principles, approaches) based on professional experience that is relevant to matter at hand. Can logically fit current policy to current situations, existing programs and services, but **difficulty explaining in exceptional circumstances.**
1	**Character evidence** relies on the integrity or personality or leadership qualities of the policy maker. The respondent relies on qualifications or experience or expertise or organizational position **rather than actual policy statements or documents**. Recalls situations where policy principles were in conflict but **no understanding of current policy** or ways of applying principles to the situation at hand. **No logical connections** drawn between policy statement and current or future course of action.

The interview guide and this instrumentation technique proved reliable for the three external judges who collected data in each country during the project. The overall Cronbach's was α = .93, and within the various domains (17 items about student assessment and evaluation policy capacity α =.95; 18 items about curriculum and materials policy α = .81; 20 items about school governance and administration policy α = .87; and 18 items about teacher quality policy α =.90).

3.2 Training

Participants were trained in a 5-day February 2016 training session; descriptions of the data collection process were central to this training and repeated in several different workshop contexts before data was actually collected. Training proceeded in stages, focusing successively and initially on the three analytic rubrics. For each rubric, the traits were described and discussed moving from the end points on the rubric to the midpoints. External judges were then assigned three-to-five interview questions from the interview guide. With this questionnaire, they conducted simulated interviews with each other and they formulated judgements using each analytic scale. Mock interview questions covered all four domains of policy-making. In plenary sessions, judgements were collated to look for adjacent and identical rating patterns. Participants suggested modifications in rubric wording to clarify the breakpoints or markers between the various criterion-scale levels. In the last stage, the three analytic rubrics were merged into a single holistic scale. Then, external judges conducted mock interviews with others in the group, and the global or holistic criteria were then further clarified on the five-point scale. Finally, particular marker characteristics were bolded in the holistic rubric to facilitate usability during actual interviews.

3.3 Analyses

The three external judges, over three to four days in March 2016, collected ratings in person from the national in-site visits. Respondents were chosen by senior ministry officials for having insight into policy making, implementation and evaluation in each of the four domains. The holistic rubric was pre-distributed to these insider respondents at least five days' ahead of interviews and a hard copy of the questionnaire was distributed to respondents immediately before interviews. The external judges' ratings of the respective respondents' answers to the questions were not provided to respondents after rating. At the end of the interview, respondents were asked to log-on to an online version, and anonymously offer their own ratings for those questions that they had been orally asked during the interview. These follow-up ratings by insiders were accomplished independently of each other; this insider respondent as direct witness data was collected within 7 days of interviews. The external judges reached a consensus rating, before submitting the own judgements in a separate field of the online instrument at the Pacific Community office in Fiji.

Given that both external judges and internally-chosen respondents as witnesses were purposefully rather than randomly sampled, non-parametric analytic techniques were adopted, using ratings as rankings in an ordinal scale. Kendall's *tau b* was adopted as the most appropriate analytic technique method because it is more statistically sensitive than Spearman's *rho*, more suitable with small samples, and generally more direct in its probability estimates.

4. Findings

In the Andromeda Islands, as with the other two countries, a halo shone over internal responses as direct witness' judgments. Mean ratings for those personnel on the ground ranged from .6 to 1.6 points higher on the 5-point scale for all domains of policy capacity than those proffered by external judges. The halo existed across all policy domains, as shown in Table 3, whether in the realm of student assessment and evaluation or in curriculum policy capacity. Conversely, these same ratings could be interpreted as the presence of substantively more somber judgements infused in the Pacific Community ratings. The median differences were particularly stark for the policy domains of school governance and administration, and for teacher quality–suggesting that insiders were decidedly more optimistic about their country's policy capacity to change educational structures and professional preparation than were outsiders' perspectives. Modal ratings can readily be over-interpreted because they exaggerate differences, due to the small number of judges and respondents involved in this pilot project. Nonetheless, they do affirm that a halo/horn effect exists in judgement about policy capacity, thus providing definition and depth to its effect's contours.

Table 3: *Descriptive statistics, Andromeda Islands, March 2016*

Outsider Ratings						**Insider Ratings**			
	Mean	**Median**	**Mode**	***SD***		**Mean**	**Median**	**Mode**	***SD***
Assessment/ evaluation	2.06	2.0	2.67	.65		3.51	3.5	3.75	.42
Curriculum	1.60	1.67	1	.54		3.44	3.33	3.33	.33
School governance/ administration	1.95	2.0	2	.67		2.95	3.0	2.5	.69
Teacher quality	1.44	1.3	1.67	.47		3.14	3.0	3	.39

Standard deviations within the ratings of outsiders and insiders suggest the halos' or horns' breadth or diameter. For most domains, the variation in ratings was noticeably similar to and/or larger within the external judges' team, than within the purposefully-chosen group of respondents, none of whom had received systematic training in rating consistency. Overall, standard deviations suggest that insider perceptions are at least as and often more consistent than those of outsiders, notwithstanding those revealed by a Cronbach's alpha.

The criterion rubric offers a descriptive view of this proclivity to paint a bright or somber picture of policy capacity. The median is the mathematically most stable measure of center with small numbers of raters. Similar to the other two countries, but in a more accentuated fashion, external raters in the Andromeda Islands found little evidence (median rating of 1 or 2) that insiders could describe policy capacity at a level much beyond personal or leadership integrity, or with circumstantial evidence. External judges found that insider respondents did demonstrate understanding of the policy but had difficulty applying it to current or exceptional circumstances. Conversely, insider raters reported (median of 3) that they relied on direct testimonials or demonstrative evidence, providing "objective" descriptions with more than one source of information in an elaborate manner. They believed they not only understood the Ministry's current policy (concepts, principles, approaches), but were also able to connect it to their professional knowledge and expertise. In this insider-witness perspective, they believed they could automatically apply it to the current situation, but were unable to explain the application in new situations. Differences in judgement between insider and outsider views, to identical questions, hinged on different conceptions of knowledge, different notions of persuasive ability, and different ideas about what constituted evidence.

A more finely-tuned look at the distance between somber appraisal and sunny ways, becomes visible through cross-tabulation of ratings, as shown in Table 4. Cross tabulations in are organized by the number of ratings offered by the judges, not by the number of raters.

Table 4: *Cross Tabulations of External Judges' vs Internal Witness Ratings of Policy Capacity, by Domain, Andromeda Islands, March 2016*

Domain			**External Judges' Ratings**							
			1	2	3	4	5			
	Internal Witness Ratings							Row total		
Assessment/ Evaluation		1	2	3	2	-	-	7		
		2	1	3	2	-	-	6		
		3	7	11	8	2	-	28		
		4	9	13	10	2	-	34		
		5	5	4	3	-	-	12		
			24	34	25	4	0	**87**		
Curriculum		1	-	-	-	-	-			
		2	3	2	-	-	-	5		
		3	12	12	3	-	-	27		

		4	13	13	4	-	-	30		
		5	1	2	2	-	-	5		
			29	29	9	0	0	**67**		
School Governance/ Administration		1	1	1	1	1	-	4		
		2	5	8	2	-	-	15		
		3	4	7	2	-	-	13		
		4	5	9	5	1	-	20		
		5	-	2	1	-	-	3		
			15	27	11	2	0	**55**		
Teacher quality		1	2	1	-	-	-	3		
		2	14	9	-	-	-	23		
		3	12	9	1	-	-	22		
		4	17	12	1	-	-	30		
		5	1	0	1	-	-	2		
			46	31	3	0	0	**80**		
										Cumulative row percentages for internal witness
Cumulative column percentages for external judges		1							5.2	
		2							19	
		3							30.2	
		4							38.4	
		5							7	
			38.4	42.4	16.8	2.4	0			

Cross tabulations in the Andromeda Islands trace the distance between insider and outsider perspectives, thus revealing many discrepancies. When looking at the diagonal for identical ratings by insiders and outsiders, the table shows only 15 out of 87 instances of identical insider and outsider judgments for student assessment and evaluation capacity, only 5 out of 67 identical judgements for curriculum. Twelve of 55 cases had identical ratings in the domain of school governance, and 12 out of 80 judgements were identical about policy capacity for teacher quality. When combining identical with adjacent appraisals, 44 of 87 ratings were in a proximate zone for student assessment and evaluation policy capacity, 26 out of 67 ratings landed in this margin for curriculum, 32 of 55 ratings in school governance and administration, and 37 out of 80 ratings for teacher preparation were in closely related. Viewed cumulatively across these core domains of Ministry policy capacity, 80% of outside ratings fell at the bottom two levels of the criterion performance described in the holistic rubric in Table 2. Only 2% of ratings were assigned the top two levels. By way of contrast, the cumulative percentages of ratings offered by insiders at these two levels on the scale pole ends were 24% low and 45 % high, respectively. Seventeen percent of outsider ratings fell at the midpoint; 30% of insider ratings were based in the middle of the criterion rubric. Because Andromeda Islands' insider ratings were normatively distributed across domains similar to those in Sagittarius, but outsiders were not, we can suggest that insider ratings were more attuned to local, endogenous perspectives, and outsider ratings were more firmly anchored in the exogenous criterion scale.

This type of analysis, however, leaves little room for making strong statements based in the content of the holistic criterion rubric, because it does not provide an analytic picture of where the differences in the appraisal of evidential, professional or persuasive ability arise. Analytic rubrics, rather than just a holistic rubric, are required for a more particularistic view. We can say that neither external nor internal raters had strong views for what constitutes direct evidence. Even when considering adjacent evidence and not just identity of ratings, no strong similar outlook was apparent on the concepts of testimonial, direct, and documentary evidence or civil servants' ability to decipher the differences or persuade others with it. Even at the lower extreme of the scale, 38% of external ratings were deemed to be based on character without the insight to understand current skills and relate them to future action, whereas only 5% of insider witness ratings were so assigned.

Table 5: *Correlation Coefficients* for External vs Internal Ratings of Policy Capacity, Andromeda Islands, March 2016*

Policy domain	**N**	**Assessment/ Evaluation**	**Curriculum**	**School Governance/ Administration**	**Teacher quality**
Assessment/ evaluation	17	-.08			
Curriculum	19		-.02**		
School governance/ administration	20			-.314	
Teacher quality	19				-.01**

**Tau b*

**Significant difference in (dis)accordant pairs at <.05

Sources for another aspect of halo versus horn effects are suggested in Table 5 for the Andromeda Islands. Correlation coefficients with *tau b* do not trace a linear co-relationship between ratings, but rather a monotonic association between insiders' and outsiders' respective rankings. That association of accordant and discordant ratings is registered on a simple scale of -1 through zero to +1. Kendall's *tau-b* ranges from -1.0 (all pairs disagree) to 1.0 (all pairs agree). A positive value indicates both variables increase together, whereas a negative value indicates that both variables decrease together. If there is no association in the ratings, the score approaches zero and therefore the ratings are autonomous of each other. Significant differences indicate no relationship between insider and outsider ratings whatsoever.

Results indicate weak to non-existent correlations between insider and outsider judgements in the Andromeda Islands. Significant differences in ratings, with a probability of occurring only by a 5% chance, are found for teacher quality and for curriculum. For all domains of policy rated in the Andromeda Islands, the coefficients are negative. The association of external raters' judgements was inversely associated to those offered by insider raters, especially for school governance and administration, though other ratings approximated zero. This suggests that insiders and outsiders were operating with different criteria, leading them in opposite directions for their ratings, notwithstanding the same criterion rubric provided to all.

5. Discussion

Sound judgements about a ministry's policy capacity hinge on whether those appraisals derive from those in the organization itself or from external observers in coordinating agencies near the country. For this project, a halo, or its more somber converse, was clearly discernable. The effect spanned all domains of educational policy-making under review, and had a measureable depth of approximately .5 to 1.7 points on a five-point scale. The distance between the internal halo and the external raters' more solemn appraisals, appears as a distinction between endogenous normative influences within Ministries and the exogenous application of criterion ratings.

The pragmatic account, which is predicated on different appraisals of evidential weight, best explains the sources for the phenomenon in this study. In all three countries in this project, there was a strong proclivity for insider raters, within the various policy areas of Ministries of Education, to rely on integrity of character, narratives or anecdotes rather than policy and less so on the protocols and formalities of official policy and procedures when addressing policy problems. On the other hand, external judges were more fully grounded in the evidential criteria. They more fully looked at the warranting strength of various kinds of evidence, at the professional knowledge levels of civil servants in the education sector, and at the persuasive competency of civil servants. Pragmatists argue that hope animates our abductive reasoning. However, “Hope is a good breakfast,” Sir Francis Bacon once wrote, “but it is a bad supper.” We need more substance at the end of the day than simply good intentions.

The inverse ratings in the Andromeda Islands results may also corroborate impression management as a source for the corona, not as a defensive strategy to counteract a potential negative appraisal, but simply as a normative belief. Normative processes inhere to any social psychological explanation. Nonetheless, we must also note that one-on-one interviews and autonomously-provided survey results, not focus groups, were the primary means of data collection for this project, undermining social psychology as an all-purpose explanation. Forty-five percent of internal ratings for Andromeda Islands’ policy capacity in its Ministry of Education were at the two top criterion levels, whereas only 2% of external ratings were these high levels. Significant differences and inverse associations in ratings suggest that insider raters had normatively more optimistic appraisals or, alternatively, that external raters were attempting to put a more “candid” assessment in this benchmarking exercise. The interesting finding that internal raters were often more internally consistent, as viewed through standard deviations, than external raters when rating capacity, suggests a more uniform posture and hence shared outlook.

This does not necessarily mean insiders have adopted a defensive stance. As individuals, we are all limited by our own scope of experience. When “insiders” have been members of organizations with a particular way of operating, that becomes their norm against their judgement of what capacity looks like. If, by that frame of reference, they perceive their organization to be operating towards the high end of a scale, it will manifest as a rosy interpretation of the situation. It is a halo effect to be sure but one not necessarily attributable to either a defensive position, nor to any firm commitment to organizational solidarity (Breuer, Nieken & Sliwka, 2013). Likewise, when the “outsider” comes from a (very) high capacity organization and set of experiences, s/he may have a skewed perception of capacity at the other extreme. Consultants often come from high-performing backgrounds – which is precisely the reason they have the credibility to put themselves forward for the work – and therefore are presumed to have expertise in judging capacity. This will result in a horned effect but perhaps not necessarily because the consultant is predisposed or expected to be critical.

This data did not support expectancy theory as halo generator. The distance between internal and external ratings is a valence measure and the instrumentality is represented in raters' use of the rubric scale. Therefore, we see that weak to negative associations between insider and outsider ratings do not support expectancy theory's hypothesis of strong multiplicative relationships between valence and instrumentality. If both the insider and the outsider are motivated by reward systems for generating positive views of performance, then both would be expected to offer optimistic accounts. Obviously, results in this pilot project will be used for actual benchmarking exercises; however, anonymity means they could not be used for personnel evaluations and rewards in the present.

Similarly, we may discount the distorting effect of affective qualities in perception, or the inability to distinguish among various criteria in this project. That there were distinct differences in perception based on organizational boundaries explicitly contradict the universal claims of a psychological mechanism. Moreover, it is unlikely such perceptual distortions would exist across all raters, in all domains, across three countries with dissimilar cultural backgrounds. Management consultants would however be well-advised to deploy both analytic and holistic rubrics so that more precise explanations can be found for discrepancies in rating the various types of evidence, the levels of professional knowledge, and the variety of persuasive strategies used by civil servants in policy development and enactment.

6. Conclusions

Regardless of task or scope of judgment, an international consultant will almost certainly encounter differences in perspective on policy capacity. An evidence-based approach should enable the consultant to proceed, whether in collecting verbal testimony or in generating quantitative information. However, there have been few scholarly efforts, to date, to clarify what kinds of evidence, with what probative value, in relation to what kinds of professional knowledge, and in consideration of what kinds of persuasive appeal, should be weighed.

Offering simple checkmarks on Likert scales without defining clearly what each scale level signifies can result in distortions of perspective stemming from different organizational rewards and dynamics within it. Without a clear view of the evidence upon which judgements ought to be based, and without more sophisticated notions of reasonable inference, organizational practitioners may be prone to abductive reasoning fallacies, reasoning to the loveliest rather than the most probable explanation.

When analyzing policy capacity, it seems reasonable to ground one's appraisals in concepts and ideas that are attuned to varying ideas of professional knowledge. We have over fifty years' worth of research into distinctions between practitioners' ordinary knowledge, which revolve around 'know how' and 'know who,' and that formal knowledge created in the academy, which revolve around 'know what' and 'know why' (Freidson, 2001). Bloom's taxonomy of knowledge is nearly 75 years old, and appears well understood by educators and civil servants in many education sectors around the world. We can hope that distinctions between the power of interpersonal persuasion, in a qualitative sense, as opposed to the measurement of effect size through computations of sampling power and probabilistic judgments that are quantitatively represented in Likert scales, is widely understood.

However, "evidence" as the central construct in evidence-informed policy making and judgement, remains dismayingly ill-defined in policy circles, in research and measurement, and even in psychology (Head, 2008). The question is 'what exactly is convincing evidence?" Is the compelling evidence an accumulating body of research; is it numeric quantities which have been rigorously marshalled; is it the testimonials of those directly engaged in the organization; is it direct or circumstantial or hearsay in nature? It follows from this that proponents and theorists of evidence-based decision making would be well-advised to turn to the discipline of law. Law offers the most elaborate body of theory for evidence, an extensively-documented public record of evidence-based decisions, and an internationally-accepted body of formal principles about evidentially- anchored decisions. Moreover, law is widely seen as one expression of official policy, thereby imparting at least as much legitimacy—when examining the probative value of evidence—as do social science or natural science principles. Rather than repeatedly turning to psychology and behavioral economics for a deeper understanding of how decision-makers' judgements may be distorted one way or another by evidence, carefully examining legal reasoning for (in)appropriate inferencing from evidence is recommended.

Whether management consultants' judgements of policy capacity bear a sunny or somber character is crucial for legitimizing the results of policy or program capacity studies. This is particularly so for international development consultants who may presumptively adopt a pessimistic stance on capacity and for those civil servants within a country who may hold overly optimistic views of their domestic expertise. Relying alone on the judgements of well-trained outsiders may be desirable for a pilot project and for establishing baselines for long-term benchmarking. We can also statistically merge or remove a bias. However, given the gap in perspectives manifest here, and the stakes involved with implementing recommendations, both insiders and outsiders' views must be better aligned. Highlighting the different perspectives promises improvement rather than positioning the consultant or external agency as antagonist. Public reporting of both perspectives can also serve as a mutual check-and-balance against tendencies to abide by the old French maxim[1] that we should feel ashamed to think badly of others.

References

Anderson, L. W., Krathwohl, D. R., & Bloom, B. S. (2001). A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives. Boston, MA: Allyn & Bacon.

Ball, S. J. (2009). Privatising education, privatising education policy, privatising educational research: Network governance and the 'competition state'. Journal of Education Policy, 24(1), 83-99.

Barnes, E. (1995). Inference to the loveliest explanation. Synthese, 103(2), 251-277.

Beckwith, N. E., & Lehmann, D. R. (1975). The importance of halo effects in multi-attribute attitude models. Journal of Marketing Research, 265-275.

Bol, J. C. (2011). The determinants and performance effects of managers' performance evaluation biases. The Accounting Review, 86(5), 1549-1575.

Bol, J. C., & Smith, S. D. (2011). Spillover effects in subjective performance evaluation: Bias

[1] Honi soit qui mal y pense.

and the asymmetric influence of controllability. The Accounting Review, 86(4), 1213-1230.

Breuer, K., Nieken, P., & Sliwka, D. (2013). Social ties and subjective performance evaluations: an empirical investigation. Review of managerial Science, 7(2), 141-157.

Congo-Poottaren, N. (2017). The influence of impression management of school leaders on followers: A case study in a secondary school in Mauritius, PEOPLE: International Journal of Social Sciences, Special Issue, 3 (1), 741- 760. https://dx.doi.org/10.20319/pijss.2017.s31.741760

Conley, J. M., O'Barr, W. M., & Lind, E. A. (1979). The power of language: Presentational style in the courtroom. Duke Law Journal, 1978(6), 1375-1399.

Czarniawska-Joerges, B. (1990). Merchants of meaning: Management consulting in the Swedish public sector. In B.A. Turner (Ed.), Organizational symbolism (pp. 139-150). Berlin: Walter de Gruyter.

Dean, D. & Guild, D. (2016). Pacific benchmarking for education results: A mid-term review. Canberra, AU: Department of Foreign Affairs and Trade. Retrieved from:
http://dfat.gov.au/about-us/Pages/about-us.aspx

Forgas, J. P. (2011). She just doesn't look like a philosopher…? Affective influences on the halo effect in impression formation. European Journal of Social Psychology, 41(7), 812-817.

Feeley, T. H. (2002). Comment on halo effects in rating and evaluation research. Human Communication Research, 28(4), 578-586.

Freidson, E. (2001). Professionalism, the third logic: On the practice of knowledge. Chicago: University of Chicago Press.

Goode, S., & Wellborn, O. G. (2007). Courtroom Evidence Handbook. St. Paul, MN: West Publishing Company.

Grund, C., & Przemeck, J. (2012). Subjective performance appraisal and inequality aversion. Applied Economics, 44(17), 2149-2155.

Halstead, D., Morash, E. A., & Ozment, J. (1996). Comparing objective service failures and subjective complaints: An investigation of domino and halo effects. Journal of Business Research, 36(2), 107-115.

Head, B. W. (2008). Three lenses of evidence-based policy. Australian Journal of Public Administration, 67(1), 1-11.

Honig, M. I., & Coburn, C. (2008). Evidence-based decision making in school district central offices: Toward a policy and research agenda. Educational Policy, 22(4), 578-608.

Isaac, R. G., Zerbe, W. J., & Pitt, D. C. (2001). Leadership and motivation: The effective application of expectancy theory. Journal of Managerial Issues, 212-226.

Kassin, S. M., Dror, I. E., & Kukucka, J. (2013). The forensic confirmation bias: Problems, perspectives, and proposed solutions. Journal of Applied Research in Memory and Cognition, 2(1), 42-52.Lapsley, I., & Oldfield, R. (2001). Transforming the public sector: Management consultants as agents of change. European Accounting Review, 10(3), 523- 543.

Laudan, L. (2007). Strange bedfellows: Inference to the best explanation and the criminal standard of proof. International Journal of Evidence and Proof, 11(4), 292-306.

Lipton, P. (2003). Inference to the best explanation. London, UK: Routledge.

Marston, G., & Watts, R. (2003). Tampering with the evidence: A critical appraisal of evidence-based policy-making. The Drawing Board: An Australian Review of Public Affairs, 3(3), 143-163.

Maas, V. S., & Torres-González, R. (2011). Subjective performance evaluation and gender discrimination. Journal of Business Ethics, 101(4), 667-681.

McKenna, C. D. (1995). The origins of modern management consulting. Business and Economic History, 24(1), 51-58.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. Review of General Psychology, 2(2), 175-230.

Nutley, S. M., Davies, H. T., & Smith, P. C. (2000). What works?: Evidence-based policy and practice in public services. Boston, MA: MIT Press.

Parsons, W. (2002). From muddling through to muddling up-evidence based policy making and the modernisation of British Government. Public Policy and Administration, 17(3), 43-60.

Peirce, C. S. (1932–1958). Collected papers of Charles Sanders Peirce, Vols. 1–8. In P. Weiss, C. Hartshorne, & A. W. Burk (Eds.). Cambridge, MA: Harvard University Press.

Ramos, W. (2017). Effects of result-based capability building program on the research competency, quality and productivity of public high school teachers. PEOPLE: International Journal of Social Sciences, Special Issue, 3(1), 109- 119. https://dx.doi.org/10.20319/pijss.2017.31.109119

Rosenzweig, P. (2007). Misunderstanding the nature of company performance: The halo effect and other business delusions. California Management Review, 49(4), 6-20.

Saint-Martin, D. (2004). Building the new managerialist state: Consultants and the politics of public sector reform in comparative perspective. Oxford UK: Oxford University Press.

Sanderson, I. (2002). Evaluation, policy learning and evidence-based policy making. Public Administration, 80(1), 1-22.

Schmidt, F. L. (1973). Implications of a measurement problem for expectancy theory research. Organizational Behavior and Human Performance, 10(2), 243-251.

Singh, P.K. & Kumar, M. (2017). A study on infrastructure and organizational learning: Rethinking knowledge performance perspective. PEOPLE: International Journal of Social Sciences, 3(2), 61-77. http://dx.doi.org/10.20319/pijss.2017.32.6177

Thomas, K. W., & Kilmann, R. H. (1975). The social desirability variable in organizational research: An alternative explanation for reported findings. Academy of Management Journal, 18(4), 741-752.

Tractinsky, N., Katz, A. S., & Ikar, D. (2000). What is beautiful is usable. Interacting with computers, 13(2), 127-145.

United Nations Development Programme. (2006). UNDP and capacity development. [PowerPoint]. Tbilisi, Georgia: Author. Retrieved from www.jposc.org/documents/workshop_georgia_UNDP_and_CD.ppt

United Nations Development Programme. (2008a). Capacity assessment methodology user's guide. New York, NY: Capacity Development Group. Bureau for Development Policy. Retrieved from http://www.undp.org/content/dam/aplaws/publication/en/publications/capacity-development/undp-capacity-assessment-methodology/UNDP%20Capacity%20Assessment%20Users%20Guide.pdf

United Nations Development Programme. (2008b). Capacity development practice note. New York, NY: Author. Retrieved from: http://www.unpcdc.org/media/8651/pn_capacity_development.pdf

United Nations Development Programme. (2010). Measuring capacity. New York, NY: Author. Retrieved from http://www.undp.org/content/dam/aplaws/publication/en/publications/capacity-development/undp-paper-on-measuring-capacity/UNDP_Measuring_Capacity_July_2010.pdf