Evaluating how powerfulness and meaningfulness affect implementation willingness

This chapter has been published as Thomann, E., Van Engen, N.A.M. & Tummers, L.G. (2018). The necessity of discretion: A behavioral evaluation of bottom-up implementation theory. Journal of Public Administration Research and Theory, early online view: doi:10.1093/jopart/muy024

ABSTRACT

The topic of discretion continues to be hotly debated in policy design and policy implementation. In top-down theories, discretion at the frontline is often seen as a control problem: discretion should be avoided as it can mean that the policy is not implemented as intended. Conversely, bottom-up theories state that discretion can help policy implementers tailor a policy to specific circumstances. However, there has been little systematic research into how the experience of having discretion motivates frontline workers to implement a policy. In this chapter, we conceptualize and test this relationship by combining public administration and motivation literature, using datasets in healthcare and education and large-N set-theoretic configurational analysis. Results robustly show that experiencing discretion is a quasi-necessary condition and, hence, a prerequisite for high implementation willingness. This finding is more in line with bottom-up than with top-down theories. Policy implementers need the freedom to adapt the program to local conditions for being motivated to implement a policy. The evidence encourages scholars and practitioners to move from the question whether frontline workers should be granted discretion to how to best make use of frontline workers' discretion instead.

5.1 INTRODUCTION

> *"The closer one is to the source of the problem, the greater is one's ability to influence it; and the problem-solving ability of complex systems depends not on hierarchical control but on maximizing discretion at the point where the problem is most immediate."*
>
> - Richard Elmore (1979)

Discretion is the freedom to decide what should be done in a particular situation. Repeatedly, research has shown that frontline workers - also referred to as public professionals or street-level bureaucrats - have an important role in the successful implementation of policies as they inevitably retain some degree of discretion (Davis, 1969; Lipsky, 1980; Hupe & Hill, 2007; Maynard-Moody & Musheno, 2012; Gofen, 2014; Barnes & Henly, 2018). However, implementation theory has always held contradictory views on the exact role of discretion (Thomann *et al.*, 2018b). Top-down perspectives treat deviations from the policy-on-paper as a control problem: room for interpretation makes it increasingly likely that policy means and ends will be mismatched (Howlett, 2004, p. 5). Conversely, bottom-up perspectives put frontline workers' discretion at the center stage of policy implementation (Lipsky, 1980; Sabatier, 1986). As the above quotation by Elmore illustrates, from this perspective frontline workers are seen as de facto policymakers. Discretion helps them to tailor a policy to specific circumstances.

Although research has moved on to hybrid, integrative frameworks, the discussion surrounding discretion at the frontline never lost its practical salience for policy design and implementation (Howlett, 2004; Hupe & Hill, 2007; Hupe, 2013). Scholars continue to discuss the reasons why frontline workers use their discretion in more or less beneficial ways for clients and public goals (e.g., Keiser, 1999; Brodkin, 2011; Maynard-Moody & Musheno, 2012; Thomann, 2015). However, little attention has been paid to the implicitly assumed link between frontline workers' discretion and the motivation to implement government policies. This is surprising, given that "research performed in ignorance of the understanding that implementing actors have about their circumstances is likely to miss important parts of the explanation" (O'Toole, 2000, p. 269).

To explore the motivational effects of discretion, this article draws on the logic of the seminal Thomas theorem: "If men define situations as real, they are real in their consequences". (Thomas, 1928, p. 572; see also Lewin, 1986). We focus on the perceived degree of discretion, instead of the objective degree of discretion, and investigate to what extent frontline workers experience discretion. We operationalize perceived discretion via the concept of powerfulness as developed in the policy alienation literature (Tummers, 2011; Loyens, 2015; Thomann, 2015; Van Engen *et al.*, 2016; Van der Voet *et al.*, 2017). Hence, powerfulness is seen as discretion as perceived by frontline workers. We define powerfulness more formally as frontline workers' perceived influence on decisions concerning the policy. The research question of this article is then: *How does powerfulness motivate frontline workers to implement policies?*

Psychologists suggest a positive link between powerfulness and motivation (Gagné & Deci, 2015). However, scholars studying policy implementation have not found a strong, consistent relation between powerfulness and implementation willingness (Tummers, 2011; Loyens, 2015; Thomann, 2015; Van Engen *et al.*, 2016). Contrary to these previous studies, we rely on an asymmetric explanation of policy implementers' motivation: the things that motivate people may be different from those that demotivate them (Herzberg *et al.*, 1959; Matzler & Renzl, 2007). Accordingly, we study two interpretations of the motivational role of powerfulness. The first interpretation argues that powerfulness is quasi-necessary, although on its own not sufficient to motivate employees (Herzberg *et al.*, 1959; Goertz & Starr, 2003; Lammers *et al.*, 2016). The second interpretation is that powerfulness is only motivating when the public policy is consistent with the frontline workers' values and, hence, perceived as meaningful (May *et al.*, 2004; Dias & Maynard-Moody, 2007; Grant & Berry, 2011).

We study these interpretations using two large samples. By doing so, this study makes two contributions to the literature. It adds to theory by clarifying a core aspect of the top-down versus bottom-up debate: is discretion beneficial for policy implementation? It does so by connecting the policy implementation literature with the motivation theory from Herzberg. Methodologically, it uses state-of-the-art tools specifically designed for capturing the hypothesized asymmetric patterns: large-N set-theoretic configurational analysis using fuzzy sets, combined with formal theory evaluation, measures of uncertainty and system-

atic robustness tests (Ragin, 2000; Schneider & Wagemann, 2012; Misangyi *et al.*, 2017).

In the next section we will introduce our theoretical framework and the hypotheses. We then introduce our methods, the research design and the data collected among 1.004 healthcare workers and 1.087 secondary school teachers in the Netherlands. After presenting the results, we conclude and discuss how our results can inform public administration scholars and practitioners.

5.2 THEORETICAL FRAMEWORK

The concept of discretion often serves as an umbrella term for different aspects of bureaucratic practice. In policy implementation research specifically, discretion concerns the extent of freedom that frontline workers have to choose among possible courses of behavior when implementing policies (Davis, 1969; Hupe, 2013). Top-down approaches emphasize the degree of freedom granted by a rule maker to an implementing actor ('discretion-as-granted'; Howlett, 2004). Contrary to this, bottom-up approaches presuppose an inevitable existence of discretion and analyze how the degree of freedom is actually used by frontline workers ('discretion-as-used'; Hupe, 2013).

Next to discretion-as-granted and discretion-as-used, we argue that there is also a key role for discretion-as-perceived: the degree to which frontline workers perceive to possess discretion. According to the Thomas theorem, people often feel and behave based on their perceptions of reality, not on the basis of reality itself (Thomas, 1928). This perspective highlights the importance of policy-related attitudes for frontline policy implementation (Ewalt & Jennings, 2004). The Thomas theorem suggests that discretion-as-used presupposes discretion-as-perceived. Frontline workers should feel that they have discretion before they can actually use it. For instance, a social worker should feel that she can grant an exception to a rule before actually doing this. Street-level bureaucracy scholars have recently begun to explore discretion-as-perceived under the heading of policy powerfulness, meaning the perceived degree of influence that frontline workers have over shaping a policy during its design and implementation (Tummers *et al.*, 2009). This power may be exercised at the strategic, tactical or

operational level. High policy powerfulness thus indicates perceived discretion; the absence of powerfulness (i.e., power*less*ness) indicates a lack of perceived discretion.

We can then connect discretion-as-perceived - here conceptualized as policy powerfulness - to implementation willingness. To actually achieve policy goals, frontline workers should be willing to implement the policy (Ewalt & Jennings, 2004; Van der Voet *et al.*, 2017). High willingness to implement means that frontline workers intend to put effort in executing the policy. Bottom-up theories assume that discretion is positively linked with successful implementation. Note, however, that what exactly success entails might differ from a bottom-up or top-down view. Conformance implementation refers to the degree to which the centrally decided blueprint is implemented from top to down ('implementation success'). From the bottom up, performance implementation means that a policy achieves outcomes that resolve the policy problem at stake ('policy success'; Barrett & Fudge, 1981). Arguably, implementation willingness matters for both conformance and performance implementation.

The positive link between discretion and implementation willingness assumes that policy powerfulness can have a motivational effect on frontline workers. Scholars agree that perceptions can, and often do, influence behavior (e.g., Ajzen & Fishbein, 1980). Experiencing powerfulness is one of the main factors stimulating employees' willingness to support a change (Greenwood *et al.*, 2002). Related to this, the policy alienation framework asserts that as frontline workers' policy powerfulness increases, their support for a policy can increase as well (Tummers *et al.*, 2009). This powerfulness can be experienced at either the national (strategic), organizational (tactical) and client (operational) level, or a combination of these. For instance, if a frontline worker has the impression she – or her colleagues or representatives of a professional organization – is able to influence the content of policies at the national level she is more likely to be motivated to implement the policy (Tummers *et al.*, 2015). This is because it is more likely then that frontline workers' interests and concerns are reflected in the content of the policy.

Next to powerfulness, policy alienation has a meaningfulness dimension. Meaningfulness concerns the perception of the frontline worker that the policy is valuable for society in general (societal meaningfulness) and for the direct clients

of the frontline worker (client meaningfulness). Perhaps contrary to expectation, in empirical tests the relation between powerfulness and implementation willingness appears either as weaker than between meaningfulness and implementation willingness (Van Engen *et al.*, 2016), as ambiguous (Loyens, 2015; Thomann, 2015), or as non-significant (Tummers, 2011).

In light of these puzzling empirical findings, we suggest two alternative interpretations of the motivational link of powerfulness on implementation willingness. Previous research has assumed symmetric effects, where the same change in implementation willingness is expected both when powerfulness is added and when it is taken away. Contrary to this, motivation theory as developed by among else Herzberg *et al.* (1959, see for recent discussions Bassett-Jones *et al.*, 2005; Matzler & Renzl, 2007; Sachau, 2007) suggests the effects of particular motivational factors are asymmetric. It is a fundamental insight from motivation theory (Herzberg *et al.*, 1959) that the things that motivate people are often different from the things that demotivate them. For instance, a low salary makes you dissatisfied. However, a high salary does not automatically make you satisfied. This means that the influence of policy powerfulness might work only, or mainly, in one direction. Thus, the change in implementation willingness might not be of the same magnitude or direction when powerfulness is added as when it is taken away. To detect such patterns, an empirical method is needed that models asymmetric effects. This is why we choose a new, set-theoretic method that enables us to model asymmetric explanatory patterns (Misangyi *et al.*, 2017).

5.2.1 Interpretation 1: Policy powerfulness is a necessary condition

The first interpretation linking powerfulness and implementation willingness builds upon the idea that discretion is a prerequisite for policy success (Matland, 1995). If this is the case, then frontline workers need to feel able to influence the policy to be willing to implement that policy; they need to feel powerful. Hence, powerfulness is a necessary condition for implementation willingness.

Policy implementation literature, especially the studies departing from the bottom-up perspective, suggests that an important factor in this willingness of frontline workers is the extent to which organizations are willing and able to delegate decision-making authority to the frontline (Meier & O'Toole, 2002; Tummers & Bekkers, 2014). This influence may be particularly pronounced in

frontline workers whose expectations of discretion and discretion contradict notions of bureaucratic control (Freidson, 2001). As we study teachers and healthcare workers, this seems to be particularly important. Maynard-Moody and Portillo (2010, p. 259) note, "Street-level workers rely on their discretion to manage the physical and emotional demands of their jobs. They also rely on their discretion to claim some small successes and redeem some satisfaction".

The enabling role of powerfulness for implementation willingness can be traced back to the human relations movement (McGregor, 1960). One of the central tenets of this movement is that employees have a right to give input into decisions that affect their working lives. Employees enjoy carrying out decisions they have helped create – as compared to decisions they have not helped create or were 'forced upon them'. As such, the human relations movement argues that when employees experience discretion during their work, this will positively influence several job indicators, such as implementation willingness, loyalty or responsibility, by fulfilling intrinsic employee needs (for more detailed discussions, see for instance Yukl & Becker, 2006). This mechanism was already proposed by Follet (1924) - her work presaged the rise of the human relations movement - who underscored the importance of leaders having the capacity to increase the sense of power among those led. So that those led, in turn, would be empowered to achieve desired changes at the organizational, community or policy level.

The above argumentation suggests that frontline workers need to feel powerful in order to be willing to implement the policy. However, feeling powerful alone may not be sufficient. Many other factors can influence the willingness of frontline workers to implement a particular policy. This can include resources available in the organization (for instance, is there enough manpower available to make a policy work) or the value of a policy for society and political processes within organizations (O'Toole, 2000; May & Winter, 2009; Thomann, 2015). Hence, frontline workers need to feel powerful, but feeling powerful is not enough. This asymmetric interpretation accounts for the fact that not all frontline workers will use their discretion to contribute to successful implementation. Contrary to a symmetric effect, we hence expect that discretion-as-perceived has an enabling effect for motivating frontline workers (Goertz & Starr, 2003).

Accordingly, we can derive the first hypothesis. In order to be motivated to implement a public policy, frontline workers need to perceive that they have the

power to influence the shaping of a policy program (powerfulness). They should experience this powerfulness at least at either the strategic, tactical or operational level in order to feel motivated for policy implementation (Van Engen *et al.*, 2016). Still, this powerfulness does not by definition result in high implementation willingness. Hence, frontline workers with high implementation willingness are a subset of those frontline workers who experience powerfulness. We hypothesize that *policy powerfulness (either strategic: SP; tactical: TP or operational: OP) is a quasi-necessary, but not sufficient condition for high implementation willingness (W)*. This is shown in Figure 5.1. To formalize this first hypothesis, the backward arrow ← means 'is necessary for' and '+' denotes the logical 'OR'.

Hypothesis 1: SP + TP + OP ← W

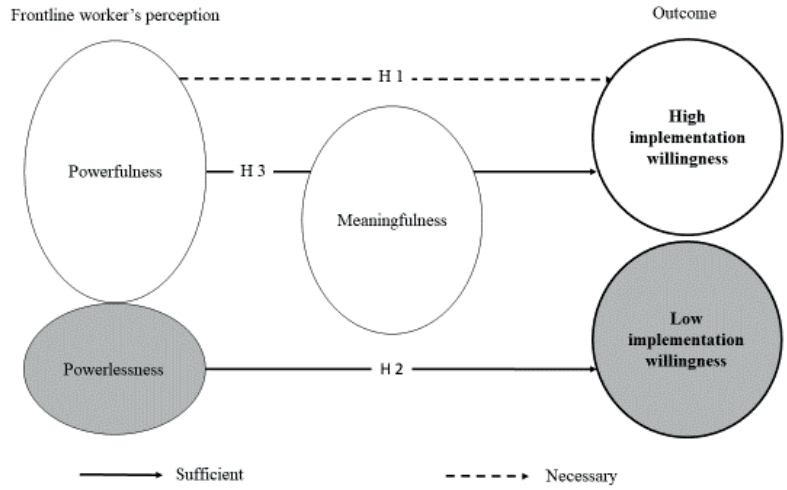

Figure 5.1 Hypotheses

Similarly, we expect that frontline workers who do not feel powerful are typically unwilling to implement government policies. As Figure 5.1 illustrates, if high implementation willingness requires the presence of powerfulness, then the frontline workers who do not feel powerful are a subset of those frontline workers with low implementation willingness. Since powerfulness is indicated by either strategic, tactical or operational powerfulness (or a combination of these three),

all three have to be absent to indicate the absence of powerfulness (Schneider & Wagemann, 2012). Our second hypothesis thus states that *the absence of the combination of strategic, tactical and operational powerfulness is quasi-sufficient for low implementation willingness.* The '*' sign denotes the logical 'AND', while the forward arrow → indicates 'is sufficient for'. The tilde sign '~' denotes the absence of a factor:

Hypothesis 2: ~SP * ~TP * ~OP → ~W

5.2.2 Interpretation 2: Policy powerfulness interplays with policy meaningfulness

The second interpretation takes into account that frontline workers often feel a desire to benefit others with their work (Dias & Maynard-Moody, 2007). They seek to help clients achieve long-term success and analyze the perceived added value of a policy for society. Meaningfulness refers to workers' perceptions of the contribution a policy makes to a greater purpose, such as societal goals (societal meaningfulness), and the added value of the policy for own clients (client meaningfulness) (Tummers *et al.*, 2009). For instance, client meaningfulness is high when a teacher believes that the policy helps her students to improve their learning outcomes. Meaningful work is of critical importance for frontline workers (May *et al.*, 2004; Grant & Berry, 2011) and numerous studies have found a strong and positive correlation between meaningfulness and implementation willingness (Tummers, 2011; Loyens, 2015; Van Engen *et al.*, 2016; Van der Voet *et al.*, 2017).

The bottom-up view acknowledges that policy changes arise from the interaction of policy and setting, and should be consonant with the values of implementing agents (Matland, 1995). If frontline workers experience discretion, they can tailor the policy to the specific situation of the clients, thereby increasing their perception of its meaningfulness. The implementing actors' perceptions, in turn, can be decisive for implementation outcomes. In summary, powerfulness adds to meaningfulness, which in turn fosters implementation willingness (Lipsky, 1980; Matland, 1995; Tummers & Bekkers, 2014).

Hence, our third hypothesis expects that frontline workers who both feel powerful and perceive the policy as meaningful are willing to implement the

policy. This hypothesis does not rule out that high implementation willingness can also result from other factors. As Figure 5.1 illustrates, it simply assumes that frontline workers who both feel powerful and find the policy meaningful are a subset of the frontline workers who are willing to implement the policy. *The combination of policy powerfulness (strategic, tactical, or operational) with policy meaningfulness (societal meaningfulness: SM, or client meaningfulness: CM) is a quasi-sufficient condition for high implementation willingness*:

Hypothesis 3: (SP + TP + OP) * (SM + CM) → W

It should be noted that these two interpretations are compatible: Powerfulness can be quasi-necessary for implementation willingness (hypothesis 1), and in combination with meaningfulness, quasi-sufficient (hypothesis 3). However, they are not identical: the first interpretation thinks of powerfulness as a prerequisite for implementation willingness (necessity), while the second one assumes that powerfulness in situations of meaningfulness typically results in high willingness to implement (sufficiency). They also represent two different variants of the bottom-up view on discretion-as-perceived. The first interpretation hypothesizes an enabling, but not automatically triggering role of powerfulness for frontline workers' willingness to implement. The second interpretation highlights the decisiveness of implementing actors' perceived meaningfulness of policies, and assumes that the degree of policy meaningfulness interacts with policy powerfulness to trigger implementation willingness. We may find that powerfulness enables, but does not always result in implementation willingness (interpretation 1 supported), while its combination with meaningfulness is not decisive for implementation willingness (interpretation 2 rejected) – or vice versa. Finally, we do not rule out that other factors than powerfulness and meaningfulness influence implementation willingness. Indeed, bottom-up perspectives highlight various factors that can impact policy implementation. Furthermore, the effects of motivating factors can differ between individuals and situations. Our goal is to clarify the motivating role of powerfulness for, rather than comprehensively explain, implementation willingness. In addition, we identify the empirical relevance of powerfulness and meaningfulness for explaining implementation willingness (Sachau, 2007).

5.3 METHOD

Above, we have theorized the role of powerfulness for implementation willingness as an asymmetric and non-linear effect. While a variety of techniques can detect non-linear effects (e.g., polynomials; see also Matzler & Renzl, 2007), we use large-N set-theoretic configurational analysis (Ragin, 1987, 2000; Schneider & Wagemann, 2012). We chose this method as it is the only available technique that models three central theoretical features of our framework (software: R packages QCA and SetMethods; Medzihorsky *et al.*, 2017; Dusa, 2018). First, set-theoretic configurational comparative methods are designed to assess subset relations like the ones hypothesized in Figure 5.1 in terms of necessity and sufficiency. Accordingly, high implementation willingness can have different causes than low implementation willingness. Second, they also provide the possibility of equifinality, meaning that various scenarios can result in high or low implementation willingness: many (but not all) roads lead to Rome. This allows for motivations to differ between individuals. Third, conjunctural explanations are possible, capturing that case-specific factors affect implementation willingness in combination rather than in isolation (Schneider & Wagemann, 2012). We need this possibility to test our third hypothesis. Configurational set-theoretic techniques can be applied to a large-N setting (Fiss, 2011). For theory-testing research designs like ours, large case numbers provide for a more robust test of the theory than small samples (Greckhamer *et al.*, 2013; Thomann & Maggetti, 2017).

Given that this method is not widely used in public administration, we shortly explain its rationale (for detailed descriptions, see Fiss, 2011; Schneider & Wagemann, 2012; Thomann & Maggetti, 2017). The set-theoretic method applied focuses on configurations of variables as sets in which cases have membership or not. The attribution of cases to sets is called calibration. Fuzzy sets allow us to account for differing degrees to which frontline workers' perceptions are present. Qualitative anchors determine the stage at which the outcome or condition is deemed fully present (fuzzy value 1), fully absent (fuzzy value 0) and an indifference (or crossover) point at 0.50. Contrary to usual measurement scales, the crossover point establishes the difference in kind. For example, fuzzy values in the set 'high implementation willingness' above 0.50 mean that implementation

willingness is quite high (W), while values below 0.50 indicate that implementation willingness is quite low (~W).

We can think of necessary and sufficient conditions as subset relations. For instance, our first hypothesis states that frontline workers with high implementation willingness are a subset of those frontline workers who feel powerful. Large-N applications integrate probabilistic elements to capture the degree to which a majority of cases correspond to the statement that X is a superset of Y (quasi-necessity; X≥Y), or a subset of Y (quasi-sufficiency; X≤Y) (Ragin, 2000). The analysis of necessity starts with identifying simple conditions that are a superset of (that is: necessary for) the outcome (here: high implementation willingness). If no simple condition proves necessary, further simple conditions can be added disjunctively until necessity is obtained (Thiem, 2014). We interpret those supersets as necessary conditions that make theoretical sense against the background of our hypotheses, and meet the criteria outlined below (cf. Schneider & Wagemann, 2012).

For the analysis of sufficiency, a 'truth table' is constructed. The rows of the truth table indicate all possible combinations. This enables us to attribute the cases accordingly to the truth table and identify empirically unobserved configurations (so-called logical remainders). If all or enough cases' fuzzy set membership in a truth table row is smaller than or equal to its membership in the outcome, then the row is identified as a sufficient configuration for the outcome. For example, if those frontline workers who partly or fully feel strategically, tactically and operationally powerful and think the policy makes sense for clients and for society are also rather or fully willing to implement the policy, then this configuration of attitudes is sufficient for high implementation willingness. The logical minimization process then identifies the shortest possible expression depicting the configurations that imply the outcome - the solution term. This is a straightforward procedure that relies on basic set theory: for example, A*B*C + A*B*~C can be reduced to A*B (Thomann *et al.*, 2018a).

To evaluate our results, we use consistency and coverage measures. The values of these fit indices can range from 0 (low) to 1 (high). Consistency is the extent to which the results are in line with the statements of necessity or sufficiency. For sufficient conditions, consistency is indicated for single truth table rows (raw consistency), for single configurations of, or for the whole solution term.

Table 5.1 Strategies to address errors and evaluate model

	Issue	Definition	Strategy	Application
Possible error sources	*Deviant case & measurement errors*	Errors related to sensitivity to one or more flawed cases	*Frequency thresholds robustness test*	Use of three different frequency thresholds; configurations without a certain frequency are treated as logical remainders
		Sensitivity to changes in raw consistency levels	*Raw consistency robustness test*	Use of three different raw consistency thresholds (criterion: PRI)
	Plausibility & tenability	Limited diversity & contradictions can trigger inferences that are implausible and/or contradictory	*Enhanced Standard Analysis*	Intermediate solution, based on directional expectations and exclusion of contradictory rows and untenable assumptions
	Accuracy	Degree to which observations correspond to set relation	*Consistency*	Necessity: ≥0.90 Sufficiency: ≥0.75
Criteria for model evaluation		Simultaneous subset relations: degree to which the same condition is not simultaneously sufficient for the negated outcome	*Proportional Reduction in Inconsistency (PRI)*	No fixed threshold
	Explanatory power	Empirical relevance of model	*Coverage & Relevance of Necessity*	Necessity: ≥0.60 RoN ≥0.60 (direct calibration) / 0.55 (recoding method) Sufficiency: verbal interpretation Low coverage indicates low explanatory power
	Random errors	Errors that are unpredictable and inconsistent in their magnitude or direction (e.g., because of estimation and personal factors in surveys)	*Probabilistic criteria*	Right-handed Z-Test for proportion of cases with X≥X (necessity), X≤ Y (sufficiency) 0.8: 'almost always'
	Limited empirical diversity	Presence of logical remainders, i.e. truth table rows without enough cases with membership > 0.5	*Limited diversity index* % remainders / logically possible configurations	Models with less limited diversity have a stronger empirical basis
	Ambiguity	Patterns in data are unclear: several equally non-redundant solutions can be derived	*Ambiguity index* (Nr. of equally plausible models)	Unambiguous models are preferred (row dominance applied)

Table 5.1 Strategies to address errors and evaluate model (continued)

	Issue	Definition	Strategy	Application
Criteria for model evaluation	*Robustness*	Terms of enhanced parsimonious solution remain robust across different models that pass consistency threshold 0.75	*Robustness index* Average % of models in which (a subset of) a term appears	More robust models are preferred
	Skewness	Skewed distributions can produce simultaneous subset relations, exacerbate limited diversity, and strongly distort parameters of fit	*Skewness statistics*	% of cases with membership >0.50 in sets is reported Skewness is problematic if the vast majority (>85%) of the cases cluster *in only one* of the four possible intersecting areas of the XY plots with two diagonals

Furthermore, the proportional reduction in inconsistency (PRI) indicates the degree to which a given configuration is not simultaneously sufficient for both the occurrence and the non-occurrence of the outcome. Coverage sufficiency depicts how well the model explains the available empirical information. Raw coverage expresses how much a single configuration covers, and unique coverage indicates how much it uniquely covers. Low coverage means that the model has a limited capacity to explain the outcome. For necessary conditions, coverage expresses their relevance in terms of the condition set not being much larger than the outcome set, and the relevance of necessity (RoN) in terms of the condition being close to a constant (all formulae in Schneider & Wagemann, 2012).

Error management is a salient issue for large-N applications of set-theoretic configurational comparative methods (Maggetti & Levi-Faur, 2013; Thomann & Maggetti, 2017). In the absence of established guidelines, we propose state-of-the-art strategies that complement the traditional parameters of fit to address possible error sources, as shown in Table 5.1. To account for different possible model specifications and to assess robustness, we calculated 54 models, using two calibration techniques (see below) and three different raw consistency and frequency thresholds. The models presented in the paper rank best on eight criteria for model evaluation, see Table 5.1. The rationale underlying the choice of different analytic thresholds and the "best" models for interpretation is outlined in detail in box 5.1.

We assess hypothesis 1 on necessary conditions in Figure 5.2. To assess our hypotheses on sufficient conditions (hypotheses 2 and 3), we apply Ragin's (1987) principles of formal set-theoretic theory evaluation, as extended by Schneider and Wagemann (2012) to account for consistency and coverage. This procedure identifies the proportion of cases that confirm, refute or extend our theoretical expectations. To this end, the scenarios expected (T) and those not expected (~T) in the hypotheses were intersected with the scenarios that were empirically (not) observed (S and ~S). This technique helps us answer three questions. First, which parts of the theory are supported by the findings (T*S and ~T*~S)? Second, in which direction should theory be expanded (~T*S)? Third, which parts of the theory need to be dropped (T*~S)? Table 5.2 summarizes the main analytic steps (on p. 96).

Table 5.2 Main steps of the large-N set-theoretic configurational analysis

Step 1	Analysis of necessity (H1)	Identify the supersets of high implementation willingness for both datasets, using two calibration strategies
Step 2	Analysis of sufficiency	Identify subsets of low and high implementation willingness, using both datasets, two calibration strategies, three different raw consistency thresholds and three different frequency thresholds
Step 3	Model evaluation, analysis of sufficiency	Identify best-performing model for each outcome, dataset and calibration strategy (for criteria, see Table 5.1)
Step 4	Model selection, sufficient conditions	Identify the models with highest explanatory power per dataset and outcome for interpretation
Step 5	Formal set-theoretic theory evaluation (H2 and H3)	Identify how results behave with respect to the hypotheses: which (parts of) the hypotheses are supported, which ones are refuted?

The data, truth tables, directional expectations, conservative and parsimonious solutions, simplifying assumptions, skewness tests, *R* codes for replication, and the results not reported in this are all provided as online supplementary material.[2]

5.3.1 Data

We used two data samples collected in the Netherlands in two sectors (healthcare and education) at two times (2010 and 2013). By analyzing these two datasets, we

2 The online Appendix and replication materials are published at dataverse, see http://dx.doi.org/10.7910/DVN/G9PYIV .

both evaluate whether our hypotheses hold for frontline workers implementing a specific policy (dataset 1), and whether the hypothesized relations hold in another policy sector and from a more general perspective (dataset 2). This allows us to adopt a comparative approach and provides a stronger empirical basis to either accept or reject the hypotheses. Still, in examining two case studies, the possibility to make general claims remains limited. This is acknowledged and will be discussed in the discussion section.

Box 5.1 Procedure for model evaluation and selection, analysis of sufficiency

Setting raw consistency thresholds is decisive for determining which conditions are sufficient. Since consistency values strongly depend on the specific dataset, truth table and case distributions, there are no fixed anchors for setting these thresholds (Schneider & Wagemann, 2012; Thomann & Maggetti, 2017). Accordingly, using standardized thresholds is widely considered bad practice (Wagemann *et al.*, 2016). Therefore, we use a context-sensitive strategy that integrates PRI values for determining raw consistency thresholds. Considering the range of PRI values in a truth table, a context-specific critical PRI value was determined. This procedure ensures that raw consistency is set such that simultaneous subset relations – when the same configuration is considered sufficient for both low and high implementation willingness – are avoided (Schneider & Wagemann, 2012). The first raw consistency threshold was set above the first row with a PRI below this critical value; the second threshold was set above the second row with a PRI below that value; and the third threshold, above the third respective row. Hence, the same principle was applied to each analysis, but considering the specificities of the respective truth table.

Tables B2-B7 in the online Appendix report all resulting models and illustrate their robustness. The 'best' models for each dataset, calibration strategy and outcome (high and low implementation willingness) were then identified according to their performance regarding consistency, PRI, coverage, statistical significance, limited diversity, ambiguity, robustness and skewness. These criteria comprehensively capture the main challenges to validity with set-theoretic techniques (Thomann & Maggetti, 2017; Table 5.1). The best model is the one whose average rank on each of these indicators is the highest amongst those models with a minimum consistency of 0.75. Below this threshold, QCA solutions are usually not considered sufficient (Schneider & Wagemann, 2012). The ranking procedure is self-explanatory for consistency, PRI, coverage, Z values and robustness. Additionally, high levels of limited diversity and model ambiguity were punished, by rewarding the lowest levels a ranking of 1; the highest level is attributed the lowest possible rank (e.g., 7 if 7 models pass the consistency threshold); then the second highest level is attributed the second worst rank, and so on. The motivation for this was that limited diversity poses serious threats to inferences with truth table analyses (Thomann & Maggetti, 2017) and model ambiguities indicate that the results are inconclusive (Baumgartner & Thiem, 2017).

This left us with six sufficient models, among which the ones with the highest explanatory power (coverage) were preferred for each outcome and dataset, reported in Table 5.4 and chosen for interpretation.[3] This procedure minimizes the weakness of many large-N set-theoretic configurational analyses, which often suffer from very limited coverage (Wagemann *et al.*, 2016).

3 No analysis of sufficiency was possible for dataset 2 using the recoding method.

Dataset 1

The 2010 study ('study 1') investigated whether Dutch mental healthcare workers felt alienated from one specific government policy program, namely, the Diagnosis-Related Group (DRG) policy, and their willingness to implement this new policy. The DRG policy was developed by the Dutch government as a means to determine the level of financial reward for mental healthcare provision by stipulating a standard rate for each disorder. The sampling frame consisted of 5.199 professionals who were members of two nationwide mental healthcare associations (see Tummers *et al.*, 2012). Using an e-mail and two reminders, 1.317 returns of the questionnaire were received (25% response). The gender composition of the respondents was 66% female. This is consistent with the Dutch average (69%) for mental healthcare professionals. The average age was slightly higher than that of the mental healthcare professional population (48 versus 44). Common reasons for not participating were a lack of time, retirement, change of occupation, or not working with the DRG policy.

Dataset 2

The 2013 study ('study 2') investigated whether Dutch teachers felt alienated from government education policies in general, and the relationship with their general willingness to implement government policies. The sampling frame consisted of a nation-wide sample of 2.863 teachers working in secondary education, selected through the pension fund for all Dutch employees in government and education (ABP) (Van Engen *et al.*, 2016). Using an e-mail and one reminder, 1.096 returns of the questionnaire were received (38% response). On average the respondents were 51 years old, and 59 percent were male. Dutch national statistics on secondary school teachers in 2013 have shown that the average age is 46, and 48 percent are male. In our sample males were therefore somewhat overrepresented, and the respondents were on average slightly older than the national average. To rule out a non-response bias, we asked the organization managing the sampling frame to analyze whether or not the respondents problematically differed from non-respondents in terms of variables such as age, gender, and occupation. For instance, the results indicated there were no significant differences between the two groups in terms of occupation (respondents with managing responsibilities: 8%; non-respondents: 9%). They also indicated that the arguments non-respon-

dents gave for not participating usually were 'no time', 'forgot the questionnaire' and 'did not open e-mail during response period'. Nevertheless, it is important to highlight that although we argue that our data is fairly representative, it is still possible that some type of response bias could have influenced our results.

5.3.2 Measures

The measures of implementation willingness, powerfulness and meaningfulness were formatted using five-point Likert scales. All measures had adequate Cronbach alphas (ranging between 0.78 and 0.97).

In dataset 1 we measured policy powerfulness (strategic, tactical and operational powerfulness: six indicators) and policy meaningfulness (societal: twelve indicators, client: four indicators) for a specific policy using the policy alienation measurement scales of Tummers (2012). In dataset 2 we measured general policy powerfulness (strategic, tactical and operational powerfulness: six indicators) and general policy meaningfulness (societal and client: four indicators) using the general policy alienation measurement scales of Van Engen *et al.* (2016). Implementation willingness was measured using five indicators corresponding to the validated scale by Metselaar (1997). If necessary, we inverted the positive and negative end of the respective scales, so that high scores always indicate high powerfulness, meaningfulness, and implementation willingness.

5.3.3 Calibration

Indicator variables were calibrated into indicator sets. Set membership requires a statement about a qualitative state: cases are either (more or less) in a set or (more or less) out of a set. The answer categories of Likert scales have a fixed qualitative meaning, which can be directly translated into set membership scores. For example, if a frontline worker answers 'disagree' (score of 2 on 1-5 scale) to the question 'In my organization, professionals could take part in conversations regarding the execution of the policy', then this means that on this item the case 'tactical powerfulness' is rather absent, but not totally absent.

The neutral answer (score of 3) poses a conceptual challenge for calibrating set membership (Wagemann *et al.*, 2016). In box 5.2 we discuss in detail the nature of this challenge and how we address it.

Box 5.2 Procedure to test for different calibration strategies

The neutral answer (score of 3) poses a conceptual challenge for calibrating set membership. Neutral answers could indicate that a frontline worker experiences neither the presence nor the absence of, say, tactical powerfulness (point of indifference). However, cases with a set membership score of 0.50 cannot be attributed to truth table rows, which results in excessive dropout rates and should therefore be avoided (Wagemann *et al.*, 2016). While Likert scales are typically acknowledged to represent ordinal rather than interval-level data (Wirth & Edwards, 2007), the status of neutral answers in the scale and hence also in the set can be disputed. One possible interpretation is that the answer 'neither agree nor disagree' indicates less agreement than 'rather agree', but more agreement than 'rather disagree' – we can treat the answers as scale. However, another possible interpretation is that 'neither agree nor disagree' indicates both 'no agreement' as well as 'no disagreement' – in other words, no presence, of, say, powerfulness at all. Hence, these cases would in fact be 'fully out' of the set of, for example, tactical powerfulness. Different calibration techniques can substantially affect the results of set-theoretic configurational analyses (Skaaning, 2011). To identify the best calibration strategy, we tested for two different commonly used calibration techniques for Likert scales. First, the direct method of calibration uses a logistic function to fit the raw data in-between the three qualitative set membership anchors (Schneider & Wagemann, 2012). This method is very popular in large-N set-theoretic configurational analyses. Typically, the crossover point is set right above the indifferent answers, resulting in set memberships extremely close to 0.50 that can hardly be interpreted in conceptual terms. As Wagemann *et al.* (2016, p. 55) point out: "This is arbitrary and should not become common practice. (…) [it] does not have much to do with a decision about set membership".
To avoid this pitfall, we interpret neutral answers as 'fully out' of the set (the cases remain in the sample, but they have a set membership of 0). Answers of 4 (agree) and 5 (fully agree) were recoded into 3 and 4 before calibration. Second, we alternatively treated the answers strictly as a scale using simple recoding technique, which involves the grouping of cases into previously defined set-membership scores (Schneider & Wagemann, 2012). Here, we followed the proposal by Emmenegger *et al.* (2014) (and slightly adapted it to account for degrees of non-membership) and used the calibration anchors shown schematically in Table 5.3.
Our results indicate that in the analysis of sufficiency, the recoding method works better for dataset 1 (the models perform better and explain more cases), while for dataset 2, the direct strategy is more feasible – recoding method leads to distorted parameters of fit that prevent a meaningful analysis of sufficiency. Importantly, however, both calibration strategies attribute indifferent answers as more out than in the set, resulting in the same conceptual meaning and attribution of cases to truth table rows. The differences in the results are thus exclusively due to changes in the parameters of fit. The results of necessity are robust regardless of the calibration strategy. Using the direct strategy for dataset 1 for sufficient conditions leads to the same overall conclusions regarding our hypotheses as with the indirect strategy. For these reasons, we adopted the recoding method for dataset 1 and the direct calibration method for dataset 2 for the results interpreted below.

In short, we conceive of indifferent values as more out than in of the set. To identify the best calibration strategy, we tested for two different commonly used calibration techniques for Likert scales. First, the direct method of calibration uses a logistic function to fit the raw data in-between the three qualitative set membership anchors (Schneider & Wagemann, 2012). Using our data, this commonly applied technique results in set membership scores of 0.05, 0.27, 0.73

and 0.95; indifferent answers were coded as 'fully out'. Second, we alternatively treated the answers strictly as a scale using a simple recoding technique. This technique involves the grouping of cases into previously defined set-membership scores (Schneider & Wagemann, 2012; Emmenegger *et al.*, 2014), see Table 5.3 for an example. Based on assessment of their performance, we adopted the recoding method for dataset 1 and the direct calibration method for dataset 2 for the results interpreted below. Both strategies attribute the same values on the Likert scale as more in/more out of the set, resulting in the same conceptual meaning, but different parameters of fit.

Table 5.3 From Likert scale to indicator sets: an example of recoding method

Likert score	**Indicator fuzzy set score**
Survey question: 'I intend to put effort into achieving the goals of the DRG policy'	*Set: 'High implementation willingness, indicator 2'*
Completely agree (5)	Highly willing (1)
Agree (4)	Mostly but not highly willing (0.8)
Neutral (3)	Rather unwilling (0.4)
Disagree (2)	Mostly but not fully unwilling (0.2)
Completely disagree (1)	Fully unwilling (0)

Missing values make it impossible to attribute cases to truth table configurations. This is a potential issue since a high share of cases has missing values on at least one indicator set in dataset 1. This is due to the fact that we gave the possibility to indicate 'don't know' for each item in dataset 1 and doing this on one out 39 items already indicates a missing value (60% in dataset 1, 7.8% in dataset 2). Excluding these cases from the analysis would result in a biased sample.

The aggregation strategy will impact the analysis. It needs to avoid such excessive dropout, while ensuring construct validity and avoiding overly skewed condition and outcome sets. The first out of three aggregation options would be building averages across the indicators. Doing so for raw values would negatively affect construct validity: the inclusion of neutral answers (score 3) leads to average values that are difficult to interpret especially since they are numerous. Calculating averages of calibrated sets is equally problematic because it can result in set memberships of 0.50, producing dropouts during truth table analysis. The second and third options are set-theoretic. Using the logical 'AND' as aggregation

strategy (minimum rule) represents a very restrictive conceptualization, as all indicators need to be present simultaneously for an attitude to be present. This results in the excessive dropouts. Moreover, it would produce highly skewed sets that make it impossible to proceed with the analysis of the outcome (Schneider & Wagemann, 2012). For example, in dataset 1, none of the aggregated sets would have more than 5% cases with membership above 0.50.

Accordingly, as the third and in our view superior option, we use the logical 'OR' to aggregate the indicators into the five condition sets. This aggregation strategy conceives of different indicators as functional equivalents that indicate the presence of an attitude (Goetz & Starr, 2003). For instance, it suffices for a frontline worker to have a score on one of the five indicator sets for implementation willingness ('W') to obtain a value for 'W' (maximum rule). This 'optimistic' measure lowers the dropout problem (final *N* for dataset 1=1.004, dropout 23.8%; for dataset 2=1.087, dropout 0.8%) and produces acceptable levels of skewness that enable an analysis of the outcome. This has consequences in terms of concept validity: the positive memberships in sets represent a wider range of functionally equivalent attitudes, which are assumed to represent the concept. This conceptualization does justice to the wide range of experiences facing frontline workers on the ground.

5.4 RESULTS

We can now test the hypotheses. Table B1 in the online Appendix displays descriptive statistics. They show that overall, the Dutch teachers (study 2) have a more positive attitude than the healthcare workers (study 1). They feel more powerful, perceive the policies as more meaningful, and have higher implementation willingness.

Regarding hypothesis 1, we indeed found that feelings of powerfulness are almost always necessary for high implementation willingness. This holds for both datasets and regardless of the calibration strategy used (see Table A1, online Appendix). This is shown in Figure 5.2. In the Dutch education sector, either strategic, tactical or operational powerfulness is needed for high implementation willingness. Among Dutch healthcare workers, the finding is even stronger:

it is enough for high implementation willingness to either feel powerful at the strategic or operational level, or alternatively, to feel powerful at the operational or tactical level. These results provide strong support for the hypothesis that powerfulness at different levels is a prerequisite for implementation willingness.

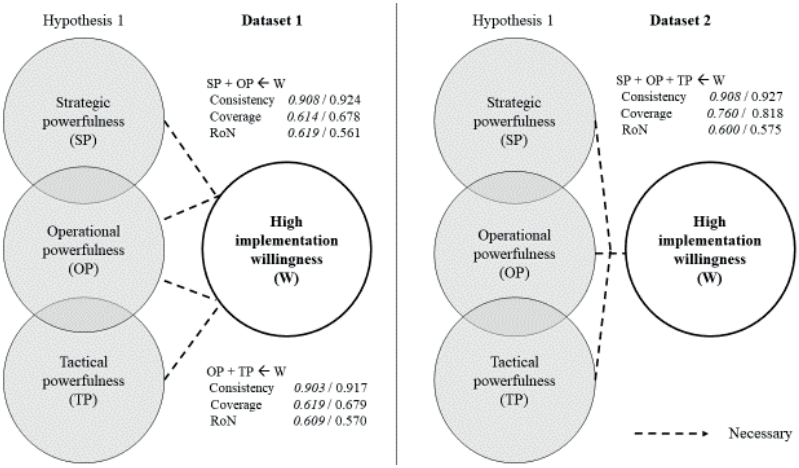

Figure 5.2 Evaluation hypothesis 1

Hypothesis 2 captured a potential consequence of the first hypothesis, namely, that a lack of powerfulness might be quasi-sufficient for low implementation willingness. Table 5.4 reveals three configurations in dataset 1, and five configurations in dataset 2, that are almost always sufficient for low implementation willingness. The Dutch health workers who are unwilling to implement the DRG policy consistently experience low levels of powerfulness and, in path 3, meaningfulness. Conversely, in the education sector, the picture is less clear at first sight: these configurations entail a mix of both positive and negative attitudes. The parameters of fit score well in dataset 1, while in dataset 2, the results are highly consistent, but have a fairly low empirical relevance (coverage).

We indicate the percentage of all cases that display these attitudes with different levels of implementation willingness, and what that means for interpreting the results. For example, in the upper left quadrant, those frontline workers that display these attitudes and have low implementation willingness support the

Table 5.4 Sufficient conditions for implementation willingness (intermediate solution), best-performing models

	Hypothesis 2: Low implementation willingness (~W)								**Hypothesis 3:** High implementation willingness (W)						
	Dataset 1 Recoding method			**Dataset 2 Direct calibration**					**Dataset 1 Recoding method**			**Dataset 2 Direct calibration**			
Configuration	**1**	**2**	**3**	**1**	**2**	**3**	**4**	**5**	**1**	**2**	**3**	**4**	**1**	**2**	**3**
High strategic powerfulness (SP)	○				○	○	○	●		●	○	●		●	
High tactical powerfulness (TP)		○				○	○	○		●	●	●			●
High operational powerfulness (OP)	○	○	○	○		○	●	○	●		●	○	●		
High societal meaningfulness (SM)			○	○	○		●	●			●	●	●	●	●
High client meaningfulness (CM)			○	●	●	●	○	○	●	●			●	●	●
Consistency	0.865	0.855	0.858	0.872	0.873	0.883	0.883	0.881	0.909	0.935	0.942	0.942	0.939	0.936	0.938
Raw coverage	0.525	0.531	0.573	0.487	0.475	0.419	0.426	0.442	0.488	0.451	0.350	0.313	0.409	0.396	0.415
Unique coverage	0.010	0.017	0.037	0.012	0.013	0.001	0.006	0.012	0.046	0.008	0.022	0.013	0.005	0.006	0.007
Solution consistency		0.818*				0.826**				0.890***				0.935***	
Solution PRI		0.627				0.360				0.727				0.841	
Solution coverage		0.640				0.567				0.565				0.436	

Explanation: This table shows the combinations of conditions (paths) that were quasi-sufficient for low implementation willingness (left-hand side) and high implementation willingness (right-hand side). Black circles indicate the presence of a condition, and white circles its absence. Blank spaces indicate the irrelevance of a condition. The different paths of one model are interchangeable. For example, in dataset one, low strategic powerfulness with low operational powerfulness typically led to low implementation willingness (path 1); alternatively, low operational powerfulness combined with feelings of client and societal meaninglessness also typically did (path 3). Table 1 explains the parameters of fit. For raw consistency thresholds, frequency thresholds and simplifying assumptions applied, see Tables B10, B11, B12, B16, B17, B18 and B19, online Appendix.

*Significant at 0.1 level, **significant at 0.05 level, ***significant at 0.01 level ('almost always sufficient'). All single paths are significantly sufficient at the 0.01 level.

hypothesis; those that have high implementation willingness are 'contradictions', that is, they separate the quasi-sufficient results from perfect sufficiency.

Using set-theoretic theory evaluation to assess hypothesis 2 formally, we find robust support that the absence of tactical, strategic and operational powerfulness implies low implementation willingness in the healthcare sector. This is shown in Table 5.5. However, quite some cases remain unexplained (lower right quadrant). In addition and compatible to what we hypothesized, the absence of operational, but not also tactical and strategic powerfulness in some situations also leads to low implementation willingness (lower left quadrant). Conversely, in the education sector, overall the empirical support for the second hypothesis is so weak that we must reject it. The contradictory cases are empirically more frequent than those instances that directly support the hypothesis (left-hand side of Table 5.5). Here, the solution term only explains a tiny fraction of the observed patterns of low implementation willingness.

Overall, the conclusion for hypothesis 2 is ambiguous. While powerfulness is a quasi-necessary condition for high willingness, the 'flipside' of this argument materializes in the healthcare, but not in the educational sector. While seemingly puzzling, this finding illustrates that the things that motivate people at the workplace can be different from those that demotivate them (see also Schneider & Wagemann, 2012).

Table 5.5 Evaluation of hypothesis 2

		Empirics	
		Detected in solution	**Not detected in solution**
Theory	**Hypothesized**	**~SP*~TP*~OP + SP*~TP*~OP** *~SM*~CM ***~SP*~TP*~OP*** *~SM*CM + ***~SP*~TP*~OP****CM* 14.1 % / *0.1 %* (~W): support theory 4.8 % / *0.6 %* (W): contradict theory & solution	Empty set ***~SP*~TP*~OP*** *~CM* + ***SP*~TP*~OP*** ***~SM*~CM*** 0 % / *4.5%* (~W): support theory 0 % / *6.6 %* (W): delimit theory
	Not hypothesized	~OP*(SP*~TP + SP*~SM*~CM + ~SP*TP + TP*~SM*~CM) *~SP*OP*~SM*CM +* *~SP*~TP*OP*SM*~CM* *+ SP*~OP*~SM*CM +* *SP*~TP*~OP*SM*~CM* *+ TP*~OP*~SM*CM +* *~SP*TP*~SM*CM* 14 % / *1.4 %* (~W): extend theory 8 % / *2.9 %* (W): empirical contradictions	OP + SP*TP*OP*CM + SP*TP*OP*SM + SP*OP + SP*TP*CM + SP*TP*SM + TP*OP *OP*~SM*~CM + TP*OP*~CM +* *OP*SM*CM + SP*OP + SP*OP*SM*CM* *+ TP*OP*SM + SP*~SM*~CM +* *SP*TP*~CM + SP*SM*CM + SP*TP*SM* *+ ~SP*TP*~OP*~CM + TP*~SM*~CM +* *TP*~CM + TP*OP*SM*CM + SP*TP*OP +* *SP*TP*SM*CM + TP*SM* 28.5 % / *21.6 %* (~W): point to overlooked explanations 30.6 % / *62.3 %* (W): support theory
Supports theory		Extends theory	Delimits theory

Based on Schneider and Wagemann (2012, p. 301).
Bold: hypothesized combinations. No italics: dataset 1 (recoding method), *italics*: dataset 2 (direct calibration). *Hypothesis 2*: **~SP*~TP*~OP → ~W**.
Explanation: This table shows how the results behave with respect to hypothesis 2. The upper left quadrant shows those attitudes that were both hypothesized and observed with a set membership > 0.5. The lower left quadrant displays those attitudes that were not expected, but observed empirically, revealing additional explanations for low implementation willingness. The upper right quadrant refers to attitudes that were expected but not observed in the solution. The lower right quadrant displays those attitudes that are neither hypothesized nor covered by the solution.

We indicate the percentage of all cases that display these attitudes with different levels of implementation willingness, and what that means for interpreting the results. For example, in the upper left quadrant, those frontline workers that display these attitudes and have high implementation willingness support the hypothesis; those that have low implementation willingness are 'contradictions', that is, they separate the quasi-sufficient results from perfect sufficiency.

Hypothesis 3 states that the combination of policy powerfulness (strategic, tactical, or operational) and policy meaningfulness (societal or client meaningfulness) is a quasi-sufficient condition for high implementation willingness. Table 5.4 indeed suggests that the combination of high powerfulness and mean-

ingfulness relate to high implementation willingness. Four configurations are very often sufficient for high implementation willingness in the Dutch healthcare sector, and three are almost always sufficient configurations in the education sector. For example, Dutch healthcare workers who feel powerful at the strategic and tactical level and to whom the DRG policy makes sense for the patients typically make efforts to implement the policy. Both models have a good consistency, while its explanatory power (coverage) is quite low in the education sector. The left-hand side and lower right quadrant of Table 5.6 lend full support to our third hypothesis. Powerfulness, in one of its three variants, combined with meaningfulness almost always results in high implementation willingness. This support is empirically stronger in study 2 (education) than in study 1 (healthcare).

However, findings also restrict the hypothesis to certain circumstances. For instance, the upper left quadrant of Table 5.6 shows that in the healthcare sector, the positive motivational role of tactical powerfulness together with meaningfulness often unfolds even in the absence of either strategic or operational powerfulness. In the education sector, regardless of the type of powerfulness typically both societal and client meaningfulness must be present. Conversely, the instances in which hypothesis 3 is rejected both datasets are negligibly rare (upper right quadrant).

In summary, both bottom-up interpretations (hypothesis 1 and 3) of how perceived discretion motivates frontline workers are indeed reflected in our data. Hypothesis 2 is supported for the first dataset (healthcare) but rejected for the second (education). However, for the second interpretation there is also room for improvement, as quite some cases are not explained (23.4% in dataset 1 and 41.7% in dataset 2 point to overlooked explanations). This is not particularly high, as we aimed to explain willingness with just a few indicators and the unexplained variance is quite low. In field studies in social sciences, we should not expect a perfect theory explaining everything. It suggests that powerfulness combined with meaningfulness is only one of several factors that explain frontline workers' high implementation willingness.

Table 5.6 Evaluation of hypothesis 3

		Empirics	
		Detected in solution	**Not detected in solution**
Theory	**Hypothesized**	**SP*SM*(TP*CM** + **TP***~OP) + **SP*CM*(OP** + **TP** + **TP***~OP***SM)** + **TP*CM*(OP*SM** + **OP) + OP*SM*(CM** + **SP*TP*CM** + ~SP***TP)** + **OP*CM** + **OP*CM*(SP*TP** + ~SP***TP*SM** + **SP* SM)** ***OP*SM*CM** + ***SP*OP*SM*CM** + ***TP*OP*SM*CM** + ***SP*SM*CM** + ***SP*TP*SM*CM** + ***TP*SM*CM*** 9.7 % / *15.7 %* (W): support theory 4.4 %/ *1.2 %* (~W): contradict theory & solution	**SP*SM*(OP***~CM + + ~TP*~CM + ~TP*~OP + **TP*OP***~CM) + **SP***~TP*~OP***CM** + ~TP***OP*SM***~CM + ~SP***TP***~OP***SM** + ~SP***TP***~OP***CM** ***OP*SM***~*CM* + ***OP***~*SM*CM* + ***SP*SM***~*CM* + ***SP***~*SM*CM* + ***TP*SM***~*CM* + ***TP***~*SM*CM*** 10.3% / *15 %* (W): support theory 6.3 % / 4.3 *%* (~W): delimit theory
	Not hypothesized	Empty set *Empty set*	~SP*~TP*~OP*~CM + ~SP*~TP*~OP + ~SP*~TP*~OP*~SM + ~SM*~CM + ~SP*~OP*~SM*~CM + ~TP*~OP*~SM*~CM *~SP*~TP*~OP*~CM + ~SP*~TP*~OP + ~SP*~TP*~OP*~SM + ~SM*~CM + ~SP*~TP*~OP*~SM*~CM* 23.4 % / *41.7 %* (W): point to overlooked explanations 46 % / *22.1 %* (~W): support theory
Supports theory		Extends theory	Delimits theory

Bold: hypothesized combinations. No italics: dataset 1 (recoding method), *italics*: dataset 2 (direct calibration). *Hypothesis 3*: OP*SM + OP*CM + SP*SM + SP*CM + TP*SM + TP*CM → W.
Explanation: This table shows how the results behave with respect to hypothesis 3. The upper left quadrant shows those attitudes that were both hypothesized and observed with a set membership >0.5. The lower left quadrant displays those attitudes that were not expected, but observed empirically. The upper right quadrant refers to attitudes that were expected but not observed in the solution. The lower right quadrant displays those attitudes that are neither hypothesized nor covered by the solution.

5.5 DISCUSSION

The main conclusion of our study is that discretion-as-perceived is a quasi-necessary condition for high implementation willingness. This aligns with Herzberg's motivation theory and suggests an enabling (but not automatically triggering) motivational effect of perceived discretion (Herzberg *et al.*, 1959; Goertz & Starr, 2003). Frontline workers need to feel that they can influence the policy – this is a necessary condition.

Secondly, we have found mixed evidence for the hypothesized more radical 'flipside' of the first interpretation. This result aligns with a classic insight from Herzberg's motivation theory: the things that make people feel satisfied and motivated on the job can be different in kind from the things that make them feel dissatisfied – and this can obviously vary between policy sectors and types of professions (Herzberg *et al.*, 1959; Bassett-Jones *et al.*, 2005; Sachau, 2007).

Thirdly, we also found that - in combination with policy meaningfulness - powerfulness is quasi-sufficient for high implementation willingness. When frontline workers felt that they had both high powerfulness and that the policy was meaningful for society, this strengthened their willingness to implement it (Maynard-Moody & Musheno, 2012; Van der Voet *et al.*, 2017).

Our results encourage scholars to rethink assumptions of implementation theory by moving from a correlational logic to the consideration of asymmetric patterns. By adapting Herzberg *et al.*'s (1959) seminal, fundamentally asymmetric two-factor theory of motivation to the context of frontline implementation, we are able to refine policy implementation theory. The important role of powerfulness could be uncovered by modeling asymmetric effects via a methodology specifically designed to test these (Ragin, 1987, 2000; Schneider & Wagemann, 2012). Our analysis sheds more light on the puzzling results of previous studies, which assumed symmetric, correlational patterns (Tummers, 2011; Van Engen *et al.*, 2016). The strong and robust asymmetric effect of powerfulness that we detected simply escaped the attention of these studies because their designs are unable to detect such asymmetric relationships (Schneider & Wagemann, 2012). This has helped us to identify discretion-as-perceived as a necessary prerequisite for high implementation willingness. Accordingly, implementation theory might fruitfully turn toward more asymmetric and complexity-oriented models of policy in practice (Raab *et al.*, 2015; Misangyi *et al.*, 2017; Thomann *et al.*, 2018a).

A number of caveats apply for this study. First, apart from powerfulness and meaningfulness, additional factors such as caseloads, interactions, and resources influence frontline workers' implementation willingness (e.g., Sabatier, 1986; O'Toole, 2000; May & Winter, 2009). Second, although we analyzed two large-N datasets, we should be careful to generalize these findings to frontline workers in other policy domains or countries. Third, while applying an 'optimistic' measure of our dependent and independent variables helped us reducing drop-out and

countering the skewness of the data, future research should study whether our results also hold applying 'pessimistic' measures, ideally using large datasets in multiple sectors and countries where cases with missing values can be completely deleted from the dataset. Fourth, although there is a fairly strong correlation between intended behavior and actual behavior (Sheeran & Orbell, 1988; Randall & Wolf, 1994; Armitage & Connor, 2001), future studies could measure behavior more directly. Fifth, it should be noted that common method bias could be a problem in our study, since we used the same data source to measure the variables under study (powerfulness, meaningfulness, implementation willingness). It is recommended that future researchers studying the relationship between powerfulness and implementation willingness apply stronger designs and techniques to establish causal inference. We recommend the use of field, lab or survey experiments.

5.6 CONCLUSIONS

Despite the fundamental theoretical debate on the role of discretion and its relevance for policy design and implementation, to date there has been little empirical research to assess the behavioral assumptions underlying this debate. Our study is the first large-N empirical illustration lending robust support to a bottom-up view on discretion as an inevitable and potentially beneficial aspect of frontline implementation. We find that possibilities to participate in and influence public policies are a *prerequisite* for frontline workers to be willing to implement the policy. However, this is not enough. It is not sufficient. Other factors, including perceiving the policy as meaningful for society and clients, are needed to truly increase the willingness to implement of frontline workers.

Our study contributes to clarifying the behavioral underpinnings of the top-down versus bottom-up debate on discretion (Sabatier, 1986; Hupe, 2013; Thomann *et al.*, 2016). The question whether frontline workers should be granted discretion continues to be hotly debated not only in research on policy implementation, but also on policy, regulatory and organizational design (e.g., Howlett 2004; Chun & Rainey, 2005). Our findings lend substantial support to a bottom-up view of street-level bureaucrats as problem-solvers who crucially

need the freedom to adapt the program to local conditions. Conversely, they lend very little support to top-down assertions that high levels of discretion often or predominantly have a negative impact on policy implementation – at least not at the perceived, motivational level.

The link between implementation willingness and actual implementation behavior - which was not analyzed here - will continue to provide fertile grounds for further exploration (see e.g., Brodkin, 1997; Chun & Rainey, 2005; Gofen, 2014). Committed implementers are a crucial factor for successful policy implementation (May & Winter, 2009). Our contribution lies in showing that the overwhelming majority of those frontline workers with high implementation willingness also experience high levels of discretion. This should encourage scholars and practitioners to move beyond the question whether frontline workers should be granted discretion: our answer to this question is yes.

The more salient question seems to be how to make best use of frontline workers' discretion to encourage behavior that eventually contributes to the achievement of policy goals. Discretion appears as a defining contextual feature of street-level bureaucratic work that changes the daily experiences shared by frontline workers. This emphasizes the importance of future research that singles out how a context of more or less discretion affects frontline workers' actual behavior, and under which specific circumstances.

Finally, systematic comparative empirical assessment of street-level bureaucracy theory like ours demonstrate the potential of large-N comparisons over different policy contexts to facilitate theoretical progress in this field (O'Toole, 2000). A micro-level perspective is useful to evaluate the underlying psychology and mechanisms of frontline implementation (Grimmelikhuijsen *et al.*, 2017). It provides valuable information to policymakers and managers engaged in shaping the macro- and meso-level contexts of street-level bureaucracy, in their continuous quest to improve public service delivery.