Do tax officials use double standards in evaluating citizen-clients?

A policy-capturing study among Dutch frontline tax officials

Nadine Raaphorst*, Sandra Groeneveld** & Steven Van de Walle***

Correspondence address: n.j.raaphorst@fgga.leidenuniv.nl

Forthcoming in *Public Administration*

* Institute of Public Administration, Leiden University, The Hague, The Netherlands

** Institute of Public Administration, Leiden University, The Hague, The Netherlands

*** Public Governance Institute, KU Leuven, Belgium

Abstract

In line with psychological and economic discrimination theories, street-level bureaucracy studies show a direct effect of citizen characteristics on officials' judgments, or show how street-level bureaucrats employ stereotypical reasoning in making decisions. Relying on sociological double standards theory, this study hypothesizes that citizen-clients' status characteristics not only directly influence officials' evaluations, but also *indirectly* and more pervasively by influencing the interpretation of other signals. By means of a policy-capturing study among Dutch frontline tax officials, this study takes a first step in testing double standards propositions in the context of official-citizen encounters. The findings support only some hypotheses, but indicate that citizen-clients' level of education could serve as a moderating context affecting the interpretation of cues. The article provides important theoretical and methodological guidelines for future research on stereotyping at the frontline.

Introduction

Street-level bureaucrats typically have considerable leeway to make judgments about citizen-clients (Lipsky 1980). Research on street-level bureaucrats, such as police officers or teachers, has shown how discretionary judgments sometimes overlap with citizens' supposed belonging to certain social groups, such as someone's race (e.g. Epp *et al.* 2014), social class (Harrits and Møller 2014), or gender (Johnson and Morgan 2013). It has been shown that, due to a lack of information, time and other resources, street-level bureaucrats develop shortcuts such as stereotypes to categorize clients (Lipsky 1980; Prottas 1979). In situations with only limited information and time pressure, the matching of citizen characteristics to stereotypes gives officials information they would otherwise not have (e.g. Maynard-Moody and Musheno 2003).

Within the public administration literature there is a lack of explanatory studies focusing on how cultural beliefs about social groups play a role in the public encounter, and affect the judgments of frontline officials (but see Andersen and Guul 2016; Harrits and Møller 2014; Schram *et al.* 2009). This is particularly interesting given the fact that frontline officials are encouraged to be flexible and to be responsive to citizens' situations when making decisions (e.g. Rice 2017). In fact, interpersonal notions as trust and collaboration have come to play an important role in frontline decisions (Bartels 2013; Yang 2005). In such contexts, officials have more room for interpretation, and leeway in using their own standards to assess who is trustworthy and who is not. Therefore, this flexibility paves the way for stereotyped images and double standards to inform judgments.

The sociological status characteristics theory holds that in situations entailing interpersonal task situations, where there is a distinction between 'failure' and 'success', evaluators look at people's status characteristics to evaluate their likely behavior and achievements (Berger *et al.*

1972). These characteristics are socially recognized attributes on which people are perceived to differ, such as ethnicity, gender or education. Status characteristics are associated with 'cultural beliefs of greater competence in those with more valued states of the characteristic' (Ridgeway 1991, p. 368). As a consequence, it is held, similar situations implying equal competences, are evaluated differently for lower status groups than for higher status groups. By testing the explanatory power of double standards theory using a policy-capturing design, this article sets out to examine how stereotyping at the frontline may be more indirect (i.e. also indirectly leading to unequal judgments) and pervasive (i.e. affecting the interpretation of other signals) than is hitherto studied within public administration research. This study thereby provides a first step in testing the explanatory potential of double standards theory in a public administration context.

In what follows, we will discuss previous research on stereotyping in frontline work more broadly. We will subsequently present our theoretical framework, describe the research setting and formulate hypotheses. Then we will describe our policy-capturing design and discuss our findings.

Stereotyping at the frontline

The literature on stereotyping at the frontline is diverse and entails different perspectives on stereotyping. Notwithstanding the differences, most of these studies focus on *direct* ways of stereotyping, i.e. how evaluations are affected by stereotypes or based on stereotypical reasoning.

In line with the economic theory of statistical discrimination, there are studies that assume that the use of stereotypes is based on statistical knowledge or prior experience to reduce uncertainty

(e.g. Gambetta and Hamill 2005; Harris 1999). Studies show how service workers in general or officials within certain professions explicitly construct types of clients that are inextricably linked to certain groupings in society. Stroshine *et al.* (2008) for example, show how police officers find black people driving in dilapidated cars in white neighborhoods suspicious. Within such studies, the mechanism of discrimination studied is *direct*: cues lead people do distinguish between social categories regardless of any other relevant characteristics. Observational studies that point to the stereotypical reasoning employed by frontline workers in reaching decisions (e.g. Maynard-Moody and Musheno 2003; Dubois 2010) also fall in this category, since they point to how differential evaluations of for instance deservingness overlap with distinctions between social groups.

Within the street-level bureaucracy literature there are only some studies that focus on indirect mechanisms of stereotyping. A study by Harrits and Møller (2014) shows how social workers' tendency to suggest interventions in similar situations is different for low and high class citizens than for middle class citizens. Drawing on the sociological literature on normality and categorization, they find some evidence that the social distance between workers and citizen-clients in interactions implicitly influences their judgments. Moreover, the experimental vignette study by Schram *et al.* (2009) on case managers' decisions to impose sanctions, shows that black welfare clients are more likely to be punished than white welfare clients when rules have been violated. They test the Racial Classification Model (RCM), a model they developed themselves, to explain how a client's race affects officials' evaluations of rule violations. The RCM posits that when cues are confirming negative racial stereotypes, this can provide expectancy confirmation, thereby reinforcing negative stereotypes in evaluators' minds. As such, that study also tested and provided evidence for an indirect mechanism of stereotyping.

Apart from these studies, there is little attention within the public administration literature for indirect mechanisms of stereotyping. Our study builds on these studies by testing propositions of the double standards theory to explain how stereotypes may also work as frames affecting officials' interpretation of similar evidence. Just like the RCM, double standards theory holds that negative cues are interpreted stricter for social groups which have a more negative status in society. The double standards theory, however, differs from the RCM in several respects. First, the double standards theory has a broader scope and not only offers explanations for stereotyping based on race, but also on other characteristics, such as gender and social class. Second, whereas the RCM only focuses on 'discrediting markers', double standards theory also offers explanations for how positive cues could be evaluated stricter for lower status groups than for higher status groups. Double standards theory, thus, has a broader applicability and offers explanations for stereotypical evaluations of both stereotype-consistent and stereotype-*in*consistent cues.

Theoretical framework

In order to test if and how status characteristics affect the interpretation of other signals, we draw on the status characteristics theory and double standards theory. These theories of status-based discrimination have their origin in the sociology of work, where they have been tested to explain why certain groups are privileged in attaining positions and rewards over other groups in society (e.g. Wagner and Berger 1993). Status characteristics theory has been depicted by Wagner and Berger (1993) as a program of interrelated theories, aimed at explaining aspects of status-based discrimination in social interaction.

Double standards theory extends status characteristics theory by proposing that status characteristics affect the standards evaluators use to determine other people's ability (Correll

and Benard 2006; Foschi 2000). The basic assumption is that standards are stricter for lower status groups than for higher status groups (Foschi 2000). As performance expectations for low status groups are lower than for high status groups, a high performance of a low status actor will be inconsistent with the expectations for lower status actors. As a result, double standards theory holds, standards will be stricter for lower status actors, i.e. they will be more critically scrutinized in this situation. A woman with three children and an outstanding CV, for example, will be looked at with suspicion, because it does not correspond to the lower expectations people have of mothers in the workplace. Employers are inclined, for instance, to look for evidence that disproves the achievements of this person. The opposite also holds: as equally high performance is consistent with performance expectations for high status actors, the consistency between observation and expectation will lead to a more lenient standard (Correll and Benard 2006). A man with three children and an outstanding CV, in this case, is not inconsistent with the high expectations people usually have of men in the workplace. As a consequence, employers will accept that person's portrayal of his CV more easily, without looking for more evidence of his competence.

Standards of both competence and incompetence can be distinguished. A strict standard for *competence* requires more evidence than a lenient one, whereas a strict standard for *incompetence* accepts less evidence of incompetence than a lenient standard (Foschi 2000). The latter means that cues indicating low competence are more easily accepted for a lower status person than a higher status person, because they are consistent with the low expectations of the lower status group's competence and *in*consistent with the high expectations of higher status group's competence. To sum up, the theory holds that both cues signaling low competence and cues signaling high competence are evaluated stricter for lower status groups than for higher status groups.

This article examines whether tax officials use double standards in evaluating various cues from entrepreneurs from different social groups. In what follows, we will describe our research setting, contextualize the theoretical propositions, and present our hypotheses.

The Dutch tax administration

This study focuses on frontline tax officials inspecting the bookkeeping records of small and medium sized enterprises. Under the heading of the so-called 'horizontal supervision' approach, the Dutch tax administration has embraced responsiveness and trust towards entrepreneurs as essential ingredient for compliance (Gribnau 2007). This horizontal policy encourages officials to assess tax returns on their acceptability, rather than their mere correctness. This means that officials are encouraged to collect 'sufficient' information to make a judgment, and have 'to do enough work, but not too much' (Belastingdienst 2016, p. 4). This practically means that officials are encouraged *not* to start their inspection with the assumption that it will probably be wrong, *not* to do their utmost to find even the smallest flaws, and *not* to enforce the maximum financial correction when it has been found that entrepreneurs just made a mistake and express their goodwill to change. As a consequence, assessments of entrepreneurs' intentions and competences are part and parcel of tax officials' judgments. The standards to assess tax returns, thus, have become less predetermined, and more dependent on officials' assessments.

Tax officials' evaluations of entrepreneurs' trustworthiness is central to this study. In determining the acceptability of entrepreneurs' tax returns, tax officials look at what is presented to them in terms of bookkeeping records, but also at whether entrepreneurs are trustworthy, in order to make inferences about the credibility of what is presented. They generally look at two aspects of trustworthiness – intentions and competences – to look at whether some sort of fraud

might be involved, or whether they are dealing with a mere fault. This, in turn, influences officials' willingness to reach a compromise, and the height of the possible fine. Within this study, we aim to cover both the evaluation of the trust that can be vested in the entrepreneur, as the evaluation of the enterprise as a whole, since these are the core evaluations tax officials make in their daily work. We are furthermore interested in tax officials' intention to more critically scrutinize the case at hand, since this is the main decision determining whether officials will intensify their inspection or not.

In this study, the focus is on the effect of status characteristics on the interpretation of signals indicating low or high quality of the bookkeeping and interaction. The status characteristics focused on are entrepreneurs' *social class* and *level of education.* A prior study on tax officials has suggested these attributes play a role in frontline tax officials' evaluations (Raaphorst and Groeneveld 2015). Whereas these characteristics tend to overlap, they are often mentioned separately by tax officials. The findings indicated that these characteristics carry along specific expectations regarding entrepreneurs' intentions and competences. These characteristics are moreover associated with more generic cultural beliefs that are shared by society at large. Lower educated people are viewed as generally less competent than higher educated people. Although level of education is generally perceived as a legitimate ground to distinguish job applicants, its relevance to street-level law enforcement is less obvious. Furthermore, lower social classes are generally perceived as less competent and in need of help (e.g. Harrits and Møller 2014). These status characteristics are likely to influence officials when they need to assess intentions and competences.

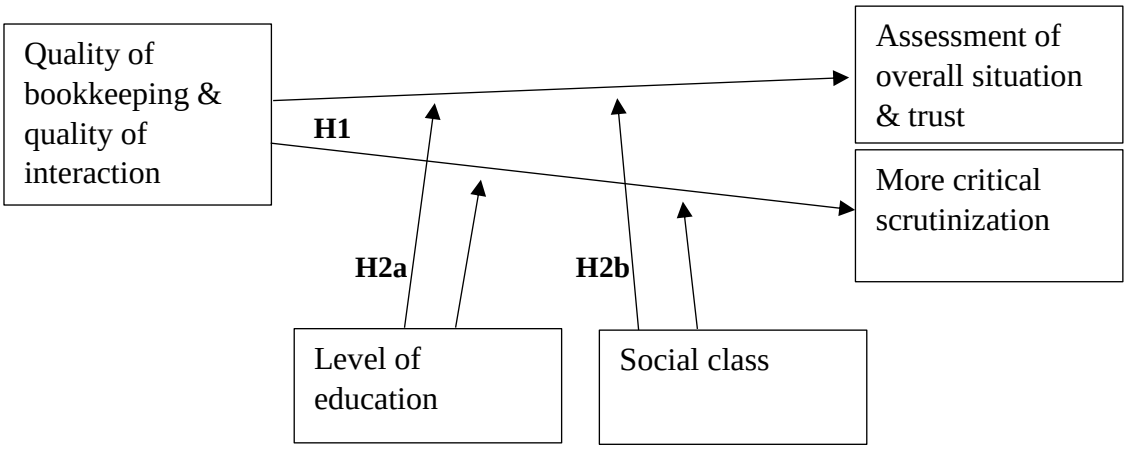
This study distinguishes two sources of attributes which serve as independent variables. Besides looking at characteristics of the bookkeeping records, and in particular how this is presented,

tax officials also take into account entrepreneurs' demeanor in the interaction to assess whether the tax return is acceptable, i.e. whether what is presented and found is credible (Raaphorst and Groeneveld 2015). For this reason we distinguish quality of the bookkeeping and quality of the interaction as determinants of officials' judgments.

Hypotheses

Within this section we formulate two sets of hypotheses. The first concerns the influence of the quality of the bookkeeping and the quality of the interaction on officials' evaluations. Secondly, we will discuss our hypotheses on the moderating effect of entrepreneurs' status characteristics on officials' evaluation of quality of bookkeeping and interaction signals. Figure 1 portrays the conceptual model and the corresponding hypotheses.

Figure 1. Conceptual model with hypotheses



The street-level bureaucracy literature and the literature on regulatory encounters provides evidence that street-level officials not only look at characteristics related to their core task, but also at how citizens behave in the interaction to make judgments (Maynard-Moody and Musheno 2003). The latter authors show how street-level bureaucrats rather respond to co-operative citizen-clients than to manipulative and over-demanding citizens. Nielsen (2007) shows that the higher the level of communication (in frequency and quality), the more lenient

an inspector is. Therefore, we expect that the higher both the quality of the bookkeeping and the interaction, the more positive officials' evaluations will be.

> H1: Cues indicating a good quality of bookkeeping and a good quality of interaction will have a more positive effect on officials' evaluation of trustworthiness and overall situation, and will have a more negative effect on officials' inclination to more critically scrutinize the entrepreneur, than cues indicating a bad quality.

Secondly, we formulate hypotheses for the indirect mechanism, which is this study's particular contribution to the public administration literature on frontline stereotyping. Based on our previous exploratory study (Raaphorst and Groeneveld 2015) we expect that frontline tax officials may use double standards to evaluate entrepreneurs. That study has suggested that differential evaluations are based on cultural beliefs about professions involving either manual or mental labor, about different levels of education and different 'classes' in society. An example mentioned within this study is the differential evaluation of 'wrongly declared turnover tax´: a 'high-level' mayor, it is held, is to blame, since he should have known, whereas a shoemaker is not to blame, because he is just incompetent. Another respondent distinguishes status groups according to their alleged intentions, and argues that residents of mobile homes cannot and *do not want* to keep their records properly, whereas manual workers simply do not have the skills. Another example is that of the lower educated entrepreneurs who are assigned bad intentions in case of wrongly kept records, whereas the intentions of a higher educated entrepreneur in a similar situation are described as good intentions that have gone bad (ibid.).

These findings thus suggest that double standards are used, but they are less straightforward about the directions in which these work. In some instances, the higher status entrepreneur is

evaluated stricter, whereas in other instances the lower status entrepreneur is evaluated stricter. This could be due to the qualitative and exploratory nature of that study, which did not allow us to keep constant research conditions. In this current study the independent variables will be manipulated, allowing us to better assess the validity of double standards theory. In line with double standards theory and the findings of our previous study, we expect that entrepreneurs' level of education and social class serve as moderating contexts, influencing the strength and possibly also the direction of the effects of signals on officials' evaluations. Our previous study has shown that a lower level of education is often associated with diminished expectations about entrepreneurs' competence (Raaphorst and Groeneveld 2015). Therefore, we expect that the same situation is evaluated stricter (i.e. more negatively) for lower educated entrepreneurs than for higher educated entrepreneurs:

> H2a: Cues of both quality of bookkeeping and quality of interaction will be evaluated stricter for the lower educated entrepreneurs than for the higher educated entrepreneurs.

Moreover, tax officials sometimes associate entrepreneurs from a lower social class not only with lower levels of competence, but also with bad intentions; i.e. entrepreneurs who try to withhold tax money (ibid.). Bookkeeping records that seem acceptable at first sight, then, could also be feigned. It is likely that such suspicions about social class influence the interpretation of other signals. For this reason, we expect a moderating impact of social class on the effect of quality of bookkeeping and quality of interaction cues as follows:

> H2b: Cues of both quality of bookkeeping and quality of interaction will be evaluated stricter for entrepreneurs from a lower social class than for entrepreneurs from a higher social class.

Based on hypotheses 2a and 2b we thus expect that similar scenarios will be more negatively evaluated for entrepreneurs with a lower level of education and from a lower social class than entrepreneurs with a higher level of education and from a higher social class.

The policy-capturing study

To examine whether officials evaluate similar evidence differently for entrepreneurs from different status groups, this study conducted a policy-capturing study. The policy-capturing design allows for studying how decision makers use information in evaluative judgments (Aiman-Smith *et al.* 2002). It involves letting respondents judge a relatively large set of hypothetical, but realistic scenarios in a row, with each scenario being composed of a distinct combination of cue values. Subsequently, respondents' evaluations are regressed on the cue values, which enables researchers to assess the relative weight of the various cues in evaluations.

We chose for a policy-capturing design because it allows for the study of stereotyping by officials in a context that resembles real-life decision making. Policy-capturing studies are typically more realistic than laboratory experiments where respondents are removed from their natural environments and typically evaluate only one scenario (Aguinis and Bradley 2014). Whereas classical experiments measure officials' first stereotypical reactions, the question remains whether these studies actually capture officials' judgments in work situations or rather first impressions they share with other people in general. The policy-capturing method has better external validity, because it allows respondents to adjust their evaluations to prior evaluations. Evaluating various cases is what tax officials do on a weekly basis. Decisions about these cases are not made in a vacuum, but compared to each other. Policy-capturing

studies thus resemble officials' actual work situations better, since such designs allow for assessments of multiple scenarios and comparisons between scenarios. Since respondents are asked to make judgments about scenarios including *multiple* cues, the policy-capturing study reduces, to some extent, the possibility for respondents to give strategic answers (Karren and Barringer 2002).

The policy-capturing design furthermore allowed us to study different combinations of stimuli and multiple decisions, whereas traditional experimental designs can only study a limited amount of decisions. Moreover, the policy-capturing design provides a relatively high degree of control over confounding factors, because of its full factorial design. Because respondents in our study evaluated all possible combinations of the different cue values, the independent effects of each value could be assessed. Within traditional experiments, there typically is more uncertainty regarding possible other explanatory factors (Aiman-Smith *et al.* 2002).

Design and scenario construction

Each scenario entailed a value of the four cues (quality of bookkeeping, quality of interaction, level of education and social class). This study employed a full factorial design, which resulted in a total amount of 36 scenarios (2x2x3x3). Each respondent was asked to evaluate 40 scenarios, including four duplicated scenarios. Whereas reliability is a necessary condition for the validity of measures, Karren and Barringer (2002) noted that few published policy-capturing studies analyzed the reliability of evaluators' judgments. The authors recommend that replicating four scenarios may serve as feasible test-retest check of the judgments. Our 9:1 scenario-to-cue ratio meets the minimum ratio of 5:1 as suggested by Cooksey (1996). The scenarios were presented in narrative form. In order not to exhaust our respondents, we constructed the scenarios in such a way as to only include the necessary information needed to

make a judgment. We undertook 10 test interviews to improve our scenarios and operationalization of cues, aiming for an optimal balance between realism and feasibility. Appendix A presents an example of a scenario used.

Cue development and operationalization

For each cue we developed several behavioral statements that represented different levels of the respective cue. The choice for these values is based on our prior in-depth study on signals of entrepreneurs' trustworthiness and untrustworthiness (Raaphorst and Groeneveld 2015), and also on 10 test interviews with tax officials. During these interviews it was assessed how statements were interpreted, which were refined or adjusted aided by respondents' input. With regards to entrepreneurs' level of education, we chose to explicitly state the level of education (either low or high) as an impression acquired during the audit, since that is typically the way officials express their sense of an entrepreneurs' cognitive abilities.

The concept of social class is broader than socioeconomic class, since it not only refers to people's economic position in society, but also more broadly to sociocultural aspects such as lifestyle and behaviors (e.g. Harrits and Møller 2014). In this study, we distinguished between a low and a middle-high social class. At first we tested a cue distinguishing between two known areas within the respective cities where the enterprise allegedly was located, of which one was known for its socioeconomic problems and the other was in the wealthier city center. However, since the areas were not known to all respondents, we had to develop other indicators. Therefore we chose to present pictures of streets where the enterprises allegedly were located. The pictures indicating a lower social class show multicultural streets with dilapidated buildings and poorly kept streets, whereas the pictures indicating a higher social class show well-kept streets, with well-maintained buildings and mainly white pedestrians.

For both quality of bookkeeping and quality of interaction we developed three levels, ranging from low to high quality. For the statistical analyses, the cues 'quality of bookkeeping' and 'quality of interaction' were dummy coded. The lowest levels of these cues were used as reference categories. For the three dependent variables – assessment of trustworthiness, overall judgment of the situation and intention to more critically scrutinize – we developed three items. See appendix B for the operationalization of all our variables. Table 1 shows the descriptive statistics for the dependent variables. The correlations of the independent and dependent variables can be found in appendix C. Although the dependent variables are highly correlated, the subsequent analyses are performed for each dependent variable separately, because they measure different judgments; 'appears okay' captures a general impression of the situation, 'trust' measures an interpersonal judgment and 'more critical scrutinization' measures a behavioral intention.

Table 1. Descriptive statistics dependent variables

	N	Mean	SD
Evaluation 'appears okay'	828	3,11	1,428
Evaluation 'trust'	828	3,26	1,297
Evaluation 'more critical scrutinization'	828	5,13	1,279

Respondent selection and data collection procedure

In line with the aim of this study, we selected respondents who work with the 'horizontal supervision' policy and have face-to-face contact with entrepreneurs as part of their work. Managers of two different tax offices in two cities in the south of the Netherlands were approached, and both were willing to cooperate with us by requesting their employees to participate in our research. 36 respondents were willing to participate. With 10 of those we conducted a test interview and with 26 we conducted the final study. For all the statistical

analyses we only included respondents who had reliable response patterns, i.e. a correlation between the replicated and original scenarios of above .50 ($p<0.10$). This resulted in a dataset with 23 respondents and 828 evaluated scenarios in total. Each row in our dataset represented an evaluated scenario. Five of the 23 respondents are female, 18 are male. Only one respondent was born in a non-western country. With regards to tenure at the time of data collection, four respondents had been in service for 3 years or less, eight respondents had been employed by the tax administration between 10 and 30 years, and 11 respondents had been in service for over 30 years.

Because the evaluation task requires respondents to invest time and effort, we decided to conduct the study within an one-on-one interview setting. In doing this, we could invest in the relationship with respondents, and enhance their motivation to participate. The first author conducted all the interviews, and the same procedure was followed within each interview. Small breaks were introduced at fixed times, to prevent respondents from getting exhausted (see online appendix for the interview procedure). After the evaluation task, respondents had the opportunity to reflect upon their experiences. This also offered us the opportunity to assess how respondents interpreted certain indicators and questions. These interviews made clear that the photos indicating low and middle-high social class were interpreted as intended.

Findings

In what follows, we will first describe the patterns of scenario evaluation found at the individual level. Second, we will test our hypotheses by multi-level analyses. Thirdly, we will use our reflection interview data to interpret the findings that were inconsistent with our hypotheses.

Individual-level exploration

In order to explore the scenario evaluations, we first conducted quantitative analyses at the individual level. IBM SPSS (version 24) was used for the analyses. We explored the direct and interaction effects on the evaluations for each respondent separately, by conducting analyses of variance. Differences across respondents were found in the patterns of direct and interaction effects involving the two status characteristics. For only five of the 23 respondents, entrepreneurs' level of education had a significant direct effect on one or several of the evaluations. No significant relations were found between social class and respondents' evaluations.

For five respondents, significant moderation effects were found. These interactions all involve a moderating effect of level of education on the relationship between a value of either quality of bookkeeping or quality of interaction with one of the evaluations. The directions of these interaction effects differed across respondents. This means that, depending on the respondent, cues of quality of bookkeeping and quality of interaction were evaluated either more negatively or more positively for the lower educated entrepreneur. We can conclude from this first exploration that for the majority of respondents no direct and interaction effects of status characteristics seemed to be at play. However, since the same analysis was repeated 23 times, the five significant interaction effects found could also have occurred by chance. Because the evaluated scenarios are nested within respondents (and observations are thus not independent), multilevel analyses were required. We estimated a maximum likelihood random intercepts, fixed slopes model. We allowed respondents to vary on the dependent variables 'trust', 'appears okay' and 'more critical scrutinization' at baseline from one another. In this model, the slopes were fixed, since we are interested in the effects of the cues (level-one units) and their interactions and not in whether these effects differ among respondents (our level-two units).

Since our explanatory variables are not defined at level two, and statistical inference is only directed at respondents in our sample, a fixed effect model is appropriate (Snijders and Berkhof 2007). Moreover, fixed effects estimates 'achieve a better control for unexplained differences between level-two units' (Snijders and Berkhof 2007, p. 143).

Multi-level analyses

Table 2 presents the results of the multilevel analyses of the direct effects of the cues and the interaction effects involving the two status characteristics on all three dependent variables. For each dependent variable, we also tested an empty model to model the random effect of respondent. For the dependent variable 'appears okay' the intraclass correlation was 0.1232 (0.251/(0.251+1.786) which indicates that around 12% of the variation in the evaluation is accounted for by the respondents. For 'trust' this correlation was 0.1892 (0.318/(0.318+1.363) which indicates that around 19% of the variation in accounted for by respondents. The intraclass correlation for 'more critical scrutinization' was 0.1053 (0.172/(0.172+1.462); around 11% of the variation is explained by respondents. For all three dependent variables, the significant estimates of variance indicate that the intercepts vary significantly across respondents. Hence, a multilevel analysis is warranted.

Model 1 added the four cues. In line with hypothesis 1, both 'missing invoices' and 'invoices in order' have a positive effect on the evaluation of the overall situation when compared to 'barely any records'. For 'more critical scrunitization' these effects are negative and also statistically significant; the results indicate that the worse the quality of the bookkeeping, the more respondents are inclined to more critically scrutinize the entrepreneur. Regarding the quality of interaction, 'to the point' has a positive effect on 'appears okay' and 'trust' when compared to 'avoids contact'. Contrary to our expectation, 'dodging around question' has a

negative effect on ‘appears okay’ and ‘trust’ when compared to ‘avoids contact’, but this effect is not significant. Again, for ‘more critical scrutinization’ these effects are reversed. This means that respondents are less inclined to more critically scrutinize the entrepreneur when s/he gives to the point answers, than when s/he avoids contact. ‘Dodging around question’ has a slightly more positive effect than ‘avoids contact’, but this effect is not significant. There were no significant direct effects of level of education and social class on each of the evaluations. For ‘appears okay’, adding the four cues accounts for 55.4% of the within respondent variability, and resulted in a significantly better fit of the model; the deviance decreased with 649,248 (df=6; $p<0.001$). For ‘trust’, 42.8% of the within respondent variability is explained by the cues. The deviance decreased significantly with 450,153 (df=6, $p<0.001$). Adding the four variables accounts for 50.4% of the within respondent variability in ‘more critical scrutinization’. The deviance decreased significantly with 564,214 (df=6, $p <0.001$).

Model 2 added the interaction effects in order to test whether values of quality of bookkeeping or quality of interaction were evaluated differently for status group entrepreneurs. Overall, one significant interaction effect was found for ‘appears okay’; ‘dodging around question’ seems to be differently evaluated for lower educated entrepreneurs than for higher educated entrepreneurs. For ‘trust’ and ‘more critical scrutinization’ no significant interaction effects were found. Contrary to our hypotheses, no significant interaction effects were found for social class. For none of the dependent variables, model 2 led to a significantly better fit of the model. In order to check whether adding the significant interaction effect alone would increase the fit of the model for ‘appears okay’, we checked whether a new model with only the direct effects and the significant interaction effect would significantly decrease variance. In this new model -2 Log Likelihood was 2218,53, and X^2-change was -3,511 compared to model 1. This model significantly improved the fit (df=1, $p<0.10$).

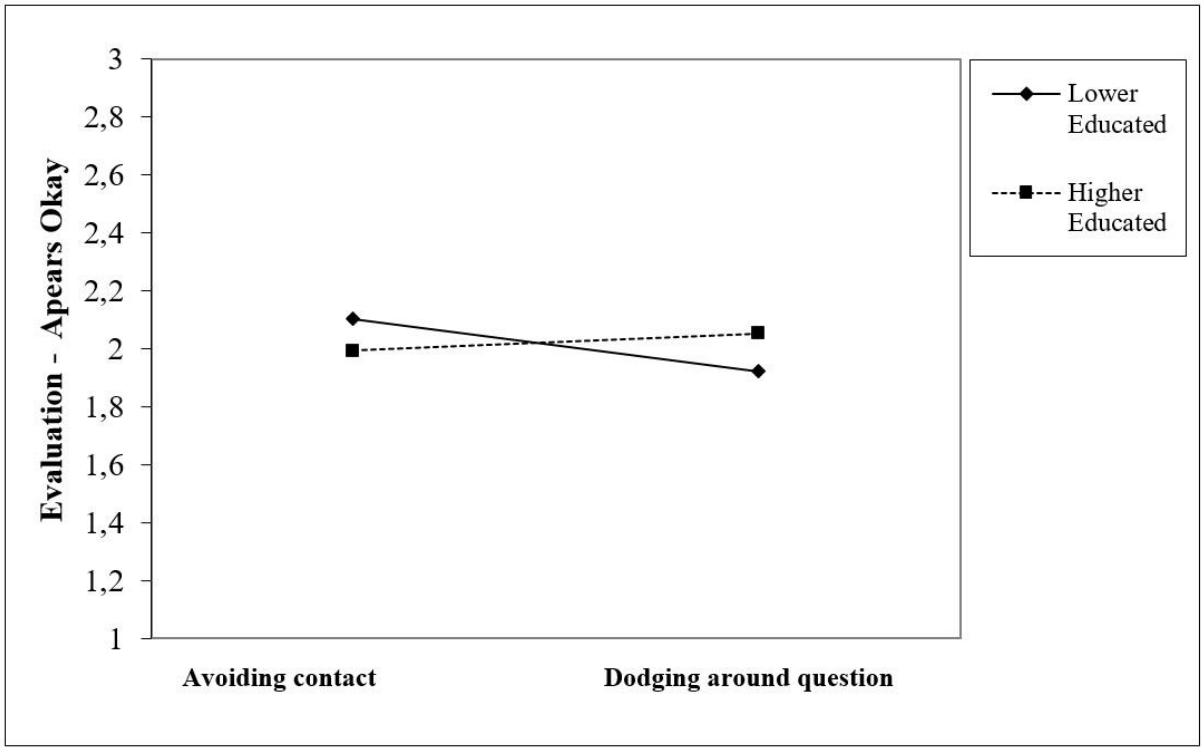
Table 2. Multilevel analyses of direct and interaction effects

	DV: Appears okay		**DV: Trust**		**DV: More critical scrutinization**	
	Model 1	Model 2	Model 1	Model 2	Model 1	Model 2
Intercept	2.030*** (0.137)	2.076*** (0.163)	2.450*** (0.145)	2.494*** (0.169)	5.978*** (0.120)	5.940*** (0.146)
Cues						
Quality of bookkeeping						
Barely any records	Ref	Ref	Ref	Ref	Ref	Ref
Invoices missing	0.406*** (0.076)	0.428*** (0.131)	0.330*** (0.075)	0.337** (0.130)	-0.283*** (0.072)	-0.348** (0.125)
Invoices in order	2.087*** (0.076)	2.072*** (0.131)	1.406*** (0.075)	1.312*** (0.130)	-1.801*** (0.072)	-0.176*** (0.125)
Quality of interaction						
Avoids contact	Ref	Ref	Ref	Ref	Ref	Ref
Dodges around question	-0.054 (0.076)	-0.192 (0.131)	-0.083 (0.075)	-0.141 (0.130)	0.083 (0.072)	0.199 (0.125)
To the point	0.775*** (0.076)	0.768*** (0.131)	0.917*** (0.075)	0.931*** (0.130)	-0.591*** (0.072)	-0.562*** (0.125)
Level of education						
Low	Ref	Ref	Ref	Ref	Ref	Ref
High	0.053 (0.062)	-0.075 (0.138)	-0.087 (0.061)	-0.138 (0.137)	-0.017 (0.059)	0.053 (0.132)
Social class						
Low	Ref	Ref	Ref	Ref	Ref	Ref
High	-0.043 (0.062)	-0.007 (0.138)	-0.014 (0.061)	-0.050 (0.137)	0.051 (0.059)	0.058 (0.132)
Two-way interactions						
Invoices missing* level of education	–	0.014 (0.152)	–	-0.036 (0.150)	–	0.087 (0.145)
Invoices in order* level of education	–	0.130 (0.152)	–	0.145 (0.150)	–	-0.094 (0.145)
Dodges around question* level of education	–	0.254 † (0.152)	–	0.138 (0.150)	–	-0.210 (0.145)
To the point* level of education	–	0.014 (0.152)	–	-0.094 (0.150)	–	0.007 (0.145)
Invoices missing* social class	–	-0.029 (0.152)	–	0.022 (0.150)	–	0.043 (0.145)
Invoices in order* social class	–	-0.101 (0.152)	–	0.043 (0.150)	–	0.022 (0.145)
Dodges around question* social class	–	0.022 (0.152)	–	-0.022 (0.150)	–	-0.021 (0.145)
To the point* social class	–	0.000 (0.152)	–	0.065 (0.150)	–	-0.065 (0.145)
-2 Log Likelihood	2222,041	2216,909	2207,449	2202,954	2137,851	2133,073
Df	*9*	*17*	*9*	*17*	*9*	*17*
X^2-change in comparison to previous model	-649,248***	-5,132	-450,153***	-5,222	-564,214***	-4,778
Variance within respondents	0.797***	0.792***	0.779***	0.775***	0.725***	0.721***
% explained variance	55.4%	55.7%	42.8%	43.1%	50.4%	50.7%
Variance between respondents	0.279**	0.279**	0.334***	0.334***	0.193**	0.193**
% explained variance	25.0%	26.1%	30.0%	30.1%	21.0%	21.1%
N (scenarios)	828	828	828	828	828	828
N (respondents)	23	23	23	23	23	23

Note: standard errors in parentheses. Significance levels: † $p \leq 0.10$; ** $p \leq 0.01$; *** $p \leq 0.001$

Figure 2 plots the significant interaction effect and shows that, in line with our hypothesis, a lower educated entrepreneur is judged slightly more negatively when dodging around a question than a higher educated entrepreneur. When an entrepreneur is avoiding contact, this is evaluated slightly more positive when s/he is a lower educated entrepreneur, than when s/he is higher educated. Whereas there is no significant direct effect of level of education, there is a significant, moderating effect of level of education. The difference is small relative to the scale on which the dependent variable is measured (smaller than .2 on a 7-point scale). However, the difference is larger when compared to the variance of 2.039 of 'Appears Okay', indicating a tight distribution of scores.

Figure 2. Graph of interaction effect Dodging around question * Level of education on 'appears okay'



Interview data

The subsequent interview allowed us to gain insight in how respondents experienced evaluating the scenarios, and how the cues and questions were interpreted. Generally, respondents experienced no difficulty in evaluating the scenarios. Some respondents noted that the scenarios looked like each other, and that reality is more complex. In reality, for instance, they also look at what people say and not only at how the interaction unfolds. Yet, the presented cues gave them sufficient grounds to make evaluations. Also, some respondents mentioned their response pattern became less extreme throughout the evaluation task.

We moreover relied on the interview data to provide possible explanations for the findings that were inconsistent with our hypotheses. Contrary to our hypothesis, we found that when an entrepreneur is avoiding contact, this is evaluated slightly more positive when s/he is a lower educated entrepreneur, than when s/he is higher educated. A statement by one of our respondents could offer an explanation for this. He argued that when a lower educated person does not seek contact this could have to do with insecurity, whereas a higher educated person has better interpersonal skills and is less insecure. As a consequence, the official starts to 'get suspicious' when a higher educated entrepreneur avoids contact. In this case, a higher expectation leads to a stricter evaluation when evidence for low competence is encountered than in case of low expectations. This could be a possible explanation for our 'reversed double standards' finding.

Moreover, some respondents mentioned they deliberately tried not to look at the photos and/or entrepreneurs' level of education. One respondent for instance argues that the photos may lead to expectations, and 'you look at it, but you try to block it'. Another respondent argued he learned to suppress his first impressions, in order to be as neutral as possible. Again other

respondents argued that one needs to be careful with presumptions, since they do not have to be true. Some say these aspects are not supposed to play a role and are not really relevant, but that they sometimes do give a first impression. One respondent mentioned he tries to be aware of his own prejudices, and always tries to postpone first impressions, but that he does not want to be naïve either. Although trying to be nonbiased, most respondents at the same time associated specific expectations to either lower or higher status groups. E.g. 'I expect more from the higher educated, and less from the lower educated', or 'the higher educated rather have a negative impact; they are more able to cheat than the lower educated'. This indicates that although some respondents learned to block their prejudices or postpone their first impressions, they can involuntary play a role. Respondents who argued they tried to not to let themselves be influenced by presuppositions or prejudices, likely also try to do this in their actual work. This may be an explanation for the nonsignificant interaction effects.

Conclusion and discussion

This study examined whether officials use double standards in evaluating entrepreneurs during inspections. It provided a first step in testing the explanatory potential of the sociological double standards theory in a public administration context. Using a policy-capturing design, this study tested whether situations involving entrepreneurs with a lower level of education and from a lower social class are evaluated more negatively than similar situations involving entrepreneurs with a higher level of education and from a higher social class. Our hypotheses were partly confirmed. Most values of quality of interaction and quality of bookkeeping, except for dodging around the question, had a significant effect on the evaluations. With regards to the double standards propositions, we found that when a lower educated entrepreneur dodges around a question this is evaluated slightly more negatively than when a higher educated entrepreneur dodges around a question. We also found evidence for the reversed practice: when a higher

educated entrepreneur avoids contact this is evaluated slightly more negatively than when a lower educated entrepreneur avoids contact. This finding underlines the importance of studying indirect mechanisms of stereotyping, especially since we did not find any direct effect of status characteristics on the evaluations.

Whereas our prior qualitative study (Raaphorst and Groeneveld 2015) suggested tax officials may use double standards, most of the interaction effects in this study were nonsignificant. When compared to the direct effects of most of the cues, the significant interaction effect is moreover only small in size. This is not surprising since quality of bookkeeping and quality of interaction are deemed essential for evaluating the acceptability of tax returns, while entrepreneurs' level of education is not. More interestingly, whereas we did not find any direct effect of level of education on the evaluations, we did find it could affect frontline evaluation in combination with other signals. These differences can have a large impact on the further evolvement of an inspection and decisions being made. It could make a difference between giving someone the benefit of the doubt or not. This frontline practice may harm equal treatment, and have lasting consequences for citizen-clients.

Our findings have several theoretical implications. First, they show that stereotyping by frontline officials could work more indirectly than is hitherto assumed within the street-level bureaucracy literature. Studies have shown that street-level bureaucrats rely on stereotypes in decision making as a way of coping with time pressures and high workloads (Lipsky 1980; Andersen and Guul 2016). These studies suggest citizen-clients' belonging to social groups serve as shortcuts to their supposed identities. Our study indicates that frontline officials employ an indirect way of stereotyping in which citizen-clients' belonging to a social group serves as frame that influences the interpretation of other signals. In fact, our analyses have shown that

entrepreneurs' level of education does not have a direct effect on the evaluations, but has an effect on one of the evaluations in combination with another signal. This subtler way of stereotyping calls for research approaches that take into account how officials interpret clusters of signals.

Our study has furthermore found evidence for the use of double standards in different directions. Findings point out that the standards can be stricter for the low status entrepreneur and more lenient for the high status entrepreneur, or the other way around. In this study, 'avoiding contact' was evaluated stricter for higher educated entrepreneurs, whereas 'dodging around question' was evaluated stricter for lower educated entrepreneurs. In line with our double standards proposition, not giving answers to questions may be interpreted stricter for lower educated entrepreneurs because it is consistent with the lower expectations officials have of their competences. A possible explanation for the finding that works in the opposite direction could be that inferences about different properties are made for the different status groups. Our qualitative data suggests that a lower educated entrepreneur who avoids contact is associated with mere incompetence in communicating, whereas this is seen as a signal for bad intentions for higher educated entrepreneurs, who are expected to have these communication skills. Foschi (2000) refers to the latter as 'reversed double standards', which has been advocated by some as a means to change the status quo. Although this might be experienced and proposed by officials as more fair, it reinforces the assumption that lower status citizen-clients cannot meet the universalistic standards and therefore have to be treated more leniently (ibid.). Either way – in receiving a stricter or more lenient treatment – lower status groups are treated as inferior.

Following up on our findings, future research should examine how organizational socialization of public officials affects their use of double standards. Especially since some respondents

suggested they have learned to block prejudices or postpone their first impressions, there are indications that organizational socialization may work to neutralize the effects of stereotypical expectations and concomitant double standards. In fact, taking into account the influence of organizational socialization, but also other background characteristics of public officials, on the use of double standards, would contribute to the development of a theory aimed at explaining the extent to which double standards are used.

Our findings also have implications for new models of governance that have come to embrace street-level officials' professional judgments as essential for decision making. Within models promoting trust between officials and citizens, officials have to work with rules and legislation that grant them more discretion to rely on their own interpretations in decision making. Within our case, the question has shifted from 'is it correct?' to 'is it acceptable?', thereby allowing officials to look at entrepreneurs' demeanor and at whether they appear trustworthy. Our study has shown that, in such a context, officials sometimes use double standards in evaluating citizen-clients. Whereas these new governance models allow frontline officials to be more responsive and – in our case – to get citizen-clients more compliant, this way of working may also have implications for consistent and equal decision making (see also Piore 2011; Rutz *et al.* 2015).

This study's approach to examining stereotyping moreover has different advantages but also drawbacks when compared to experimental research designs using control and treatment groups. Recent experimental studies have found evidence for direct effects of stereotypes, such as ethnicity, on decision making (e.g. Andersen and Guul 2016). We did not find such direct effects. Rather than making statements about which findings are more true, it is more fruitful to reflect on the implications of using different methods. Whereas the classical experiments do

primarily measure officials' first stereotypical reactions, the question remains whether these studies actually capture officials' judgments in work situations, or their first impressions as human beings. Policy-capturing studies probably resemble officials' actual work situations better, since such designs allow for assessments of multiple scenarios and comparisons between scenarios. As such, respondents have more opportunity to reflect on their first impressions and adjust their responses accordingly. However, this seems to accord with officials' daily practice in which they try not to rely on their prejudices. An interesting venue for future research would be to analyze whether and how officials try to make their decisions consistent with prior decisions, by specifically looking at carry-over effects.

This study also has some limitations. First, this study does not allow for generalization to a larger population. We only had a small sample that was not selected on grounds of representativeness for a larger population. Yet, our main aim was to theoretically generalize: we tested the validity of the double standards theory in a new context, that is, street-level decision making. It is highly likely that our main finding that in some occasions officials use double standards is generalizable to comparable frontline domains where rules and guidelines have become less clear-cut and there is more room for officials' interpretation. Second, because we had many conditions and only a small sample, we could not control for possible order effects. Therefore we kept the scenario order constant for each respondent. By using larger samples and less conditions future research could disentangle cue effects from possible order effects by randomizing the order of scenarios.

Third, the way cues were operationalized could have impacted our findings. Level of education as a signal for competence, for example, was given as an impression acquired through the inspection, and not measured by more implicit indicators. This could have raised respondents'

awareness about the focus of our study. Using more fine-grained indicators for level of education could have resulted in better identifying interaction effects. Our cue of social class, as a signal for intentions, furthermore, portrayed not only indicators of wealth and maintenance of streets, but also of ethnicity. While these often tend to go together, they are not the same. Our cue thereby grasped a broader stereotype around social class. Future research could disentangle these indicators and measure the effects of social class and ethnicity separately.

Fourth, because respondents were asked to evaluate a fairly large amount of scenarios, respondents learned about their own response patterns and the manipulated cues, and could have adjusted their responses accordingly. Although this learning effect may indeed have occurred, this probably resembles tax officials' daily practice where they have to inspect multiple cases on a monthly and sometimes weekly basis, and compare cases to make consistent decisions. Hence, within an experimental research design where respondents only evaluate one scenario, it is likely that there would be more and stronger evidence for the use double standards. Yet, findings of such experiments are less generalizable to real-life settings, where officials attempt to make consistent and fair decisions. Moreover, since our study still found evidence for the use of double standards, it is likely that the amount of information in vignettes made it difficult to give strategic answers and that the trust established in the one-on-one setting made respondents feel comfortable in making honest evaluations. Future studies on frontline stereotyping could compare different methods, such as policy-capturing and experiments with treatment and control groups, to study similar research questions. In doing this, the specific contributions of each method to the study of stereotyping could be assessed and compared.

This study has shown the added value of using a policy-capturing design to examine officials' implicit use of stereotypes in decision making without stripping it of the broader decision

making context. However, while the study resembles real-life settings, the scenarios are still hypothetical and compromise the complexity of real-life frontline decision making. Scholars interested in studying indirect stereotyping could consider conducting field experiments, which typically have better external validity. However, such studies are more difficult to conduct. Either way, this study has suggested that citizen-clients' status characteristics may affect the standards officials use to interpret information, without necessarily affecting their evaluations directly. This finding calls for research approaches and methods that are able to grasp this indirect, but pervasive, form of stereotyping.

Acknowledgements

This work was supported by the Dutch Organization for Scientific Research (NWO) [Vidi grant no. 452-11-011]. We are indebted to the respondents for their time and willingness to talk openly about their work. We thank Noortje de Boer and Machiel van der Heijden for their methodological advice, and the anonymous reviewers for their helpful comments on our manuscript. Any remaining shortcomings are ours.

References

Aguinis, H. and Bradley, K.J. 2014. 'Best Practice Recommendations for Designing and Implementing Experimental Vignette Methodology Studies', *Organizational Research Methods,* 17, 4, 351-371.

Aiman-Smith, L., Scullen, S.E. and Barr, S.H. 2002. 'Conducting Studies of Decision Making in Organizational Contexts: A Tutorial for Policy-Capturing and Other Regression-Based Techniques', *Organizational Research Methods,* 5, 4, 388-414.

Andersen, S.C., and Guul, T.S. 2016. *Minority Discrimination at the Front Line. Combined Survey and Field Experimental Evidence.* Paper prepared for presentation at the 2016 Annual Meeting of the Southern Political Science Association, January 7-9, 2016, San Juan, Puerto Rico.

Bartels, K.P.R. 2013. 'Public Encounters: The History and Future of Face-to-Face Contact Between Public Professionals and Citizens', *Public Administration,* 91, 469-483.

Belastingdienst, 2016. *Controleaanpak Belastingdienst (CAB): De CAB en zijn modellen toegepast in toezicht.* Retrieved from: https://www.belastingdienst.nl/wps/wcm/connect/bldcontentnl/themaoverstijgend/brochures_en_publicaties/controleaanpak_belastingdienst

Berger, J., Cohen, B.P. and Zelditch, M. 1972. 'Status Characteristics and Social Interaction', *American Sociological Review,* 37, 3, 241-255.

Cooksey, R.W. 1996. *Judgment Analysis: Theory, Methods and Applications.* San Diego, CA: Academic Press.

Correll, S.J. and Benard, S. 2006. 'Biased estimators? Comparing status and statistical theories of gender discrimination', in S.R. Thye and E.J. Lawler (eds), *Social Psychology of the Workplace (Advances in Group Processes Volume 23).* New York, NY: Elsevier Science, pp. 89-116.

Dubois, V. 2010. *The Bureaucrat and the Poor: Encounters in French Welfare Office.* Farnham and Burlington, VT: Ashgate.

Epp, C.R., Maynard-Moody, S. and Haider-Markel, D.P. 2014. *How Police Stops Define Race and Citizenship.* Chicago and London: The University of Chicago Press.

Foschi, M. 2000. 'Double Standards for Competence: Theory and Research', *Annual Review of Sociology,* 26, 21-42.

Gambetta, D. and Hamill, H. 2005. *Streetwise: How Taxi Drivers Establish Their Customers' Trustworthiness.* New York: Russell Sage Foundation.

Gribnau, H. 2007. 'Soft Law and Taxation: The Case of the Netherlands', *Legisprudence*, 1, 291-326.

Harris, D.A. 1999. 'The Stories, the Statistics and the Law: Why 'Driving While Black' Matters', *University of Minnesota Law Review*, 84, 2, 265-326.

Harrits, G.S. and Møller, M.Ø. 2014. 'Prevention at the Front Line. How Home Nurses, Pedagogues, and Teachers Transform Public Worry into Decisions on Special Efforts', *Public Management Review*, 16, 4, 447-480.

Johnson, R.R. and Morgan, M.A. 2013. 'Suspicion Formation among Police Officers: An International Literature Review', *Criminal Justice Studies,* 26, 1, 99-114.

Karren, R.J. and Barringer, M.W. 2002. 'A Review and Analysis of the Policy-Capturing Methodology in Organizational Research: Guidelines for Research and Practice', *Organizational Research Methods,* 5, 4, 337-361.

Lipsky, M. 1980. *Street-Level Bureaucracy: Dilemmas of the Individual in Public Services.* New York: Russell Sage Foundation.

Maynard-Moody, S. and Musheno, M. 2003. *Cops, Teachers, Counselors: Stories from the Front Lines of Public Service.* Ann Arbor: The University of Michigan Press.

Nielsen, V. L. 2007. 'Differential Treatment and Communicative Interactions: Why the Character of Social Interaction is Important', *Law and Policy*, 29, 2, 257-283.

Prottas, J.M. 1979. *People-Processing: The Street-Level Bureaucrat in Public Service Bureaucracies.* Massachusetts and Toronto: D.C. Heath and Company Lexington.

Raaphorst, N. and Groeneveld, S. 2015. *How do Tax Officials Assess Citizen-Clients' Trustworthiness? The Role of Signals and Status Characteristics.* Paper prepared for presentation at the 2015 Annual Conference of the European Group for Public Administration (EGPA), August 26-28, Toulouse, France.

Rice, D. 2017. 'How Governance Conditions Affect the Individualization of Active Labour Market Services: An Exploratory Vignette Study', *Public Administration.* doi: 10.1111/padm.12307

Ridgeway, C. 1991. 'The Social Construction of Status Value: Gender and Other Nominal Characteristics', *Social Forces*, 70, 2, 367-386.

Schram, S.F., Soss, J., Fording, R.C. and Houser, L. 2009. 'Deciding to Discipline: Race, Choice, and Punishment at the Frontlines of Welfare Reform', *American Sociological Review*, 74, 3, 398-422.

Snijders, T.A.B. and Berkhof, J. 2007. 'Diagnostic Checks for Multilevel Models' in J. de Leeuw and E. Meijer (eds), *Handbook of Multilevel Analysis.* New York: Springer, pp. 139-173.

Stroshine, M., Alpert, G. and Dunham, R. 2008. 'The Influence of "Working Rules" on Police Suspicion and Discretionary Decision Making', *Police Quarterly*, 11, 3, 315-337.

Wagner, D.G. and Berger, J. 1993. 'Status Characteristics Theory: The growth of a Program' in J. Berger and M. Zelditch, Jr. (eds), *Theoretical Research Programs: Studies in the Growth of Theory.* Stanford, California: Stanford University Press, pp. 23-63.

Yang, K. 2005. ‘Public Administrators’ Trust in Citizens: A Missing Link in Citizen Involvement Efforts’, *Public Administration Review*, 65, 3, 262-275.

Appendix A – Scenario example

The undermentioned inspection is part of a random sample. The entrepreneur does not have an advisor. Read the description and answer the statements for the described inspection.

You get the task to conduct an inspection at an entrepreneur with a clothing store. It's an one-man business, situated in the street you see in the picture. Your preparation did not yield any particularities. The respective entrepreneur avoids contact with you during the introductory meeting. During the inspection you notice that some invoices are missing from the records. You have the impression that the entrepreneur is lower educated.

	1 Totally disagree	2 Disagree	3 Somewhat disagree	4 Neither disagree, nor agree	5 Somewhat agree	6 Agree	7 Totally agree
It seems fine here	☐	☐	☐	☐	☐	☐	☐
I think the entrepreneur can be trusted	☐	☐	☐	☐	☐	☐	☐
I would more critically look at this entrepreneur	☐	☐	☐	☐	☐	☐	☐

10

Appendix B – Operationalization

Cues – behavioral statements and pictures

Quality of bookkeeping

1. You notice that hardly any records are kept
2. You notice that some invoices are missing from the records
3. You notice that the invoices in the records are numbered consecutively and continuously

Quality of interaction

1. The entrepreneur avoids contact with you
2. The entrepreneur talks around your questions
3. The entrepreneur answers your questions to the point

Level of education

1. You've the impression that the entrepreneur is lower educated
2. You've the impression that the entrepreneur is higher educated

Social class[*]

1. Photo 1, 2, 3 & 4
2. Photo 5, 6, 7 & 8

Photo 1

Photo 2

[*] Photo 4 and 8 have been downloaded from the website Flickr and are royalty free. The other photos have been bought at a website that allows use for non-commercial purposes.

Photo 3

Photo 4

Source photo: Flickr, made by FaceMePLS

Photo 5

Photo 6

Photo 7

Photo 8

Source photo: Flickr, made by Stipo Team for Urban Development

Dependent variables – items (7-point Likert scale: totally disagree – totally agree)

Trust evaluation

I think the entrepreneur can be trusted

Overall evaluation

It seems fine here

Intended behavior

I would more critically look at this entrepreneur

Appendix C – Correlation matrix

	V1	V2	V3	V4	V5	V6	V7	V8	V9
V1: Appears okay	–								
V2: Trust	,812**	–							
V3: More critical scrutinization	,794**	,724**	–						
V4: Dummy missing invoices	,211**	,136**	,228**	–					
V5: Dummy invoices in order	,622**	,451**	,612**	,500**	–				
V6: Dummy dodge around question	,146**	,197**	,140**	,000	,000	–			
V7 Dummy to the point	,265**	,348**	,233**	,000	,000	,500**	–		
V8: Level of education	,019	-,034	-,007	,000	,000	,000	,000	–	
V9: Social class	-,015	-,006	,020	,000	,000	,000	,000	,000	–

** $p < 0.01$.

Online appendix – Interview procedure

Step 1 – introduction and background questions

Introduction

a) Introducing myself and general topic of research
b) Guaranteeing anonymity of data processing and confidentiality
c) Explanation of procedure

Background questions

a) When started as tax official? How?
b) What kind of job before that?

Instructions given

a) The scenarios describe audits. Although they resemble real audits, they are different because there is less information. We believe that inspectors are able to make assessments based on these scenarios. The scenarios look alike, but are different. Please read them carefully and look at the pictures.
b) Because there is only concise information, we don't ask you to make a final judgment. It's rather a provisional assessment based on your first impression/feeling. We know there are other aspects you would commonly further investigate that could shed a whole different light on the case. We are *not* interested in that. Only take the mentioned information into consideration.
c) We want to emphasize that we are really interested in your first impression, and not in what other people might expect, or in what you think you should do. We're looking for honest answers. We're not testing whether you do something good or wrong in this research.
d) Please fill out the scenarios yourself. We can discuss possible questions or doubts afterwards. If you doubt about something, try to fill out the questions based on your own impression. Halfway, we'll stop for 5 minutes and I'll ask you some background questions.
e) Try not to think too long before giving your answers; we're interested in your first impression.

Step 2 – first 20 scenarios

Researcher distanced herself, and made notes on:

a) Atmosphere of interview (open/closed; signals of fatigue)
b) Time respondents took to fill out first 20 scenarios
c) Questions and remarks respondent had (were only answered and discussed during reflection)

Step 3 – background questions (around 5 minutes)

Background questions

a) Function? Specialization?
b) What kind of taxes?
c) Projects?

Step 4 – last 20 scenarios

Researcher distanced herself, and made notes on:

a) Atmosphere of interview (open/closed; signals of fatigue)
b) Time respondents took to fill out last 20 scenarios
c) Questions and remarks respondent had (were only answered and discussed during reflection)

Step 5 – reflection and disclosing more about study

Reflection

a) How is evaluation task experienced? Difficult/easy?
    i. Researcher made notes on how respondent interpreted certain cues/questions.

Disclosing of cues

a) How to rank-order these cues so it reflects the importance these aspects play in getting a first impression in still uncertain situations (as described in scenarios)?
    i. Researcher made notes when sensing possible desirable answering to scenarios.

Step 6 – small questionnaire and wrapping up

Small questionnaire

a) Propensity to trust; Level of education; Country of origin

Wrapping up

a) Thank you and small thank you gift
b) Informing about presentation of findings