HERMAN DELEECK
CENTRE FOR SOCIAL POLICY

Marjolijn De Wilde & Peter Goos

The Implementation of Social Policy: A Factorial Survey Approach

WORKING PAPER

No. 17.06

April 2017

University of Antwerp
Herman Deleeck Centre for Social Policy
centrumvoorsociaalbeleid.be

The Implementation of Social Policy: A Factorial Survey Approach

Marjolijn De Wilde & Peter Goos

Working Paper No. 17 / 06
April 2017

ABSTRACT

The implementation of social policies has a multidimensional character. We present an experimental method (factorial survey) by means of which one can gather data from a large number of respondents from several agencies and across countries. A concrete research example involves a survey among Belgian social assistance case managers, who were asked to predict the likelihood of experimentally varied hypothetical clients being sanctioned. The data had a multi-level structure (3=organisation [n=79]; 2=respondent [n=594]; and 1=client descriptions [n=4855]).

We empirically show how the method is useful for studies on issues such as conditionality (client level), discretion (social worker and organisation level), decentralisation (municipality/region level) and international policymaking (country level). Our recommendations for the use of factorial surveys with regard to social policy implementation research are: asking for expected and not for preferred treatment, adding a questionnaire about respondents and their organisation, stratified sampling of respondents and using multi-level techniques for analysis.

Keywords: Methodology, Social Policy, Factorial Survey, Experimental Design, Discretion, Decentralisation, Conditionality, Social Assistance, Multi-level Governance

Corresponding author:

Marjolijn De Wilde & Peter Goos
Herman Deleeck Centre for Social Policy (CSB)
Faculty of Political and Social Sciences
University of Antwerp
Email: marjolijn.dewilde@uantwerpen.be

1 Introduction

The implementation of social policy legislation has a multidimensional character. It is contingent on several levels, for example, the country, the organisation, the social worker and the client (Priem, Walters, & Li, 2011; Rice, 2012). First of all, implementation depends on existing legislation (country) and the characteristics of clients who are deemed eligible for the specific policy (client). Second, the extent to which decision-making power is decentralised to subnational levels (region, municipality, organisation, team) and the degree to which these subnational levels use their policy-making and implementing discretion[i] result in treatment variation across implementing organisations and municipalities, and thus on the actual treatment of eligible clients. Furthermore, as social policies are often implemented by local case managers, the degree of discretion offered to or used by these individual professionals is a third overarching and influential factor (Evans and Harris 2004; Lipsky 2010). A final level that, while interesting, is not discussed further in this paper, is the European coordination of national policies. It is in this context that the term 'multi-level-governance' is often used to refer to the distribution of decision-making power over several actors both vertically (national, supranational and subnational) and horizontally (several actors at one level) (Benz 2000; Hooghe 1996; Hooghe and Marks 2003; Piattoni 2009; Stephenson 2013). It is noteworthy that, to the best of our knowledge, the decision-making power of individual case managers is seldom mentioned in the multi-level-governance literature.

The aim of this paper is to show that it is particularly difficult to account for all of these levels using one research method. This is due partly to data-gathering limitations, but also to paradigm differences between research fields. In this paper, we demonstrate how factorial surveys have the potential to overcome methodological limitations, and, in doing so, provide opportunities to study the multi-level nature of social policy implementation.

In factorial surveys, respondents are presented with experimentally varied hypothetical stories and asked to judge the situations. We illustrate our claim that factorial surveys have much potential with an example in which almost 600 case managers from 79 social assistance agencies in Flanders (Belgium) were presented with unique sets of nine client descriptions. In this example, we focused on the question whether labour market activation is as intense among clients who have children, including sick children, as it is among clients who do not have children. We were able to demonstrate that the answer to this question depends both on the client characteristics in terms of 'parenthood' and on the characteristics of the case manager treating the client, as well as on several characteristics inherent to the organisation and the municipality in which the case manager works. Furthermore, we observed extensive variation between case managers from the same organisation and only little variation between organisations.

The remainder of the paper consists of three parts. The first part reviews the literature on the conditionality of social assistance, the decentralisation of decision-making power to subnational levels and individual case managers' discretion. We illustrate that, because one

or more of the levels described above were ignored in published work, the conclusions in the literature are in fact biased. In the second part, we outline the factorial survey approach, arguing that it offers the potential to overcome the shortcomings of other methods provided certain requirements are met. In the third part, we demonstrate the positive impact of using a factorial survey by means of an example.

2 Existing methods in social policy implementation research

For a long time, policy makers and researchers believed that policies were decided upon at national level and implemented at lower levels as intended by the legislators (Tabin & Perriard, 2016). Two different evolutions have altered this perception. The first relates to the growing importance of other policy-making levels (Europe, regions, municipalities, etc.). The second relates to research on street-level bureaucracy, which acknowledges the impact of street-level workers in policy implementation and in policy-making.

At present, two methods are commonly used to study and compare the national legislation and implementation of social assistance policies (Tabin & Perriard, 2016). The first method involves the study of national spending on both cash welfare expenditure and activation-related spending (e.g. active labour market programmes or ALMPs) (Champion & Bonoli, 2011). A comparison of the two types of spending provides information about governmental choices and, indirectly, about policy implementation. Although still widely used and valuable to a certain extent, this approach has serious limitations. We discern two main problems. The first concerns the classification of spending packages. Several studies have shown that, by classifying the same programs differently, the results change considerably (De Deken & Kittel, 2007; Kittel & Obinger, 2003). Second, these studies provide no information about how programs are actually implemented. The same program may be implemented entirely differently by different organisations or case managers.

The second method focusing on national legislations, involves interviewing country experts about standard families and about the interpretation of the country's legislation when it comes to families with a given composition and a particular set of problems (Eardley, Bradshaw, Ditch, Gough, & Whiteford, 1996; Kazepov, 2010; Kazepov & Barberis, 2012; Marchal, Marx, & Van Mechelen, 2014; Marchal & van Mechelen, 2017; Van Mechelen, Marchal, Goedemé, Marx, & Cantillon, 2011). This powerful method provides insight into similarities and differences in legislation across countries and over time within countries. However, the method ignores the way in which policies are embedded locally. Marchal and Van Mechelen (2017) circumvent this problem in countries with high levels of local discretion by interviewing extra experts from local municipalities or states (e.g., a stakeholder in the city of Antwerp in the case of Belgium). Yet, the authors do conclude that a limitation of their study is that they need to rely on one particular case to draw conclusions about an entire country or region (Marchal & van Mechelen, 2017). In this sense, both methods have the drawback that it is difficult to study local embedment, which is particularly relevant. Some authors

suggest, indeed, that, in comparison to other forms of policy, social policy is one of the domains in which the importance of the local level is high (Ripley & Franklin, 1982 as cited in Hasenfeld & Brock, 1991). This is mainly due to the fact that national legislation merely provides a framework and does not have immediate applicability. To a large extent, discretion is left to local organisations and its social workers (Wallander, 2012).

This is where decentralisation comes in. The interplay between the national and local policy levels is very country-specific. Kazepov (2010) describes four different means of organising social policy-making: "(1) countries with strong local autonomy which is centrally framed; (2) countries with a strong national/central frame; (3) countries with strong regional (or federal) frame; (4) countries with mixed frames in transition from one frame to another" (Kazepov & Barberis, 2012). Differences between these systems are evident both in the aspect of legislative power and in the degree of financial autonomy at each level. Besides Italy, there are very few countries in Europe without national framework (Kazepov & Barberis, 2012). Yet, the details of each set of regulations differ from country to country and may range from establishing the right to social assistance to detailed regulations concerning eligibility criteria, eligibility duration and sanction measures. Furthermore, the regulations may be implemented by locally embedded organisations that depend on the state level, or by individual municipalities themselves, at local level. Decentralisation dynamics are often studied with the aim of establishing how implementation is organised (Hölsch & Kraus, 2006; Minas, Whrigth, & Van Berkel, 2012; van Berkel, 2006). To date, however, very few studies have examined the effect of such decentralisation on actual implementation or actual client treatment (Carpentier, 2016).

Furthermore, recent history teaches us that, in many branches of social policy, the degree of discretion that organisations and case managers are permitted to use has increased due to the new focus on integration and activation rather than provision of financial support (Kazepov & Barberis, 2013; Vando Borghi & van Berkel, 2007). In order to decide whether a client is in financial need, a social worker is required to perform a means test. The norms for this test can be established at national level. However, decisions about clients' integration trajectories are likely to be tailored more closely to the needs and abilities of the client in question. National guidelines concerning activation are more difficult to formulate than norms for means tests. Consequently, the allocation of resources and activation measures depends to a large extent on the individual social worker (Lipsky, 2010).

Several methods are currently being used to study the implementation of social policies by case managers and local organisations. The most straightforward of these is the observation method, as it is less biased by the case manager's or client's view than survey research. A disadvantage of observation, however, is that it is very time-consuming, meaning that researchers need substantial human and financial resources to overcome the main weakness of existing studies, which is that they typically involve only a few municipalities (Thoren, 2008).

More generally, most studies that address the territorial dimension, by interviewing either field workers or clients themselves, are qualitative in nature. When the interviews are held face-to-face and the questions are open-ended, a substantial amount of information can be collected (Hermans, 2005; Nybom, 2012). This information can result in meaningful hypotheses about the reasons behind specific ways of implementation, but – because it is not quantitative – it does not allow for these hypotheses to be tested. Qualitative research on social policy implementation is limited to comparisons of agencies or municipalities within single countries (Hermans, 2005; Nybom, 2013; Thoren, 2008) or among small numbers of municipalities in multiple countries (Evans, 2007; Saraceno, 2002). Therefore, no general view exists on the treatments received by clients.

Some studies have collected quantitative local data, however. The most promising of these relied on registered data. Bargain and colleagues (2012) used a Finnish survey to measure the use of social assistance. The survey in question collects administrative data on 0.5% of the population of Finland. It allows the researchers to carry out a systematic check of how many citizens with low incomes are receiving benefits. Carpentier and Neels (Carpentier, Neels, & Van den Bosch, 2014) used a Belgian database that collects registration data from various institutions, including information on demographics, unemployment benefits, social assistance benefits and health insurance. They had access to five years of data on one third of the individuals who started receiving social assistance benefits in 2004. This type of data makes it possible to follow clients over long assistance pathways, without having to interview them multiple times. It is, however, unclear whether the data registered is suitable for investigating all aspects of policy implementation. Information about clients' motivation and life experiences is often missing. Nybom (2013) has attempted to overcome this shortcoming by combining administrative data (300 client records from four municipalities) with interviews. Again, this is a very time-consuming method and does not permit extensive general investigations.

Another problem with large-scale registration or survey data is that the selection bias caused by differences in client profiles across municipalities impedes the comparison of local treatment policies. A strong activation practice in a certain municipality may be caused either by local policy or by specific claimant characteristics. Moreover, certain types of clients might never occur in certain municipalities, thus hindering comparison across municipalities (Blommesteijn, van Geuns, Groenewoud, & Slotboom, 2012). Finally, most registration data are country-specific, which prevents comparisons between countries.

It should be clear that each of the existing methods has major weaknesses. We believe that using the factorial survey approach would mean a step forward, because it overcomes some of the shortcomings of the approaches described above.

3 The factorial survey approach

3.1 Introduction to factorial surveys

Factorial surveys incorporate the positive features of experimental research. In a traditional comparative experiment, the effect of one factor on a dependent variable is tested. In a factorial experiment, however, at least two factors are varied at once, enabling the researchers to investigate the effect of multiple factors and their 'interaction effects' on the dependent variable[ii]. Each factor has two or more levels or categories (e.g. female and male for gender). When all possible combinations of all levels across all factors are considered, a full factorial design is obtained. For example, when an experiment consists of three factors and all these factors have two possible levels, then the full factorial design consists of eight possible factor-level combinations (2×2×2).

In a factorial survey, the experiment requires each respondent to read a story about a hypothetical person or situation and rate the person or the situation according to well-defined dependent variables (see appendix for an example of a vignette and related dependent variables). This story or situation is usually referred to as a vignette. Like every experimental test in a factorial experiment, every story involves only one level of each factor. In some cases, all respondents rate all possible stories (= the full factorial universe). Most commonly, however, a sample is taken from the vignette population. Samples can be taken at random, but a more appropriate approach – especially if few respondents are involved in the survey – is to select a D-efficient sample. A D-efficient sample ensures that maximum information is obtained about the effects of the experimental factors. If there are no practical constraints in the selection of a sample, then the D-efficient sample consists of an orthogonal design in which, for each factor, the levels occur equally often (level balance). An orthogonal design ensures that a linear regression or ANOVA model relating the factors to the ratings does not suffer from multicollinearity (Atzmüller & Steiner, 2010; Dülmer, 2007, 2016) and allows the impact of the experimental factors on the dependent variable to be quantified with maximum precision. Practical constraints, such as the fact that certain combinations of factor levels are unrealistic and can therefore not be used in the survey, or the fact that only limited numbers of vignettes are practically feasible, may make it impossible to construct a perfectly orthogonal design. The D-efficient survey design approach is flexible, however, in the sense that it allows for the generation of a survey with maximum information content under the given constraints. The resulting D-efficient sample is then as close as possible to being orthogonal.

Typically, a factorial survey is sent out to various respondents, and each respondent evaluates only a subset of all of the vignettes in the D-efficient sample. An attractive feature of the D-efficient survey design approach is that it allows researchers to determine the best possible allocation of vignettes to respondents, in the sense that the influence of both the experimental factors and the respondent characteristics can be quantified with maximum

precision. This allocation of vignettes to respondents is known as 'blocking' in the experimental design literature (see, e.g., Goos & Jones, 2011).

Using the factorial survey approach has the potential to solve some of the problems in social policy implementation research listed above. It makes it possible to target a large number of respondents in multiple municipalities (and countries), thus allowing for generalisation without ignoring local processes (Aguinis & Bradley, 2014). Further, the possibility to manipulate the independent variables (i.e. the vignette attributes) and measure their independent effects on the dependent variable allows for conclusions about causal relationships. Lastly, randomisation of the assignment of vignette blocks to respondents prevents selection bias. This might be compared to studies on registered real data where selection bias constitutes a problem, as in reality some client problems might be over- or underrepresented in certain areas. The possibility to randomise cases over professionals and over organisations thus makes local policies truly comparable.

3.2 Studying social policy implementation with factorial surveys

Studying social policy implementation is closely related to what is called 'measuring professional judgment' (Jasso, 2006; Taylor, 2006; Wallander, 2012), which is concerned with the way in which social workers make decisions in their everyday practice. The primary aim is to unravel professional agreements concerning specific cases or problems. Because these agreements are likely to be based on experience in social work practice or in personal life, rather than on scientific knowledge, they may be difficult to detect without experimental methods. One of the most important authors in this field, Wallander (2012), has made a case for studying professional disagreement too, but this has not yet been picked up in existing research. At present, in the literature we see intense examination of client characteristics as the cause of specific treatments (Morley, 2010; Ortega, Baz, & Sánchez, 2012; Skarlicki & Turner, 2014; Stokes & Schmidt, 2012; Webster, O'Toole, O'Toole, & Lucal, 2005), while characteristics of client managers and agencies are largely overlooked. The factorial survey may help us to obtain a fuller picture of social policy implementation. Below, we outline four valuable additions that we suggest to the existing guidelines of Wallander (2012) and Taylor (2006).

3.3 Additions to traditional factorial survey research

First, the dependent variables (or at least one of them) should relate to plausible acts or treatments (e.g. starting an activation programme with a client) and not merely to an opinion about the situation of the client (e.g. the client is ready for the regular labour market). Further, these treatments should be framed in the context of the organisation, meaning that the respondent is asked to disclose what should happen to a certain client in the organisation where she is working. The respondent should not be asked what she would personally prefer

or what she would personally do, but should be encouraged to consider all elements that could influence the decision concerning treatment. We are not aware of a single study that has framed the survey in this broader perspective on treatment. In our view, this means that existing studies neglect the decision-making process in organisations, a process which is rarely straightforward. A teacher's decision, for example, to report the abuse of a pupil by her parents might be influenced by the situation surrounding the abuse (vignette attributes) and by the intuitions of the teacher (= respondent), but also by the fact that the teacher must discuss her decisions with a supervisor. If the respondent is not asked to take her organisation's decision-making process into account, the researcher runs the risk of missing the information she is looking for.

Secondly, in order to be able to account for local decision-making processes or policies, researchers should include a questionnaire about the respondent and the organisation where she works (Wallander, 2012). This questionnaire might collect information about the partners involved in making decisions about the treatment. This may simply be the respondent herself, but could potentially involve the head of the team, the entire team or an external committee. Such decision structures reveal a great deal about the discretion a social worker has in deciding what treatment is offered. The structure can vary substantially depending on the organisation in question. It should be noted that, whereas the factorial survey enables collecting comparable data on the vignette / client level, the collection of data on the respondent or organisation level suffers from the same comparison problems as traditional surveys.

The third valuable addition to existing guidelines would be to perform appropriate sampling of respondents so that the discretion used by an organisation and/or its social workers can be measured properly (Aguinis & Bradley, 2014). Respondents should be distributed evenly over the region being investigated and it should also be possible to compare agencies and respondents (with different characteristics). This means that a stratified or a clustered sample is required (Bryman, 2012; Groves et al., 2009; Heeringa & Berglund, 2010). Researchers considering the use of a factorial survey to investigate social policy implementation will mostly opt for a multi-stage sampling technique that combines clustered and stratified sampling on several levels. The theoretical literature on professional judgment research (Taylor, 2006; Wallander, 2012) draws a range of conclusions about sampling. Taylor (2006) suggests taking random samples from the respondent population. Wallander (2012) pays more attention to the organisational level and suggests sampling multiple respondents from several workplaces to be able to account for disagreement in treatment practice across organisations. While random sampling is frequently used in existing research, it remains uncommon to consider the local level by clustering or stratifying.

Fourth, multi-level regression or ANOVA models are appropriate for quantifying and testing how respondents' ratings relate to the experimental factors, respondent and agency characteristics. Multi-level modelling takes into account statistical dependencies between ratings given by the same respondent (Aguinis & Bradley, 2014). When implementing the above recommendations on the acquisition of information about the agency and the use of

clustered/stratified sampling, respondents are clustered into agencies. This leads to a second kind of dependency in the individual vignette ratings: ratings given by two different respondents from the same agency tend to be correlated and are therefore not independent. This is because every agency is attempting to implement certain local policies and because all of the respondents from a given agency are subject to similar unobserved environmental/cultural/political influences. A multi-level regression model allows for the inclusion of known organisational or municipal features, such as population parameters or number of clients with the problem being studied (e.g. people on social assistance). Further, it is possible to study the unexplained variance at all levels included in the model. For instance, when the results indicate large differences between agencies or municipalities, this means that much of the discretionary freedom is being used at agency level, with local regulations or evolved treatment traditions guiding the decisions of social workers.

3.4 Obstacles

The vignette method does have a number of drawbacks, however. Most of them concern validity: questions arise as to whether or not the method measures what it sets out to. Internal validity, meaning the ability to draw causal conclusions, is high since the method operates on an experimental basis (Taylor, 2006; Wallander, 2012) and ensures that no systematic errors occur. External validity, in all its forms, is more problematic. Most of the problems and corresponding solutions have already been discussed in the existing literature, and, due to space limitations, we refer the reader to other articles. Examples of external validity problems are: the risk that vignettes might lack realism (Aguinis & Bradley, 2014; Aiman-Smith, Scullen, & Barr, 2002; Karren & Barringer, 2002; Wallander, 2009); the risk that the respondent's answers lack realism (Armacost, Hosseini, Morris, & Rehbein, 1991; Eifler, 2010; Kirwan, Chaput de Saintonge, Joyce, & Currey, 1983; Langley, Tritchler, Llewellyn-Thomas, & Till, 1991; Mohan et al., 2014; Peabody, Luck, Glassman, Dresselhaus, & Lee, 2000); and the potential over- and under-complexity of the vignettes (Aiman-Smith et al., 2002; Auspurg, Hinz, Liebig, & Sauer, 2015; Auspurg & Jäckle, 2015; Caussade, Ortúzar, Rizzi, & Hensher, 2005; DeShazo & Fermo, 2002; Johnson, 2006; Taylor, 2006; Wallander, 2009). The type of factorial survey used to study social policy implementation involves the same problems as traditional factorial survey research, with one problem even becoming bigger. As stated in Section 3.3, the respondent should be asked about what would happen in her organisation and not about her own individual decisions. This increases the risk of response error. The results therefore need to be treated with caution and interpreted in relation to other types of research, such as analysis of administrative data or interviews with client managers and other actors in the agency. As a contribution to existing methodological research, it might be an interesting avenue for future research to compare factorial survey data on social policy with administrative data on the same topic. It is unlikely, however, that cases with the same features in both types of research are wholly comparable. Furthermore, each vignette could be followed by two questions, one of which asks the professional what she would personally

advise while the other asks what she thinks that would happen in reality. As this makes the questionnaire longer and hereby less straightforward to answer, this might ultimately be an unsolvable problem. However, intuition suggests that responses to experimentally created stories should be closer to reality than those given in interviews directly by respondents (Armacost et al., 1991; Kirwan et al., 1983; Peabody et al., 2000).

4 Example

In this section, we demonstrate the potential of the factorial survey method with an example. The policy being investigated in the example is social assistance in Flanders. We asked almost 600 case managers working in 79 social assistance organisations to respond to a unique[iii] set of nine descriptions of social assistance claimants. The main question was whether clients with and without children are treated similarly with regard to activation requirements (= direction towards the labour market). The dependent variable under investigation concerned the likelihood (measured on a 7-point Likert scale) that a client would lose her social assistance benefits if she refused a job or activation offer that started at 5 o'clock in the morning. In each of the following three paragraphs, we address one of the claims made above. First, we demonstrate the effects of the client characteristic 'parenthood', as well as the case manager's and organisational characteristics. Second, we demonstrate the usefulness of the multi-level technique. In the third paragraph, we elaborate on the sampling procedure.

4.1 Accounting for effects on three levels

The vignettes used in this study portrayed single clients who had no work, no money, no savings and no contact with close relatives. The clients in the vignettes differed with regard to 14 client characteristics, including gender, nationality, level of education, language abilities, mental health, addiction, motivation to work, work and activation experiences and attitude (see appendix for a detailed outline of the vignette characteristics). The characteristic that interests us most in this paper is parenthood. We included three levels: a client with no children, a client with a healthy two-year-old child and a client with a sick two-year-old child (immunity disorder, meaning that the child falls ill often and unpredictably). Based on the literature (Eardley et al., 1996; Kazepov, 1999; Lødemel & Trickey, 2001; Lorentzen, Dahl, & Harsløf, 2012; Oorschot, Uunk, & Jeene, 2008; Rice, 2012) and on interviews with social assistance stakeholders and academics about the factors influencing eligibility and sanction decisions, we expect that clients with children are less likely to lose their benefit. Likewise, we expect this likelihood to drop even further for clients with sick children. In Belgian legislation (Wet op maatschappelijke integratie, 2002), the condition 'parenthood' as such does not exempt from job or activation acceptance. Claimants younger than 25 years are required to cooperate in the activation process, unless fairness or health reasons apply. Such fairness reasons are not described in legislation. Whether being a parent (of sick children) is considered as a fairness reason might depend on tendencies, across case managers and municipalities, to

do so. If such tendencies are present, we should see a strong effect of including the client characteristic 'parenthood' in the regression model. The fairness assessment could, however, also be influenced by individual considerations of case managers (i.e., discretion) or by local legislation or practice (i.e., decentralisation). To test this, we used multi-level models (see the section below) and we include explanatory variables of the client, the case manager and the organisation level.

In Model 1, we included all 14 client characteristics, but, in Table 1, we focus on the parenthood variable. The effect of having a healthy or a sick child on the likelihood of being sanctioned after refusing a job or activation offer that starts early in the morning is clearly seen in Table 1 in the 'Model 1' column. Having a child reduces the likelihood of being sanctioned, as measured on the Likert scale, by around one point out of seven; this means a reduction of 13 to 18 percentage points. Although this is not shown in the table, whether or not having a child is the client characteristic that had the largest estimated effect on the dependent variable.

Up to this point, the method involves nothing new. This is the traditional way in which factorial surveys are used: to detect the effect of vignette compartments on the dependent variable of interest. In Model 2, however, we introduce explanatory variables related to the respondents rating the vignettes. These were 594 case managers from 79 Flemish social assistance offices. We wanted to enter two sorts of respondent characteristics that might influence the respondents' approach to sanctioning clients with children. The first characteristic was parenthood: we asked whether the respondent had children herself (binary variable). To check whether respondents with children reacted differently when treating clients who also had children, we included an interaction term between the two variables. We also included the age of the respondent, because the direct effects of parenthood might actually be caused by age effects, as older respondents are more likely to have children. The second variable added to the model was an opinion question. We asked the respondents to rate the following statement from 1 (disagree) to 5 (agree): "welfare clients should be sanctioned more often if they do not comply with agreements". This variable was drawn from a broader standardised opinion questionnaire filled in by all respondents and should not normally be used on its own. As a matter of fact, the two variables entered in the model are only given for the purpose of illustration. In a thorough analysis, we would include a larger number of variables. It is true that the effects measured in this paper might have been influenced by variables we did not take into account, but, for clarity of exposition, we have chosen to keep the analyses simple.

The results of the modified model are shown in the 'Model 2' column in the table 2. There was no direct effect of respondent parenthood, but there did seem to be an age effect. Older respondents predicted less sanctioning of clients than their younger colleagues did. Furthermore, case managers whit children predicted less sanctioning of clients with sick children than their childless colleagues did. The case manager's attitude towards sanctioning in general also influenced the results. The more the case manager believed that additional

sanctioning is positive, the more likely she was to predict that the hypothetical clients would be sanctioned.

In Model 3, we added variables related to the organisation and the municipality in which it is located. First, we entered a variable indicating the availability of childcare, based on figures from the organisation governing childcare provision in Flanders (Kind en Gezin). The figures reflect the percentage of children aged between 0 and 3 for whom childcare is available in the municipality. Across the 79 municipalities surveyed in this research, the figures ranged from 19% to 69% with a mean of 42%. We combined this variable with the client parenthood variable in an interaction term. We also added each organisation's mean opinion of sanctioning (see respondent variable described above). In this way, we were able to test whether the common culture in an organisation influences the predictions of sanctioning and potentially overrules the rather strong effect of case managers' personal opinions.

No significant, meaningful results were found here. Childcare availability is significant, but not in the way that we expected: municipalities with high childcare availability seem to sanction less. This unexpected result may be due to one or more other characteristics not included in our analysis here. Also the interaction with client parenthood – the variable that interested us most – was not found to be significant. The same was true for the mean opinion on sanctioning in the municipality.

This example, which demonstrates the inclusion of variables from three of the levels that are deemed decisive for social assistance implementation, shows that factorial surveys are a powerful method. In principle, it should also be possible to investigate several countries using one set of vignettes. Country-dummies could be added to the multi-level regression model, additively and in interaction terms with variables from the first or second level, to detect country-specific effects. If enough countries are included in the study, country characteristics can be added too. This would result in a degree of comparability that is unique in social sciences.

Table 1.
Five multi-level models including, step by step, fixed effects (Model 1 – 3) and random effects (Model 4) based on an online survey in 79 Flemish social assistance organisations among 492 case managers, 2015

	Model 0 = multi-level model with no predictor variables	**Model 1** = Model 0 + client characteristics as fixed effects	**Model 2** = Model 1 + respondent characteristics as fixed effects	**Model 3** = Model 2 + organisation and municipality characteristics as fixed effects	**Model 4** = Model 3 + randomisation of client characteristic at the second level
Intercept	3,82 (0,1)***	4,06 (0,2)***	3,71 (0,1)***	4,64 (0,3)***	4,63 (0,5)***
Parenthood client	No child				
	Healthy child	-0,7 (0,1)***	-0,75 (0,1)***	-1,0 (0,1)***	-1,0 (0,1)***
	Sick child	-0,95 (0,1)***	-1,04 (0,1)***	-1,35 (0,1)***	-1,41 (0,1)***
Respondent age			-0,02 (0,0)**	-0,02 (0,0)**	-0,02 (0,0)**
Parenthood respondent	No child				
	Child(ren)		-0,12 (0,1)	-0,13 (0,1)	-0,1 (0,1)
Interaction parenthood client & parenthood respondent	Client healthy child & respondent child(ren)		-0,13 (0,1)	-0,23 (0,1)	-0,18 (0,1)
	Client sick child & respondent child(ren)		-0,23 (0,1)*	-0,23 (0,1)*	-0,26 (0,1)*
Opinion concerning sanctioning of welfare clients in general			0,28 (0,1)***	0,27 (0,1)***	0,26 (0,1)***
Availability of childcare				-0,02 (0,0)**	-0,02 (0,0)*
Interaction parenthood client & availability of childcare	Client healthy child * availability of childcare			0,01 (0,0)	0,01 (0,0)
	Client sick child * availability of childcare			0,01 (0,0)	0,01 (0,0)
Opinion concerning sanctioning of welfare clients (mean in the organisation)				0,82 (0,5)	0,82 (0,5)
Variance components					
Level 3 - municipality	0,19 (0,1)*	0,17 (0,1)*	0,17 (0,1)*	0,09 (0,1)	0,12 (0,1)*
Level 2 - respondent	1,6 (0,1)***	1,46 (0,1)***	1,33 (0,1)***	1,35 (0,1)***	1,49 (0,1)***
cov(cons\healthy child)					-0,53 (0,1)***
var(healthy child)					1,2 (0,1)***
cov(cons\sick child)					-0,56 (0,1)***
cov(healthy child\sick child)					1,3 (0,1)***
var(sick child)					1,37 (0,1)***
Level 1 - vignette	1,49 (0,0)***	1,29 (0,0)***	1,29 (0,0)***	1,29 (0,0)***	1,01 (0,0)***
N	4838	4838	4838	4838	4838

Standard errors in parentheses
* p<0.05, ** p<0.01, *** p<0.001

4.2 The advantages of multi-level models

As pointed out, all of the analyses presented in Section 4.1 involved multi-level models. A major advantage of multi-level modelling is that it offers the possibility to investigate which level (organisation – case manager – client) influences treatment most. In multi-level models, the unexplained variance of the regression model is divided into as many parts as there are levels in the model. The resulting variance components are shown in the bottom part of Table

1. The null model (column 3) provides insight into the division of the variance if no explanatory variables have been included yet. As we can see, most of the unexplained variance in the null model appears at the first and second level, namely at the client and respondent level. Only 6% of all unexplained variance appears at the municipality level. Most of the unexplained variance in the null model is due to differences between case managers within a municipality (49%). A similar amount of variance is due to predicted treatment differences for client vignettes rated by a given case manager (45%). Before evaluating these percentages, it is important to know whether the variation in the responses is substantial. As each possible value from 1 to 7 is chosen almost equally often by the respondents in our example, that variation is certainly large. The fact that the variation in responses is so high, and situated at the respondent and client levels, means that predicted treatment mainly depends on client characteristics (as should be expected) and on case manager preferences. If equal treatment for similar clients is targeted, this is an important result which cannot be established without using multi-level techniques.

A second advantage of using multi-level techniques is that they enable us to check the impact of adding explanatory variables (fixed effects) to the model on the unexplained variance at each level. As we can see in the above table, the unexplained variance at the first level (the vignette) decreased by 0.2, which is a reduction of 13%. In total, 11% of the original total unexplained variance is explained by adding the client characteristics. Six percent of this decline is situated at the vignette level. The remaining 5% is situated at the respondent (4%) and the municipality (1%) level. This means that adding the client characteristics explains four percentage points of the variation among case managers. This is probably a result of blocking the vignettes into respondent-specific vignette sets. All of the respondents received a different set of vignettes. D-efficient sampling of vignettes in unique decks guarantees as balanced a blocking process as possible. However, fairly small decks of nine vignettes are no guarantee that no method-effects occur. Some respondents might see no extreme cases, where others do, which possibly influences the ratings. So part of the variation among respondents might be explained by this blocking[iv]. Of the 11% total decline in unexplained variance due to entering the client variables in the model, 5% is due to the parenthood variable (not shown in Table 1). This means that one client variable explains almost as much as the 13 other client variables.

When we added the variables at the respondent level, the unexplained variance at the second level decreased by 9% and the total unexplained variance decreased by 5% (from Model 1 to Model 2), with regard to the total unexplained variance of the null model, this decline represents 4% (from Model 0 to Model 2). This decrease is located entirely at the second level, which means that the respondent characteristics selected do not influence the results at the other levels. Lastly, the addition of the organisation and municipality variables causes a reduction of almost half of the unexplained variance at the first level. However, this only represents 1% of the total variance.

A final advantage of multi-level techniques is shown in the last column of Table 1 (Model 4): the ability to add random slopes. Adding random slopes for the client parenthood variable, for instance, at the second level (the respondent) means we do not expect the effect of client parenthood to be the same for all respondents. After all, given that the preferences of the case managers do matter (see variances in the null model), the likelihood that clients with children will be sanctioned less than clients without children probably varies from respondent to respondent. When calculating the mean unexplained variance at the second level for respondents rating clients whit children, we do indeed see more unexplained variance than when respondents are rating clients who do not have children, especially when the child is sick. This means that clients with children are sanctioned less than clients who do not have children (see above for client effect), but that case managers differ in their degree of sanctioning.

4.3 The advantages of stratified sampling

We selected the 79 municipalities and almost 600 case managers surveyed in Flanders in 2015 using multi-stage stratified sampling. The first stratification level was composed by grouping the municipalities (each municipality in Belgium has one social assistance organisation). The strata were based on the number of inhabitants (three levels: <50,000; 50,000–100,000; and >100,000), the percentage of social assistance claimants among all inhabitants (four levels: <0.5%; 0.5–1%; 1–2%; and >2%) and an index based on socio-economic and socio-demographic characteristics (Belfius, 2007). We established 36 strata in total. Next, guided by power analyses based on a pilot study, we randomly selected a third of the municipalities in each stratum. This resulted in 90 municipalities. If a social assistance organisation's contact person declined the invitation to participate, we replaced the corresponding municipality with another one from the same stratum. In total, we invited 104 municipalities, 90 of which actually participated in our study. In 13 of these municipalities, only one person eventually responded. In two municipalities there was only one case manager working in the particular position. These municipalities were kept in the analyses. We excluded the other municipalities from our analysis, because having only one respondent at the second of three levels makes multi-level analyses more difficult.

In each municipality, we sampled one third of the case managers involved in the decision-making process concerning eligibility for benefits. This means that social workers specialised in debt counselling or activation trajectories were not surveyed. We asked the social assistance organisations to provide lists of these case managers and to include information about gender, age group (4 levels) and position (team manager or not). Based on these characteristics, we constructed 16 strata, from which we drew a proportional one-third sample, while making sure that at least one person was selected from each stratum available in each municipality. In the event a certain respondent did not participate in the study after two reminders, we approached a similar case manager from the same municipality. In total,

839 case managers received the questionnaire, 681 started the survey, 610 completed it and 594 were used in the analyses.

In the remainder of this section, we demonstrate how our results would have been biased if we had not used random or stratified samples. Paying insufficient attention to the respondent and municipality levels tends to give rise to two rather common practices, both of which cause bias in the results. The first of these is only studying some of the organisations (generally the largest in size), while the second is the surveying of only those professionals who volunteer to participate. Here, we wish to demonstrate that, even if a researcher only wants to take the effect of vignette characteristics into account, he or she should take a random or stratified sample of all possible organisations and a random or stratified sample of respondents from those organisations.

In order to demonstrate the potential bias, we added two new models to our study (see Table 2). Model 5 adds several variables to Model 1 (only client characteristics): seven dummy variables indicating the largest municipalities in the study (more than 80,000 inhabitants = Aalst, Antwerp, Bruges, Ghent, Hasselt, Leuven and Ostend) and the interaction of these dummies with the client parenthood variable. As shown in Table 2, some of the municipalities entered as dummies have a direct effect on the likelihood of sanctioning. Case managers in Municipality 1 tend to predict sanctioning more often than those in smaller municipalities and in the six other large municipalities. Case managers in municipality 6, on the other hand, seem to sanction less. Some of the municipalities also display a specific effect with regard to client parenthood. For Municipalities 3 and 6, for instance, the negative effect of having children on the likelihood of being sanctioned is less strong than for the other municipalities. These results show that, should a researcher only survey these large cities, the results would turn out to be biased.

Table 2

Two multi-level models (Model 5 and 6) showing the significant effect of specific municipalities / organisations or groups of respondents and their interaction with the client's

parenthood based on an online survey in 79 Flemish social assistance organisations among 494 case managers, 2015

		Model1 = multi-level model with client characteristics as fixed effects	Model 5 = Model 1 + dummy variables for 7 municipalities + interactions of these dummies with client parenthood	Model 6 = Model 1 + respondent-strata-variable + interactions of this variable with client parenthood
Intercept		4,06 (0,1)***	4,13 (0,1)***	4,12 (0,2)***
Parenthood client	No child			
	Healthy child	-0,7 (0,1)***	-0,76 (0,1)**	-0,81 (0,1)***
	Sick child	-0,95 (0,1)***	-1,1 (0,1)***	-1,08 (0,1)***
Municipalities – dummies	Municipality 1		0,36 (0,2)*	
	Municipality 2		-0,24 (0,2)	
	Municipality 3		-0,14 (0,3)	
	Municipality 4		-0,4 (0,3)	
	Municipality 5		0,09 (0,5)	
	Municipality 6		-0,94 (0,2)***	
	Municipality 7		0,33 (0,6)	
Interactions municipalities * parenthood client	M1 * C healthy child		-0,09 (0,1)	
	M1 * C sick child		-0,05 (0,1)	
	M2 * C healthy child		0.02 (0,2)	
	M2 * C sick child		-0,25 (0,2)	
	M3 * C healthy child		0,22 (0,2)	
	M3 * C sick child		0,47 (0,2)*	
	M4 * C healthy child		0,27 (0,2)	
	M4 * C sick child		0,21 (0,2)	
	M5 * C healthy child		-0,09 (0,3)	
	M5 * C sick child		-0,31 (0,3)	
	M6 * C healthy child		0,32 (0,1)**	
	M6 * C sick child		0,61 (0,1)***	
	M7 * C healthy child		0,51 (0,4)	
	M7 * C sick child		0,58 (0,4)	
Type of respondent	Base = R = female - 30 to 40 years – not team manager			
	R = male – 40 to 50 years – not team manager			-0,77 (0,3)*
Interaction type of respondent * parenthood client	R = female - -30 years – not team manager * C healthy child			0,25 (0,1)*
	R = female - -30 years – not team manager * C sick child			0,25 (0,1)*
	R = male - -30 years – not team manager * C healthy child			0,48 (0,2)*
	R = male - -30 years – not team manager * C sick child			0,5 (0,3)*
	R = male – 30 - 40 years – not team manager * C healthy child			0,38 (0,2)*
	R = male – 30 - 40 years – not team manager * C sick child			0,46 (0,2)*
	R = male – 30 - 40 years –team manager * C sick child			0,88 (0,4)*
	R = male – 40 - 50 years – not team manager * C healthy child			0,6 (0,3)*
	R = male – 40 - 50 years – team manager * C healthy child			0,77 (0,3)*

In Model 6, we added a different sort of variable to demonstrate how a non-random or non-stratified sample of case managers might bias the results. As described above, we stratified the respondent population (= all case managers working on social assistance eligibility files in the 79 selected municipalities) by three characteristics, namely age (four groups), gender and position (two groups each), resulting in 16 strata. One of these strata was absent in both the population and the sample, namely male team managers younger than 30. Most case managers in the respondent population were female, aged between 30 and 40 years, and did not hold a management position. In Model 6, we treated this group as the reference group and entered a dummy variable indicating the 14 other strata. Furthermore, we entered an interaction term combining these strata with the client parenthood variable. If researchers were to rely either on voluntary cooperation or on non-stratified sampling, it is very likely that

some of these strata would be absent in the actual sample. We show in Column 5 of Table 2 that this would bias the results. In the table, in the interest of space, we present only the significant interaction terms. The male respondents and the respondents who were one decade older than the reference group differ significantly from the female reference group: they predicted less sanctioning. The bottom part of Table 2 shows several additional significant interaction terms indicating that the parenthood variable has an impact that differs across the strata. If a researcher were to miss these specific groups of case managers, due to incomplete sampling methods, the results would be biased. Because these additional interactions are positive, the effect of having sick children would be overestimated.

5 Conclusion

In this paper, we argue that the factorial survey approach is a suitable method for studying social policy implementation. As this implementation is affected by the country, region, organisation, social worker and client, as well as by the cultural, political, economic, demographic and social environments, it needs to be studied as a multidimensional topic. Existing research suffers from a lack of generalisability, mostly due to a qualitative or single-country approach. Large-scale research, mostly conducted using administrative data, suffers from a selection bias. In reality, clients are not distributed randomly over municipalities, which makes it difficult to detect what might affect treatment: the client's characteristics or the agency where the client is assisted. The factorial survey may help overcome these problems by disseminating vignettes across a large number of respondents in multiple agencies, municipalities and various countries, using modern, inexpensive communication channels. Furthermore, the D-efficient sampling of the vignette universe and assignment tot the respondents ensures that the analysis of client characteristics is impacted as little as possible by differences between respondents.

The proposed method for studying social policy implementation exhibits similarities with the recent approach to investigating professional judgment. The latter focuses on detecting the typical treatment of clients with certain characteristics, whereas we are also interested in the typical treatment of clients in specific types of municipalities or by respondents with certain characteristics. This means that, for our kind of study, a questionnaire about the respondent and the organisation she works for is necessary, in addition to a set of vignettes to rate. As shown in the example concerning the likelihood of sanctioning social assistance clients who refuse a job offer starting early in the morning, not only the client's parenthood itself mattered to explain the reduced likelihood of being sanctioned in the event the client had children. When entering respondent characteristics such as the case manager's parenthood or her opinion on sanctioning in general, these characteristics proved to be significant and provided valuable information about the rationale behind the implementation of sanction measures.

Furthermore, the sampling of respondents is of major importance. Respondents should be distributed evenly over municipalities with specific characteristics and selected at random or

stratified from the group of employees in each municipality. In our example, we show that if researchers focus on large municipalities or on the voluntary contribution of the case managers, it is very likely that the results will be biased. Some large Flemish municipalities displayed very specific results, concerning both sanctioning in general and sanctioning clients who do or do not have children. Furthermore, for the group of case managers that is overrepresented in the population (female case managers aged between 30 and 40 years), the effect of having children on the likelihood of being sanctioned is substantially larger than for other groups of case managers. With non-random or non-stratified sampling, some of these other small groups would not appear in the sample, which would bias the results.

A final recommendation formulated in this paper was to use multi-level regression methods. Researchers studying social policy implementation should take into account the clustered nature of the data collected by surveying case managers (vignettes nested within respondents) from several agencies (respondents clustered into agencies). Multi-level methods allow for explanatory variables from different levels to be included simultaneously in the regression model: client characteristics (= vignette dimensions), respondent characteristics and agency characteristics. Some additional results can be obtained from these analyses. We showed that the large variation in the responses could be explained mainly by differences between case managers from the same municipality and between the different clients rated by a given case manager. Entering the dummy variables from some municipalities showed that certain large municipalities displayed particular results (see the previous paragraph), but, overall, differences between municipalities (79) were not decisive in explaining variation. This is a valuable result as it contradicts the literature on decentralisation, which expects to see large differences among local organisations, and confirms – quantitatively – the street level bureaucracy research emphasizing the extensive policy-making power of local case managers. Second, we were able to calculate the explanatory effect of specific fixed effects on the unexplained variance. We saw that the client variable parenthood reduced the unexplained variance by 5%, while the respondent variables entered led to a decrease of 4% in unexplained variance. This means that the respondent variables have an almost equal effect on the final results as the client characteristics. Third, by entering random slopes in the model, we showed that respondents not only differ significantly in their general predictions of sanctioning, but also in their predictions of the effect of parenthood on the likelihood of being sanctioned. The variation among respondents in predicting sanctions was higher when evaluating clients with children.

6 Appendix

This appendix shows an example of a vignette, the questions (dependent variables) asked about that vignette and the different categories per attribute / characteristic in the vignette. The vignettes are experimentally varied, with one category from each attribute in each vignette, in order to produce unique vignettes.

Vignette (The words or words groups between brackets are the attributes that change from vignette to vignette):

> X is a 22-year-old single [woman]. She has no income or savings and no debts.
>
> X [and her parents were born in Belgium]. She [speaks Dutch well]. She has [secondary vocational education qualifications].
>
> X [lives in an apartment in your municipality which she rents at a reduced price because of a government subsidy. She has enough space]. She has no contact with her parents and says there isn't really anyone she can count on. She has [no children].
>
> X has no physical health problems [and no known mental problems]. She has [no known addiction problems]. [X's childhood was marked by violence, abuse and poverty.] [She has spent two months in prison for repeated shoplifting.]
>
> X [wants to work]. She [does not yet have any work experience]. She [does not yet have any experience with activation projects]. She [has thus far been very conscientious about keeping appointments with your agency].

Question 1: What is the likelihood that X would receive social assistance in the office where you are working? (1-7 scale)

Question 2: What is the likelihood – in the office where you are working – that X would lose social assistance if refusing an activation or job offer that started at 5 o'clock in the morning? (1-7 scale)

Categories of the 14 attributes included in the vignettes:

Attribute	Category1	Category2	Category2
Gender	Female	Male	
Nationality background	Client and parents born in Belgium	Client born in Belgium – parents born in Morocco	Syrian refugee – one year in Belgium. Registered
Command of the national language	Good	Limited	Poor
Level of education	Primary education	Secondary education	Bachelor's degree
Parenthood	No children	One healthy child aged two years	One sick child aged two years (immune disorder)
Housing situation	Stable housing in a subsidized apartment	Lives with a friend – lots of quarrels	Homeless – staying with various friends
Mental health	Good	Depressed – no treatment	Lack of intellectual ability – not diagnosed by a doctor
Drug addiction	None	Slightly addicted – refuses treatment	Severely addicted – refuses treatment
Adverse life experiences	*No information included*	Youth with violence and abuse	Lost a baby one year ago
Alternative ways of earning money	*No information included*	Been in prison for shoplifting	Has been prostituting her/himself
Aspirations	Wants to work	Wants to study	Is not motivated to work or study
Employment experience	No experience	Positive experience – made redundant	Negative experience – quarrels with superior
Activation experience	No experience	One negative experience – not diligent	Several negative experiences – not diligent
Attitude	Diligent	Missed one appointment	Missed several appointments

7 Bibliography

Aguinis, H., & 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., & 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.

Armacost, R. L., Hosseini, J. C., Morris, S. A., & Rehbein, K. A. (1991). An Empirical Comparison of Direct Questioning, Scenario, and Randomized Response Methods for Obtaining Sensitive Business Information. *Decision Sciences*, *22*(5), 1073–1090.

Atzmüller, C., & Steiner, P. M. (2010). Experimental Vignette Studies in Survey Research. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *6*(3), 128–138.

Auspurg, K., Hinz, T., Liebig, S., & Sauer, C. (2015). The Factorial Survey as a Method for Measuring Sensitive Issues. In *Improving Survey Methods. Lessons from recent Research*.

Auspurg, K., & Jäckle, A. (2015). First Equals Most Important? Order Effects in Vignette-Based Measurement. *Sociological Methods & Research*.

Bargain, O., Immervoll, H., & Viitamäki, H. (2012). No claim, no pain. Measuring the non-take-up of social assistance using register data. *Journal of Economic Inequality*, *10*(3), 375–395.

Belfius. (2007). Lokale financiën. Sociaaleconomische typologie van de gemeenten. Belfius.

Benz, A. (2000). Two types of multi-level governance: Intergovernmental relations in German and EU regional policy. *Regional & Federal Studies*, *10*(3), 21–44.

Blommesteijn, M., van Geuns, R. C., Groenewoud, M., & Slotboom, S. T. (2012). *Vakkundig aan de slag. Een onderzoek naar vakmanschap in de gemeentelijke re-integratiesector* (p. 74). Amsterdam: Regioplan.

Bryman, A. (2012). *Social research methods*. Oxford; New York: Oxford University Press.

Carpentier, S. (2016). *Lost in Transition? Essays on Socio-Economic Trajectories of Social Assistance Beneficiaries.* Universiteit Antwerpen, Antwerpen.

Carpentier, S., Neels, K., & Van den Bosch, K. (2014). How Do Exit Rates from Social Assistance Benefit in Belgium Vary with Individual and Local Agency Characteristics? In S. Carcillo, H. Immervoll, S. Jenkins P., S. Königs, & K. Tatsiramos (Eds.), *Safety Nets and Benefit Dependence* (pp. 151–187). Emerald Group Publishing Limited.

Caussade, S., Ortúzar, J. de D., Rizzi, L. I., & Hensher, D. A. (2005). Assessing the influence of design dimensions on stated choice experiment estimates. *Transportation Research Part B: Methodological*, *39*(7), 621–640.

Champion, C., & Bonoli, G. (2011). Institutional fragmentation and coordination initiatives in western European welfare states. *Journal of European Social Policy*, *21*(4), 323–334.

De Deken, J., & Kittel, B. (2007). Social expenditure under scrutiny: the problems of using aggregate spending data for assessing welfare state dynamics. In J. Clasen & N. A. Siegel, *Investigating Welfare State Change: The 'Dependent Variable Problem' in Comparative Analysis*. Cheltenham, Northampton: Edward Elgar.

DeShazo, J. R., & Fermo, G. (2002). Designing Choice Sets for Stated Preference Methods: The Effects of Complexity on Choice Consistency. *Journal of Environmental Economics and Management*, *44*(1), 123–143.

Dülmer, H. (2007). Experimental Plans in Factorial Surveys Random or Quota Design? *Sociological Methods & Research*, *35*(3), 382–409.

Dülmer, H. (2016). The Factorial Survey: Design Selection and its Impact on Reliability and Internal Validity. *Sociological Methods & Research*, *45*(2), 304–347.

Eardley, T., Bradshaw, J., Ditch, J., Gough, I., & Whiteford. (1996). *Social Assistance in OECD Countries: Synthesis Report, Department of Social Security Research Report* (No. 46). London: HMSO.

Eifler, S. (2010). Validity of a Factorial Survey Approach to the Analysis of Criminal Behavior. *Methodology: European Journal of Research Methods for the Behavioral and Social Sciences*, *6*(3), 139–146.

Evans, P. M. (2007). (Not) taking account of precarious employment: Workfare policies and lone mothers in Ontario and the UK. *Soc Policy Admin*, *41*(1), 29–49.

Evans, T., & Harris, J. (2004). Street-Level Bureaucracy, Social Work and the (Exaggerated) Death of Discretion. *British Journal of Social Work*, *34*(6), 871–895.

Goos, P., & Jones, B. (2011). *Optimal Design of Experiments: A Case Study Approach* (1st ed.). West-Sussex: Wiley.

Groves, R. M., Jr, F. J. F., Couper, M. P., Lepkowski, J. M., Singer, E., & Tourangeau, R. (2009). *Survey Methodology*. John Wiley & Sons.

Heeringa, S., & Berglund, P. A. (2010). *Applied survey data analysis*. Boca Raton, FL: Chapman & Hall/CRC.

Hermans, K. (2005). *De actieve welvaartsstaat in werking. Een sociologische studie naar de implementatie van het activeringsbeleid op de werkvloer van de Vlaamse OCMW's.* KULeuven, Leuven.

Hölsch, K., & Kraus, M. (2006). European schemes of social assistance: an empirical analysis of set-ups and distributive impacts. *International Journal of Social Welfare*, *15*(1), 50–62.

Hooghe, L. (1996). *Cohesion Policy and European Integration*. Oxford: Oxford University Press.

Hooghe, L., & Marks, G. (2003). Unraveling the central state, but how? Types of multi-level governance. *American Political Science Review*, *97*(2), 233–243.

Jasso, G. (2006). Factorial Survey Methods for Studying Beliefs and Judgments. *Sociological Methods & Research*, *34*(3), 334–423.

Johnson, F. R. (2006). Comment on 'Revealing Differences in Willingness to Pay Due to the Dimensionality of Stated Choice Designs: An Initial Assessment'. *Environmental and Resource Economics*, *34*(1), 45–50.

Karren, R. J., & 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.

Kazepov, Y. (1999). At the Edge of Longitudinal Analysis. Welfare Institutions and Social Assistance Dynamics, *33*(3), 305–322.

Kazepov, Y. (2010). *Rescaling Social Policies: Towards Multilevel Governance in Europe*. Ashgate Publishing, Ltd.

Kazepov, Y., & Barberis, E. (2012). Social Assistance Governance in Europe: Towards a Multilevel Perspective. In I. Marx & K. Nelson, *Minimum Income Protection in Flux* (pp. 217–248).

Kirwan, J. R., Chaput de Saintonge, D. M., Joyce, C. R., & Currey, H. L. (1983). Clinical judgment in rheumatoid arthritis. I. Rheumatologists' opinions and the development of 'paper patients'. *Annals of the Rheumatic Diseases*, *42*(6), 644–647.

Kittel, B., & Obinger, H. (2003). Political parties, institutions, and the dynamics of social expenditure in times of austerity. *Journal of European Public Policy*, *10*(1), 20–45.

Langley, G. R., Tritchler, D. L., Llewellyn-Thomas, H. A., & Till, J. E. (1991). Use of written cases to study factors associated with regional variations in referral rates. *Journal of Clinical Epidemiology*, *44*(4–5), 391–402.

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

Lødemel, I., & Trickey, H. (2001). *'An Offer You Can't Refuse': Workfare in International Perspective*. Bristol: The Policy Press.

Lorentzen, T., Dahl, E., & Harsløf, I. (2012). Welfare risks in early adulthood: A longitudinal analysis of social assistance transitions in Norway. *International Journal of Social Welfare*, *21*(4), 408–421.

Marchal, S., Marx, I., & Van Mechelen, N. (2014). The Great Wake-Up Call? Social Citizenship and Minimum Income Provisions in Europe in Times of Crisis. *Journal of Social Policy*, *43*(02), 247–267.

Marchal, S., & van Mechelen, N. (2017). A New Kid in Town? Active Inclusion Elements in European Minimum Income Schemes. *Social Policy & Administration*, *51*(1), 171–194.

Minas, R., Whrigth, S., & Van Berkel, R. (2012). Decentralization and centralization: Governing the activation of social assistance recipients in Europe. *International Journal of Sociology and Social Policy*, *32*(5/6), 286–298.

Mohan, D., Fischhoff, B., Farris, C., Switzer, G. E., Rosengart, M. R., Yealy, D. M., … Barnato, A. E. (2014). Validating a Vignette-Based Instrument to Study Physician Decision Making in Trauma Triage. *Medical Decision Making*, *34*(2), 242–252.

Morley, C. P. (2010). The effects of patient characteristics on ADHD diagnosis and treatment: a factorial study of family physicians. *BMC Family Practice*, *11*(1), 11.

Nybom, J. (2013). Activation and 'coercion' among Swedish social assistance claimants with different work barriers and socio-demographic characteristics: What is the logic? *Int. J. Soc. Welf.*, *22*(1), 45–57.

Oorschot, W. J. H. van, Uunk, W. J. G., & Jeene, M. D. (2008). Who should get what and why, under which conditions: Descriptions and explanations of public deservingness opinions.

Ortega, E. G., Baz, B. O., & Sánchez, F. L. (2012). Professionals' Criteria for Detecting and Reporting Child Sexual Abuse. *The Spanish Journal of Psychology*, *15*(03), 1325–1338.

Peabody, J. W., Luck, J., Glassman, P., Dresselhaus, T. R., & Lee, M. (2000). Comparison of vignettes, standardized patients, and chart abstraction: a prospective validation study of 3 methods for measuring quality. *JAMA: The Journal of the American Medical Association*, *283*(13), 1715–1722.

Piattoni, S. (2009). Multi-level Governance: a Historical and Conceptual Analysis. *Journal of European Integration*, *31*(2), 163–180.

Priem, R. L., Walters, B. A., & Li, S. (2011). Decisions, Decisions! How Judgment Policy Studies Can Integrate Macro and Micro Domains in Management Research. *Journal of Management*, *37*(2), 553–580.

Rice, D. (2012). Street-Level Bureaucrats and the Welfare State: Toward a Micro-Institutionalist Theory of Policy Implementation. *Administration & Society*, *45*(9), 1038–1062.

Saraceno, C. (2002). *Social Assistance Dynamics in Europe: National and Local Poverty Regimes*. Bristol: The Policy Press.

Skarlicki, D. P., & Turner, R. A. (2014). Unfairness begets unfairness: Victim derogation bias in employee ratings. *Organizational Behavior and Human Decision Processes*, *124*(1), 34–46.

Stephenson, P. (2013). Twenty years of multi-level governance: 'Where Does It Come From? What Is It? Where Is It Going?' *Journal of European Public Policy*, *20*(6), 817–837.

Stokes, J., & Schmidt, G. (2012). Child Protection Decision Making: A Factorial Analysis Using Case Vignettes. *Social Work*, *57*(1), 83–90.

Tabin, J.-P., & Perriard, A. (2016). Active social policies revisited by social workers. *European Journal of Social Work*, *19*(3–4), 441–454.

Taylor, B. J. (2006). Factorial Surveys: Using Vignettes to Study Professional Judgement. *British Journal of Social Work*, *36*(7), 1187–1207.

Thoren, K. H. (2008). *'Activation Policy in Action': A Street-level Study of Social Assistance in the Swedish Welfare State*. Växjö University, Växjö.

van Berkel, R. (2006). The decentralisation of social assistance in The Netherlands. *International Journal of Sociology and Social Policy*, *26*(1/2), 20–31.

Van Mechelen, N., Marchal, S., Goedemé, T., Marx, I., & Cantillon, B. (2011). The CSB-Minimum Income Protection Indicators dataset (CSB-MIPI). *CSB-Workingpaper*.

Wallander, L. (2009). 25 years of factorial surveys in sociology: A review. *Social Science Research*, *38*(3), 505–520.

Wallander, L. (2012). Measuring social workers' judgements: Why and how to use the factorial survey approach in the study of professional judgements. *Journal of Social Work*, *12*(4), 364–384.

Webster, S. W., O'Toole, R., O'Toole, A. W., & Lucal, B. (2005). Overreporting and underreporting of child abuse: Teachers' use of professional discretion. *Child Abuse & Neglect*, *29*(11), 1281–1296.

[i] In this paper we understand the term 'discretion' as 'individual' or 'locally based' decision making. Such discretion can be studied in two different ways. First, it can be studied as the room that is allowed to lower-level actors (e.g. case managers) by higher management levels (e.g. managers). This sort of discretion is called *de jura* discretion by Evans (2012). However, the factorial survey only enables us to test the extent to which case managers differ in their treatment proposals and thus use their discretion, which is *de facto* discretion (T. Evans, 2012).

[ii] The term 'interaction effect' refers to a situation in which the impact of one factor depends on the level of one or more other factors. For example, overall, offensive behaviour may have a minor impact on the frequency of support. However, if we study the interaction between offensive behaviour and gender, it might be that women who exhibit offensive behaviour are supported as frequently as other women, but that men who exhibit offensive behaviour are supported significantly less frequently than other men. Thus, offensive behaviour has an impact on support treatment, but only among men.

[iii] In total, there were 400 unique questionnaires (decks), spread over almost 600 case managers. Each deck was rated at least once.

[iv] To test this claim, we introduced two variables that captured whether a respondent rated a deck containing an extreme case. An extreme case was a vignette with four of the characteristics that might make life difficult: having a child or a sick child, having mental problems, having a severe addiction or having a difficult life trajectory. Other extreme cases were vignettes with four characteristics linked to attitude problems: not motivated to work, negative work experience, negative activation experience and often too late at appointments. Inclusion of two dummies, representing a deck with an extreme case, revealed that the dummy for extremely difficult life circumstances had a positive effect on sanctioning. This seems to suggest that respondents who had to evaluate a deck with extreme cases (maximum one for each respondent) are, relative to their colleagues, more likely to sanction clients with a less difficult life.