Journal of Methods and Measurement in the Social Sciences,
Vol. 7, No.2, 52-94, 2016

Designing Valid and Reliable Vignette Experiments for Survey Research: A Case Study on the Fair Gender Income Gap

Peter M. Steiner
University of Wisconsin-Madison

Christiane Atzmüller
University of Vienna, Austria

Dan Su
University of Wisconsin-Madison

In survey research, vignette experiments typically employ short, systematically varied descriptions of situations or persons (called vignettes) to elicit the beliefs, attitudes, or behaviors of respondents with respect to the presented scenarios. Using a case study on the fair gender income gap in Austria, we discuss how different design elements can be used to increase a vignette experiment's validity and reliability. With respect to the experimental design, the design elements considered include a confounded factorial design, a between-subjects factor, anchoring vignettes, and blocking by respondent strata and interviewers. The design elements for the sampling and survey design consist of stratification, covariate measurements, and the systematic assignment of vignette sets to respondents and interviewers. Moreover, the vignettes' construct validity is empirically validated with respect to the real gender income gap in Austria. We demonstrate how a broad range of design elements can successfully increase a vignette study's validity and reliability.

Keywords: Vignette experiment, factorial survey, confounded factorial design, anchoring vignettes, ANOVA, multilevel modeling

Experiments in survey research have gained increasing attention over the last decades because the experiment's internal validity is augmented by the survey's external validity (Gaines, Kuklinski & Quirk, 2007; Schlüter & Schmidt, 2010; Sniderman & Grob, 1996). In particular, vignette experiments embedded in surveys—also called factorial surveys—are now becoming more popular, though they had been introduced to sociology by Peter Rossi more than five decades ago (Atzmüller & Steiner, 2010; Auspurg & Hinz, 2015; Dülmer, 2007; Jasso, 2006; Nock & Rossi, 1978; Rossi 1979; Rossi et al., 1974a; Rossi & Anderson; 1982; Sauer et al., 2011; Steiner & Atzmüller, 2006). A vignette experiment consists of a collection of vignettes, that is, a set of systematically varied descriptions of subjects, objects, or situations in order to elicit respondents' beliefs, attitudes, or intended behaviors with respect to the presented vignettes. The vignettes used in a vignette experiment are typically generated by factorially combining the levels of factors considered as relevant for the study.

In comparison to traditional survey questions, vignettes have several advantages. First, since vignettes are multivalent representations of subjects or situations, the corresponding questions are embedded in a concrete, realistic context. Thus, vignette questions are more realistic and less abstract than conventional survey questions. Second, the multivalent character of vignettes allows for a simultaneous investigation of the factors varied in the vignette experiment—interaction effects among vignette factors can be estimated and tested. Third, using an experimental design for the vignette experiment guarantees a high internal validity. Fourth, vignettes are very flexible, they can be used in different formats and for different purposes. For instance, Cook (1979) used text vignettes to investigate the willingness of Americans to support programs for social groups in need of aid; Atzmüller and Kromer (2013) investigated peer violence among adolescents using short video vignettes; Atzmüller and Kromer (2014) used audio vignettes mimicking radio news on crimes; and Lim (2013) explored vignettes in form of sketches (symbolic maps) in order to investigate why Laotian people choose grazing areas for their cattle close to tiger habitats. Fifth, due to the vignettes' flexibility they can be used as a projective technique for avoiding socially desirable or politically correct answers when dealing with sensitive topics. Sixth, respondents view vignettes frequently as a welcome relief from monotonous survey questions. All these advantages contribute to a vignette study's internal validity, construct validity, and reliability. However, the internal validity and reliability of vignette experiments can be increased considerably by using additional design elements like anchoring vignettes or blocking the vignette experiment by respondent strata and interviewers.

By *internal validity* we mean the validity of inferences about the cause-effect relationship between the presented vignette stimuli and respondents' reaction to the stimuli (Shadish, Cook & Campbell, 2002). Internal validity is established by experimental control which allows researchers to uniquely assess the vignette factors' causal effect on the outcome variable—that is, the effect estimates are free of any bias. However, the specifics of a concrete vignette experiment restrict the *external validity* of inferences drawn from the experiment. The very specific choice of vignette factors and factor levels, mode of presentation, vignette questions, and selection of respondents does not warrant inferences to other factors and factor levels, to different settings and outcome measures, and to a different sample of respondents. However, embedding the vignette experiment in a survey (with random sampling) extends a vignette experiment's external validity at least to the survey's target population. Though, the realistic, multivalent character of vignettes might restrict the external validity, it is frequently assumed that highly contextualized vignettes increase the *construct validity*, that is, the degree to which the vignettes measure what we intend to measure. Finally, by a vignette study's *reliability* we refer to the study design's ability to control for *measurement error*, *experimental error*, and *sampling error* (Hinkelmann & Kempthorne, 1994). A reliable vignette study is characterized by reliable vignette measurements, a balanced and blocked experimental design, and a stratified respondent sample of sufficient size. Thus, a reliable study design results in precise (i.e., efficient) effect estimates and sufficiently powered hypothesis tests which is important for securing a high *statistical conclusion validity*.

In this article, we use a case study on the fair gender income gap for discussing the rationale and implementation of several design elements for strengthening a vignette experiment's validity and reliability. The goal of the case study was to assess whether Austrians think that female and male employees should receive the same or a different income for doing the same job (Steiner, Atzmüller & Wroblewski, 2009). Directly asking respondents whether female employees should earn the same income as male employees would have most likely resulted in answers affected by social norms (i.e., there should be no gender discrimination in the labor market: female and male employees with identical occupations, occupational experience, and education should get the same income). Thus, we used a vignette experiment in order obtain assessments of the fair income for virtual female and male employees and, consequently, an estimate of the fair gender gap.[1] In order to rule out politically correct answers we implemented employees' gender as a between-subjects factor in the vignette experiment, that is, respondents either received vignettes of female employees or male employees—making it impossible to guess the goal of the study and to adjust answers according to social norms. Another advantage of using vignettes is that we can directly estimate the magnitude of the fair gender income gap, which is almost impossible with traditional survey questions. Simultaneously, we can also assess how strongly the other vignette factors (education, occupational experience, industry, parental leave) affect fair income. In addition to the fair income, respondents also had to assess the actual income of the virtual employees (i.e., the real income that a corresponding Austrian employee actually gets for the described job). This allowed us to probe the vignette experiment's construct validity by comparing the actual gender gap estimated from the

[1] In letting respondents directly assign a fair income to virtual employees, our study differs from other studies on the fair gender pay gap which let respondents judge whether a certain income is fair or unfair (e.g., Jasso & Meyersson Milgrom, 2008; Jasso & Webster, 1997, 1999; Sauer, 2014). They then analyze the justice evaluations instead of the underlying incomes (which represent systematically varied levels of the income factor). For a discussion of the comparative advantages of indirect and direct income assessments, see Jasso (2012) and Markovsky & Eriksson (2012).

vignette experiment with the real gender income gap estimated from Austrian register data. To obtain more reliable measurements of the fair income and to control for potential biases, we introduced respondent-specific anchoring vignettes. That is, each respondent was required to fill in a blank vignette according to his own factor attributes and to assign himself a fair income. Then, the fair incomes for the other vignettes had to be assigned with reference to the respondent's own fair income.

Using this case study, we highlight the importance of different design elements for increasing a vignette experiment's internal validity, construct validity, external validity and reliability (statistical conclusion validity). The design elements which we address in detail are the following:

(1) *Confounded factorial design for the vignette experiment:* Controls the experimental error and avoids a confounding of main and two-way interaction effects with set effects.
(2) *Gender as between-subjects factor:* Rules out biases due to socially desirable or politically correct answer behaviors.
(3) *Ranking & rating of vignettes*. Helps in obtaining logically consistent and more reliable assessments of the fair income.
(4) *Anchoring respondent vignettes:* Aim at increasing the reliability of assessments and controls for a potential confounding of the between-subjects factor (i.e., the gender income gap).
(5) *Covariate measurements*: Helps in reducing the error variance and explaining effect heterogeneities.
(6) *Pilot studies and power analysis*: Helps in increasing the measurements' validity and reliability and in determining the required sample size.
(7) *Stratified respondent sample*: Reduces the sampling error and helps in ensuring external validity.
(8) *Blocking by respondent strata and interviewers*: Reduces the experimental error and increases the experimental design's efficiency.
(9) *Systematic assignment of vignette sets to respondent strata and interviewers*: Rules out random and systematic stratum and interviewer effects in the estimated gender income gap.
(10) *Empirical validation of vignettes based on the actual income*. Allows for probing the construct validity of vignettes.

We will argue that together all these design elements considerably increase the validity and reliability of the vignette experiment. However, in order to take full advantage of those design elements they need to be accurately reflected in the statistical models for analyzing the data. Failing to do so might result in biased and less efficient effect estimates. We will briefly discuss two main models for analyzing vignette data: the analysis of variance (ANOVA) and multilevel modeling.

While many of these design elements are well known, vignette studies rarely take advantage of a broad set of design elements. They frequently rely on a few design elements only, with an emphasis either on the experimental design or on the survey design, but rarely on both designs together. Moreover, even when design elements like blocks, strata, and vignette sets are present, they are regularly not included in the analytic model resulting in incorrect standard errors and type I error rates. Thus, many vignette studies do not fully utilize the advantages of vignette experiments in survey research. The aim of this case study on the gender income gap, therefore, is to demonstrate the implementation of a broad range of design elements for increasing the study's validity and reliability and to show how the design elements are correctly accounted for in the statistical analysis. To the best of our knowledge, two design elements have never been discussed or implemented before: respondent-specific anchoring vignettes and the systematic assignment of vignette sets to respondent strata and interviewers.

In discussing the case study, we emphasize not only the strengths but also the weaknesses and limitations of our study. This is particularly important because the limitations directly restrict the validity, reliability, and generalizability of the conclusions drawn from the study. All too often publications fail to critically address the study's limitations (because the authors are frequently not aware of all plausible threats to validity). In discussing our case study, we demonstrate that valid and reliable inferences can only be made if the limitations are as clearly addressed as the strengths. This does not require that one reflects on *all* possible threats to validity (which is not even possible), but it requires the assessment of the *most plausible* threats and a discussion of how the study dealt with the threat.

It is also important to bear in mind that a vignette experiment‘s validity depends on the design's extent to guard against *potential* rather than actual threats (Shadish, Cook & Campbell, 2002). A study's validity cannot depend on how it deals with actually operating threats because one can rarely convincingly demonstrate which threats are present and which ones are absent. Thus, what matters is whether a study design is able to neutralize or detect a threat if the threat would be present. This can frequently be assessed on pure theoretical grounds and does not require an empirical evaluation (e.g., via a split ballot design). Thus, adding non-redundant design elements increases a vignette experiment's validity, provided that they do not have unintended side effects. For each design element of our case study, we will discuss which potential threats it addresses and how it rules out or detects the threat if present.

This article is organized as follows. We begin with a thorough description of the experimental vignette design and the ranking and rating tasks. Then we briefly describe the questionnaire that was administered after the vignette experiment. The section on the sampling design discusses the power analyses, the stratified respondent sampling, the blocking of the vignette experiment by respondent strata and interviewers, and the systematic assignment of vignette sets to strata and interviewers. The implementation section describes interviewer recruitment, procedural aspects of the vignette experiment, and some additional study characteristics. The section on the analytic strategies discusses the statistical methods used for analyzing the vignette data. In the results section we then discuss the findings from the vignette experiment. Finally, we present an empirical evaluation of our vignette experiment's construct validity and conclude with the discussion section.

Experimental Vignette Design

Confounded Factorial Design with a Between-Subjects Factor

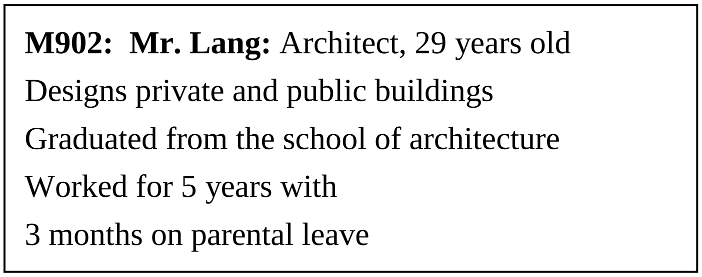
Choice of factors and factor levels. The first step in designing a vignette experiment is the choice of factors and factor levels that are systematically varied to produce a whole population of vignettes. We derived the most important factors that presumably determine an employee's income from Mincer's wage equation, which is one of the most widely used economic models (Lemieux, 2006). Mincer's wage equation states that the log hourly wage rate is determined as a linear combination of education (years of schooling), occupational experience in years, and the squared term of occupational experience. This equation is frequently estimated for subpopulations of female and male employees or for different occupations or industries in order to investigate wage discrimination and heterogeneity in labor markets (Altonji & Blank, 1999; Blau & Kahn, 1996, Blinder 1973, García, Hernándes & López-Nicolás, Oaxaca, 1973). Corresponding models have been estimated for Austria by Böheim, Hofer & Zulehner (2005) and Zweimüller & Winter-Ebmer (1994). In addition to Mincer's wage equation we also considered the results from vignette experiments on the fair income published by Alves & Rossi (1978), Jasso (1992), Jasso & Meyersson Milgrom (2004), and Jasso & Webster (1997, 1999). Based on Mincer's wage equation and the published vignette experiments, we then chose five factors to characterize a virtual employee in our vignette experiment: (a) Gender (G), (b) highest educational degree attained (E), (c) occupational experience in years (Y), (d) industry (I), and (e) parental leave in months (L). Except for gender, all factors have three levels (Table 1). To keep the total number of possible vignettes low, we restricted the number of factor levels to three. While the choice of levels for education, occupational experience and parental leave was driven by the corresponding distributions among Austrian employees, we chose the three industries so that one industry is dominated by female employees (health & care), one dominated by male employees (construction), and one industry is balanced with respect to the distribution of female and male employees (business-related services).

Table 1
Vignette Factors and Factor Levels

Factor	Factor levels	
Gender (G)	2 levels	male / female
Educational degree (E)	3 levels	apprenticeship training / high school / college
Occupational experience (Y)	3 levels	5 years / 20 years / 35 years
Industry (I)	3 levels	health & care / construction / business-related services
Parental leave (L)	3 levels	0 months / 3 months / 24 months

Vignettes & vignette population. The factorial combination of all five factors results in a population of $2\times3^4 = 162$ virtual employees which were presented to respondents via text vignettes. Figure 1 shows an example vignette of a 29-year-old male employee who works as an architect in the construction business, has a college degree, five years of occupational experience, and spent three months on parental leave.

Figure 1. Example of a male vignette (M902: male vignette set number 9, vignette number 02)



M902: Mr. Lang: Architect, 29 years old
Designs private and public buildings
Graduated from the school of architecture
Worked for 5 years with
3 months on parental leave

As the example vignette shows, we included the employee’s occupation (architect) instead of the industry (construction). Table 2 shows that we selected three occupations for each industry. We chose the occupations in accordance with the three educational levels apprenticeship training, high school, and college. For instance, for the construction industry we chose (a) draftsman because it typically requires an apprenticeship degree, (b) construction engineer because it requires a high school degree, and (c) architect because of its required college degree. Since the nine occupations in Table 2 reflect the nine industry×education combinations, occupation does not constitute its own factor. Nonetheless, we chose to include employees' occupations in the vignette descriptions because in every-day conversations people tend to relate variations in income to variations in occupations rather than the combination of industry and educational degree. The inclusion of employees' occupations creates a more realistic context and, thus, presumably results in more reliable assessments of the actual and fair incomes. For the same reasons we also included employees' age in addition to the vignette factors. We computed the age values by adding six years of preschool, the years of schooling required for obtaining the educational degree, the years of occupational experience, and the years of parental leave.

Table 2
Vignette occupations by educational degree and industry

Highest Educational Degree	Industry		
	Construction	Health & care	Business related services
Apprenticeship training	Draftsman	Elderly care nurse	Administrative assistant
High school	Construction engineer	Hospital nurse	Bank employee
College	Architect	Clinical psychologist	Tax accountant

The vignette in Figure 1 also indicates that we personalized the vignettes by using concrete names for the virtual employees (e.g., Mr. Lang). In order to avoid effects due to the variation in names we only used traditional Austrian names so that income effects are not confounded with potential effects of ethnic groups. Moreover, we only used nine different names that varied with the nine occupations but held them constant across variations of the other vignette factors (since each respondent had nine vignettes to judge, every respondent was exposed to the same set of names, though combined with different factor levels in occupational experience and parental leave). Only after implementing the study did we realize that the complete confounding of the names with the occupations could threaten the experiment's validity because some names (particularly those etymologically related to occupations) can evoke specific associations (Mutz, 2011). For instance, we used the very common names Weber and Schmidt, which connote weaver and smith. Though the confounding of occupations with names might have affected the measurement of the actual and just income, we believe that the effects are at best negligibly small because the occupational connotations of the chosen names were unrelated to the nine occupations of our vignette experiment.

Vignette sets. Estimating all main and interaction effects of the five vignette factors (up to the five-way interaction effects) would require that each respondent assesses all 162 vignettes. However, this is neither useful—because respondents would get tired of the repetitive assessment of 162 vignettes—nor required—because three-way and higher-order interaction effects are rarely of importance. Thus, it is advisable to systematically partition the entire vignette population into small and mutually exclusive sets such that all main effects and two-way interaction effects remain estimable and only a few of the higher-order interaction effects are confounded with the set effect. The set effect represents potential assessment differences across sets due to set-specific context effects, that is, the assessment of vignettes contained in a specific set might depend on the context created by these vignettes (Su & Steiner, 2016).

In order to create vignette sets, we first split the vignette population by gender to form two subpopulations of 81 vignettes each—one consisting of vignettes describing female employees (referred to as female vignettes) and the other consisting of male vignettes. This splitting makes gender a between-subjects factor and is discussed in more detail below. Then, using a randomized block confounded factorial design (RBCF-3^4, Kirk, 1995; Atzmüller & Steiner, 2010; Steiner & Atzmüller, 2006; Wu & Hamada, 2009), we partitioned both subpopulations into nine sets of nine vignettes each. We chose a set size of nine for three reasons: First, nine vignettes can be judged by each respondent without getting tired or frustrated over the repetitive task. Second, respondents can compare and assess the nine vignettes simultaneously instead of sequentially—they can put the vignettes next to each other and then rank and rate them. This helps to avoid sequence and carry over effects and, thus, increases the reliability of vignette measurements. Third, the systematic confounding according to the RBCF-3^4 design still allows us to estimate all main and two-way interaction effects without any confounding. Using a set size of nine vignettes, we partially confounded the three-way interaction effects I×E×Y, I×E×L, I×Y×L, and E×Y×L with the set effect and the corresponding four-way interactions effects I×E×Y×G, I×E×L×G, I×Y×L×G, and E×Y×L×G with the interaction effect of set and gender. All other four- and five-way interaction effects remain unconfounded (but they are estimated across sets because they are not completely contained in each set). With less than nine vignettes per set (i.e., a larger number of sets) some of the two-way interaction effects would have necessarily been confounded with the set effect. Keeping two-way interaction effects free of any confounding was crucial to our study because we expected several two-way interaction effects to be important.

Splitting the overall vignette population into female and male subpopulations resulted in vignette sets that either contained only female vignettes or male vignettes. Consequently, a single respondent never got female and male vignettes simultaneously, implying that the vignette factor gender is a *between-subjects factor* while all other vignette factors are within-subjects factors. We intentionally created gender-specific sets because we aimed at avoiding socially desirable or politically correct answers with regard to the question about the fair income (i.e., whether female and male employee's with identical vignette factors should earn the same income). If respondents would have gotten both female and male vignettes, they most likely would have guessed the goal of the study and then adjusted their answers according to social norms (i.e., that female and male employees in the same job should get equally paid).

There are three threats to the internal validity and reliability of a between-subjects design. First, differential effects due to different set partitions of the female and male vignette population could bias the gender income gap. Second, bias in the gender income gap might also result if the respondents assessing female vignettes are in some characteristics different to respondents assessing male vignettes. This aspect needs particular attention when respondents are neither randomly selected nor randomly assigned to vignette sets but deliberately selected by interviewers as it was the case in our study. Third, the power for testing the gender income gap drastically diminishes.

In our study we addressed these threats as follows. In order to avoid a biased gender income gap due to differential set effects across the two subpopulations of female and male vignette sets, we used the same partitions for the female and male vignette populations (i.e., the same confounding contrast for the RBCF-3^4 design). Thus, for each female vignette set exists an equivalent male vignette set containing exactly the same vignettes (i.e., the same factor level combinations, except for gender, of course). Potential set effects are then most likely identical for female and male vignette sets and, thus, do not confound the gender income gap. We tried to circumvent the second and third validity threat (bias due to respondent differences and diminished power) by (a) systematically assigning vignette sets to blocks of interviewers and respondent strata, (b) introducing respondent-specific anchoring vignettes, and (c) measuring respondent characteristics. All these design features are described in more detail below.

It is important to note that we used a confounded factorial design for partitioning the vignette population. Alternatively, we could have used a random selection strategy for creating vignette sets (Jasso, 2006; Rossi & Anderson, 1982). Though randomly generating vignette sets is easier to implement than a confounded factorial design, it might result in random confounding and a loss in efficiency (Atzmüller & Steiner, 2010; Su &

Steiner, 2016). However, with more complex mixed designs that include many more factors than our vignette experiment and that have different numbers of factor levels, one can employ a D-, A- or I-optimal design (Montgomery, 2013; Dülmer, 2007; Su & Steiner, 2016). Such designs are generated by software packages like *jmp* from SAS (SAS Institute Inc., 2012) or *AlgDesign* in R (Wheeler, 2014).

Ranking and Rating Tasks, Anchoring Respondent Vignettes, and Vignette Layout

Given our interest in the actual and fair income of employees, respondents were required to report the actual and fair income for each vignette of the set. In a first step, respondents had to assess the actual incomes of the virtual employees (without respondent-specific anchoring vignettes). In a second step, they had to assess the fair incomes with reference to the fair income of their own vignette.

Ranking and rating with respect to the actual income. The nine vignettes of a set had to be first ranked according to respondents' beliefs about the employees' *actual* income (i.e., the real income of a corresponding Austrian employee) and then rated by assigning actual monthly net incomes (in €) to the vignettes. In letting respondents first rank and then rate the vignettes, we intended to make the income assessments less prone to order effects and logical inconsistencies (i.e., an inconsistent assignment would occur if a respondent assigns a higher income to employee A than employee B, though he would rank B before A). Thus, the preceding ranking aimed at increasing the reliability of vignette measurements. The simultaneous presentation and ranking of vignettes also helps in mitigating order effects (Su & Steiner, 2016).

Creation of respondent-specific anchoring vignette. After the ranking and rating of vignettes with respect to the actual income, respondents had to create their own vignette by filling in the blank vignette shown in Figure 2. Upon completion of the blank vignette, respondents were asked to report their own income. By including the respondent's own vignette in the vignette experiment we obtained almost complete data for respondents' actual income and the other variables on the anchoring vignette. Importantly, the respondent-specific vignette then served as a reference for ranking and rating the vignettes with respect to the fair income.

Ranking and rating with respect to the fair income. As for the actual income, we required respondents to rank the vignettes according to the fair income, that is, the income they considered as being fair for the employees described in the vignettes. However, the ranking and rating of vignettes had to be done in relation to the respondents' own vignette. That is, the respondent's own vignette became a part of the vignette experiment by serving as an anchoring vignette for assessing the fair income.

Figure 2. Respondent vignette

B110: Occupation: ……………………………..
Age: …………………… Years
Education: ………………………………..
Employed/working for …………… years
Parental leave: ……………… months

Introducing respondent-specific anchoring vignettes made the assessment of the fair income more realistic and reliable because people tend to assess the fairness of a third person's income in comparison to their own income. Without the use of respondent-specific anchoring vignettes, the respondent-level variance might be larger because (a) a respondent's fair income level may depend on the specificities of the vignette judged first (i.e., order effects), and (b) respondents may use absurdly high or low incomes more frequently. We observed the latter in our pilot studies which did not use anchoring vignettes.

In planning our study, we also considered anchoring vignettes that do not depend on individual respondent characteristics. For instance, we thought about using a single anchoring vignette describing a unique employee for all respondents. Such an anchoring vignette produces for each respondent a reference measurement on a stimulus that is identical across all respondents. Though a single unique anchoring vignette has some advantages over the respondent-specific anchoring vignettes (Grol-Prokopczyk, 2014), we would have needed a female anchoring vignette for respondents judging female vignettes and a male anchoring vignette for respondent judging male vignettes. But this would have not only anchored the fair incomes but also the fair income gap on the two specific anchoring vignettes (i.e., the fair income gap would have been strongly predetermined by the choice of the female and male anchoring vignettes rather than the experimentally varied vignettes).[2]

[2] Note that we introduced the anchoring vignettes primarily for reducing random measurement error rather than for making respondents' fair income judgments fully comparable as classical anchoring vignettes try to do (Grol-Prokopczyk, 2014; King et al., 2004; King & Wand, 2007). For our research question, we did not need fully comparable fair income judgments because we were interested only in the relative differences (in percent) between income ratings rather than the absolute values of the fair incomes (see the analysis section). By using respondent-specific anchoring vignettes we allowed each respondent to choose his very own level of the fair income scale (i.e., the fair income that the respondent accepts for his own job) which then served as a reference for all other

Key-word vignettes. We aimed at increasing the reliability of vignette measurements also by using key-words printed on small cards such that nine vignettes could be easily arranged by the respondent on a table or a ranking cardboard with vignette holders (in case no table was available when conducting the interview). In using small key-word vignettes, it was easy for respondents to quickly grasp the differences between the portrayed employees and to bring the vignettes in the desired rank order. A ranking sheet was used by the interviewer to first record the rank order of vignettes and then the income assigned to each vignette. Though an electronic presentation of vignettes and an electronic questionnaire would have been more convenient for collecting the data, the use of physical vignette cards provided the respondents greater flexibility in sorting and re-ordering the nine vignettes. The increased flexibility presumably contributed to the reliability of vignette measurements.

Questionnaire & Covariate Measurements

After the vignette experiment, which was conducted as a face-to-face interview, respondents had to fill in a questionnaire with 15 questions consisting of a total of 45 questionnaire items. The questionnaire covered standard survey questions on actual and fair income issues (38 items) and on sociodemographic characteristics (7 items, including citizenship, number children, working hours, or industry of occupation). In addition, interviewers were instructed to ask questions about respondents' income, age, and occupation in case they did not completely fill in their respondent vignette or did not state their actual income in the vignette experiment. Further respondent information came from the beginning of the interview where interviewers asked for respondents' sex, age, employment status and educational degree. The collection of (almost) complete sociodemographic measures is important for the analysis of vignette data because they help in investigating heterogeneous response behaviors and in reducing the error variance at the respondent level. This was of particular importance to our study because we designed the gender income gap as a between-subjects factor. Overall, we collected measures of

income assessments. Thus, we aimed at a high response consistency between the anchoring vignette and the vignettes in the set rather than the equivalence of anchoring vignettes. Using a single male (or female) anchoring vignette would not have worked because respondents judging female (male) vignettes could have adjusted their fair income assessments according to social norms. Using a male anchoring vignettes for respondents judging male vignettes and a corresponding female vignette for respondents judging female vignettes would have anchored the fair gender income gap at the specificities of the anchoring vignette (different choices of anchoring vignettes may result in different fair gender income gaps).

15 sociodemographic variables (including respondents' actual and just income).

Sampling Design

Pilot Studies

We conducted two small pilot studies with 34 and 49 respondents in order to test the practicability of our vignette experiment and to obtain empirical data for determining the required sample size in a power simulation. For the fair income, we aimed at a minimum detectable gender gap of 3% (i.e., the difference in female and male incomes expressed as percentage of the male income). We implemented the power analysis as a simulation study with a multilevel data-generating process and an analytic outcome model which both reflected the experimental vignette design—but without the anchoring vignettes. The data-generating model was based on parameter settings we obtained from the two pilot studies. The power simulations indicated that with nine vignettes per set we would have needed 1,300 respondents to obtain a power of 0.8 for inferring a significant gender gap of 3% in the fair income (with a type-I error of .05). Not surprisingly, the required sample size is large because the gender income gap is estimated between respondents (between-subjects factor). However, due to budget constraints we were not able to interview more than 1,000 respondents. But with 1,000 respondents the minimum detectable gender gap would have been about 4%. Thus, in order to obtain the desired power of 0.8 for a minimum detectable income gap of 3%, we aimed at reducing the error variance in the actual and fair income measurements by (a) introducing anchoring respondent vignettes, (b) measuring respondent covariates as described above, (c) stratifying the respondent sample by sex and age, and (d) blocking the experiment by respondent strata and interviewers. The analysis of the main study suggests that the anchoring vignettes, and to a lesser extent the strata, blocks, and additional covariates, successfully reduced the error variance in the fair income (see the Results section). Variance reduction in the actual income was less successful but we had no power issues here to begin with (because of the larger gender gap).

Stratified Sampling of Respondents

Target population. The target population consisted of the work force in Vienna, Austria in 2008. We restricted the population to people between the ages of 18 and 65 years that were either employed, unemployed, or on parental or educational leave at the time of the interview. From this population, we drew a quota sample instead of a stratified random sample because we did not have access to a representative sampling frame that would have allowed us to draw a random sample. It is important to realize that the impossibility to draw a clean random sample restricts the external validity of our vignette experiment (at least on formal statistical grounds).

Quota sample. In order to obtain a sample of 1,000 respondents, interviewers could deliberately choose eligible respondents in Vienna within the prescribed quota for each respondent stratum. We determined the respondent strata according to the bivariate sex×age distribution of the target population in Vienna. We defined three age strata based on the terciles of the age distribution: 18-33 years, 34-44 years, and 45-65 year. Using age terciles guaranteed that we obtained a self-weighting sample with respect to the age distribution. Because the age terciles were almost identical for the female and male target population, the quota sample is approximately self-weighting also with respect to female and male subpopulations. We did not aim for a self-weighting sample with respect to the sex distribution because we were interested in judgment differences between female and male respondents. We sampled the same number of female and male respondents though about 40% of Vienna's work force is female and 60% male. We chose sex and age as stratum variables because we assumed that these variables are predictive of the actual and fair vignette incomes and, thus, would lead to more efficient estimates of the gender-income gaps.

Blocking by Respondent Strata and Interviewers, and Assignment of Vignette Sets to Interviewers and Respondents

In blocking the vignette experiment by respondent strata and interviewers we tried to balance the frequency with which each female and male vignette set is measured across interviewers and sampling strata. In doing so, we aimed at reducing the experimental error by eliminating random and systematic variations in the vignette measurements across interviewers and respondent strata. Since we were not able to completely block the experiment by interviewers (this would have required that each interviewer conducts 108 interviews), we implemented an incomplete interviewer block design by systematically assigning predefined packages of twelve vignette sets to each interviewer. Table 3 shows the assignment of vignette sets to sampling strata and interviewers. The first two columns represent the sex×age respondent strata. For each stratum, the double-row contains the nine female vignette sets (f1 to f9 in the first row) and the corresponding male vignette sets (m1 to m9 in the second row). Consider interviewer package #1 in the grey rectangular box in Table 3. This package of 12 vignette sets (f1, m1, f2, m2, f3, m3, f1, m1, f2, m2, f3, m3) is formed by three female (f1, f2, f3) and the corresponding three male sets (m1, m2, m3), where female and male vignette sets with the same number (e.g., f1 and m1) are identical except for vignette gender. Note that the package contains each female and male vignette set twice. This allowed us to present the female and corresponding male vignette sets (e.g., f1 and m1) to both female and male respondents in the same age stratum (18-33 years). Because the male and corresponding female vignette sets are presented by the *same interviewer* to respondents in the *same stratum*, potential interviewer effects or stratum effects with respect to the gender income gap are differenced out by design. And because we used the same set assignment for female and male respondent strata within an interviewer package, the comparison of gender income gaps across female and male respondents is unconfounded by any interviewer or age stratum effects and tested with maximum efficiency.

Table 3

Assignment of vignette sets to respondent strata and interviewers

Sample strata		Interviewer packages								
Sex	Age	#1	#2	#3	#4	#5	#6	#7	#8	#9
Female	18-33	f1 m1	f4 m4	f7 m7	f2 m2	f5 m5	f8 m8	f3 m3	f6 m6	f9 m9
Female	34-44	f2 m2	f5 m5	f8 m8	f3 m3	f6 m6	f9 m9	f4 m4	f7 m7	f1 m1
Female	45-65	f3 m3	f6 m6	f9 m9	f4 m4	f7 m7	f1 m1	f5 m5	f8 m8	f2 m2
Male	18-33	f1 m1	f4 m4	f7 m7	f2 m2	f5 m5	f8 m8	f3 m3	f6 m6	f9 m9
Male	34-44	f2 m2	f5 m5	f8 m8	f3 m3	f6 m6	f9 m9	f4 m4	f7 m7	f1 m1
Male	45-65	f3 m3	f6 m6	f9 m9	f4 m4	f7 m7	f1 m1	f5 m5	f8 m8	f2 m2

Note. The first column (grey box) shows the vignette sets contained in package #1. The second column shows the vignette sets contained in package #2, and so on. f1 to f9 represent female vignette sets and m1 to m9 represent the corresponding male vignette sets.

We restricted the number of vignette sets (= number of interviews) per package to 12 because interviewers were strongly encouraged to return completed interviews of an entire package (instead of a fraction only, which would have overthrown the balanced set assignment). Interviewers typically took on multiple packages in which vignette sets were systematically varied. In assigning packages to interviewers, we started with package #1 and then sequentially assigned one or multiple packages to one interviewer after the other. After the first nine packages #1 to #9 were assigned, we continued assigning packages beginning again with package number #1, and so on. If each interviewer would have taken on three packages we would have obtained a complete blocking by interviewers because each of the three groups of packages {#1, #2, #3}, {#4, #5, #6}, and {#7, #8, #9} contains all 18 vignette sets (all nine female and nine male vignette sets). But since not all interviewers took three or a multiple of three packages blocking was incomplete. However, the systematic variation of vignette sets across the three groups of packages guaranteed a complete blocking of the vignette experiment by the sex×age strata. This becomes apparent by looking at a single sampling stratum, say female employees of age 18-33: the sets across all nine packages (#1 to #9) fully cover all 18 vignette sets and thus all 162 possible vignettes.

The systematic assignment of vignette sets to interviewer packages allowed us (a) to collect vignette data that are balanced across sampling strata (i.e., each female and male vignette is measured with the same frequency in each stratum), and (b) to take care of potential interviewer and stratum effects with respect to the gender income gap. Ideally, we would have perfectly balanced the assignment of interviewers to packages, that is, each interviewer should have gotten a complete set of nine packages (#1 to #9) or at least three packages that completely exhaust all vignette sets. But due to the limited pool of interviewers and their restricted willingness to take on three or even nine packages we were not able to do so.

Overall, the systematic assignment of vignette sets resulted in a vignette experiment with a complete blocking by the six respondent strata and an incomplete blocking by interviewer. While the respondent strata clearly represent a blocking factor and, thus, need to be reflected in the statistical model, it is less clear for the interviewers. However, because of the systematic assignment of vignettes to packages and packages to interviewers, the experimental design still suggests the inclusion of interviewer effects in the analytic models. It is also important to note that the inclusion of sampling strata and interviewer effects in the statistical model is not only justified by the experimental vignette design but also by the sampling design because respondents were sampled within the sex×age strata and then interviewed by 20 interviewers (resulting in a nesting of respondents within interviewers).

In order to avoid systematic selection effects with respect to the gender income gap, interviewers would ideally assign the vignette sets to respondents at random (within respondent strata). However, since random assignment is hard to control in face-to-face interviews with physically (rather than electronically) administered vignettes, we required interviewers to systematically assign vignette sets according to the assignment plan in Table 3. This means that, for each respondent stratum, interviewers had to assign the female set first and then the male vignette set (beginning with the vignette sets of the first assigned vignette package, then the second, and so on). We do not know whether the interviewers strictly adhered to this procedure, but if interviewers would have intentionally assigned female and male vignette sets to respondents with different characteristics, then the estimated gender income gaps would be biased only if the selection characteristics also affected the actual and just income assessments. Moreover, the systematic selection process would need to be similar across interviewers, otherwise interviewer-specific selection differences would at least partially cancel out. Since there is no plausible explanation why interviewers should have intentionally assigned female and male vignette sets to respondents with different characteristic, the systematic assignment of vignette sets to respondents constitutes an unlikely threat to internal validity. Moreover, the descriptive results and the empirical validation of vignettes do not show any signs of a systematic assignment of sets to respondents (see the Results and Empirical Evaluation sections).

Implementation of the Vignette Study

Interviewer Recruitment

We recruited 20 graduate students with prior interviewing experience. Given our goal of 1,000 interviews, we expected each interviewer to conduct about 50 interviews, that is, 4 or 5 packages of vignette sets. Interviewers received a monetary compensation for each interview. An interview lasted about 30 minutes (vignette experiment and questionnaire together). Though the number of interviews per interviewer is rather high, we kept the number of interviewers low because finding qualified interviewers was not easy and the half-day interviewer training was intensive (ideally, each interviewer should have obtained not more than three packages).

Implementation of the Vignette Experiment

In face-to-face interviews, respondents first assessed the actual income and then the fair income for each virtual employee of the set. For assessing the *actual income*, a respondent received all nine vignettes (in random order) and was then asked to rank the nine vignettes according to *actual* income (i.e., the real income of a corresponding Austrian employee). After the ranking, the respondent assigned an income to each vignette, that is, an estimate for the actual monthly net income (in €). Then, the respondent received a blank vignette (Figure 2) which he had to fill in according to his own vita, including his own income.

For assessing the *fair income*, the interviewer put the respondent vignette to the other nine vignettes, shuffled them, and handed them again to the respondent with the request to rank the vignettes according to the *fair income*. The interviewer encouraged the respondent to rank the vignettes with reference to his own vignette, that is, to begin with his own vignette and then to judge which employees of the vignette set justifiably deserve a lower or higher monthly net income. After the ranking, the respondent assigned a fair income (in €) to its own vignette first and then to each of the ranked vignettes. At the end of the vignette experiment, respondent were required to fill in the short questionnaire.

Study Characteristics and Effective Sample Size

The study took place between July and September 2008 in Vienna, Austria. During July and August, the 20 interviewers collected data from 980 respondents (aged between 18 and 65 years). We had to remove 24 interviews because of missing responses on crucial variables like the actual and fair vignette income. It also turned out that interviewers did not completely adhere to the quota plan. Thus, in order to account for the resulting imbalance in the design of the vignette experiment, we collected another 53 interviews in September 2008. Overall, we obtained 1009 complete interviews. Since the data of one interviewer significantly deviated from the data of the other interviewers, we decided to delete this interviewer's data (also because the interviewer could not satisfactorily explain the observed inconsistencies). This left us with an effective sample size of 910 interviews. However, due to the systematic assignment of vignette sets to interviewers and sampling strata, the balance of the experimental vignette design and the distribution of interviews across respondent strata was only marginally affected.[3] Table 4 shows the effective distribution of vignette sets across respondent strata.

Analytic Strategies

Imputation of Missing Values

Of the 910 × 9 = 8190 vignette assessments only ten actual incomes from two respondents and three fair incomes from one respondent were missing. Since these missing values occurred in the outcome variables, we deleted the corresponding vignette data from our data set. For the 15 sociodemographic variables we had overall 2.3% missing values. Three

[3] We analyzed the data also without deleting the unreliable interviews and obtained results that neither significantly nor substantially differed from the main results.

covariates had no missing values, 10 covariates had less than 3% and only 2 covariates had more than 3% missing values (respondent's actual income 7% and parental leave 10%). Two of the 15 covariates also contained implausible or incorrect values (0.13% in respondent's actual income and one value in number of children). We set those implausible values to missing and then imputed them together with the other missing values.

Since interviewers were instructed to ask respondents at the end of the interview about unreported items in the anchoring respondent vignette we were able to impute most of these missing data from the respondents' own questionnaire data (cold deck imputation). For the remaining missing values, we used a multivariate hot deck imputation procedure (via chained equations as implemented in the *mice* package in R; van Buuren &

Table 4
Distribution of vignette sets across respondent strata (i.e., respondent sex and age)

	Female				Male				
Set	18-33	34-44	45-65	Sum	18-33	34-44	45-65	Sum	Total
f1	9	11	8	28	8	9	8	25	53
f2	8	8	6	22	8	7	9	24	46
f3	9	9	8	26	9	7	8	24	50
f4	9	11	7	27	8	8	8	24	51
f5	8	8	10	26	9	8	9	26	52
f6	10	8	8	26	8	8	7	23	49
f7	9	9	8	26	9	9	9	27	53
f8	8	7	9	24	10	8	9	27	51
f9	12	8	9	29	9	8	7	24	53
m1	9	8	8	25	10	8	9	27	52
m2	8	8	9	25	8	9	9	26	51
m3	11	7	8	26	9	7	9	25	51
m4	8	9	7	24	9	9	7	25	49
m5	8	9	9	26	8	7	9	24	50
m6	9	8	7	24	9	9	7	25	49
m7	8	9	10	27	7	10	7	24	51
m8	10	8	9	27	8	7	9	24	51
m9	9	9	6	24	8	8	8	24	48
Total	162	154	146	462	154	146	148	448	910

Groothuis-Oudshoorn, 2011). Given the small number of missing values, we decided to use single instead of multiple imputation because the effects on the standard errors can be expected to be negligibly small. This particularly holds for the results of the design-based models presented below because they do not include any respondent-level covariates as statistical controls (except for respondents' just income which only contains six hot-deck imputed values).

Analytic Methods

Since the experimental vignette design and sampling design affect the collection and generation of vignette data, the analysis needs to directly reflect (a) the four within-subjects factors (industry, education, occupational experience, and parental leave), (b) the between-subjects factor (gender), (c) the set effect of the RBCF design, (d) respondents as random effects because they served as blocks in the RBCF design, (e) the respondent-specific anchoring vignette for the analysis of the fair income, (f) the six respondent strata used for blocking and stratified sampling, and (g) the interviewer effect. If one or several of these design elements were omitted from the analysis, standard errors and type I error rates would be incorrect (most likely standard error would be overestimated and significance tests overly conservative). In addition to the income of the anchoring vignette and the stratum variables, we also considered the inclusion of several respondent covariates. Though their inclusion is not directly justified by the experimental vignette and sampling design, respondent covariates can help in reducing the error variance at the respondent level and, thus, increase the power for the test of the gender income gap.

We briefly discuss two methods for analyzing vignette data: *Analysis of variance* (ANOVA) with random respondent effects and *multilevel modeling* with vignettes as level-one units and respondents as level-two units (for other analytic approaches see Jasso, 2006). Given that we designed the vignette experiment according to a RBCF-3^4 design with a between-subjects factor, ANOVA is the natural choice for the analysis because the deliberate confounding of the three-way and corresponding four-way interaction effects has been based on ANOVA's orthogonal variance decomposition. For the multilevel model, the confounding structure of parameter estimates depends on the coding scheme of the vignette factors. For instance, while deviation coding maintains the RBCF design's confounding structure (i.e., only the parameters of the three-way and corresponding four-way interactions are confounded with the set effect and the set×gender interaction, respectively), dummy coding results in a confounding of many more parameters, including the main effects. In our study, dummy coding would lead to a confounding of main and two-way interaction effects with three- and four-way interaction effects. Moreover, since both deviation and dummy coding are non-orthogonal coding schemes, the corresponding predictors are no longer independent, resulting in less efficient estimates and, thus, the need for model selection (Wu & Hamada, 2009). Despite these disadvantages of multilevel modeling, it directly provides parameter estimates, in particular an estimate of the gender income gap, and corresponding significance tests (ANOVA tests variance components instead of parameter estimates). Multilevel modeling also deals more naturally with unbalanced data that originate from an imperfect implementation of the design or due to single missing vignette measurements (Raudenbush & Bryk, 2002; Snijders & Bosker, 2012). Unbalanced data are more problematic for ANOVA models because the decomposition of variance (type I sums of squares) and hypothesis tests then depend on the inclusion order of factors in the model (Kirk, 1995; Searle, 1987; Speed, Hocking & Hackney, 1978; Venables, 1998). However, slight imbalances in the data only have small effects on the variance decomposition, particularly if the imbalances are caused by random rather than systematic processes. Thus, we will use type I sums of squares but include the gender effect at a position in the model where the experimental vignette effects (i.e., the set effect and the main effects of the other vignette factors) are eliminated before estimating the gender income gap. All other effects, that is, the error-control effects (i.e., stratum and interviewer effects) and interaction effects, are eliminated only after estimating the gender income gap. Alternatively, we could have used type III sums of squares that partial out all the other effects in the model, but this type of analysis is done anyway by the multilevel analysis (though within the framework of maximum likelihood estimation rather than variance decomposition).

In all analyses we use the log of the actual and fair income according to Mincer's wage equation (both income distributions are right-skewed). Parameter estimates can be conveniently interpreted in terms of percentage changes. Here, we only present the design-based models that reflect data-generating design elements of the study, but they can easily be extended to include additional covariates, interaction terms, or random effects.

Analysis of variance. Following the RBCF design of the vignette experiment with gender as a between-subjects factor, the ANOVA model for the *actual income* is given by

$$\begin{aligned}
&\log(Y_{ijklmnopqz}) = \mu + \zeta_z + o_i + \delta_j + \rho_k + \iota_l + \alpha_m + \beta_n + \gamma_o + (\beta\gamma)_{no} + \eta_q + \pi_{p(mnoqz)} + (\zeta\alpha)_{zm} + \\
&(o\delta)_{ij} + (o\rho)_{ik} + (\delta\rho)_{jk} + (o\iota)_{il} + (\delta\iota)_{jl} + (\rho\iota)_{kl} + (o\alpha)_{im} + (\delta\alpha)_{jm} + (\rho\alpha)_{km} + (\iota\alpha)_{lm} + \\
&(o\delta\rho)_{ijk} + (o\delta\iota)_{ijl} + (o\rho\iota)_{ikl} + (\delta\rho\iota)_{jkl} + (o\delta\alpha)_{ijm} + (o\rho\alpha)_{ikm} + (\delta\rho\alpha)_{jkm} + (o\iota\alpha)_{ilm} + \\
&(\delta\iota\alpha)_{jlm} + (\rho\iota\alpha)_{klm} + (o\delta\rho\iota)_{ijkl} + (o\delta\rho\alpha)_{ijkm} + (o\delta\iota\alpha)_{ijlm} + (o\rho\iota\alpha)_{iklm} + (\delta\rho\iota\alpha)_{jklm} + \\
&(o\delta\rho\iota\alpha)_{ijklm} + (o\delta\rho\iota \times \pi)_{ijklp(mnoqz)} + \varepsilon_{ijklp(mnoqz)}
\end{aligned} \tag{1}$$

where $\log(Y_{ijklmnopqz})$ is the logarithm of the actual vignette income for respondent p, μ is the grand mean across all the treatment combinations, sampling strata, and interviewers. ζ_z is the set effect for sets z = 1, ..., 9. The within-subjects effects o_i, δ_j, ρ_k, and ι_l (*i, j, k, l* = 1, 2, 3) refer to the vignette factors industry, education, occupational experience, and parental leave, respectively. The gender income gap is given by the between-subjects effect α_m of the vignette factor gender (m = 1, 2). The effects of the six sampling strata are represented by the main effects for respondents' sex (β_n, n = 1, 2) and age (γ_o, o = 1, 2, 3), and the two-way interaction effect $(\beta\gamma)_{no}$. The interviewer effects are modeled as fixed effects and given by η_q (q = 1, 2, ...,19). We did not model them as random effects because interviewers were not randomly drawn from an underlying target population of interviewers. $\pi_{p(mnoqz)}$ is the random effect of respondents which is independent and identically normally distributed with mean zero and variance $\sigma_\pi^2, NID(0,\sigma_\pi^2)$.

All the two-way and higher-order interaction effects of the vignette factors are included in the model. Note that, the three-and four-way interaction effects $(o\delta\rho)_{ijk}$, $(o\delta\iota)_{ijl}$, $(o\rho\iota)_{ikl}$, $(\delta\rho\iota)_{jkl}$, $(o\delta\rho\alpha)_{ijkm}$, $(o\delta\iota\alpha)_{ijlm}$, $(o\rho\iota\alpha)_{iklm}$, and $(\delta\rho\iota\alpha)_{jklm}$ only represent the unconfounded part of the three- and four-way interactions (i.e., each effect is estimated with only 6 degrees of freedom (df), 2 df are absorbed by the set effects due to partial confounding). While the confounded part of the three-way interactions is included in the set effect ζ_z (they cannot be separately estimated from each other), the confounded part of the four-way interactions is included in the interaction effect between set and gender $(\zeta\alpha)_{zm}$ which models set effect differences between female and male vignette sets. Finally, $(o\delta\rho\iota \times \pi)_{ijklp(mnoqz)}$ is the joint interaction effect of the vignette factors and respondents which is a $NID(0,\sigma^2_{o\delta\rho\iota\pi})$ distributed random effect that is independent of $\pi_{p(mnoqz)}$. $\varepsilon_{ijklmnopz}$ is a $NID(0,\sigma_\varepsilon^2)$ distributed error term which is independent of $\pi_{p(mnoqz)}$. $(o\delta\rho\iota \times \pi)_{ijklp(mnoqz)}$ cannot be estimated separately from $\varepsilon_{ijklp(mnoqz)}$. Note that we modeled the gender effect α_m after the set and within-subjects effects but before all other effects. Thus, using type I sums of squares, the gender income gap is estimated after eliminating the set effect and the main effects of all other vignette factors but before eliminating any effects related to the strata, interviewers, and interactions.

The ANOVA model for the *fair income* is equivalent to Equation (1) except for the inclusion of respondents' fair income which we obtained from the anchoring respondent vignette. We included the respondent's fair income as a factor with ten levels (deciles) because the fair respondent income and fair vignette income were not linearly related to each other (even after taking the log). The respondents' fair income allows us to remove potential bias in the gender income gap due to assessment differences of respondents judging female vignettes and respondents judging male vignettes. Since all vignette assessments of the fair income were anchored on respondents' own fair income, we included respondents' fair income between the grand mean and the set effect (i.e., using type I sums of squares, the anchoring vignette effect is eliminated before estimating all other effects).

Multilevel model. The random intercept model for the *actual income* is given by

$$\begin{aligned}\log(Y_{ij}) &= \beta_{0j} + \mathbf{X}'_{ij}\boldsymbol{\beta} + r_{ij} \\ \beta_{0j} &= \gamma_{00} + \mathbf{set}'_j\boldsymbol{\gamma}_{01} + \gamma_{02}\mathrm{vgen}_j + \gamma_{03}\mathrm{rsex}_j + \mathbf{rage}'_j\boldsymbol{\gamma}_{04} + \mathrm{rsex}\times\mathbf{rage}'_j\boldsymbol{\gamma}_{05} + \mathbf{int}'_j\boldsymbol{\gamma}_{06} + u_{0j}\end{aligned} \qquad (2)$$

where $\log(Y_{ij})$ is the logarithm of the actual vignette income for vignette *i* judged by respondent *j*. β_{0j} is the random intercept for respondent *j*. $\mathbf{X}_{ij}$ is the design vector containing the levels of the vignette factors (industry, education, occupational experience, parental leave) and selected interaction terms; **β** is the corresponding coefficient vector. The error term r_{ij} is independent and identically normal distributed with mean zero and variance σ_r^2, $NID(0,\sigma_r^2)$. In the level-two equation, γ_{00} represents the average intercept across respondents. $\mathbf{set}_j$ is the vector of set predictors and $\boldsymbol{\gamma}_{01}$ the corresponding coefficient vector. The gender-income gap is given by γ_{02} where vgen$_j$ represents vignette gender. The six sampling strata are given by one predictor for respondents' sex (rsex$_j$), two predictors for respondents' age (**rage**$_j$) and two interaction terms (**rsex**×**rage**$_j$). The respective coefficients for the sampling strata are γ_{03}, $\boldsymbol{\gamma}_{04}$ and $\boldsymbol{\gamma}_{05}$. The 18 predictors for the interviewer effects are given by the vector $\mathbf{int}_j$, with $\boldsymbol{\gamma}_{06}$ being the corresponding coefficient vector. The multilevel model for the *fair income* differs from the model in Equation (2) only by the inclusion of the predictor vector $\mathbf{j\_rinc}_j$ for the fair respondent income (i.e., nine predictors for the ten categories of the fair respondent income).

For all predictors (**X**, **set**, vgen, rsex, **rage**, **int**, and **j_rinc**) we chose deviation coding because it guarantees that no additional confounding in parameter estimates is introduced. However, though the effects are unconfounded, the predictors of main and interaction effects are not orthogonal (i.e., they are mutually dependent). For this reason, we selected a parsimonious model that includes all main effects but only significant interaction effects of vignette factors (Wu & Hamada, 2009). Interaction effects were selected according to a stepwise backward and forward search using the likelihood ratio test as selection criteria.

Statistical inference. Since we neither randomly sampled respondents nor randomly assigned vignette sets to respondents, the statistical inference with ANOVA and the multilevel model requires some remarks. The lack of randomly sampling respondents restricts the external validity of results because quota sampling does not directly license an inference to the target population of the Viennese work force. We nonetheless argue that we can cautiously generalize the results to our target population of the Viennese work force because (a) our quota sample is presumably not too different from the target population and (b) there is no reason to believe that the data-generating model for the sample (which we mostly controlled by design) would strongly differ from the corresponding model for the entire target population. However, if one does not want to rely on these assumptions one can refrain from statistical inference and safely read the results as descriptive parameter estimates for the sample. Note that the very same issue frequently occurs with random samples as well, for instance, if a considerable portion of the sampled respondents refuses to participate or gets discarded from the data set because of too many missing or implausible data. A similar issue arises due to the systematic assignment of vignette sets to respondents (within strata). However, the lack of random assignment mostly affects the vignette experiment's internal validity as already discussed.

Results

Respondent Differences

We first investigate whether respondents judging female vignettes significantly differ from respondents judging male vignettes. This is important because the gender gap in the actual and fair income is assessed as a between-subjects factor and, thus, prone to biases caused by differences between the two respondent groups. This is particularly crucial for our study because respondents were neither randomly selected nor randomly assigned to vignette sets. Instead, interviewers deliberately selected respondents and assigned vignette sets within respondent strata which could have created systematic selection bias in the gender income gap. Bias in the fair gender income gap might also be due to context effects because the choice of the fair income levels could have been influenced by the prior assessment of actual incomes. Such context effects are plausible because respondents judging female vignettes typically assigned lower actual incomes than respondents judging male vignettes and, thus, might have chosen a lower fair income for their own anchoring vignette and, consequently, also for the other vignettes. Thus, if respondents judging female vignettes assigned lower fair incomes because of context effects the estimate for the fair gender income gap would be biased.

Table 5 shows the corresponding differences for all 15 respondent covariates. Since all covariate differences are insignificant, systematic bias in the actual and fair gender income gap is unlikely. However, note that the average fair anchoring income of respondents judging female vignettes is €79.9 lower than the average fair income of respondents judging male vignettes (second row in Table 5).

Though the difference in the fair respondent income is not significant (*p*-value = .22) it might nonetheless be due to context effects or selection effects. Thus, including the fair respondent income in our analyses of the fair income removes potential biases due to context and selection effects but also reduces error variance at the respondent level.

Table 5
Respondent differences between respondents judging female vignettes and respondents judging male vignettes

	Female vignettes	Male vignettes	Difference	*p*
Actual respondent income (€)	1783.3	1792.4	-9.1	.89
Fair respondent income (€)	2183.5	2263.2	-79.7	.22
Sex (%)				.90
Female	51.1	50.4	0.7	
Male	48.9	49.6	-0.7	
Age (years)	37.9	37.7	0.2	.84
Occupation (%)				.62
Soldier	0.2	0.0	0.2	
Officer	4.1	5.1	-1.0	
Academic	17.2	15.9	1.3	
Engineer	17.5	21.0	-3.5	
Secretary/clerk	34.1	30.5	3.6	
Service occupation	12.4	12.8	-0.4	
Agriculture	0.9	0.2	0.7	
Craftsman	7.0	6.4	0.6	
Mechanical engineer	2.0	3.1	-1.1	
Unskilled worker	4.6	4.9	-0.3	
Industry (%)				.69
Agriculture/production	12.4	11.5	0.9	
Trade/traffic	20.1	24.3	-4.2	
Service	31.2	29.6	1.6	
Administration/education	13.1	13.9	-0.8	
Health/care	12.0	10.4	1.6	
Other	11.1	10.2	0.9	
Education (%)				.16
Compulsory school	5.9	8.0	-2.1	
Apprenticeship training	16.4	20.1	-3.7	
VET school	8.1	6.4	1.7	
Academic high school	18.3	15.9	2.4	
Vocational high school	18.8	15.0	3.8	
College/university	28.2	27.7	0.5	
Other post-secondary	4.4	6.9	-2.5	
Occup. experience (years)	15.4	16.1	-0.7	.37

Table is continued on the next page.

Table 5 continued

	Female vignettes	Male vignettes	Difference	*p*
Parental leave (months)	7.7	7.6	0.1	.92
Number of children	0.9	0.9	0.0	.60
Working hours per week	37.0	38.3	-1.3	.10
Citizenship				.94
Austrian	87.6	87.2	0.4	
Other	12.4	12.8	-0.4	
Place of residence				.84
Vienna	76.2	77	-0.8	
Other	23.8	23	0.8	
Age of youngest child				.19
No children	52.4	50.4	2.0	
<6 years	10.5	15	-4.5	
6-18 years	21	20.8	0.2	
19+ years	16.2	13.7	2.5	
Employment status				.18
Employed	95.4	93.1	2.3	
Other	4.6	6.9	-2.3	

Note. *p*-values are based on two-sample *t*-tests for continuous variables and χ^2-tests for categorical variables.

Analysis of (Co)Variance

Figures 3 and 4 show the observed marginal means and the gender gaps in the actual and fair income for each of the four within-subjects factors. For both the actual and fair income, respondents assigned a lower income to female employees than to male employees. But the gender gap in the fair income (Figure 4) is clearly smaller than the gender gap in the actual income (Figure 3). Moreover, while the magnitude of the gender gap in the actual income slightly varies with an employee’s industry and educational level, the fair income gap is nearly constant across different levels of occupation, education, occupational experience, and parental leave.

Figure 3. Marginal means of the actual income by vignette gender, industry, educational degree, occupational experience, and parental leave

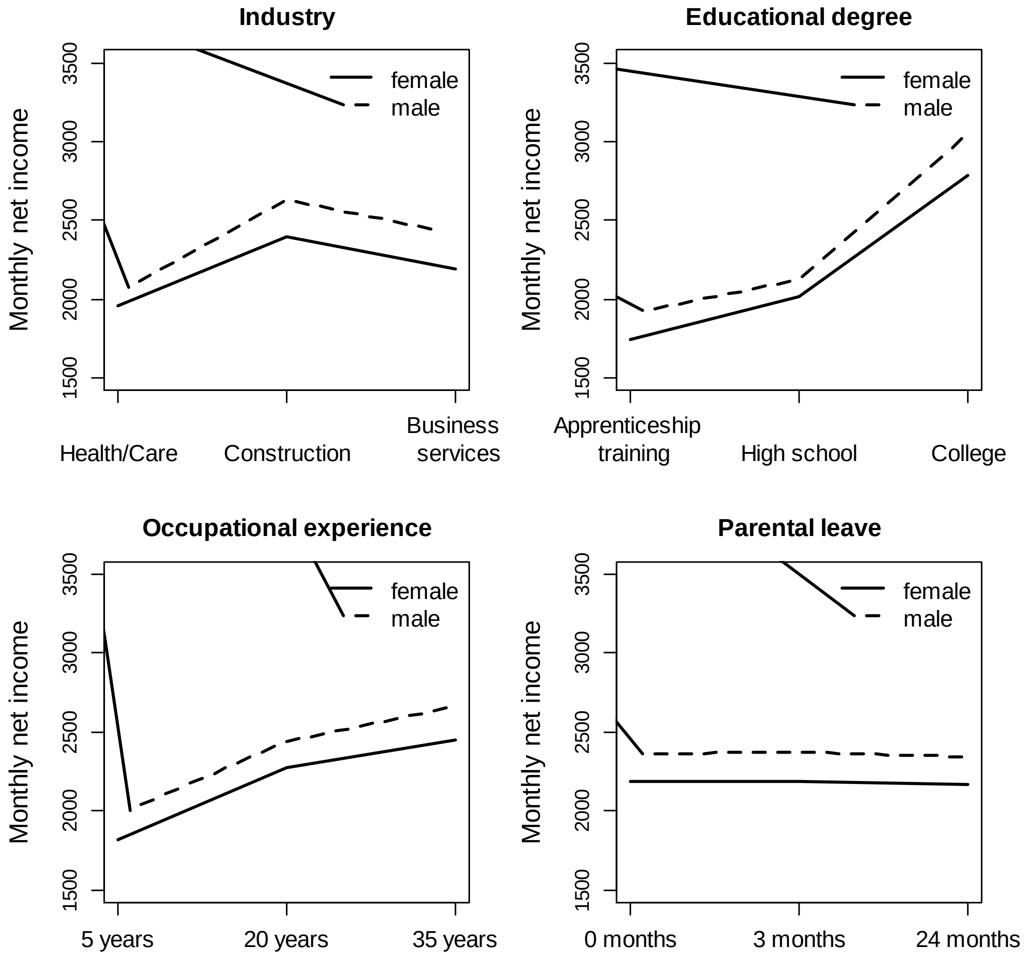


Figure 4. Marginal means of the fair income by vignette gender, industry, educational degree, occupational experience, and parental leave

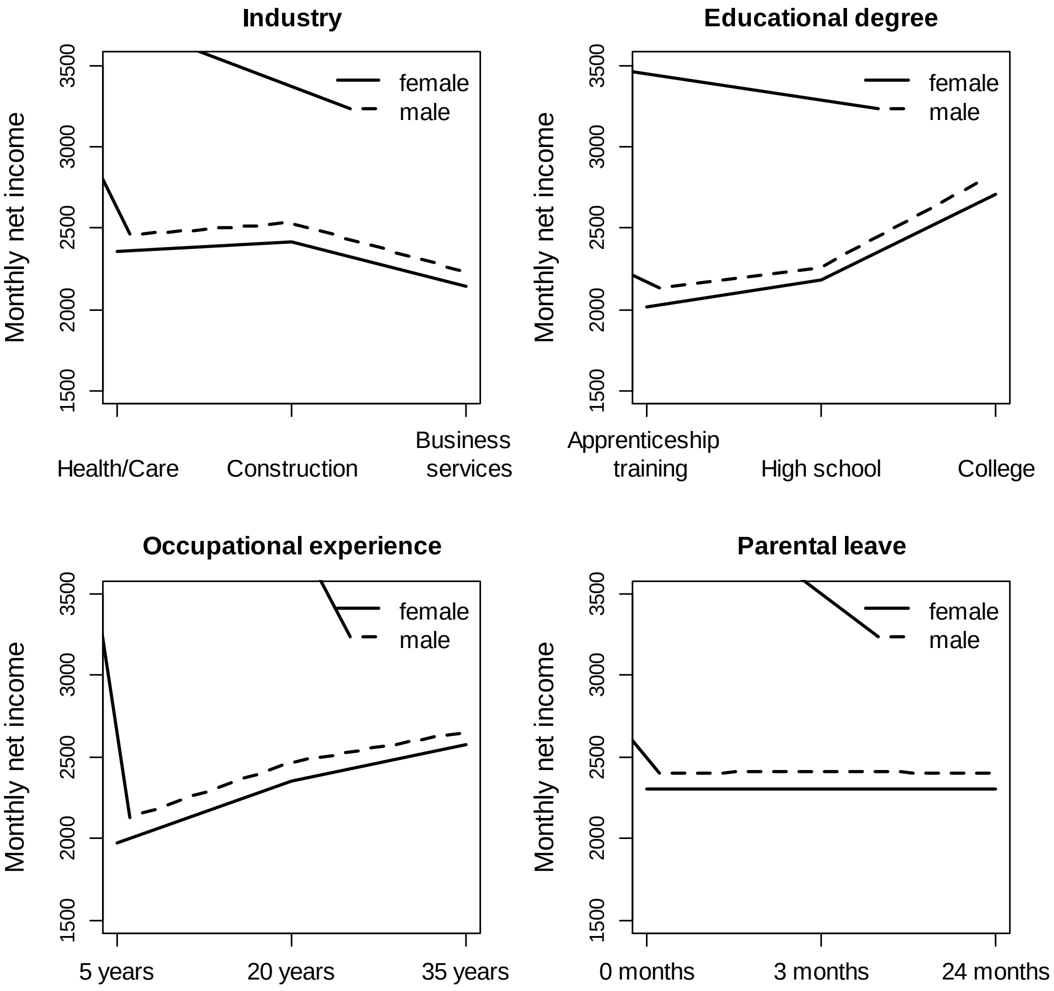


Table 6 presents the ANOVA results for the actual and fair log incomes. For both incomes, the gender gap is significant but for the fair income the gender gap's p-value is closer to the α-level of .05 ($p = .018$). The ANOVA analyses also indicate that the gender gap for the actual income varies with the levels of industry and education, that is, the two-way interaction effects industry×gender (I×G) and education×gender (E×G) are significant. Not surprisingly, the actual and fair incomes strongly depend on an employee's industry, education, and occupational experience. The duration of parental leave has no effect on both the actual and fair income. While some of the two-way interaction effects are significant (e.g., industry×education or industry×occupational experience) none of the unconfounded three-way or higher-order interaction effects are significant. The *F*-tests for the unconfounded part of the partially confounded three way interactions (I×E×Y, I×E×L, I×Y×L, E×Y×L) indicate that some are significant. However, due to the partial confounding they are not directly interpretable. Also the set effect which is confounded with the three-way interaction effects and the set×gender effect which is confounded with the four-way interaction effects are insignificant.

Table 6
ANOVA tables for actual and fair income models

	Actual log income				***Fair log income***			
	df	*SS*	*F*	*p*	*df*	*SS*	*F*	*p*
Between-subjects effects								
Fair respondent income					9	117.7	47.2	.00
Set	8	2.3	0.8	.59	8	3.1	1.4	.19
Gender income gap								
Vignette gender (G)	1	11.3	32.5	.00	1	1.6	5.7	.02
Stratification variables								
Respondent sex	1	1.3	3.7	.05	1	0.2	0.6	.43
Respondent age	2	15.6	22.4	.00	2	2.2	4.0	.02
Respondent sex×age	2	0.3	0.4	.68	2	0.0	0.1	.93
Interviewer	18	24.5	3.9	.00	18	15.3	3.1	.00
Set×gender	8	2.9	1.1	.39	8	2	0.7	.65
Residuals	853	297.4			839	232.5		
Mean square		0.35				0.28		
Within-subjects effects								
Vignette factors								
Industry (I)	2	79.0	718.0	.00	2	21.9	206.4	.00
Education (E)	2	303.2	2756.3	.00	2	117.3	1106.7	.00
Occ. experience (Y)	2	122.4	1112.3	.00	2	86.3	814.2	.00
Parental leave (L)	2	0.1	1.0	.35	2	0.0	0.2	.83
Two-way interactions								
I×E	4	10.8	49.1	.00	4	0.9	4.1	.00
I×Y	4	3.0	13.8	.00	4	0.9	4.0	.00
E×Y	4	0.2	1.1	.36	4	0.4	2.1	.08
I×L	4	0.3	1.2	.30	4	0.7	3.4	.01
E×L	4	0.1	0.5	.74	4	0.1	0.6	.63
Y×L	4	0.2	0.9	.48	4	0.3	1.5	.20
I×G	2	1.1	10.3	.00	2	0.0	0.4	.66
E×G	2	0.6	5.4	.00	2	0.0	0.3	.72
Y×G	2	0.1	0.9	.40	2	0.2	2.1	.12
L×G	2	0.0	0.2	.80	2	0.0	0.0	.99

Table is continued on the next page.

Table 6 continued

	Actual log income				***Fair log income***			
	df	*SS*	*F*	*p*	*df*	*SS*	*F*	*p*
Three-way interactions								
I×E×Y[a]	6	1.9	5.7	.00	6	0.5	1.7	.11
I×E×L[a]	6	0.3	0.9	.50	6	0.1	0.3	.91
I×Y×L[a]	6	0.9	2.6	.02	6	0.9	2.7	.01
E×Y×L[a]	6	0.4	1.1	.36	6	0.2	0.5	.83
I×E×G	4	0.4	1.8	.12	4	0.1	0.5	.77
I×Y×G	4	0.2	0.7	.58	4	0.2	0.9	.48
E×Y×G	4	0.2	1.1	.36	4	0.1	0.4	.80
I×L×G	4	0.1	0.6	.63	4	0.3	1.3	.26
E×L×G	4	0.2	0.8	.54	4	0.1	0.5	.73
Y×L×G	4	0.1	0.5	.73	4	0.1	0.7	.61
Higher-order interactions								
I×E×Y×L	16	0.7	0.8	.70	16	0.5	0.6	.92
I×E×Y×G[a]	6	0.3	0.8	.56	6	0.2	0.5	.78
I×E×L×G[a]	6	0.1	0.2	.97	6	0.3	0.9	.49
I×Y×L×G[a]	6	0.2	0.6	.71	6	0.2	0.6	.75
E×Y×L×G[a]	6	0.3	1.0	.44	6	0.1	0.3	.93
I×E×Y×L×G	16	0.6	0.7	.80	16	0.5	0.6	.91
Residuals	7127	392.0			7132	377.8		
Mean square		0.06				0.05		

Note: *df* are the degrees of freedom, *SS* the sums of squares, *F* is the *F*-test statistic, and *p* is the *p*-value. Variance decomposition is based on type I sums of squares. Between-subjects effects of within-subjects factors are not shown in the table—they are due to the slight unbalancedness of the vignette data and are not significantly different from zero. Vignette factors: Industry (I), Educational degree (E), Occupational experience (Y), Parental leave (L), Gender (G).

[a] These three- and four-way interaction effects are partially confounded with the set effect and the set×gender interaction, respectively, and thus are not directly interpretable. Without confounding, each of these three-way interaction effects would have 8 df, but due to the partial confounding with the set effect and the set×gender interaction they only have 6 df.

From a methodological point of view, the results for the fair income demonstrate the importance of the respondent-specific anchoring vignettes for assessing the between-subjects factor gender. The predictors of the fair respondent income explain 31.2% (*SS* = 117.7) of the variance across respondents. Omitting the predictors of the fair respondent income (everything else held constant), the *F* statistic for the fair gender income gap would have been 1.6 / [(232.5 + 117.7) / 848] = 3.87 which corresponds to a *p*-value of .049. Under the presumption that the variance of the fair income measurements would have been considerably larger without using anchoring vignettes (as we actually saw from the pilot study), the *p*-value would have been larger than .05. This suggests that the anchoring vignettes successfully reduced the variability across respondents and, thus, considerably increased the power for testing the fair gender income gap. Also note that the 19 interviewers and the six sampling strata explain some variance in the actual and fair income across respondents. While the interviewer effect explains 6.8% and 4.1% of the between-subjects variance of the actual and fair income, the respondent strata only explain 4.8% and 0.6%, respectively. Also the inclusion of further respondent-level covariates—respondent's occupation, industry, occupational experience, number of children and weekly working hours—would have helped in reducing the residual sum of squares and, thus, in increasing the power for testing the gender income gap (*p* = 0.015 for the fair income gap; results are not shown).

Multilevel Analysis

Table 7 presents the results of the multilevel analyses for the actual and fair log income. The estimated coefficients are based on deviation coding and were multiplied by 100 (thus, they can directly be interpreted as percentages). For the gender gap in the actual income we get a significant estimate of -7.56 (= 2 × -3.78, due to deviation coding), that is, with respect to our population of virtual employees, females' actual monthly net income is on average 7.56% lower than males' income.[4] Regarding the fair income, we get a significant gender gap of -2.74% (= 2 × -1.37; *t* = -2.33).

Though the assessed gender gap for the fair income is much smaller than for the actual income, it is still significant, suggesting that male employees should earn a slightly higher income than female employees. Interestingly, the gender gap in the fair income is constant across industries, educational degrees, occupational experience, and parental leave (none of the interaction effects with vignette gender has been significant; see also Figure 4). It is also worth noting that the fair income differences between the different levels of industry, education, and occupational experience are less pronounced than for the actual income. The corresponding estimates for the fair income, which indicate the deviations from the grand mean, are on average smaller than for the actual income. For instance, the fair income of employees in the construction industry is 4.5% above the average income and for employees in business-related services 7.21% below the average income. Deviation coding then implies that the fair income of employees in health & care should earn 2.71% (= -4.50 + 7.21) more than the average employee. For the actual

[4] Given the discrete nature of the sex factor, the correct estimate would be (exp(.0756)-1)×100=7.85%. However, since the differences are negligibly small we report and discuss untransformed coefficients only.

income, the corresponding effects are on average greater: 10.81%, 2.15%, and -12.96 (= -10.81 – 2.15), respectively.

Table 7

Multilevel analysis for actual and fair income models

	Actual log income			***Fair log income***		
	B	*SE B*	*t*	*B*	*SE B*	*t*
Intercept	764.97	0.67	1150.1	773.86	2.76	280.1
Fair respondent income (€)						
Income1 (400, 573]				-38.8	9.62	-4.0
Income2 (573, 821]				4.37	9.55	0.5
Income3 (821, 1188]				-27.98	4.37	-6.4
Income4 (1188, 1703]				-17.89	2.95	-6.1
Income5 (1703, 2441]				-3.46	2.91	-1.2
Income6 (2441, 3533]				7.73	2.96	2.6
Income7 (3533, 5065]				25.27	3.76	6.7
Income8 (5065, 7259]				33.68	6.62	5.1
Income9 (7259, 10509]				-14.26	16.07	-0.9
Set indicators						
Set1	1.57	1.85	0.8	-0.72	1.65	-0.4
Set2	0.68	1.90	0.4	1.23	1.69	0.7
Set3	-0.56	1.87	-0.3	0.1	1.66	0.1
Set4	-0.84	1.88	-0.4	-2.4	1.67	-1.4
Set5	0.30	1.85	0.2	2.69	1.66	1.6
Set6	-0.10	1.88	-0.1	1.16	1.68	0.7
Set7	1.20	1.83	0.7	1.65	1.64	1.0
Set8	-1.14	1.85	-0.6	-1.23	1.64	-0.8
Vignette predictors						
Industry–construction (I1)	10.81	0.37	29.4	4.50	0.36	12.5
Industry–business (I2)	2.15	0.37	5.8	-7.21	0.36	-20.0
Educat.–high school (E1)	-6.82	0.37	-18.6	-4.90	0.36	-13.6
Educat.–college (E2)	26.25	0.37	71.5	16.48	0.36	45.9
Occ. experience–20y (Y1)	4.74	0.37	12.9	3.01	0.36	8.4
Occ. experience–35y (Y2)	12.02	0.37	32.7	10.80	0.36	30.1
Parental leave–3m (L1)	0.31	0.37	0.8	0.21	0.36	0.6
Parental leave–24m (L2)	-0.53	0.37	-1.4	-0.05	0.36	-0.1
Gender income gap						
Vignette gender-female (G1)	-3.78	0.66	-5.8	-1.37	0.59	-2.3

Table is continued on the next page.

Table 7 continued

	Actual	*log income*		*Fair log income*		
	B	*SE B*	*t*	*B*	*SE B*	*t*
Stratification variables						
Respondent sex (female)	-1.21	0.66	-1.8	0.56	0.59	0.9
Respondent age (age34-44)	0.35	0.93	0.4	-1.96	0.84	-2.3
Respondent age (age45+)	5.01	0.93	5.4	-0.55	0.89	-0.6
female×age (34-44)	-0.32	0.93	-0.3	0.10	0.83	0.1
female×age (45+)	-0.58	0.93	-0.6	-0.46	0.84	-0.5
Industry×Education						
I1×E1	1.59	0.52	3.1	0.85	0.51	1.7
I2×E1	1.85	0.52	3.6	0.10	0.51	0.2
I1×E2	-4.21	0.52	-8.1	1.02	0.51	2.0
I2×E2	-3.01	0.52	-5.8	-0.37	0.51	-0.7
Industry×Occup. experience						
I1×Y1	0.04	0.52	0.1	0.97	0.51	1.9
I2×Y1	0.62	0.52	1.2	-0.53	0.51	-1.0
I1×Y2	1.58	0.52	3.0	0.53	0.51	1.0
I2×Y2	1.37	0.52	2.6	0.78	0.51	1.5
Industry×Gender						
I1×G1	-0.87	0.37	-2.4			
I2×G1	-0.79	0.37	-2.2			
Education×Gender						
E1×G1	1.20	0.37	3.3			
E2×G1	-0.77	0.37	-2.1			
Industry×Parental leave						
I1×L1				1.61	0.51	3.2
I2×L1				-1.62	0.51	-3.2
I1×L2				-1.02	0.51	-2.0
I2×L2				0.81	0.51	1.6
Interviewer (the 18 effects are not shown; some are significant)						
SD random intercept	18.0			15.7		
SD level-one	23.5			23.0		

Note. *B* represents the deviation-coded effect estimate, *SE B* the coefficients' standard error, and *t* the *t* test statistic (we do not report *p*-values because the *t* test statistic is only roughly *t* distributed). Income1 to Income9 are indicator variables for nine income deciles (the boundaries are given in parentheses).Interviewer effects are not shown in table (some are significant). All coefficients and standard errors are multiplied by 100 such that they can be interpreted in terms of percentages.

Deviation coded predictors of vignette factors: Industry (I): I1 construction, I2 health & care; Educational degree (E): E1 apprenticeship training, E2 high school; Occupational experience (Y): Y1 5 years, Y2 20 years; Parental leave (L): L1 0 months, L2 3 months; Gender (G): G1 male.

As in the ANOVA case, the fair respondent income of the anchoring vignettes helped in increasing the power for testing the gender gap in the fair income. Including the predictors of the anchoring income reduced the standard error of the gender gap from .70 to .59. Adding the significant respondent-level covariates (actual income, occupation, weekly working hours) further decreased the standard error but only to 0.57 (not shown in the table). Importantly, the predictors of the fair respondent income (anchoring vignette) removed a potential bias in the fair gender income gap due to assessment and selection differences between respondents judging female vignettes and respondents judging male vignettes. Without including the fair respondent income, we would have obtained a gender gap in the fair income of -4.34% (instead of -2.74%).

Empirical Evaluation of the Vignette Experiment's Construct Validity

A crucial question with respect to our vignette experiment is whether it is a valid design for inferring the gender gap in the fair income. More specifically, do the vignettes and the between-subjects design actually measure the fair income gap we intended to measure? Though we cannot directly investigate this question for the fair income, we can at least probe the construct validity with respect to the actual income by comparing the vignette results against the real gender income gap in Austria (i.e., the actually existing gender income gap). Together with Statistics Austria we estimated the gender income gap from a register data base that linked the Austrian microcensus 2005 to income data from the Austrian wage withholding tax statistics 2005. The overall sample size of the data set was about 20,000 respondents. In order to obtain the gender income gap we essentially estimated the same Mincer wage equation as for the vignette data except that we used the net hourly wage rate (hwage) instead of the actual income and included education (edu) and occupational experience (oexp) as continuous variables (in years) instead of categorical variables:

$$\log(\text{hwage}_i) = \beta_0 + \beta_1 \text{sex}_i + \beta_2 \text{edu}_i + \beta_3 \text{oexp}_i + \beta_4 \text{oexp}_i^2 + \varepsilon_i .$$

We estimated this equation separately for full time employees in the three industries construction, business related services, and health & care. To approximately maintain comparability between the vignette data and the register data, we restricted the register data to those occupational groups that contained the occupations we used in our vignette design (see Table 2).

The results in table 8 indicate that the actual gender income gaps estimated from the vignette experiment are very close to the real gender gaps in Austria. Overall, the income gaps from the vignette experiment slightly underestimate the real gender gaps but, except for the industry of business-related services, the differences are not significant (note that the differences might be due to the specific choice of occupations in the vignette experiment which do not perfectly represent the occupational groups in the register data). This suggests—but does not prove—that our vignette experiment possesses construct validity for inferring the gender income gap in the actual income but very likely also in the fair income.

Table 8

Gender gaps (in %) in the actual income estimated from vignette and register data

Data source	Construction	Business-rel. services	Health/care	Total
Vignette data	-9.3 (1.5)	-9.2 (1.5)	-4.2 (1.5)	-7.6 (1.3)
Register data	-9.9 (5.8)	-13.4 (1.0)	-7.2 (1.9)	-10.2 (2.1)

Note. Standard errors are given in parentheses.

Discussion

Using a case study on the fair gender income gap, we demonstrated the benefits of several design elements for increasing a vignette experiment's validity and reliability. We used a confounded factorial design to keep all main and two-way interaction effects free of any confounding, and implemented employee's gender as between-subjects effect in order to avoid social desirability bias. For ruling out systematic selection and interviewer biases in the fair gender income gap, we introduced respondent-specific anchoring vignettes and systematically assigned vignette sets to interviewer packages. In aiming at high reliability, we stratified the respondent sample by sex and age, blocked the vignette experiment by respondent strata and interviewers, introduced the anchoring vignettes, tested the practicability of the vignette experiment in two pilot studies, determined the required sample size for a sufficiently powered test of the gender income gap, measured additional respondent covariates, and let the respondents rank the vignettes before they rated them. Finally, we probed the vignette experiment's construct validity by comparing the actual gender income gap estimated from the vignette experiment to the gender income gap estimated from register data. Overall, the chosen design elements were able to guard against many plausible threats to validity and to increase the vignette experiment's reliability (as partially seen from the results). This case study also showed that between-subjects factors (gender income gap) can be estimated from vignette data with high validity and reliability, provided that the respondent-level variance is successfully reduces by design and statistical control (stratified sample, blocking by strata and interviewers, systematic assignment of vignette sets to interviewers, anchoring vignettes, covariate measures). Two of the case study's design elements have been suggested and used for the first time: respondent-specific anchoring vignettes and the systematic assignment of vignette sets to respondent strata and interviewers.

In presenting ANOVA and multilevel models for analyzing the vignette data, we highlighted that the statistical models must correctly reflect the underlying data-generating mechanism. Analyzing the vignette data requires the inclusion of set, stratum, and interviewer effects but also an adequate modeling of respondents' anchoring vignette income. Though ANOVA and multilevel models rely on different statistical principles (decomposition of variance and maximum likelihood) the results are very similar—the significance patterns are essentially identical. However, for multilevel models it is important to use deviation or an orthogonal coding scheme because dummy coding results in confounded main and interaction effects. The results from both models indicate that the gender gap in the fair income is significant. The multilevel model provides an estimate of 2.74% for the fair gender gap which is in line with some but not all results of other studies on the gender income gap (Jasso & Webster, 1997, 1999; Sauer, 2014).

Though we used a series of design elements to ensure our study's validity and reliability, we were not able to rule out all plausible threats to validity. First, respondents were deliberately selected by interviewers instead of randomly drawn from a stratified sampling frame. Moreover, we neither know the portion nor the characteristics of persons who were asked to participate but refused to do so. Both selection processes limit the vignette experiment's external validity. Second, vignette sets were not randomly assigned to respondents (within respondent strata), which diminishes the internal validity of the experiment if a systematic selection related to the income assessments would have taken place. Third, the confounding of the virtual employees' names with the occupations in the vignette description could have biased the effect estimates. Forth, the vignette experiment was incompletely blocked by interviewers because not every interviewer could take three or nine packages of vignette sets. As discussed, we were not able to implement the randomizations and the complete blocking by interviewer because of practical restrictions. Fifth, we also had implementation issues (incomplete adherence of interviewers to quota and set assignments, incomplete or unreliable interviews, a few missing data) which resulted in slightly imbalanced data and a marginal decrease in efficiency. As we argued for each of the five limitations in more detail in the description of the case study, we do not think that they strongly affected the internal and external validity of the study. The empirical validation with respect to the actual income seems to support this belief. Again, a better (more valid and reliable) study would have tried to avoid the vulnerability to these validity threats. However, highlighting both the study's strengths and weaknesses allows the reader to critically judge the study's validity and reliability and to draw correspondingly cautious conclusions about the findings.

It is important to mention that the design elements discussed were tailored to the vignette experiment on the fair income. For vignette experiments on a different research question, some of the design elements discussed might be completely inappropriate while other design elements we did or could not use might become highly relevant. For instance, if the main goal of our study would have been on respondents' perception of the gender gap in the actual income (rather than the fair income), we would not have designed gender as a between-subjects factor. Or, the ranking of vignettes is more easily implemented in face-to-face interviews with physical vignettes than in online surveys with electronic vignettes. But an electronic administration of the vignette experiment would have allowed for an automated random assignment of vignette sets to respondents. Thus, the choice of design elements always depends on the specific research question, the interview mode and vignette presentation, and the corresponding threats to validity. Also budget and feasibility restrictions influence the choice of design elements.

Nonetheless, researchers using vignette experiments in survey research should always thoroughly consider the implementation of various design elements that guarantee a valid and reliable estimation of effects. In doing so, they can follow fundamental principles in experimentation and survey research. In designing a vignette experiment, the three fundamental principles *blocking*, *randomization*, and *replication* can be used to increase the experiments validity and reliability (Hinkelmann & Kempthorne, 1994; Wu & Hamada, 2009): Block what you can. Randomize what you cannot control by design. Replicate the experiment, that is, have multiple measurements for each vignette within each block such that the error variance can be reliably estimated. Similar principles apply to sampling and survey designs (Kish, 1987): *Stratification*, *representativeness (randomization)*, *replication*, and *statistical control*. Stratify the target population and randomly sample subjects within strata in order to obtain a probabilistically representative sample. Again, replication (within strata) is required to reliably estimate the error variance and achieve the desired power for testing the hypotheses of interest. Finally, use statistical controls (covariate measurements) for what cannot be controlled by design. Vignette studies that make extensive use of these fundamental principles in experimentation and survey research allow for more valid and reliable inferences.

Author Notes: We thank Courtney Hall for helpful comments. This study was supported by the Jubilee Fund Grant 12320 from the Austrian National Bank (OeNB).

References

Altonji, J., & Blank, R. (1999). Race and gender in the labor market. *Handbook of Labor Economics, Vol. 3,* 3143-3259.

Alves, W. M., & Rossi, P. H. (1978). Who should get what? Fairness judgments of the distribution of earnings. *American Journal of Sociology, 84,* 541-564.

Atzmüller, C., & Kromer, I. (2013). *Peer Violence: Gewalt unter Jugendlichen aus der Perspektive von Mädchen und Burschen*. Research report for the Federal Ministry of Science, Research and Economy, Vienna, Austria.

Atzmüller, C., & Kromer, I. (2014). *Peer Delinquency: Wahrnehmung und Bewertung typischer Jugenddelikte aus der Sicht Jugendlicher als Grundlage für Präventionsmaßnahmen*. Research report for the Federal Ministry for Transport, Innovation and Technology, Vienna, Austria.

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,* 128-138.

Auspurg, K., & Hinz, T. (2015). *Factorial survey experiments*. Thousand Oaks: Sage.

Berk, R. A. (2004). *Regression analysis: A constructive critique*. Thousand Oaks: Sage Publications.

Blau, F. D., & Kahn, L. M. (1996). Wage structure and gender earnings differentials: An international comparison. *Economica, 63*, 29-62.

Blinder, A. S. (1973). Wage discrimination: Reduced form and structural estimates. *Journal of Human Resources, 18,* 436-455.

Böheim, R., Hofer, H., & Zulehner, C. (2005). *Wage differences between men and women in Austria: Evidence from 1983 and 1997.* IZA Discussion Paper series No.1554.

Cook, F. L. (1979). *Who should be helped? Public support for social services*. London: Sage Publications.

Dülmer, H. (2007). Experimental plans in factorial surveys. Random or quota design? *Sociological Methods & Research, 35,* 382-409.

Gaines, B. J., Kuklinski, J. H., & Quirk, P. J. (2007). The logic of the survey experiment reexamined. *Political Analysis, 15*, 1-20

García, J., Hernández, P. J., & López-Nicolás, A. (2001). How wide is the gap? An investigation of gender wage differences using quantile regression. *Empirical Economics, 26*, 149–67.

Grol-Prokopczyk, H. (2014). Age and sex effects in anchoring vignette studies: Methodological and empirical contributions. *Survey Research Methods*, *8*, 1-17.

Hinkelmann, K., & Kempthorne, O., (1994). *Design and analysis of experiments: Vol. 1. Introduction to experimental design.* New York: Wiley.

Jasso, G. (1992). Assessing individual and group differences in the sense of justice: Framework and application to gender differences in the justice of earnings. *Social Science Research, 23*, 368-406.

Jasso, G. (2006). Factorial survey methods for studying beliefs and judgments. *Sociological Methods & Research, 34,* 334-423.

Jasso, G. (2012). Safeguarding justice research. *Sociological Methods & Research*, *41*, 217-239.

Jasso, G., & Meyersson Milgrom, E. M. (2008). Distributive justice and CEO compensation. *Acta Sociologica, 51*, 123-143.

Jasso, G., & Webster, M. J. (1997). Double standards in just earnings for male and female workers. *Social Psychology Quarterly*, *60*, 6-78.

Jasso, G., & Webster, M. J. (1999). Assessing the gender gap in just earnings and its underlying mechanisms. *Social Psychology Quarterly, 62*, 367-380.

King, G., Murray, C., Salomon, J., & Tandon, A. (2004). Enhancing the validity and cross-cultural comparability of survey research. *American Political Science Review*, *98*, 191-207.

King, G., & Wand, J. (2007). Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Analysis*, *15*, 46-66.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Pacific Grove: Brooks/Cole Publishing Company.

Kish, L. (1987). *Statistical design for research*. Hoboken, NJ: Wiley & Sons.

Lemieux, T. (2006). The 'Mincer equation' thirty years after schooling, experience, and earnings. In S. Grossbard (ed.), *Jacob Mincer: A pioneer of modern labor economics* (pp. 127-145). New York: Springer.

Lim, A. (2014). *People and tigers: Finding solutions to human-carnivore conflict in Nam Et-Phou Louey National Protected Area, Northern Lao PDR*. Paper presented at The Fourth International Conference on Lao Studies, Madison, Wisconsin.

Montgomery, D. C. (2013). *Design and analysis of experiments*. New York: Wiley.

Nock, S. L., & Peter H. R. (1978). Ascription versus achievement in the attribution of family social status. *American Journal of Sociology, 84*, 565-590.

R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Raudenbush, S. W., & Bryk A. S. (2002). *Hierarchical linear models. Applications and data analysis methods. Second Edition.* London: Sage Publications.

Rossi, P. H. (1979). Vignette analysis: Uncovering the normative structure of complex judgments. In R. K. Merton, J. S. Coleman & P. H. Rossi (Eds.), *Qualitative and quantitative social research. Papers in honor of Paul F. Larzarsfeld* (pp. 176-186). New York: Free Press.

Rossi, P. H., & Anderson, A. B. (1982). The factorial survey approach: An introduction. In P. H. Rossi & S. L. Nock (Eds.), *Measuring social judgments: The factorial survey approach*. Beverly Hills: Sage Publications.

Rossi, P. H., Sampson, W. A., Bose, C. E., Jasso, G., & Passel, J. (1974a). Measuring household social standing. *Social Science Research, 3*, 169-190.

Rossi, P. H., Waite, E., Bose, C. E., & Berk, R. E. (1974b). The seriousness of crimes: Normative structures and individual differences. *American Sociological Review, 39*, 224-237.

SAS Institute Inc. (2012). *Using JMP 10*. Cary, NC: SAS Institute Inc.

Sauer, C. (2014). *A just gender pay gap? Three factorial survey studies on justice evaluations of earnings for male and female employees*. SFB 882 Working Paper Series, 29. Bielefeld: DFG Research Center (SFB).

Sauer, C., Auspurg, K., Hinz, T., & Liebig, S. (2011). The application of factorial survey in general population surveys: The effects of respondent age and education on response times and response consistency. *Survey Research Methods, 5*, 89-102.

Schlüter, E., & Schmidt, P. (2010). Special issue: Survey experiments. *Methodology European Journal of Research Methods for the Behavioral and Social Sciences, 6*, 93-95.

Searle, S. R. (1987). *Linear models for unbalanced data*. New York: Wiley & Sons.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin Company.

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). London: Sage Publishers.

Speed, F. M., Hocking, R. R., & Hackney, O. P. (1978): Methods of analysis of linear models with unbalanced data. *Journal of the American Statistical Association, 73*, 105-112.

Steiner, P. M., & Atzmüller, C. (2006). Experimentelle Vignettendesigns in faktoriellen Surveys. *Kölner Zeitschrift für Soziologie und Sozialpsychologie, 58(1)*, 117-146.

Steiner, P. M., Atzmüller, C., & Wroblewski, A. (2009). *Wahrnehmung, Bewertung und Messung geschlechtsspezifischer Einkommensunterschiede: Eine methodologische Untersuchung mittels Vignettenexperiment, einem traditionellen Fragebogen und Registerdaten*. Research report for the Austrian National Bank (OeNB), Vienna, Austria.

Sniderman, P. M., & Grob, D. B. (1996). Innovations in experimental design in attitude surveys. *Annual Review of Sociology, 22*, 377-399.

Su, D., & Steiner, P. M. (2016). *An evaluation of experimental designs for constructing vignette sets in factorial surveys* (Unpublished manuscript). University of Wisconsin, Madison.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*, 1-67. http://www.jstatsoft.org/v45/i03/.

Venables, W. N. (1998). *Exegeses on linear models*. Paper presented at the S-Plus User's Conference, Washington DC.

Markovsky, B., & Eriksson, K. (2012). Comparing direct and indirect measures of just rewards. *Sociological Methods & Research, 41*, 199–216.

Wheeler, B. (2014). *AlgDesign: Algorithmic experimental design*. R package version 1.1-7.2. http://CRAN.R-project.org/package=AlgDesign.

Wu, C. J., & Hamada, M. S. (2009). *Experiments: Planning, analysis, and optimization* (Vol. 552). John Wiley & Sons.

Zweimüller, J., & Winter-Ebmer, R. (1994). Gender wage differentials in private and public sector jobs. *Journal of Population Economics, 7*, 271-285.