sociological science

Supplement to:

Forster, G. Andrea and Martin Neugebauer. 2024. "Factorial Survey Experiments to Predict Real-World Behavior: A Cautionary Tale from Hiring Studies" Sociological Science 11: 886-906.

SUPPLEMENTARY MATERIALS

This document contains additional information on the experimental design, materials and variables used as well as on additional analyses and robustness checks to the paper: *Forster, A.G. and M. Neugebauer (2024). Factorial survey experiments to predict real-world behavior: A cautionary tale from hiring studies.*

Contents

1 Experimental Design and Materials

1.1 Occupations

We sent applications to job advertisements from a list of 24 occupations in four occupational fields (see Table S1). These fields were chosen to represent a broad range of occupations across different labor market segments. The selection of professions was also guided by the following criteria:

1. Correspondence to study subjects: The selected occupations should have a correspondence to the study subjects of mathematics and German studies, as these subjects were selected for the profiles of applicants who dropped out of their studies
2. Social relevance: the selected occupations should be as important as possible on the labor market, i.e. there should be as many filled and advertised positions as possible.
3. Typicality: The selected occupations should be popular among high school graduates with Abitur and HE dropouts and represent a typical career choice for these two categories of applicants. Therefore, only professions with a quota of at least 33 percent Abitur graduates were selected within the four occupational fields.

Table S1: List of Occupations

Occupational field	Classification code (KldB)	Apprenticeship occupation
Mechatronic and electronic occupations	26122	Electronics technician for automation technology
	26322	Microtechnologist
	26312	Electronics technician for information & systems technology
	26332	Aircraft electronics technician
	26312	IT systems electronics technician
Laboratory occupations	41212	Biology laboratory technician
	41212	Dairy laboratory technician
	41322	Chemical laboratory technician
	41312	Pharmaceutical technician
	41412	Physics laboratory technician
	41422	Materials tester
Administration occupations	73202/73282	Administrative assistant
	73212	Specialist for labor market services
	73212	Social insurance clerk
	73222	Management assistant in healthcare
		Specialist for media and information services
	73312	- Specialization: Archives, Image agency
	73322	- Specialization: Library
	73332	- Specialization: Information & Documentation
	73342	- Specialization: Medical documentation
Advertisement, marketing, and media occupations	92112	Management assistant for marketing communication
	92122	Management assistant for dialog marketing
	92302	Management assistant for audiovisual media
	92302	Digital & Print Media Specialist
	62512	Bookseller

Note: KldB: German Classification of Occupations 2010, translated from German

1.2 Application Materials

This section provides an illustrative comparison between the FE and the 'adapted' FS application materials. Table S2 shows example cover letters for FE and FS next to each other. To obtain realistic cover letters and CVs for the FE, we reviewed a large number of real apprenticeship applications and several online guidelines that give recommendations to applicants. We then designed four sets of cover letters for the four different occupational groups, tailoring the content (e.g. specific interests and hobbies) to the occupational field. The structure of the letters was kept the same across the four occupational fields. Using the FE cover letters as a base, we constructed shorter cover letters for the vignettes in the FS. We made sure that the formulations in the FS were similar enough to the FE to convey the same content but different enough to minimize the risk that respondents realized that the FE applications were part of an experiment. We conducted cognitive pretest interviews with recruiters from the four occupational fields and job counselors to evaluate our application materials and adjusted the materials to suggestions of these recruiters.

In the FE, we included a short note on place of living in the cover letter. This was necessary as the apprenticeship market is very local in Germany due to the age of the applicants (usually 16-19 years old). However, for the technical setup of the experiment, the applicants all had to live at the same address in Germany (in the town of Siegen). Therefore, we constructed a story that explained the reasons for relocating to the municipality of the apprenticeship position. This story was evaluated as being plausible by all the employers in qualitative pretest interviews. As this story would not have been plausible if we applied close to Siegen, we excluded vacancies in a radius of 25km around Siegen from our experiment. Additional checks of our data showed that the relocation distance did not influence the invitation probability.

Table S2: Example of cover letter in FE and FS for laboratory occupations (Translated from German, original available upon request)

FE	FS
Application for an Apprenticeship Position as Chemical Laboratory Technician	Application for an Apprenticeship Position as Chemical Laboratory Technician
Dear Sir or Madam,	Dear Sir or Madam,
I read the advertisement with the reference number [PLATFORM REF NUMBER] on the job board of the federal employment agency with great interest. With the apprenticeship for a chemical laboratory technician, you offer exactly the professional perspective that I imagine for my future.	I saw your advertisement with great interest. It offers me exactly the career prospects that I imagine.
After graduating from high school, I started studying mathematics. During this orientation phase, I discovered that my strengths lie more in practice than in theory. That's why I finished my studies and decided to do an apprenticeship that combines my interests with my skills.	I recently started studying mathematics, but ended my studies early because it didn't offer me enough practice.
I started carrying out small experiments early on. What helps me most is my conscientiousness. I was able to build on this during my school days and always found scientific	I am very interested in science and conscientiously carry out small experiments and nature observations.

questions the most interesting. In my free time I investigate processes in nature. What I particularly like about this job is that it is very varied and you can always solve new tasks and problems. I have already heard a lot of good things about your company and would be happy to complete my training with you. After I last lived in Siegen because my parents temporarily moved, I'm moving back to [MUNICIPALITY OF WORKPLACE] next summer so that I can live and work in the place where I feel at home again. That's why the position with you is very attractive to me. I look forward to convincing you in a personal conversation. Sincerely, Anna Schmidt	 The variety that this job offers really appeals to me. I have already heard a lot of positive things about you as an employer. […] Sincerely, Julia Fischer

For the CV's we used similar processes. We first constructed realistic CVs for the FE and then designed shorter but sufficiently similar CVs for the vignettes in the FS. Table S3 shows the information contained in these CV's. In the FS some of the information was kept more general. For example, we did not include specific locations of the schools or names of the internship companies. It might seem peculiar to include information on the applicants' parents in the CV but this is still common practice in Germany when applying to apprenticeship positions and this information gave us the opportunity to signal applicants' social class background.

Table S3: Example of CV's in FE and FS (Translated from German, original available upon request)

	FE	FS
Personal Data	Name: Julia Fischer Address: Köpfchenstraße 32, 57072 Siegen Telephone: 0151/67579358 Email: julia_fischer@posteomail.net Date of birth: 08/09/2003 in Frankfurt Marital status: single, no children Parents: Sandra Fischer (dental care assistant) Andreas Fischer (bank clerk)	Name: Anna Schmidt Address: - Telephone: - Email: - Date of birth: 10/09/2003 Marital status: - Parents: Stefanie Schmidt, management assistant real estate Andreas Schmidt, dental technician
Education		

	10-2021 – 01-2022 | Bachelor's program in Mathematics at the University of Siegen 08-2018 – 06-2021 | Comprehensive school in Siegen (upper secondary level) 08-2013 – 07-2018 | Grammar school in Frankfurt (lower secondary level) 08-2009 – 07-2013 | Elementary school in Frankfurt	10-2021-08-2022 | Bachelor's program in Mathematics 06-2021 | High school diploma (Abitur) -- | -- Until 06-2013 | Elementary school
Practical Experience	03.2019 | Student internship at BGH Edelstahl Siegen GmbH in Siegen	02.2019 | Internship at a steel construction company
Personal Skills	Very good MS Office skills Driving license (class B), obtained 2021 English (fluent) French (basic knowledge)	-
Hobbies Interests	Volleyball	Basketball

Next to cover letters and CV's we also included school leaving certificates in our applications as this is common practice in Germany. For the FE we used a full school leaving certificate from a comprehensive school in Siegen that allowed us to use these materials for the experiment. We altered these certificates to contain the names and birth dates of our fictitious applicants and specific school grades for the main subjects (German, Mathematics, English, Physics). For the FS, we constructed short excerpts from fictitious school leaving certificates, containing the same information but making them considerably shorter (e.g. we only included the school grades for main subjects).

As it is common practice in Germany to include a photo with application materials, a total of 10 portrait photos were required to provide application documents and vignettes with different photos. The photos show five male- and five female-connotated applicants. Since there was no good scientific data source for the specific requirements for the photos, a larger pool of 30 similar stock photos was initially selected online by several evaluators. These all showed people who have a similar facial expression and look as if they are around 19 years old. The photos were edited so that they all showed a similar image section. The photos were then evaluated in a pre-test using a sample from the online access panel Prolific (n=100) for attractiveness, age and other characteristics. The "most similar" 10 photos were selected for the experiments. Two of the photos

were reserved for the field experiment. All applicants were photoshopped to wear the same plain black button-down shirt so that they were as similar as possible also regarding clothing.

1.3 Measures

Dependent Variable

We took great care to align the decision-making process and the outcome variables of the FE and the FS.

In the FE, we recorded the rate by which the applicants received interview invitations or other positive reactions from employers as an outcome variable. First, all reactions of employers to our application by email, phone, and mail were recorded and coded into twelve different categories. In a second step, these responses were categorized in positive and negative/neutral callbacks. As positive reactions, we count: "invitation for interview, "invitation for (online) test, "invitation for assessment center", and "invitation for internship, and "invitation for try-out". As negative or neutral we categorize: "confirmation of receipt", "call without message and follow up", "rejection", "request for additional documents", "request for callback", "other request", and "no reply" in this way, we receive a dichotomous variable for invitations. Sensitivity analyses with different categorizations of the dependent variable did not lead to different results (see section 4.7 of this supplement).

In the FS, we attempted to closely mimic the actual process of decision-making where recruiters first view all vignettes, forming first opinions about them before making a final decision about invitations. After each of the eight vignettes respondents were asked to indicate on a scale from 0 to 100 percent (in steps of 10) how likely they were to invite the candidate for the second step in the hiring process. After they had seen all eight vignettes, they saw the entire pool of applicants again on one page with key information and their own previous rating. On this page, they could give a final voting on which candidates they wanted to invite. Our qualitative pre-studies confirmed that this process of ranking applicants first before making a final decision closely resembles the way, recruiters make hiring decisions in the real world.

We used the dichotomous evaluation on each applicant from the final page as our dependent variable to make it comparable to the FE. A sensitivity analysis using the percentage measure shows the same results (see section 4.7 of this supplement).

Experimental Dimensions and further variables

On the application materials, we varied a number of characteristics of the applicants in order to create the experimental manipulation for our study. In both experiments these characteristics were educational level, field of study, ethnic background and gender of the applicants. An overview of the levels can be found in Table 1 in the main text. Additionally, in the FS, we varied school achievement and socioeconomic background (see Table S4). In both experiments, we also varied other characteristics of the applicants in the FS to make the vignettes less repetitive. For these characteristics we picked values that did not vary in their level. These characteristics were

phone numbers, email addresses, photos (see details above in section 1.2), hobbies and internship companies.

Table S4: Overview of Additional Dimensions in the factorial survey

	Field experiment	Factorial Survey
School achievement	Signaled by grades in German, Mathematics and English as well as by final school GPA Levels: 2 = intermediate (satisfactory)	Signaled by grades in German, Mathematics and English as well as by final school GPA Levels: 1 = low (sufficient) 2 = intermediate (satisfactory) 3 = high (good)
Socioeconomic background	Signaled via the occupation of parents on the resume of the applicant. Levels: 2 = skilled worker, EGP V/VI (e.g.bank clerk)	It is signaled via the occupation of parents on the resume of the applicant. Levels: 1 = unskilled worker, EGP VII (e.g. cleaning personnel) 2 = skilled worker, EGP V/VI (e.g. management assistant) 3 = graduate worker EGP I//II: (e.g. construction engineer)

The experimental dimensions were randomized in the FE and FS and are therefore not correlated with each other. The correlation matrices for both experiments are shown in Tables S5 and S6. They show very minimal correlations between the variables. The non-zero correlations are due to a small number of non-completed vignette sets and invalid FE applications.

Table S5: Correlation Matrix Factorial Survey

	Gender	Migration	Education	SES	Achievement
Gender	1.00				
Migration	0.00	1.00			
Education	-0.01	0.01	1.00		
SES	0.01	-0.01	0.01	1.00	
Achievement	0.00	0.00	-0.01	0.04	1.00

Table S6: Correlation Matrix Field Experiment

	Gender	Migration	Education
Gender	1.00		
Migration	-0.01	1.00	
Education	-0.00	-0.01	1.00

2 Full Tables of Analysis Regarding H1

Table S7 provides the full regression results for the analysis concerning Hypothesis 1 in the main text. These regression results are also the basis of Figure 3 in the paper. The significance test between the FE and FS models was estimated by pooling the data of the two experiments (N=6,842) and estimating interaction effects between all independent variables and a dummy variable for experimental condition.

Table S7: Regression Results on the Association between Applicant Characteristics and Invitation Probability

	FE		FS		Difference FE and FS	
Applicant Education (Ref = Abitur)						
Intermediate HS	--		-0.122	***	--	
	--		(0.018)		--	
Abitur + some college	0.028		-0.050	**	-0.080	**
	(0.018)		(0.016)		(0.026)	
Applicant Gender (Ref = Male)						
Female	-0.006		0.029	*	0.036	
	(0.018)		(0.012)		(0.022)	
Applicant Ethnic Background (Ref = German)						
Turkish	-0.070	***	-0.001		0.070	**
	(0.018)		(0.011)		(0.023)	
Occupational Field (Ref = Electronics Technician)						
Laboratory Technician	-0.000		0.042		0.038	
	(0.032)		(0.055)		(0.062)	
Administration Clerk	0.106	***	0.039		-0.076	
	(0.024)		(0.046)		(0.050)	
Media Clerk	-0.020		-0.049		-0.018	
	(0.028)		(0.051)		(0.056)	
Achievement (Ref = Intermediate grades)						
Low grades			-0.329	***	--	
			(0.021)		--	
High grades			0.119	***	--	
			(0.017)		--	
Socio-economic status (Ref = intermediate SES)						
Low SES			0.018		--	
			(0.015)		--	
High SES			0.022		--	
			(0.015)		--	
Wave (Ref = Spring 2022)						
Fall 2022	0.032		-0.018		-0.050	
	(0.019)		(0.028)		(0.033)	
Intercept	0.510	***	0.687	***	--	
	(0.026)		(0.049)		--	
Number of observations	3,002		3,840		6,842	
R squared	0.02		0.16		0.02	

Standard errors in parentheses. Significance levels: *** p<.001, ** p<.01, * p<.05

3 Full Tables and Additional Analyses regarding Hypotheses 2 and 3

Tables S8, S9 and 10 contain the full regression results for the analyses regarding Hypotheses 2 and 3. Table S8 shows the results for Hypothesis 2, regarding social desirability as respondent characteristic. This table is the basis for Figure 4 in the main text.

Table S8: Regression Results – Factorial Survey Sample Split by Disposition for Social Desirability

	FE		FS low soc. desirability		FS intermed. soc. desirability		FS high soc. desirability	
Applicant Education (Ref = Abitur)								
Intermediate HS	--		-0.164	***	-0.091	**	-0.080	*
	--		(0.029)		(0.032)		(0.033)	
Abitur + some college	0.028		-0.055	*	-0.036		-0.052	
	(0.018)		(0.023)		(0.029)		(0.034)	
Applicant Gender (Ref = Male)								
Female	-0.006		0.034		0.057	**	-0.014	
	(0.018)		(0.019)		(0.021)		(0.020)	
Applicant Ethnic Background (Ref = German)								
Turkish	-0.070	***	-0.002		0.008		-0.005	
	(0.018)		(0.017)		(0.019)		(0.023)	
Occupational Field (Ref = Electronics Technician)								
Laboratory Technician	-0.000		0.047		0.030		0.023	
	(0.032)		(0.080)		(0.088)		(0.112)	
Administration Clerk	0.106	***	-0.027		0.090		0.011	
	(0.024)		(0.068)		(0.074)		(0.090)	
Media Clerk	-0.020		-0.185	**	0.098		0.006	
	(0.028)		(0.070)		(0.082)		(0.112)	
Achievement (Ref = Intermediate grades)								
Low grades			-0.287	***	-0.329	***	-0.404	***
			(0.032)		(0.038)		(0.042)	
High grades			0.150	***	0.090	**	0.102	**
			(0.026)		(0.030)		(0.035)	
Socio-economic status (Ref = intermediate SES)								
Low SES			0.015		0.055		-0.027	
			(0.023)		(0.029)		(0.026)	
High SES			0.001		0.062	*	-0.009	
			(0.023)		(0.026)		(0.029)	
Wave (Ref = Spring 2022)								
Fall 2022	0.032		-0.092	*	0.097		-0.018	
	(0.019)		(0.043)		(0.054)		(0.052)	
Intercept	0.510	***	0.794	***	0.497	***	0.784	***
	(0.026)		(0.066)		(0.082)		(0.100)	
Number of observations	3,002		1,600		1,272		968	
R squared	0.02		0.18		0.16		0.21	

Standard errors in parentheses. Significance levels: *** p<.001, ** p<.01, * p<.05

In the main text, we only showed the analysis regarding the effect of ethnic background in Figure 4. Figure S1 shows the same analysis for the effect of HE non-completion. We compare the effect of HE non-completion in the FE (orange square) to the same effect in three subgroups of the FS sample: respondents with low SDB (green diamond), intermediate SDB (green triangle) and high SDB (green circle). We do see that for no level of respondent SDB the effect of HE non-completion is similar to that in the FE. This analysis leads to the same conclusion as the analysis on the effect of ethnic background in the main text: We have to reject hypothesis 2 that the results between FE and FS would align better if disposition for SDB is low.

Figure S1: Coefficient plot of the effect of HE non-completion (vs. Abitur) for the FE and for three different levels of disposition for social desirability in the FS

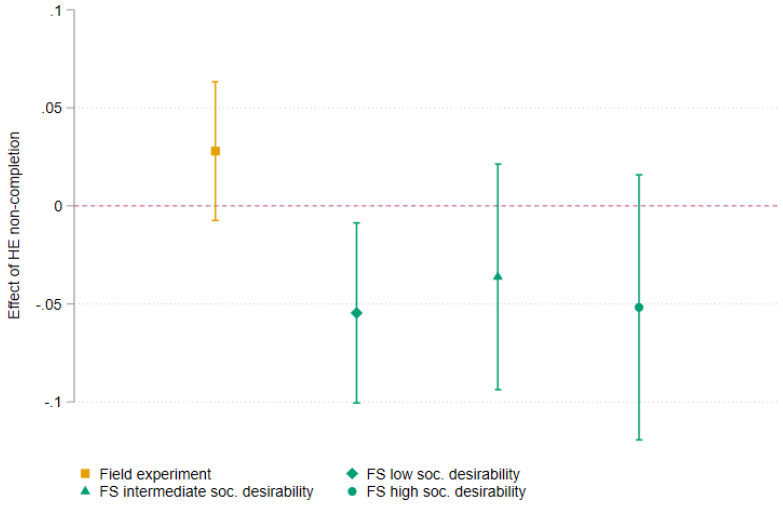


Table S9 shows the results for Hypothesis 3, regarding effort as respondent characteristic – using response time as a measure for effort. This table is the basis for Figure 5 in the main text.

Table S9: Regression Results – Factorial Survey Sample Split by Vignette Response Time

	FE		FS low response time		FS intermediate response time		FS high response time	
Applicant Education (Ref = Abitur)								
Intermediate HS	--		-0.069	*	-0.099	***	-0.189	***
	--		(0.034)		(0.029)		(0.031)	
Abitur + some college	0.028		-0.012		-0.078	**	-0.047	
	(0.018)		(0.027)		(0.026)		(0.028)	
Applicant Gender (Ref = Male)								
Female	-0.006		0.021		0.034		0.022	
	(0.018)		(0.021)		(0.021)		(0.019)	
Applicant Ethnic Background (Ref = German)								
Turkish	-0.070	***	-0.009		0.010		-0.011	
	(0.018)		(0.019)		(0.018)		(0.019)	
Occupational Field (Ref = Electronics Technician)								
Laboratory Technician	-0.000		-0.035		0.024		0.136	
	(0.032)		(0.117)		(0.097)		(0.074)	
Administration Clerk	0.106	***	-0.035		0.006		0.144	*
	(0.024)		(0.083)		(0.080)		(0.061)	
Media Clerk	-0.020		-0.148		-0.025		0.040	
	(0.028)		(0.088)		(0.086)		(0.081)	
Achievement (Ref = Intermediate grades)								
Low grades			-0.247	***	-0.391	***	-0.345	***
			(0.036)		(0.035)		(0.038)	
High grades			0.045		0.128	***	0.198	***
			(0.029)		(0.027)		(0.031)	
Socio-economic status (Ref = intermediate SES)								
Low SES			0.004		0.029		0.042	
			(0.026)		(0.026)		(0.027)	
High SES			0.009		0.054	*	0.011	
			(0.025)		(0.025)		(0.027)	
Wave (Ref = Spring 2022)								
Fall 2022	0.032		0.005		-0.084		0.038	
	(0.019)		(0.055)		(0.046)		(0.045)	
Intercept	0.510	***	0.702	***	0.768	***	0.561	***
	(0.026)		(0.088)		(0.088)		(0.071)	
Number of observations	3,002		1,280		1,264		1,296	
R squared	0.02		0.08		0.22		0.25	

Standard errors in parentheses. Significance levels: *** p<.001, ** p<.01, * p<.05

For Hypothesis 3, in the main text, we only showed the analysis regarding the effect of ethnic background in Figure 5. Figure S2 shows the same analysis for the effect of HE non-completion. We compare the effect of HE non-completion in the FE (orange square) to the same effect in three subgroups of the FS sample: respondents with short response time (green diamond), intermediate response time (green triangle) and high response time (green circle). We do see that for no level of response time the effect of HE non-completion is positive as in the FE. We had hypothesized that higher response time would lead to a closer alignment of the FS with the FE. However, if anything, we see that the effect in the low response time group (-0.012) is closer to the FE (0.028) than in the intermediate (-0.078) and high response time group (-0.047)

This analysis leads to the same conclusion as the analysis on the effect of ethnic background in the main text: We have to reject Hypothesis 3.

Figure S2: Coefficient plot of the effect of dropout (vs. Abitur) for the FE and for three different levels of response time in the FS

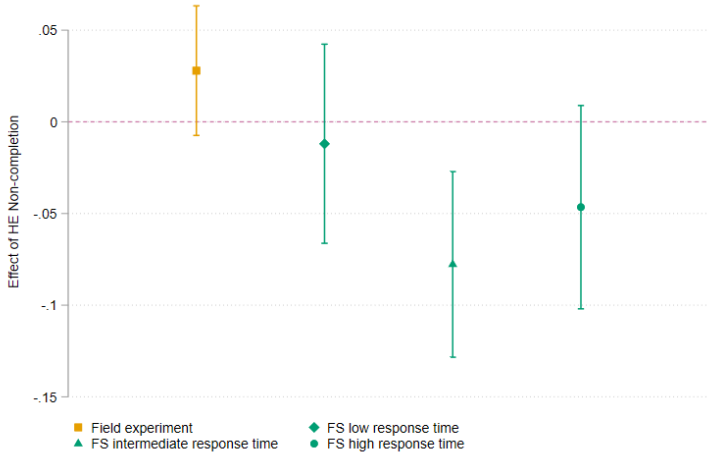


Table S10 shows the results for Hypothesis 3, regarding effort as respondent characteristic – using attitudes towards surveys as a measure for effort. This table is the basis for Figure 6 in the main text.

Table S10: Regression Results – Factorial Survey Sample Split by Appreciation of Surveys

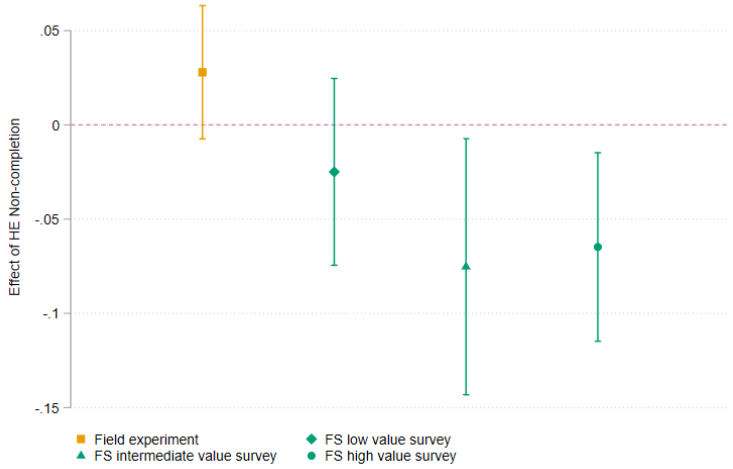
	FE		FS low value		FS intermed. value		FS high value	
Applicant Education (Ref = Abitur)								
Intermediate HS	--		-0.090	***	-0.173	***	-0.130	***
	--		(0.026)		(0.041)		(0.032)	
Abitur + some college	0.028		-0.025		-0.075	*	-0.065	*
	(0.018)		(0.025)		(0.034)		(0.025)	
Applicant Gender (Ref = Male)								
Female	-0.006		0.026		0.039		0.026	
	(0.018)		(0.016)		(0.024)		(0.023)	
Applicant Ethnic Background (Ref = German)								
Turkish	-0.070	***	-0.002		0.007		-0.004	
	(0.018)		(0.016)		(0.025)		(0.018)	
Occupational Field (Ref = Electronics Technician)								
Laboratory Technician	-0.000		0.179	*	0.031		-0.182	
	(0.032)		(0.074)		(0.107)		(0.108)	
Administration Clerk	0.106	***	0.106		0.045		-0.083	
	(0.024)		(0.064)		(0.091)		(0.079)	
Media Clerk	-0.020		0.018		-0.054		-0.161	
	(0.028)		(0.071)		(0.103)		(0.094)	
Achievement (Ref = Intermediate grades)								
Low grades			-0.349	***	-0.300	***	-0.325	***
			(0.031)		(0.045)		(0.037)	
High grades			0.109	***	0.160	***	0.103	***
			(0.024)		(0.038)		(0.031)	
Socio-economic status (Ref = intermediate SES)								
Low SES			0.029		0.012		0.006	
			(0.024)		(0.029)		(0.025)	
High SES			0.016		0.064	*	-0.002	
			(0.023)		(0.030)		(0.026)	
Wave (Ref = Spring 2022)								
Fall 2022	0.032		-0.036		0.008		-0.012	
	(0.019)		(0.042)		(0.056)		(0.050)	
Intercept	0.510	***	0.588	***	0.667	***	0.864	***
	(0.026)		(0.067)		(0.104)		(0.087)	
Number of observations	3,002		1,720		936		1,184	
R squared	0.02		0.18		0.18		0.16	

Standard errors in parentheses. Significance levels: *** p<.001, ** p<.01, * p<.05

Also for survey attitudes, we only showed the analysis regarding the effect of ethnic background in the main text (Figure 6). Figure S3 shows the same analysis for the effect of HE non-completion. We compare the effect of HE non-completion in the FE (orange square) to the same effect in three subgroups of the FS sample: respondents with low valuation of surveys (green diamond), intermediate valuation of surveys (green triangle) and high valuation of surveys (green circle). We do see that for no level of survey valuation the effect of HE non-completion is positive as in the FE. We had hypothesized that higher valuation of surveys would lead to a closer alignment of the FS with the FE. However, if anything, we see that the effect in the low valuation group (-0.025) is closer to the FE (0.028) than in the intermediate (-0.075) and high valuation group (-0.065)

This analysis leads to the same conclusion as the analysis on the effect of ethnic background in the main text: We have to reject Hypothesis 3.

Figure S3: Coefficient plot of the effect of dropout (vs. Abitur) for the FE and for three different levels of survey appreciation



4 Robustness Checks

4.1 Restriction of the Sample to the First Vignette

In our FS design each employer received a set of eight vignettes while in the FE only one application was sent out per employer. One could suspect that through this repetition in the FS, respondents were more likely to discover the topic behind the experimental manipulation (e.g., ethnic background), which may have increased social desirability bias. To test if this was the case, we ran an analysis where we only use the first vignette for each FS respondent.

Table S11 shows the FS results for the full vignette sample (Model 1) and for the sample of first vignettes (Model 2). For this analysis, we have to use the percentage measure of invitation that was collected after each vignette was shown as only in this measure respondents were still oblivious to the following vignettes when deciding. The table shows that the effect of ethnic background is insignificant in both models, indicating that there is no closer alignment between FE and FS if only the first vignette is used. The effect of ethnic background is even more discrepant in the sample that only uses first vignettes. The same holds for the effect of higher education non-completion. The effect of having some college compared to only Abitur is negative in both samples of the FS while it is positive in the FE. If anything, using only the first vignette leads to a bigger difference between FE and FS concerning the effect of HE non-completion. We do not find indications that using only the first vignette would change our substantive conclusions made in the paper.

Table S11: Restricting Sample of the FS to only Using the First Vignette (DV=Percentage measure of invitation)

	FS (all 8 vignettes)		FS (1st vignette only)	
Applicant Education (Ref = Abitur)				
Intermediate HS	-7.491	***	-8.270	*
	(0.960)		(3.479)	
Abitur + some college	-1.499		-2.302	
	(0.790)		(2.888)	
Applicant Gender (Ref = Male)				
Female	0.992		0.010	
	(0.614)		(2.471)	
Applicant Ethnic Background (Ref = German)				
Turkish	-0.257		1.233	
	(0.556)		(2.480)	
Occupational Field (Ref = Electronics Technician)				
Laboratory Technician	-9.024	**	-8.459	
	(3.203)		(4.544)	
Administration Clerk	-4.131		-5.120	
	(2.309)		(3.620)	
Media Clerk	-12.944	***	-12.034	**
	(2.688)		(4.127)	
Achievement (Ref = Intermediate grades)				
Low grades	-18.352	***	-20.901	***
	(1.067)		(3.030)	
High grades	9.962	***	10.281	***
	(0.865)		(2.811)	
Socio-economic status (Ref = intermediate SES)				
Low SES	0.125		0.032	

	(0.677)		(2.902)	
High SES	1.233		3.283	
	(0.710)		(3.118)	
Wave (Ref = Spring 2022)				
Fall 2022	2.137		2.350	
	(1.826)		(2.579)	
Intercept	79.330	***	79.861	***
	(2.242)		(4.492)	
Number of observations	3,840		480	

Standard errors in parentheses. Significance levels: *** p<.001, ** p<.01, * p<.05

4.2 Same Recruiter in FE and FS

One plausible source of measurement error arises from cases where the FS respondent was different from the person reviewing the real-world application. If hiring decisions vary between managers within firms this may lead to disagreement between FS and FE. To test this concern, we restricted the sample to respondents that indicated that they were the only ones responsible for hiring apprentices in their firm and did not make decisions together with colleagues. For these cases we can more safely assume that the hiring manager in the FS and in the FE were the same person.

Table S12 shows the results for the FE (Model 1), for the full FS sample (Model 2) and for the FS sample restricted to those respondents who made decisions alone ((N=632 vignettes, N=79 employers, Model 3). We do see that the effect of ethnic background is still insignificant and now slightly positive in Model 3. The effect of having some college instead of only Abitur is still negative and significant in the restricted sample. Restricting the sample in this way does not lead to a closer alignment of the FS with the FE and therefore does not change our substantive conclusions.

Table S12: One person responsible for hiring only

	FE		FS full sample		FS solely responsible	
Applicant Education (Ref = Abitur)						
Intermediate HS	--		-0.122	***	-0.124	**
	--		(0.018)		(0.045)	
Abitur + some college	0.028		-0.050	**	-0.094	*
	(0.018)		(0.016)		(0.040)	
Applicant Gender (Ref = Male)						
Female	-0.006		0.029	*	0.032	
	(0.018)		(0.012)		(0.024)	
Applicant Ethnic Background (Ref = German)						
Turkish	-0.070	***	-0.001		0.025	
	(0.018)		(0.011)		(0.030)	
Occupational Field (Ref = Electronics Technician)						
Laboratory Technician	-0.000		0.042		-0.036	
	(0.032)		(0.055)		(0.143)	
Administration Clerk	0.106	***	0.039		0.022	
	(0.024)		(0.046)		(0.128)	
Media Clerk	-0.020		-0.049		-0.097	

	(0.028)		(0.051)		(0.140)	
Achievement (Ref = Intermediate grades)						
Low grades			-0.329	***	-0.287	***
			(0.021)		(0.044)	
High grades			0.119	***	0.097	*
			(0.017)		(0.044)	
Socio-economic status (Ref = intermediate SES)						
Low SES			0.018		0.026	
			(0.015)		(0.043)	
High SES			0.022		0.039	
			(0.015)		(0.033)	
Wave (Ref = Spring 2022)						
Fall 2022	0.032		-0.018		0.071	
	(0.019)		(0.028)		(0.077)	
Intercept	0.510	***	0.687	***	0.649	***
	(0.026)		(0.049)		(0.133)	
Number of observations	3,002		3,840		632	
R squared	0.02		0.16		0.14	

Standard errors in parentheses. Significance levels: *** p<.001, ** p<.01, * p<.05

4.3 Realism of Applicants

It is possible that not all employers perceived our applicant profiles as realistic, which may have led to biased responses. To test this, we restricted the sample to recruiters who indicated that our applicant profiles were in close correspondence with real-world applicants. This information was directly collected from employers in the background questionnaire of the FS by asking: *"how closely do the applicants from this survey correspond to typical applicants for an apprenticeship in [occupation title]".*

Table S13 shows the results for the FE (Model 1), for the full sample of the FS (Model 2) and for the FS sample restricted to those vignettes (N=2,344) for which recruiters deemed applicants to be realistic (Model 3). We do see that the effect of ethnic background is still insignificant in the restricted FS sample. The effect of having some college instead of only Abitur is still negative and significant. Restricting the sample in this way does not lead to a closer alignment of the FS with the FE and therefore does not change our substantive conclusions.

Table S13: Restrict FSE Sample to Employers who confirm that our Vignettes are "Typical Applicants"

	FE		FS Full sample		FS Realistic applicants	
Applicant Education (Ref = Abitur)						
Intermediate HS	--		-0.122	***	-0.129	***
	--		(0.018)		(0.023)	
Abitur + some college	0.028		-0.050	**	-0.063	**
	(0.018)		(0.016)		(0.021)	
Applicant Gender (Ref = Male)						
Female	-0.006		0.029	*	0.034	*
	(0.018)		(0.012)		(0.015)	
Applicant Ethnic Background (Ref = German)						
Turkish	-0.070	***	-0.001		0.008	
	(0.018)		(0.011)		(0.014)	
Occupational Field (Ref = Electronics Technician)						

Laboratory Technician	-0.000		0.042		0.034	
	(0.032)		(0.055)		(0.079)	
Administration Clerk	0.106	***	0.039		0.075	
	(0.024)		(0.046)		(0.069)	
Media Clerk	-0.020		-0.049		-0.056	
	(0.028)		(0.051)		(0.074)	
Achievement (Ref = Intermediate grades)						
Low grades			-0.329	***	-0.370	***
			(0.021)		(0.028)	
High grades			0.119	***	0.136	***
			(0.017)		(0.021)	
Socio-economic status (Ref = intermediate SES)						
Low SES			0.018		0.016	
			(0.015)		(0.019)	
High SES			0.022		0.045	*
			(0.015)		(0.019)	
Wave (Ref = Spring 2022)						
Fall 2022	0.032		-0.018		-0.026	
	(0.019)		(0.028)		(0.034)	
Intercept	0.510	***	0.687	***	0.681	***
	(0.026)		(0.049)		(0.076)	
Number of observations	3,002		3,840		2,344	
R squared	0.02		0.16		0.22	

Standard errors in parentheses. Significance levels: *** p<.001, ** p<.01, * p<.05

4.4 Same level of achievement and SES in FE and FS

In the FS we varied two additional dimensions: socioeconomic status and achievement. This additional variation could have altered the evaluation of the profiles. To address this concern, we restricted the FS sample to vignettes that show applicants with intermediate SES and achievement as we did in the FE.

Table S14 shows the results from the FE (Model 1), from the full FS sample (Model 2) and from the FS sample restricted to vignettes with intermediate SES and achievement as applicant characteristics (Model 3). We do see that the effect of ethnic background is still insignificant in the restricted FS sample. The effect of having some college instead of only Abitur is still negative and significant. Restricting the sample in this way does not lead to a closer alignment of the FS with the FE and therefore does not change our substantive conclusions.

Table S14: Restrict sample to vignettes with intermediate SES and achievement

	FE		FS Full sample		FS only intermediate SES and achievement	
Applicant Education (Ref = Abitur)						
Intermediate HS	--		-0.122	***	-0.239	***
	--		(0.018)		(0.064)	
Abitur + some college	0.028		-0.050	**	-0.132	*
	(0.018)		(0.016)		(0.052)	
Applicant Gender (Ref = Male)						
Female	-0.006		0.029	*	0.086	

	(0.018)		(0.012)		(0.047)	
Applicant Ethnic Background (Ref = German)						
Turkish	-0.070	***	-0.001		-0.004	
	(0.018)		(0.011)		(0.044)	
Occupational Field (Ref = Electronics Technician)						
Laboratory Technician	-0.000		0.042		-0.067	
	(0.032)		(0.055)		(0.093)	
Administration Clerk	0.106	***	0.039		0.043	
	(0.024)		(0.046)		(0.078)	
Media Clerk	-0.020		-0.049		-0.118	
	(0.028)		(0.051)		(0.089)	
Achievement (Ref = Intermediate grades)						
Low grades			-0.329	***		
			(0.021)			
High grades			0.119	***		
			(0.017)			
Socio-economic status (Ref = intermediate SES)						
Low SES			0.018			
			(0.015)			
High SES			0.022			
			(0.015)			
Wave (Ref = Spring 2022)						
Fall 2022	0.032		-0.018		0.018	
	(0.019)		(0.028)		(0.054)	
Intercept	0.510	***	0.687	***	0.743	***
	(0.026)		(0.049)		(0.088)	
Number of observations	3,002		3,840		418	

Standard errors in parentheses. Significance levels: *** p<.001, ** p<.01, * p<.05

4.5 Different Specifications of DV

We explored different specifications of the dependent variable in the FE and in the FS. Table S15 shows different specifications of callback in the FE. Model 1 shows the analysis from the main text where we only counted positive reactions as a callback (i.e. interviews, online tests, assessment center). In Model 2, callback is defined more loosely. Here, all reactions are counted that are not negative i.e. in addition to those reactions in Model 1, we also counted clarification questions and requests for documents. In Model 3, we exclude invitations for (online) tests as in some cases we suspect that the invitation to a test is extended to all applicants and is therefore no real invitation to a second round in the application process. Across all different operationalizations, we do see a negative and significant effect of ethnic background and a positive effect of HE non-completion. The HE non-completion effect is significant for the loose definition of callback but not for the other two specifications. None of the differences does change the substantive conclusions we draw regarding our hypotheses.

Table S15: Field experiment different specifications of DV

	FE Callback strict definition	FE Callback loose definition	FE Callback without tests
Applicant Education (Ref = Abitur)			

Dropout	0.028		0.046	**	0.013	
	(0.018)		(0.018)		(0.018)	
Applicant Gender (Ref = Male)						
Female	-0.006		-0.010		-0.005	
	(0.018)		(0.018)		(0.018)	
Applicant Ethnic Background (Ref = German)						
Turkish	-0.070	***	-0.089	***	-0.086	***
	(0.018)		(0.018)		(0.018)	
Occupational Field (Ref = Electronics Technician)						
Laboratory Technician	-0.000		0.044		-0.035	
	(0.032)		(0.031)		(0.031)	
Administration Clerk	0.106	***	0.114	***	-0.061	**
	(0.024)		(0.023)		(0.024)	
Media Clerk	-0.020		0.040		-0.019	
	(0.028)		(0.027)		(0.027)	
Wave (Ref = Spring 2022)						
Fall 2022	0.032		0.014		-0.079	***
	(0.019)		(0.018)		(0.019)	
Intercept	0.510	***	0.578	***	0.531	***
	(0.026)		(0.026)		(0.026)	
Number of observations	3,002		3,002		3,002	

Standard errors in parentheses. Significance levels: *** p<.001, ** p<.01, * p<.05

Table S16 shows different specifications of the DV in the factorial survey. In Model 1, the dichotomous invitation variable, i.e. the answer from the overview page after the respondents already have seen all 8 vignettes, is used as in the main text. In Model 2, the percentage variable of probability for invitation is used as recorded on each vignette page while the respondents are seeing the vignettes for the first time. In Model 3, a dichotomized version of the percentage variable is used where those vignettes with 100% invitation probability are coded as 1 and all other vignettes are coded as 0. In all three models the effect of ethnic background is insignificant (the effect size is larger in Model 2 due to the different scaling of the percentage measure). The effect of HE non-completion is only significant if the dichotomous measure is used but in all three models, the effect size is negative (and thus different from the significantly positive effect in the FS). Even if the effect does not reach significance in Model 2 and 3, we tend to conclude that this does not mean that the effect in the FE is replicated as the sign of the effect does differ. Therefore, none of the effects here changes the substantive conclusions, we draw regarding our hypotheses.

Table S16: Factorial Survey different specifications of DV

	FS dichotomous DV		FS percentage DV		FS percentage DV dichotomized	
Applicant Education (Ref = Abitur)						
Intermediate HS	-0.122	***	-7.491	***	-0.039	**
	(0.018)		(0.960)		(0.015)	
Dropout	-0.050	**	-1.499		-0.024	
	(0.016)		(0.790)		(0.013)	
Applicant Gender (Ref = Male)						
Female	0.029	*	0.992		0.015	

	(0.012)		(0.614)		(0.010)	
Applicant Ethnic Background (Ref = German)						
Turkish	-0.001		-0.257		0.005	
	(0.011)		(0.556)		(0.009)	
Occupational Field (Ref = Electronics Technician)						
Laboratory Technician	0.042		-9.024	**	-0.090	
	(0.055)		(3.203)		(0.053)	
Administration Clerk	0.039		-4.131		-0.003	
	(0.046)		(2.309)		(0.046)	
Media Clerk	-0.049		-12.944	***	-0.175	***
	(0.051)		(2.688)		(0.048)	
Achievement (Ref = Intermediate grades)						
Low grades	-0.329	***	-18.352	***	-0.152	***
	(0.021)		(1.067)		(0.015)	
High grades	0.119	***	9.962	***	0.177	***
	(0.017)		(0.865)		(0.017)	
Socio-economic status (Ref = intermediate SES)						
Low SES	0.018		0.125		-0.013	
	(0.015)		(0.677)		(0.013)	
High SES	0.022		1.233		-0.019	
	(0.015)		(0.710)		(0.012)	
Wave (Ref = Spring 2022)						
Fall 2022	-0.018		2.137		0.007	
	(0.028)		(1.826)		(0.029)	
Intercept	0.687	***	79.330	***	0.319	***
	(0.049)		(2.242)		(0.047)	
Number of observations	3,840		3,840		3,840	

Standard errors in parentheses. Significance levels: *** p<.001, ** p<.01, * p.05

4.6 Sample Selection Bias

Another threat to external validity that is independent of behavioral validity but nevertheless of utmost importance for FS research is sample selection bias. If we compare FS results to a behavioral real-world benchmark, it is important to distinguish whether bias has been introduced by a lack of behavioral validity or by selectivity issues of the sample at hand.

Table S17: Restrict Sample to respondents who participated in both experiments

	FE Full sample		FE restricted to FS sample		FS	
Applicant Education (Ref = Abitur)						
Intermediate HS	--		--		-0.122	***
	--		--		(0.018)	
Abitur + Some College	0.028		-0.018		-0.050	**
	(0.018)		(0.043)		(0.016)	
Applicant Gender (Ref = Male)						
Female	-0.006		-0.066		0.029	*
	(0.018)		(0.043)		(0.012)	
Applicant Ethnic Background (Ref = German)						
Turkish	-0.070	***	-0.072		-0.001	
	(0.018)		(0.044)		(0.011)	
Occupational Field (Ref = Electronics Technician)						

Laboratory Technician	-0.000		-0.062		0.042	
	(0.032)		(0.085)		(0.055)	
Administration Clerk	0.106	***	0.084		0.039	
	(0.024)		(0.067)		(0.046)	
Media Clerk	-0.020		-0.112		-0.049	
	(0.028)		(0.077)		(0.051)	
Achievement (Ref = Intermediate grades)						
Low grades					-0.329	***
					(0.021)	
High grades					0.119	***
					(0.017)	
Socio-economic status (Ref = intermediate SES)						
Low SES					0.018	
					(0.015)	
High SES					0.022	
					(0.015)	
Wave (Ref = Spring 2022)						
Fall 2022	0.032		0.014		-0.018	
	(0.019)		(0.046)		(0.028)	
Intercept	0.510	***	0.708	***	0.687	***
	(0.026)		(0.070)		(0.049)	
Number of observations	3,002		480		3,840	
R Squared	0.02		0.04		0.16	

Standard errors in parentheses. Significance levels: *** p<.001, ** p<.01, * p<.05

When samples of real employers are used, response rates are often as low as 10 to 20 percent and this is also the case for our study. With such a high level of non-response it is likely that the recruiters that do reply to the request for survey participation differ in systematic ways from those that do not reply. Therefore, selective non-response is a serious concern for these types of FS experiments.

As we carried out both experiments on the same sample, we can to a certain extent evaluate what influence nonresponse in the FS has on the results. The FE – by design – is not affected by nonresponse. For this analysis, we restrict the sample of the FE to those respondents who also participated in the FS so that we look at exactly the same respondents in both experiments. The results are shown in Table S17. This leads to a significantly reduced sample size of N=480 for the FE where we only have one observation per respondent. While this naturally decreases the statistical power to find significant differences, we can still compare the effect sizes across the two experiments. For the effect of ethnic background, we still see a negative effect of 7.2 percentage points in the field experiment compared to the null effect in the factorial survey. So, also with the exact same sample, we do not find a correspondence between the results in the FE and the FS. For the effect of dropout, we observe that now the effect in the FE is negative (-0.018) while it was slightly positive before (0.027). While still not significant and smaller than in the FS, we do see a somewhat closer correspondence with the dropout effect in the FS (-0.050) now. This is, however,

not enough to conclude that sample selection bias can explain the divergence in results between FS and FE. Therefore, this analysis does not change the conclusions we draw in the main text.

4.7 Different Model Specifications: Logistic Regression and Random-effects Model

We test if different statistical model specifications would have led to different conclusions. Table S18 shows the results of a logistic regression model with the same samples. We do see the same patterns of effect sizes and statistical significance as in the main models. Table S19 compares the results of the FS between the OLS regression model with clustered standard errors that was used in the main text and a random effects model using ID as the cluster variable with the same sample. The results are very similar. Therefore, these different model specifications do not alter our conclusions drawn in the main text.

Table S18: Logistic Regression

	FE		FS		Diff. FE and FS
Applicant Education (Ref = Abitur)					
Intermediate HS	--		-0.601	***	
	--		(0.092)		
Dropout	0.115		-0.255	**	**
	(0.074)		(0.082)		
Applicant Gender (Ref = Male)					
Female	-0.024		0.144	*	
	(0.074)		(0.058)		
Applicant Ethnic Background (Ref = German)					
Turkish	-0.288	***	-0.007		**
	(0.074)		(0.055)		
Occupational Field (Ref = Electronics Technician)					
Laboratory Technician	-0.001		0.210		
	(0.128)		(0.273)		
Administration Clerk	0.434	***	0.194		
	(0.098)		(0.225)		
Media Clerk	-0.078		-0.236		
	(0.111)		(0.249)		
Achievement (Ref = Intermediate grades)					
Low grades			-1.393	***	
			(0.097)		
High grades			0.615	***	
			(0.091)		
Socio-economic status (Ref = intermediate SES)					
Low SES			0.088		
			(0.074)		
High SES			0.109		
			(0.074)		
Wave (Ref = Spring 2022)					
Fall 2022	0.133		-0.089		
	(0.077)		(0.139)		
Intercept	0.041		0.808	***	
	(0.107)		(0.241)		

Table S19: Random Effects Model

	FS Main analysis		FS Multilevel model	
Applicant Education (Ref = Abitur)				
Intermediate HS	-0.122	***	-0.122	***
	(0.018)		(0.017)	
Dropout	-0.050	**	-0.050	***
	(0.016)		(0.014)	
Applicant Gender (Ref = Male)				
Female	0.029	*	0.029	*
	(0.012)		(0.012)	
Applicant Ethnic Background (Ref = German)				
Turkish	-0.001		-0.001	
	(0.011)		(0.012)	
Occupational Field (Ref = Electronics Technician)				
Laboratory Technician	0.042		0.042	
	(0.055)		(0.051)	
Administration Clerk	0.039		0.039	
	(0.046)		(0.041)	
Media Clerk	-0.049		-0.049	
	(0.051)		(0.047)	
Achievement (Ref = Intermediate grades)				
Low grades	-0.329	***	-0.331	***
	(0.021)		(0.015)	
High grades	0.119	***	0.123	***
	(0.017)		(0.015)	
Socio-economic status (Ref = intermediate SES)				
Low SES	0.018		0.023	
	(0.015)		(0.015)	
High SES	0.022		0.020	
	(0.015)		(0.015)	
Wave (Ref = Spring 2022)				
Fall 2022	-0.018		-0.018	
	(0.028)		(0.028)	
Intercept	0.687	***	0.685	***
	(0.049)		(0.043)	
Number of observations	3,840		3,840	
R squared	0.16			
Variance (between)			0.069	
			(0.006)	
Variance (within)			0.133	
			(0.003)	

Standard errors in parentheses. Significance levels: *** p<.001, ** p<.01, * p<.05