JAMA Network Open

Original Investigation | Neurology

Association Between Hormone-Modulating Breast Cancer Therapies and Incidence of Neurodegenerative Outcomes for Women With Breast Cancer

Gregory L. Branigan, BS; Maira Soto, PhD; Leigh Neumayer, MD, MS; Kathleen Rodgers, PhD; Roberta Diaz Brinton, PhD

Abstract

IMPORTANCE The association between exposure to hormone-modulating therapy (HMT) as breast cancer treatment and neurodegenerative disease (NDD) is unclear.

OBJECTIVE To determine whether HMT exposure is associated with the risk of NDD in women with breast cancer.

DESIGN, SETTING, AND PARTICIPANTS This retrospective cohort study used the Humana claims data set from January 1, 2007, to March 31, 2017. The Humana data set contains claims from private-payer and Medicare insurance data sets from across the United States with a population primarily residing in the Southeast. Patient claims records were surveyed for a diagnosis of NDD starting 1 year after breast cancer diagnosis for the duration of enrollment in the claims database. Participants were 57 843 women aged 45 years or older with a diagnosis of breast cancer. Patients were required to be actively enrolled in Humana claims records for 6 months prior to and at least 3 years after the diagnosis of breast cancer. The analyses were conducted between January 1 and 15, 2020.

EXPOSURE Hormone-modulating therapy (selective estrogen receptor modulators, estrogen receptor antagonists, and aromatase inhibitors).

MAIN OUTCOMES AND MEASURES Patients receiving HMT for breast cancer treatment were identified. Survival analysis was used to determine the association between HMT exposure and diagnosis of NDD. A propensity score approach was used to minimize measured and unmeasured selection bias.

RESULTS Of the 326 485 women with breast cancer in the Humana data set between 2007 and 2017, 57 843 met the study criteria. Of these, 18 126 (31.3%; mean [SD] age, 76.2 [7.0] years) received HMT, whereas 39 717 (68.7%; mean [SD] age, 76.8 [7.0] years) did not receive HMT. Mean (SD) follow-up was 5.5 (1.8) years. In the propensity score–matched population, exposure to HMT was associated with a decrease in the number of women who received a diagnosis of NDD (2229 of 17 878 [12.5%] vs 2559 of 17 878 [14.3%]; relative risk, 0.89; 95% CI, 0.84-0.93; $P < .001$), Alzheimer disease (877 of 17 878 [4.9%] vs 1068 of 17 878 [6.0%]; relative risk, 0.82; 95% CI, 0.75-0.90; $P < .001$), and dementia (1862 of 17 878 [10.4%] vs 2116 of 17 878 [11.8%]; relative risk, 0.88; 95% CI, 0.83-0.93; $P < .001$). The number needed to treat was 62.51 for all NDDs, 93.61 for Alzheimer disease, and 69.56 for dementia.

CONCLUSIONS AND RELEVANCE Among patients with breast cancer, tamoxifen and steroidal aromatase inhibitors were associated with a decrease in the number who received a diagnosis of NDD, specifically Alzheimer disease and dementia.

JAMA Network Open. 2020;3(3):e201541. doi:10.1001/jamanetworkopen.2020.1541

Key Points

Question Is hormone-modulating therapy associated with neurodegenerative disease in women with breast cancer?

Findings In this cohort study of 57 843 perimenopausal- to postmenopausal-aged women with breast cancer, exposure to hormone-modulating therapy (tamoxifen and aromatase inhibitors, especially exemestane) was associated with a significant decrease in the number of women who received a diagnosis of neurodegenerative disease, most specifically Alzheimer disease.

Meaning With the increased life expectancy seen after treatment, therapy selection for breast cancer should include a careful discussion of the risks and benefits of each treatment option that may be associated with a reduced risk of neurodegenerative disease.

+ Supplemental content

Author affiliations and article information are listed at the end of this article.

Introduction

Worldwide, breast cancer is the second most common cancer in women (after skin cancer).[1] Approximately 12.8% of women will receive a diagnosis of breast cancer during their lifetime.[1] As of 2019, more than 268 000 new cases of breast cancer were diagnosed, representing 15.2% of all new cases of cancer.[2] As of 2016, 3 477 866 women were estimated to be living with breast cancer in the United States,[3] and the number of breast cancer cases continues to increase.[2] Although the rate of death from breast cancer has decreased, the mean 5-year relative survival rate is 89.9% and ranges from 98.8% to 27.4%.[2] As the number of women with a diagnosis of breast cancer increases and survival rates improve, the number of women living with breast cancer who are at risk for other diseases will escalate. Thus, the potential additional risks and benefits of the therapies to reduce breast cancer recurrence will have increasing importance.

One potential factor associated with the increase in new cases of breast cancer is an aging population. Age remains a major risk factor for breast cancer, with 61 years as the median age at diagnosis.[1] In parallel, age is the greatest risk factor for developing age-associated neurodegenerative diseases (NDDs).[4,5] Two age-associated NDDs have a greater prevalence among women: Alzheimer disease (AD) and multiple sclerosis (MS).[6,7] Women are at a 2-fold greater lifetime risk than men for developing AD, and MS is 2.8 times more prevalent among women than men.[7,8]

Currently, AD affects 1 in 9 persons in the US older than 65 years,[9] two-thirds of whom are women. In this age group, breast cancer is projected to increase and will account for almost one-third of all cancers among women in 2019.[1,2] The improvement in survival after breast cancer is associated with the long-term use of antiestrogen therapies. Often these therapies are associated with subjective reports of diminished cognitive function, an early indicator of AD risk.[10,11]

Therapies for breast cancer include surgery, radiotherapy, hormonal modulation, biologics, and chemotherapy.[12] Most breast cancers express estrogen and progesterone receptors (hormone positive) and generally respond well to surgery with or without radiotherapy and to systemic therapy with hormone modulation.[12] Given the prevalence of hormone-positive breast cancer, there are multiple breast cancer therapies targeting the estrogen receptor or the production of estrogen.[12,13] Hormone-modulating therapies (HMTs) include the selective estrogen receptor modulators (SERMs; tamoxifen and raloxifene) and aromatase inhibitors (steroidal, exemestane; nonsteroidal, anastrozole and letrozole). These drugs have been used for the treatment of estrogen receptor–positive breast cancers and have been shown to decrease estrogen's effects at the level of the breast tissue.[13] Tamoxifen is used in both the treatment and the prevention of estrogen receptor–positive breast cancer and is a common therapy for premenopausal women and an option for postmenopausal women.[12]

Analyses reported herein were designed to determine potential associations between HMT cancer therapies that affect estrogen's action and the incidence of 4 age-associated NDDs: AD, MS, Parkinson disease, and amyotrophic lateral sclerosis. Our study was conducted using a US-based population electronic medical record data set and a substantially larger number of women with breast cancer than previously reported.[14-19] Furthermore, we investigated the risk of developing multiple age-associated NDDs that occur sporadically within an aging population. We report the association of individual hormonal modulators and their drug families within the HMT category with the risk of development of age-associated NDDs.

Methods

Data Source

The Humana data set is an insurance claims data set that serves the United States, with a population primarily residing in the Southeastern region. PearlDiver is for-fee research software that facilitates interaction with individual commercial, state-based Medicaid, Medicare stand-alone prescription drug plan, group Medicare Advantage, and individual Medicare Advantage data sets.[20] The Humana

data set contains patient demographic characteristics, prescription records, and numerous other data points for patients with *Current Procedural Terminology*, *International Classification of Diseases, Ninth Revision*, and *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision* codes. As of June 2018, Humana represented 25 million patients with claims, including prescription records, from January 1, 2007, through March 31, 2017. This report follows the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) reporting guideline. This study was approved by the University of Arizona Institutional Review Board. Requirements for informed consent were waived because the data were deidentified.

Study Variables

The outcome variable was defined as the occurrence of the first NDD diagnosis for each outcome of interest based on *International Classification of Diseases, Ninth Revision, Clinical Modification* and *International Statistical Classification of Diseases and Related Health Problems, Tenth Revision, Procedure Coding System* codes in the patient's medical claims data. The HMT exposure group is defined as patients having at least 1 medication charge occurring after the diagnosis of breast cancer. Age is defined by the age at diagnosis of breast cancer. Neurodegenerative diseases included AD, dementia, Parkinson disease, MS, and amyotrophic lateral sclerosis (eTable 4 in the Supplement). Special attention was given to comorbidities known to be associated with NDD outcomes: stroke, hypertension, cardiovascular disease, type 2 diabetes, and chronic kidney disease (eTable 4 in the Supplement). For the chemotherapy analysis, intravenous therapeutics were excluded (eTable 3 in the Supplement).

Statistical Analysis

Statistical analyses were conducted between January 1 and 15, 2020. Patient demographic statistics and incidence statistics were analyzed using unpaired 2-tailed *t* tests or χ^2 tests, as appropriate, to test the significance of the differences between continuous and categorical variables. In all analyses, a 2-sided $P < .05$ was considered statistically significant.

After the unadjusted analysis, a propensity score–matched population was generated using the Bellwether-PearlDiver Interface and analyzed again using the Fisher exact test. Specifically, the association of HMT with NDD and each subtype was estimated in the unadjusted populations. To minimize confounding by indication, we used propensity score analysis to examine the association between HMT and subsequent NDD (or subtype). For the propensity score matching, using logistic regression, we first estimated for each participant the probability (ie, the propensity) of receiving HMT based on age, race/ethnicity, comorbidities of interest (**Table 1**), and Charlson Comorbidity Index score. Next, we modeled the associations between NDD and HMT, weighted by the inverse propensity score, after adjusting for stroke and chronic obstructive pulmonary disease (COPD) as statistically significant values obtained from the linear regression analysis. To examine the effect of weighting, we compared the covariates before and after adjustment for propensity score.

Kaplan-Meier curves were created using the propensity score–matched population generated using the Bellwether-PearlDiver Interface. Medication possession ratios were used to calculate the median adherence rates for each HMT type.

Results

Of the 326 485 patients with breast cancer in the Humana database, 57 843 met the inclusion and exclusion criteria and the claims enrollment period requirements for our study (**Figure 1**). An index date 1 year after the diagnosis of breast cancer was selected to rule out any diagnosis likely associated with chemotherapy or other interventions administered immediately after diagnosis before the start of HMT. Patient groups were defined according to the therapeutic intervention used. Of the 57 843 patients enrolled in the study, 18 126 (mean [SD] age, 76.2 [7.0] years) received HMT, whereas 39 717

individuals (mean [SD] age, 76.8 [7.0] years) were not treated with HMT (Figure 1). Hormone-modulating therapy was started a mean (SD) 133 (134) days after the diagnosis of breast cancer. The mean number of filled prescription days was 1078 (interquartile range, 540-1560). The drugs defined as HMT, the number of patients, and the median adherence rate for each drug are reported in eTable 1 in the Supplement. The generic drug codes used within the PearlDiver database are included in eTable 2 in the Supplement. These patient groups were then followed up for the duration of their claims data entries and surveyed for any diagnosis of NDD. The mean (SD) follow-up was 5.5 (1.8) years. The median (SD) time to diagnosis of NDD was 2.8 (2.3) years in the non-HMT exposure cohort and 2.9 (2.3) years in the HMT exposure cohort. The median (SD) time to diagnosis of AD was 3.1 (2.4) years in the non-HMT exposure cohort and 3.3 (2.2) years in the HMT exposure cohort.

The ages of patients included in the analysis ranged between 45 and 90 years of age or older, which was associated with significant differences in the age of patients receiving HMT vs no HMT. Most patient data records in the study were from women aged 55 to 69 years and women aged 80 to 89 years (Table 1). Furthermore, there were significant differences between the white patients who received HMT and white patients who did not receive HMT (13 642 of 18 126 [75.3%] vs 29 261 of 39 717 [73.7%]). Comorbidities were significantly different between patients who received HMT and

Table 1. Baseline Characteristics for Unadjusted Enrolled and Propensity Score–Matched Patients With or Without HMT Exposure

	Unadjusted cohort			Propensity score–matched cohort[a]		
	Patients, No. (%)			Patients, No. (%)		
Characteristic	HMT (n = 18 126)	No HMT (n = 39 717)	*P* Value	HMT (n = 17 878)	No HMT (n = 17 878)	*P* value
Age, y						
45-49	660 (3.6)	1701 (4.3)	<.001	647 (3.6)	820 (4.6)	<.001
50-54	709 (3.9)	1871 (4.7)		686 (3.8)	833 (4.7)	
55-59	913 (5.0)	2184 (5.5)		893 (5.0)	966 (5.4)	
60-64	1123 (6.2)	2504 (6.3)		1103 (6.2)	1140 (6.4)	
65-69	4618 (25.5)	10040 (25.3)		4557 (25.5)	4552 (25.5)	
70-74	4426 (24.4)	8493 (21.4)		4373 (24.5)	3852 (21.6)	
75-79	2930 (16.2)	6083 (15.3)		2899 (16.2)	2709 (15.2)	
80-84	1699 (9.4)	3758 (9.5)		1686 (9.4)	1647 (9.2)	
85-89	320 (1.8)	767 (1.9)		314 (1.8)	333 (1.9)	
≥90	728 (4.0)	2316 (5.8)		720 (4.0)	1026 (5.7)	
Race/ethnicity						
Unknown	2051 (11.3)	4762 (12.0)	.04	1995 (11.2)	2208 (12.4)	.01
White	13 642 (75.3)	29 261 (73.7)		13 443 (75.2)	13 105 (73.3)	
Black	2000 (11.0)	4574 (11.5)		1969 (11.0)	2085 (11.7)	
Other	151 (0.8)	318 (0.8)		149 (0.8)	140 (0.8)	
Asian	95 (0.5)	213 (0.5)		93 (0.5)	99 (0.6)	
Hispanic	200 (1.1)	481 (1.2)		198 (1.1)	203 (1.1)	
North American Native	32 (0.2)	63 (0.2)		31 (0.2)	38 (0.2)	
Comorbidities						
Type 2 diabetes	1079 (6.0)	2188 (5.5)	.03	998 (5.6)	1030 (5.8)	.48
CVD	374 (2.1)	995 (2.5)	.01	370 (2.1)	423 (2.4)	.06
Hypertension	2459 (13.6)	5500 (13.9)	.37	2423 (13.6)	2438 (13.6)	.83
CKD	391 (2.2)	987 (2.5)	.02	388 (2.2)	417 (2.3)	.32
Stroke	355 (2.0)	904 (2.3)	.02	349 (2.0)	399 (2.2)	.07
COPD	184 (1.0)	569 (1.4)	<.001	183 (1.0)	235 (1.3)	.01
Charlson Comorbidity Index						
0-4	14 174 (78.2)	30 654 (77.2)	<.001	13 982 (78.2)	14 536 (81.3)	<.001
5-10	3647 (20.1)	8077 (20.3)		3599 (20.1)	3068 (17.2)	
≥11	280 (1.5)	869 (2.2)		297 (1.7)	274 (1.5)	

Abbreviations: CKD, chronic kidney disease; COPD, chronic obstructive pulmonary disease; CVD, cardiovascular disease; HMT, hormone-modulating therapy.

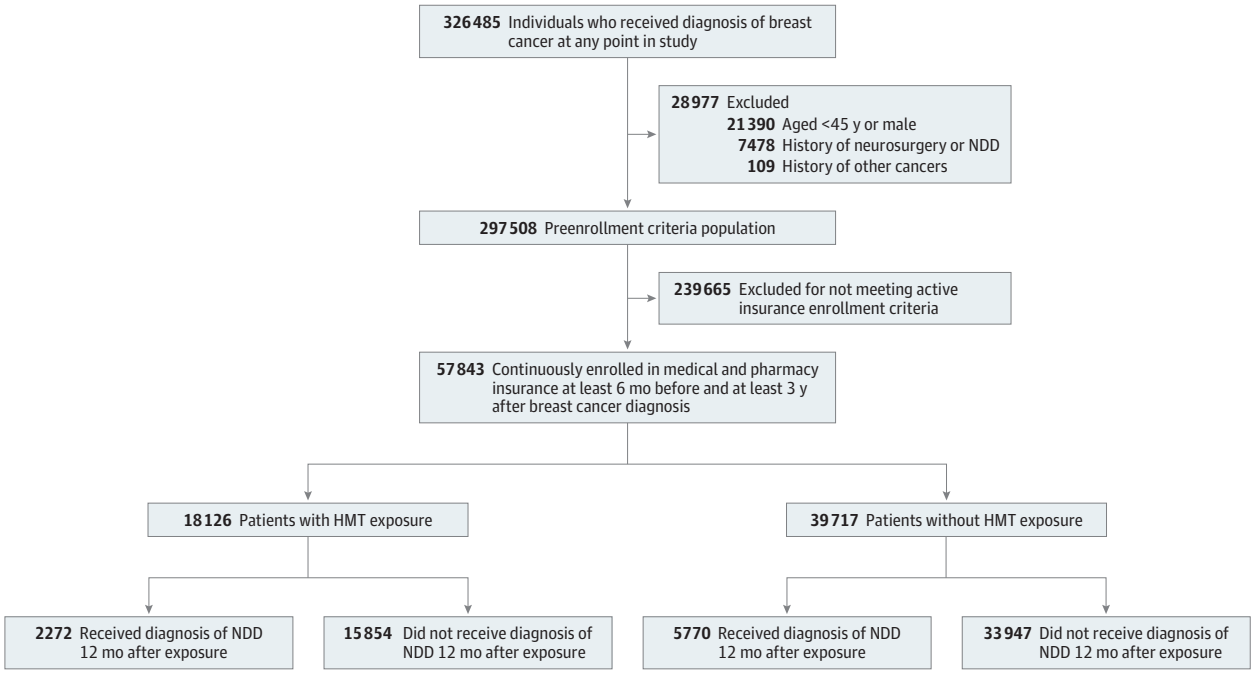
[a] Adjusted for history of stroke and COPD before the diagnosis of breast cancer.

those who did not (diabetes, 1079 of 18 126 [6.0%] vs 2188 of 39 717 [5.5%]; cardiovascular disease, 374 of 18 126 [2.1%] vs 995 of 39 717 [2.5%]; chronic kidney disease, 391 of 18 126 [2.2%] vs 987 of 39 717 [2.5%]; stroke, 355 of 18 126 [2.0%] vs 904 of 39 717 [2.3%]; COPD, 184 of 18 126 [1.0%] vs 569 of 39 717 [1.4%]). Last, there were significant differences between patients who received HMT and those who did not in Charlson Comorbidity Index score categories of 0 to 4 (14 174 of 18 126 [78.2%] vs 30 654 of 39 717 [77.2%]) and 11 or more (280 of 18 126 [1.5%] vs 869 of 39 717 [2.2%]). The linear regression of the comorbidities of interest and therapy selection indicated significant differences in therapy selection for those with a diagnosis of COPD and stroke. To address the association of comorbidities, propensity score matching was performed to create representative groups that controlled for COPD and stroke history. The demographic characteristics of the populations generated by propensity matching appear in Table 1 and show the same statistical differences between treatment groups as in the unadjusted data.

Analyses of unadjusted population data indicated that HMT exposure compared with no HMT exposure was associated with a significant decrease in the incidence of AD (900 of 18 126 [5.0%] vs 2436 of 39 717 [6.1%]; relative risk [RR], 0.81; 95% CI, 0.75-0.87; $P < .001$), dementia (1894 of 18 126 [10.4%] vs 4892 of 39 717 [12.3%]; RR, 0.85; 95% CI, 0.80-0.89; $P < .001$), non-AD dementia (1079 of 18 126 [6.0%] vs 2657 of 39 717 [6.7%]; RR, 0.89; 95% CI, 0.83-0.89; $P < .001$), and all NDDs (2272 of 18 126 [12.5%] vs 5770 of 39 717 [14.5%]; RR, 0.86; 95% CI, 0.82-0.90; $P < .001$) (**Table 2**). The outcomes of multiple sclerosis and Parkinson disease were not significantly different among those with HMT exposure. Although not significant, the incidence of amyotrophic lateral sclerosis appeared to be increased among patients exposed to HMT. No change was observed in the association of risk reduction with HMT in the unadjusted population after the removal of patients who received intravenous chemotherapy, which indicates that the association is likely due to the presence of the HMT and not to the cytotoxic effects of chemotherapy (eTable 3, eTable 5, and eFigure 2 in the Supplement).

In the propensity score–matched population, AD- and dementia-associated outcomes were specifically analyzed because these diagnoses were statistically significant in the unadjusted

Figure 1. Study Design and Patient Breakdown



HMT indicates hormone-modulating therapy; NDD, neurodegenerative disease.

populations (Table 2). The results of the χ^2 analysis in the matched patient group indicated that the significant decreases in the numbers of patients with a diagnosis of AD (877 of 17 878 [4.9%] vs 1068 of 17 878 [6.0%]; RR, 0.82; 95% CI, 0.75-0.90), dementia (1862 of 17 878 [10.4%] vs 2116 of 17 878 [11.8%]; RR, 0.88; 95% CI, 0.83-0.93; *P* < .001), and all NDDs ((2229 of 17 878 [12.5%] vs 2559 of 17 878 [14.3%]; RR, 0.89; 95% CI, 0.84-0.93; *P* < .001) who received HMT were sustained. The analysis of non-AD dementia outcomes, such as vascular dementia or Lewy body dementia, was no longer significant in the matched population. The propensity score–matched population was then used to generate Kaplan-Meier survival curves for NDD-free survival for each NDD subtype to evaluate the rate and percentage of the population who developed each disease (eFigure 1 in the Supplement). Changes in the rate of disease incidence between patients receiving HMT and patients not receiving HMT mirror the results seen in the χ^2 analysis.

In the propensity score–matched population, groups were stratified by age (65-69, 70-74, 75-79, and 80-84 years) to determine a potential age-specific association in overall NDD and AD risk outcomes (**Figure 2**; eFigure 3 in the Supplement). For patients 65 to 69 years of age, there was no significant difference between patients receiving HMT and patients not receiving HMT in the risk of NDD or AD; SDs overlapped in the 5-year analysis. In contrast, increasing age was associated with a greater reduction of risk for all NDDs in women receiving HMT; SDs did not overlap in the 5-year analysis and were divergent. This association between age and reduced incidence of NDD was replicated in the survival curves specifically for AD (Figure 2; eFigure 3 in the Supplement).

To address the potential selectivity of the action of HMT to reduce the incidence of AD, analysis of the incidence of AD by therapeutic mechanism of action and tissue specificity was conducted. Patients receiving HMT were divided into 3 groups based on therapeutic mechanism (**Figure 3**): tamoxifen (n = 5335), raloxifene (n = 1972), or aromatase inhibitors (n = 16 032). Tamoxifen showed the strongest associated decreased risk for each disease (RR, 0.84; 95% CI, 0.80-0.88; *P* < .001). Raloxifene, while also a SERM, had no significant association with the RR for any NDD (RR, 1.04; 95% CI, 0.93-1.16; *P* = .54). The aromatase inhibitors, known to block the enzyme responsible for the conversion of testosterone and androstenedione to estrogen,[21] also was associated with a reduction in the RR for the development of the NDDs of interest (RR, 0.83; 95% CI, 0.76-0.89; *P* < .001). Thus, the reduced RR seen in the HMT-treated population is primarily associated with patients receiving tamoxifen or aromatase inhibitors (Figure 3).

Table 2. Relative Risk of Unadjusted and Propensity Score–Matched Patients Developing NDDs After Receiving HMT

Characteristic	All NDDs	AD	Dementia	Non-AD Dementia	MS	PD	ALS
Unadjusted cohort							
Patients who received HMT, [a] No. (%)	2272 (12.5)	900 (5.0)	1894 (10.5)	1079 (6.0)	129 (0.7)	328 (1.8)	15 (0.1)
Patients who did not receive HMT, [b] No. (%)	5770 (14.5)	2436 (6.1)	4892 (12.3)	2657 (6.7)	306 (0.8)	755 (1.9)	22 (0.1)
Relative risk (95% CI)	0.86 (0.82-0.90)	0.81 (0.75-0.87)	0.85 (0.80-0.89)	0.89 (0.83-0.89)	0.92 (0.75-1.13)	0.95 (0.84-1.08)	1.49 (0.78-2.85)
NNT	50.17	85.61	53.53	137.21	1702	1094	3655
P value	<.001	<.001	<.001	<.001	.46	.47	.29
Propensity score–matched cohort[c]							
Patients who received HMT, [a] No. (%)	2229 (12.5)	877 (4.9)	1862 (10.4)	1040 (5.8)	NA	NA	NA
Patients who did not receive HMT, [b] No. (%)	2559 (14.3)	1068 (6.0)	2116 (11.8)	1106 (6.2)	NA	NA	NA
Relative risk (95% CI)	0.89 (0.84-0.93)	0.82 (0.75-0.90)	0.88 (0.83-0.93)	0.94 (0.87-1.02)	NA	NA	NA
NNT	62.51	93.61	69.56	255.4	NA	NA	NA
P value	<.001	<.001	<.001	.15	NA	NA	NA

Abbreviations: AD, Alzheimer disease; ALS, amyotrophic lateral sclerosis; HMT, hormone-modulating therapy; MS, multiple sclerosis; NA, not applicable; NDD, neurodegenerative disease; NNT, number needed to treat; PD, Parkinson disease.

[a] Unadjusted cohort, 18 126 patients; propensity score–matched cohort, 17 878 patients.

[b] Unadjusted cohort, 39 717 patients; propensity score–matched cohort, 17 878 patients.

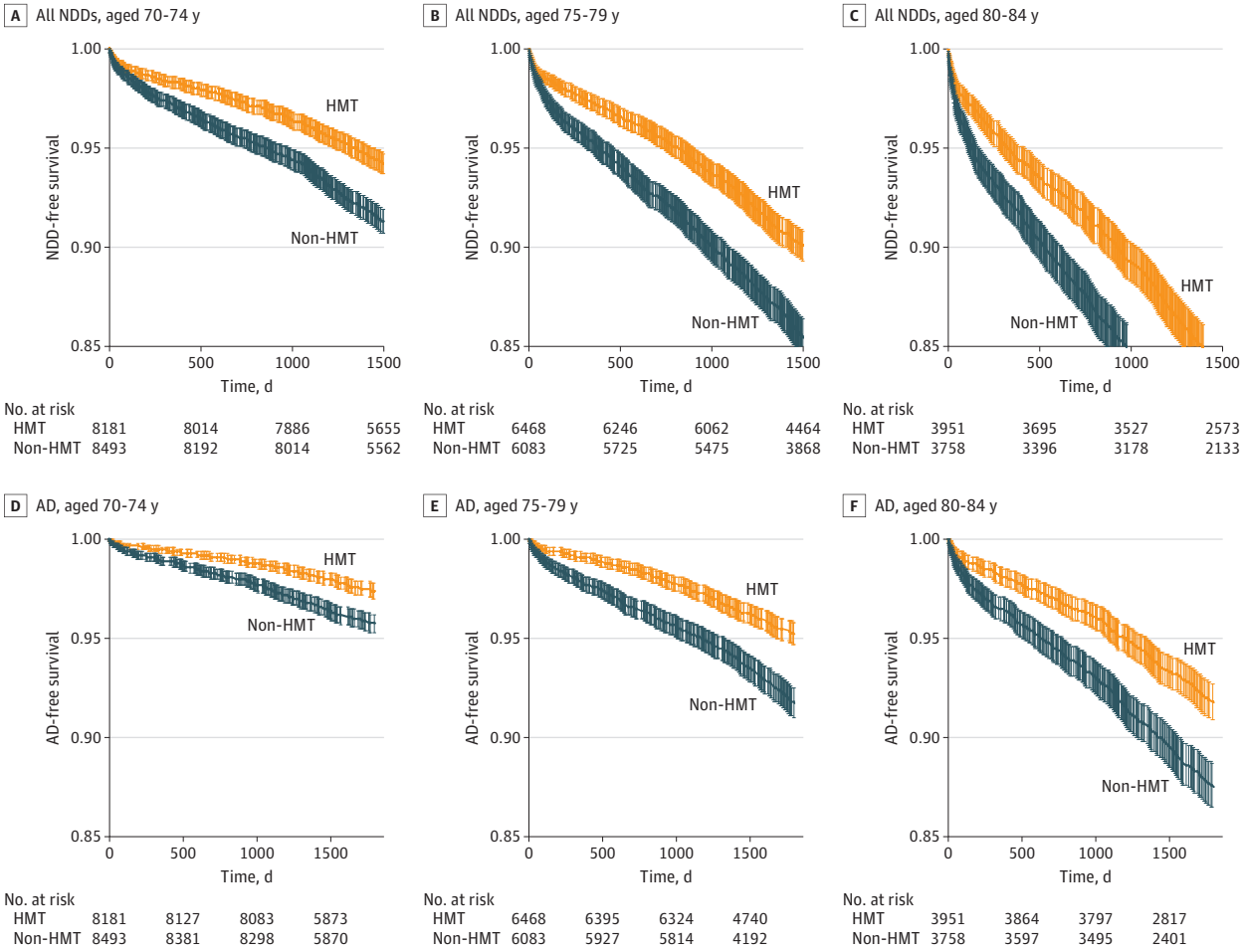
[c] Adjusted for history of stroke and chronic obstructive pulmonary disease before the diagnosis of breast cancer.

Aromatase inhibitors are either nonsteroidal (anastrozole and letrozole) or steroidal (exemestane), which led to questions surrounding possible differences in their potential protective effects seen in AD and dementia (eFigure 4 in the Supplement). To address this question, analysis of the individual drugs in these classes was undertaken to determine potential differences in their association with the reduction of NDD risk. Outcomes of this analysis indicated that patients receiving exemestane, the steroidal aromatase inhibitor, had a statistically significant decrease in the incidence of AD and dementia compared with patients receiving the nonsteroidal drugs anastrozole and letrozole (eFigure 4 in the Supplement). However, both types of aromatase inhibitors exerted a protective association compared with patients with breast cancer who were not receiving any HMT.

Discussion

The first goal of the National Plan to Address Alzheimer Disease is to prevent AD by 2025.[22] The short time horizon for achieving this goal is challenging but not insurmountable. One potentially effective strategy is to identify populations of individuals at risk for AD who have received therapeutic interventions that modify the risk of a diagnosis of AD. Although complex, multiple epidemiologic studies indicate that estrogen hormone therapy is associated with a reduced risk of AD.[19,23-32] The loss of estrogen in the brain can be a factor associated with the 2-fold greater lifetime risk of

Figure 2. Age-Dependent Reduction in Risk for All Neurodegenerative Diseases (NDDs) and Alzheimer Disease (AD) Associated With Hormone-Modulating Therapy (HMT) Exposure



A significantly decreased risk of diagnosis of both overall NDDs and, more specifically, AD was observed for patients treated with HMT vs those not treated with HMT.

developing AD.[33] The patient group in our study is one such group, women receiving antiestrogen therapies in midlife and later life.

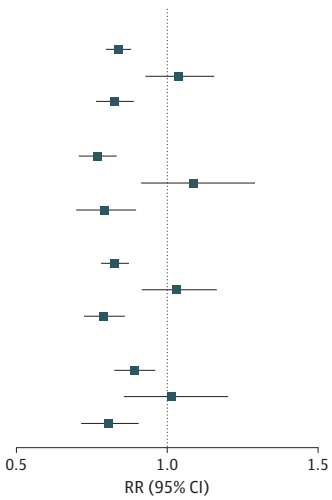
Treatments for age-associated NDDs remain an unmet need and challenge because each NDD has a complex and multifaceted pathophysiology, resulting in few therapeutic options for treatment or prevention. For AD, HMTs (such as estrogen therapy) have been shown to be associated with the onset of AD,[34] whereas estrogen therapy has been shown to be ineffective as a treatment for those with a diagnosis of AD.[35-43] Similarly, the failure of trials evaluating the use of SERMs in the treatment of AD could be due to the fact that the intervention was started past the therapeutic window for HMTs.[18,44-46] Here, we show the beneficial effects of exposure to HMT as a prophylactic treatment for the potential prevention of AD.

Although our preliminary results show a decrease in the number of NDDs overall in patients receiving HMTs, the predominant association was a significant decrease in the RR of dementia-related outcomes. The lack of significance in nondementia NDDs is likely owing to a decreased incidence of MS, Parkinson disease, and amyotrophic lateral sclerosis in our study population as well as in the general population overall. In the dementia-associated diseases, there was a larger association observed for AD outcomes despite the disease being less prevalent than non-AD dementias. This trend was also evident in the propensity score–matched populations, in which, after controlling for cerebrovascular and respiratory disease, the protective effects of HMT occurred exclusively for AD outcomes. This finding points to a potential specific biological mechanism associated with estrogen loss in the brain in the pathophysiology of AD. Alternatively, the results from the propensity score–matched populations could be due to a stronger association of cerebrovascular and respiratory disease with non-AD dementia.

Previous reports have typically focused on a single drug and disease outcome, which limits the scope of the translational outcomes. We have included all HMTs used in breast cancer treatment as well as multiple age-associated NDDs. These treatments generally fall into 2 classes: SERMs and aromatase inhibitors. Tamoxifen and aromatase inhibitors exhibited the strongest association with reducing the incidence of AD and related dementia. The protection associated with the SERMs was exclusively due to tamoxifen and not to raloxifene. This finding might explain why previous studies using SERMs were not found to be effective because raloxifene was the focus of several of these AD trials.[18,47] Mechanistically, tamoxifen and raloxifene are known to act in a tissue-specific manner. Tamoxifen and raloxifene are known estrogen receptor antagonists in breast tissue but show

Figure 3. Relative Risk (RR) of All Neurodegenerative Disease, Alzheimer Disease, and Dementia Outcomes With Aromatase Inhibitors, Raloxifene, and Tamoxifen

Outcome	RR (95% CI)	*P* Value
Neurodegenerative disease		
Aromatase inhibitor	0.84 (0.88-0.80)	<.001
Raloxifene	1.04 (1.16-0.93)	.54
Tamoxifen	0.83 (0.89-0.76)	<.001
Alzheimer disease		
Aromatase inhibitor	0.77 (0.84-0.71)	<.001
Raloxifene	1.09 (1.29-0.91)	.37
Tamoxifen	0.79 (0.90-0.70)	<.001
Dementia		
Aromatase inhibitor	0.83 (0.87-0.78)	<.001
Raloxifene	1.03 (1.17-0.92)	.62
Tamoxifen	0.79 (0.86-0.73)	<.001
Non-Alzheimer dementia		
Aromatase inhibitor	0.89 (0.96-0.83)	.01
Raloxifene	1.02 (1.20-0.86)	.90
Tamoxifen	0.81 (0.91-0.72)	<.001



Tamoxifen and the aromatase inhibitors were associated with significantly reduced RRs for all neurodegenerative diseases, Alzheimer disease, dementia, and non-Alzheimer dementia in the hormone-modulating therapy treatment groups. More specifically, the steroidal aromatase inhibitor (exemestane) had a greater association than nonsteroidal aromatase inhibitors (anastrozole and letrozole) with reduced RR of neurodegenerative disease outcomes.

divergent actions in uterine tissue[48-51] and brain tissue.[52,53] Aromatase inhibitors are known to act systemically to decrease the amount of estrogen.[21] However, a recent study suggests that there also may be divergent actions of aromatase inhibitors in specific brain regions.[54] Alternatively, upstream precursors of estrogen, such as testosterone and androstenedione, which may be increased by aromatase inhibitors, can be associated with cognition.[55,56] If tamoxifen and aromatase inhibitors are acting to increase estrogen-related actions in brain tissue, the argument for the protective association of estrogen with AD-related outcomes is strengthened.

Limitations

This analysis has several limitations. First, it is a retrospective analysis of a claims database. The patients included may have obtained services outside of those included in this database. Second, there could be factors, known and unknown, that even with propensity matching may not be adequately addressed. Third, the rate of women with a diagnosis of breast cancer who were exposed to HMT is seemingly low (approximately one-third of the sample). Although this proportion seems low, there are other data that show that, while adherence to endocrine therapy in clinical trials is high, adherence in clinical practice is substantially lower, with only about 50% of women completing 5 years of therapy.[57] The factors associated with nonadherence include perception of a low risk of recurrence, adverse effects (perceived or real), costs, suboptimal patient-physician communication, and lack of social support.[57] Moreover, in the age-stratified data (Figure 2; eFigure 3 in the Supplement), we show that a low level of HMT exposure was associated with the population younger than 70 years (8237 of 26 627 patients aged <70 years received HMT [31.2%]), whereas patients 70 years or older received HMT at a rate of 50% (18 600 of 36 934). In addition, HMT exposure is assessed by filled prescription charges to Humana, indicating that a drug has been picked up by a patient; however, data on specific breast pathologic condition, on contraindications for therapy, and on therapeutics actually prescribed for a patient cannot be assessed in this data set.

Conclusions

This study found that among patients with breast cancer, tamoxifen and steroidal aromatase inhibitors were associated with a decrease in the number who received a diagnosis of NDD, specifically AD and dementia. As we advance in our abilities to prevent, treat, and cure cancer, discussions around optimal care will need to include understanding the long-term outcomes of therapy selection for age-related NDDs. The fact that breast cancer is the second most common cancer in women (after skin cancer) and that women are disproportionately affected by AD and related dementia provides us with an opportunity to reduce the global disease burden of NDDs.[58]

ARTICLE INFORMATION

Accepted for Publication: January 26, 2020.

Published: March 24, 2020. doi:10.1001/jamanetworkopen.2020.1541

Corresponding Author: Roberta Diaz Brinton, PhD, Center for Innovation in Brain Science, University of Arizona, BSRL, 1230 N Cherry Ave, Tucson, AZ 85721 (rbrinton@email.arizona.edu).

Author Affiliations: Center for Innovation in Brain Science, University of Arizona, Tucson (Branigan, Soto, Rodgers, Brinton); Department of Pharmacology, University of Arizona College of Medicine, Tucson (Branigan, Soto, Rodgers, Brinton); MD-PhD Training Program, University of Arizona College of Medicine, Tucson (Branigan); Department of Surgery, University of Arizona College of Medicine, Tucson (Neumayer); Department of Obstetrics and Gynecology, University of Arizona College of Medicine, Tucson (Neumayer); Department of Neurology, University of Arizona College of Medicine, Tucson (Brinton).

Author Contributions: Mr Branigan and Dr Soto had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis. Mr Branigan and Dr Soto are co–first authors.

Concept and design: Branigan, Soto, Rodgers, Brinton.

Acquisition, analysis, or interpretation of data: Soto, Neumayer, Rodgers, Brinton.

Drafting of the manuscript: Branigan, Soto, Brinton.

Critical revision of the manuscript for important intellectual content: All authors.

Statistical analysis: Branigan, Soto, Rodgers.

Obtained funding: Brinton.

Administrative, technical, or material support: Soto, Rodgers.

Supervision: Soto, Neumayer, Rodgers, Brinton.

Conflict of Interest Disclosures: Dr Brinton reported receiving grants from the Women's Alzheimer's Movement and the National Institute on Aging during the conduct of the study. No other disclosures were reported.

Funding/Support: This study was supported by grants from the Women's Alzheimer's Movement and National Institute on Aging grants P01AG026572 (Perimenopause in Brain Aging and Alzheimer's Disease), T32AG061897 (Translational Research in Alzheimer's Disease and Related Dementias [TRADD]), and R37AG053589 (Aging and Estrogenic Control of the Bioenergetic System in Brain) to Dr Brinton.

Role of the Funder/Sponsor: The National Institutes of Health and the Women's Alzheimer's Movement Foundation had no role in the design and conduct of the study; collection, management, analysis, and interpretation of the data; preparation, review, or approval of the manuscript; and decision to submit the manuscript for publication.

REFERENCES

1. Siegel RL, Miller KD, Jemal A. Cancer statistics, 2019. *CA Cancer J Clin*. 2019;69(1):7-34. doi:10.3322/caac.21551

2. Surveillance, Epidemiology, and End Results Program. SEER cancer statistics review (CSR) 1975-2016. Accessed January 1, 2020. https://seer.cancer.gov/csr/1975_2016/

3. Rojas K, Stuckey A. Breast cancer epidemiology and risk factors. *Clin Obstet Gynecol*. 2016;59(4):651-672. doi:10.1097/GRF.0000000000000239

4. Johnson IP. Age-related neurodegenerative disease research needs aging models. *Front Aging Neurosci*. 2015;7:168. doi:10.3389/fnagi.2015.00168

5. Niccoli T, Partridge L. Ageing as a risk factor for disease. *Curr Biol*. 2012;22(17):R741-R752. doi:10.1016/j.cub.2012.07.024

6. Riedel BC, Thompson PM, Brinton RD. Age, *APOE* and sex: triad of risk of Alzheimer's disease. *J Steroid Biochem Mol Biol*. 2016;160:134-147. doi:10.1016/j.jsbmb.2016.03.012

7. Wallin MT, Culpepper WJ, Campbell JD, et al; US Multiple Sclerosis Prevalence Workgroup. The prevalence of MS in the United States: a population-based estimate using health claims data. *Neurology*. 2019;92(10):e1029-e1040. doi:10.1212/WNL.0000000000007035

8. Noonan CW, Kathman SJ, White MC. Prevalence estimates for MS in the United States and evidence of an increasing trend for women. *Neurology*. 2002;58(1):136-138. doi:10.1212/WNL.58.1.136

9. Alzheimer's Association. 2018 Alzheimer's disease facts and figures. *Alzheimers Dement*. 2018;14(3):367-429. doi:10.1016/j.jalz.2018.02.001

10. Biro E, Kahan Z, Kalman J, et al. Cognitive functioning and psychological well-being in breast cancer patients on endocrine therapy. *In Vivo*. 2019;33(4):1381-1392. doi:10.21873/invivo.11615

11. Nead KT, Gaskin G, Chester C, Swisher-McClure S, Leeper NJ, Shah NH. Association between androgen deprivation therapy and risk of dementia. *JAMA Oncol*. 2017;3(1):49-55. doi:10.1001/jamaoncol.2016.3662

12. National Comprehensive Cancer Network. Practice guidelines in oncology: breast cancer: version 2.2017. Accessed September 15, 2019. https://www.nccn.org/professionals/physician_gls/pdf/breast_blocks.pdf

13. Chen WY. Selective estrogen receptor modulators and aromatase inhibitors for breast cancer prevention. UpToDate. Accessed July 31, 2019. https://www.uptodate.com/contents/selective-estrogen-receptor-modulators-and-aromatase-inhibitors-for-breast-cancer-prevention

14. Sun LM, Chen HJ, Liang JA, Kao CH. Long-term use of tamoxifen reduces the risk of dementia: a nationwide population-based cohort study. *QJM*. 2016;109(2):103-109. doi:10.1093/qjmed/hcv072

15. Liao KF, Lin CL, Lai SW. Nationwide case-control study examining the association between tamoxifen use and Alzheimer's disease in aged women with breast cancer in Taiwan. *Front Pharmacol*. 2017;8:612. doi:10.3389/fphar.2017.00612

16. Latourelle JC, Dybdahl M, Destefano AL, Myers RH, Lash TL. Risk of Parkinson's disease after tamoxifen treatment. *BMC Neurol*. 2010;10:23. doi:10.1186/1471-2377-10-23

17. Kesler SR, Rao V, Ray WJ, Rao A; Alzheimer's Disease Neuroimaging Initiative. Probability of Alzheimer's disease in breast cancer survivors based on gray-matter structural network efficiency. *Alzheimers Dement (Amst)*. 2017;9:67-75. doi:10.1016/j.dadm.2017.10.002

18. Henderson VW, Ala T, Sainani KL, et al. Raloxifene for women with Alzheimer disease: a randomized controlled pilot trial. *Neurology*. 2015;85(22):1937-1944. doi:10.1212/WNL.0000000000002171

19. Tang M-X, Jacobs D, Stern Y, et al. Effect of oestrogen during menopause on risk and age at onset of Alzheimer's disease. *Lancet*. 1996;348(9025):429-432. doi:10.1016/S0140-6736(96)03356-9

20. PearlDiver. Healthcare research. Accessed February 11, 2020. http://www.pearldiverinc.com

21. Miller WR. Aromatase inhibitors: mechanism of action and role in the treatment of breast cancer. *Semin Oncol*. 2003;30(4)(suppl 14):3-11. doi:10.1016/S0093-7754(03)00302-6

22. Office of the Assistant Secretary for Planning and Evaluation. National plan to address Alzheimer's disease: 2018 update. Accessed January 1, 2020. https://aspe.hhs.gov/report/national-plan-address-alzheimers-disease-2018-update

23. Cholerton B, Gleason CE, Baker LD, Asthana S. Estrogen and Alzheimer's disease: the story so far. *Drugs Aging*. 2002;19(6):405-427. doi:10.2165/00002512-200219060-00002

24. Fillit H, Cummings J; Alzheimer's Disease (AD) Managed Care Advisory Council. Practice guidelines for the diagnosis and treatment of Alzheimer's disease in a managed care setting: part II—pharmacologic therapy. *Manag Care Interface*. 2000;13(1):51-56.

25. Fillit HM. The role of hormone replacement therapy in the prevention of Alzheimer disease. *Arch Intern Med*. 2002;162(17):1934-1942. doi:10.1001/archinte.162.17.1934

26. Janicki SC, Schupf N. Hormonal influences on cognition and risk for Alzheimer's disease. *Curr Neurol Neurosci Rep*. 2010;10(5):359-366. doi:10.1007/s11910-010-0122-6

27. Merlo S, Spampinato SF, Sortino MA. Estrogen and Alzheimer's disease: still an attractive topic despite disappointment from early clinical results. *Eur J Pharmacol*. 2017;817:51-58. doi:10.1016/j.ejphar.2017.05.059

28. Mulnard RA, Cotman CW, Kawas C, et al. Estrogen replacement therapy for treatment of mild to moderate Alzheimer disease: a randomized controlled trial: Alzheimer's Disease Cooperative Study. *JAMA*. 2000;283(8):1007-1015. doi:10.1001/jama.283.8.1007

29. Simpkins JW, Perez E, Wang X, Yang S, Wen Y, Singh M. The potential for estrogens in preventing Alzheimer's disease and vascular dementia. *Ther Adv Neurol Disord*. 2009;2(1):31-49. doi:10.1177/1756285608100427

30. Villa A, Vegeto E, Poletti A, Maggi A. Estrogens, neuroinflammation, and neurodegeneration. *Endocr Rev*. 2016;37(4):372-402. doi:10.1210/er.2016-1007

31. Zandi PP, Carlson MC, Plassman BL, et al; Cache County Memory Study Investigators. Hormone replacement therapy and incidence of Alzheimer disease in older women: the Cache County Study. *JAMA*. 2002;288(17):2123-2129. doi:10.1001/jama.288.17.2123

32. Zhao L, O'Neill K, Brinton RD. Estrogenic agonist activity of ICI 182,780 (Faslodex) in hippocampal neurons: implications for basic science understanding of estrogen signaling and development of estrogen modulators with a dual therapeutic profile. *J Pharmacol Exp Ther*. 2006;319(3):1124-1132. doi:10.1124/jpet.106.109504

33. Brinton RD, Yao J, Yin F, Mack WJ, Cadenas E. Perimenopause as a neurological transition state. *Nat Rev Endocrinol*. 2015;11(7):393-405. doi:10.1038/nrendo.2015.82

34. Fontes F, Pereira S, Castro-Lopes JM, Lunet N. A prospective study on the neurological complications of breast cancer and its treatment: updated analysis three years after cancer diagnosis. *Breast*. 2016;29:31-38. doi:10.1016/j.breast.2016.06.013

35. Sherwin BB. Estrogen and cognitive functioning in women: lessons we have learned. *Behav Neurosci*. 2012;126(1):123-127. doi:10.1037/a0025539

36. Shumaker SA, Legault C, Rapp SR, et al; WHIMS Investigators. Estrogen plus progestin and the incidence of dementia and mild cognitive impairment in postmenopausal women: the Women's Health Initiative Memory Study: a randomized controlled trial. *JAMA*. 2003;289(20):2651-2662. doi:10.1001/jama.289.20.2651

37. McCarrey AC, Resnick SM. Postmenopausal hormone therapy and cognition. *Horm Behav*. 2015;74:167-172. doi:10.1016/j.yhbeh.2015.04.018

38. Rapp SR, Espeland MA, Shumaker SA, et al; WHIMS Investigators. Effect of estrogen plus progestin on global cognitive function in postmenopausal women: the Women's Health Initiative Memory Study: a randomized controlled trial. *JAMA*. 2003;289(20):2663-2672. doi:10.1001/jama.289.20.2663

39. Chen S, Nilsen J, Brinton RD. Dose and temporal pattern of estrogen exposure determines neuroprotective outcome in hippocampal neurons: therapeutic implications. *Endocrinology*. 2006;147(11):5303-5313. doi:10.1210/en.2006-0495

40. Breuer B, Anderson R. The relationship of tamoxifen with dementia, depression, and dependence in activities of daily living in elderly nursing home residents. *Women Health*. 2000;31(1):71-85. doi:10.1300/J013v31n01_05

41. Vogelvang TE, Mijatovic V, van der Mooren MJ, et al. Effect of raloxifene and hormone therapy on serum markers of brain and whole-body cholesterol metabolism in postmenopausal women. *Maturitas*. 2005;50(4):312-320. doi:10.1016/j.maturitas.2004.08.004

42. Chen X, He X, Tao L, et al. The working memory and dorsolateral prefrontal-hippocampal functional connectivity changes in long-term survival breast cancer patients treated with tamoxifen. *Int J Neuropsychopharmacol*. 2017;20(5):374-382. doi:10.1093/ijnp/pyx008

43. Le Rhun E, Delbeuck X, Lefeuvre-Plesse C, et al. A phase III randomized multicenter trial evaluating cognition in post-menopausal breast cancer patients receiving adjuvant hormonotherapy. *Breast Cancer Res Treat*. 2015;152(3):569-580. doi:10.1007/s10549-015-3493-1

44. Barron TI, Connolly R, Bennett K, Feely J, Kennedy MJ. Early discontinuation of tamoxifen: a lesson for oncologists. *Cancer*. 2007;109(5):832-839. doi:10.1002/cncr.22485

45. Hochner-Celnikier D. Pharmacokinetics of raloxifene and its clinical application. *Eur J Obstet Gynecol Reprod Biol*. 1999;85(1):23-29. doi:10.1016/S0301-2115(98)00278-4

46. Nickelsen T, Lufkin EG, Riggs BL, Cox DA, Crook TH. Raloxifene hydrochloride, a selective estrogen receptor modulator: safety assessment of effects on cognitive function and mood in postmenopausal women. *Psychoneuroendocrinology*. 1999;24(1):115-128. doi:10.1016/S0306-4530(98)00041-9

47. Yaffe K. Estrogens, selective estrogen receptor modulators, and dementia: what is the evidence? *Ann N Y Acad Sci*. 2001;949:215-222. doi:10.1111/j.1749-6632.2001.tb04024.x

48. Cano A, Hermenegildo C. The endometrial effects of SERMs. *Hum Reprod Update*. 2000;6(3):244-254. doi:10.1093/humupd/6.3.244

49. Hu R, Hilakivi-Clarke L, Clarke R. Molecular mechanisms of tamoxifen-associated endometrial cancer. [review]. *Oncol Lett*. 2015;9(4):1495-1501. doi:10.3892/ol.2015.2962

50. Polin SA, Ascher SM. The effect of tamoxifen on the genital tract. *Cancer Imaging*. 2008;8(1):135-145. doi:10.1102/1470-7330.2008.0020

51. Rey JR, Cervino EV, Rentero ML, Crespo EC, Alvaro AO, Casillas M. Raloxifene: mechanism of action, effects on bone tissue, and applicability in clinical traumatology practice. *Open Orthop J*. 2009;3:14-21. doi:10.2174/1874325000903010014

52. O'Neill K, Chen S, Brinton RD. Impact of the selective estrogen receptor modulator, raloxifene, on neuronal survival and outgrowth following toxic insults associated with aging and Alzheimer's disease. *Exp Neurol*. 2004;185(1):63-80. doi:10.1016/j.expneurol.2003.09.005

53. O'Neill K, Chen S, Diaz Brinton R. Impact of the selective estrogen receptor modulator, tamoxifen, on neuronal outgrowth and survival following toxic insults associated with aging and Alzheimer's disease. *Exp Neurol*. 2004;188(2):268-278. doi:10.1016/j.expneurol.2004.01.014

54. Gervais NJ, Remage-Healey L, Starrett JR, Pollak DJ, Mong JA, Lacreuse A. Adverse effects of aromatase inhibition on the brain and behavior in a nonhuman primate. *J Neurosci*. 2019;39(5):918-928. doi:10.1523/JNEUROSCI.0353-18.2018

55. Barrett-Connor E, Goodman-Gruen D. Cognitive function and endogenous sex hormones in older women. *J Am Geriatr Soc*. 1999;47(11):1289-1293. doi:10.1111/j.1532-5415.1999.tb07427.x

56. Barrett-Connor E, Goodman-Gruen D, Patay B. Endogenous sex hormones and cognitive function in older men. *J Clin Endocrinol Metab*. 1999;84(10):3681-3685. doi:10.1210/jc.84.10.3681

57. Chlebowski RT, Kim J, Haque R. Adherence to endocrine therapy in breast cancer adjuvant and prevention settings. *Cancer Prev Res (Phila)*. 2014;7(4):378-387. doi:10.1158/1940-6207.CAPR-13-0389

58. Cummings J, Lee G, Ritter A, Zhong K. Alzheimer's disease drug development pipeline: 2018. *Alzheimers Dement (N Y)*. 2018;4:195-214.

SUPPLEMENT.

Auditing: A Journal of Practice & Theory
Vol. 33, No. 4
November 2014
pp. 197–219

American Accounting Association
DOI: 10.2308/ajpt-50792

Do Big 4 Auditors Provide Higher Audit Quality after Controlling for the Endogenous Choice of Auditor?

John Daniel Eshleman and Peng Guo

SUMMARY: Recent research suggests that Big 4 auditors do not provide higher audit quality than other auditors, after controlling for the endogenous choice of auditor. We re-examine this issue using the incidence of accounting restatements as a measure of audit quality. Using a propensity-score matching procedure similar to that used by recent research to control for clients' endogenous choice of auditor, we find that clients of Big 4 audit firms are less likely to subsequently issue an accounting restatement than are clients of other auditors. In additional tests, we find weak evidence that clients of Big 4 auditors are less likely to issue accounting restatements than are clients of Mid-tier auditors (Grant Thornton and BDO Seidman). Taken together, the evidence suggests that Big 4 auditors do perform higher quality audits.

Keywords: Big 4 auditor; audit quality; propensity-score matching; audit quality proxies.

JEL Classifications: M41; M42.

Data Availability: All data are publicly available from sources identified in the text.

INTRODUCTION

One of the earliest theories in the audit literature is that Big 4[1] auditors, due to their larger size and better training programs, provide higher audit quality than other auditors. The argument is that larger audit firms have more reputation to lose by sacrificing their independence on any given audit engagement (DeAngelo 1981). In addition, larger audit firms have

John Daniel Eshleman is an Assistant Professor at Oklahoma State University, and Peng Guo is a Ph.D. Student at Louisiana State University.

We thank Donald J. Stokes, two anonymous reviewers, Qiang Cheng, Neil Bhattacharya, Jae Bum Kim, Chee Yeow Lim, Jeff Ng, Tharindra Ranasinghe, Ken Reichelt, Jared Soileau, Yoonseok Zang, and all workshop participants at Singapore Management University. All errors that remain are our own.

Editor's note: Accepted by Donald J. Stokes.

Submitted: February 2013
Accepted: April 2014
Published Online: April 2014

[1] We use the term Big 4 to refer to the Big 5 or Big 4 accounting firms.

more resources to invest in training programs, resulting in better trained auditors. Early literature provided evidence consistent with DeAngelo's theory. For example, Francis and Krishnan (1999) find that Big 4 auditors exhibit greater conservatism when issuing audit reports. There is also evidence that investors place more weight on the earnings of a firm audited by a Big 4 auditor, consistent with investors viewing the earnings as being of higher quality (Teoh and Wong 1993). In addition, it has been shown that clients of Big 4 auditors exhibit higher earnings quality in the form of a lower magnitude of discretionary accruals (Becker, DeFond, Jiambalvo, and Subramanyam 1998; Francis, Maydew, and Sparks 1999). To the extent that discretionary accruals capture opportunistic earnings management, this implies that Big 4 auditors tolerate less earnings management than other auditors.

However, there is a potential endogeneity problem with the research discussed above. Firms select their auditors and auditors decide if they will accept the firm as their client. Audit firms will tend to prefer less risky clients with higher earnings quality.[2] Thus, it is not clear from the early research (Teoh and Wong 1993; Becker et al. 1998) that Big 4 auditors have higher audit quality. Lawrence, Minutti-Meza, and Zhang (2011) correct for this endogenous choice of the firm's auditor using a propensity-score matching approach. After correcting for the endogeneity inherent in client selection, the authors provide evidence that suggests that clients of Big 4 auditors exhibit audit quality similar to the matched sample of non-Big 4 clients.[3] In a related study, Boone, Khurana, and Raman (2010) find that Big 4 auditors and Mid-tier auditors (i.e., Grant Thornton and BDO Seidman) exhibit similar audit quality, where audit quality is measured using the absolute value of discretionary accruals. The authors also find "weak evidence that the Big 4 have a higher propensity to issue going concern audit opinions for distressed companies" (Boone et al. 2010, 330). Boone et al. (2010) conclude that, among Big 4 and Mid-tier audit firm clients, there is little difference in actual audit quality.

The purpose of this paper is to re-examine whether Big 4 auditors deliver higher audit quality after controlling for the endogenous choice of auditor. In this study, we choose an audit quality proxy, which we believe better captures whether the client engaged in non-GAAP reporting. Our proxy is the likelihood of a firm issuing an accounting restatement. We construct our sample by matching non-Big 4 clients with Big 4 clients via a propensity-score matching model similar to that used by Lawrence et al. (2011) and Boone et al. (2010). We first replicate the results from Lawrence et al. (2011) and find similar results for our sample, implying that clients of the Big 4 do not receive higher quality audits than clients of the non-Big 4. However, our restatement analysis tells a different story. We find that clients of non-Big 4 auditors are significantly more likely to subsequently issue an accounting restatement than are clients of the Big 4. This result holds after controlling for a set of innate firm characteristics known to affect the likelihood of issuing a restatement. This is consistent with non-Big 4 auditors allowing a higher frequency of material misstatements than Big 4 auditors. In additional analyses, we present evidence that clients of Big 4 auditors are significantly less likely to be sanctioned by the SEC for an Accounting and Auditing Enforcement Release (AAER) than are clients of other auditors.

We then use the same propensity-score matching approach to construct a matched sample of Big 4 and Mid-tier auditors. We find that, when using the matched sample, clients of the Big 4 are significantly less likely to subsequently issue an accounting restatement than are clients of Mid-tier auditors. However, these results do not hold when using the full (non-matched) sample.

[2] In addition, small auditors are likely unable to audit the largest companies, due to capacity constraints and litigation costs.

[3] Specifically, the authors find that, among Big 4 and non-Big 4 clients, there is little difference in the magnitude of discretionary accruals, analyst forecast accuracy, and the cost of equity capital.

We also construct a matched sample of Mid-tier and small auditors. We find no evidence that Mid-tier auditors provide higher audit quality than the small audit firms. Finally, we construct a matched sample of Big 4 and small auditors. We find that clients of small auditors are significantly more likely to subsequently issue an accounting restatement than are clients of the Big 4. Taken together, the evidence suggests a hierarchy of audit firms, with Big 4 auditors providing the highest audit quality, small auditors providing the lowest level of audit quality, and Mid-tier auditors providing audit quality in between the Big 4 and the small auditors.

The evidence presented in this paper is of interest to managers, audit committees, investors, creditors, and regulators. Managers and audit committees would like to know whether the Big 4 actually do provide higher quality audits. This information will help them choose an auditor. Given that Big 4 auditors earn a fee premium (Ireland and Lennox 2002), managers and audit committees must decide whether the services they receive from the auditor are worth the premium. Investors and creditors will also be interested in our results, as this will help them assess the credibility of firms' financial reports. Regulators are also interested in whether the Big 4 accounting firms actually provide higher quality audits. Recently, some have suggested that smaller firms can provide audits of similar quality. For example, the U.S. Chamber of Commerce has recently suggested that "all parties should actively encourage public companies to consider high quality firms outside the Big 4" (U.S. Chamber of Commerce 2006, 18).

This paper contributes to the audit literature by providing evidence of superior audit quality at Big 4 audit firms. Recent research by Lawrence et al. (2011) and Boone et al. (2010) suggests that differences in audit quality among Big 4 and non-Big 4 clients may be attributable to the different clienteles of Big 4 and non-Big 4 auditors. Our study complements and extends these two studies by using a different proxy for audit quality. The fact that we draw different inferences using a different audit quality proxy underscores the importance of using caution when selecting audit quality proxies. We believe focusing on tangible outputs of the audit, such as accounting restatements, provides better tests of audit quality.[4]

The rest of the paper is organized as follows. The following section contains a literature review and our hypotheses, the third section contains the research design, the fourth contains the sample selection, and the fifth reports the replication of Lawrence et al. (2011). The sixth section reports the results for the Big 4 versus non-Big 4 analysis, the seventh section reports the results for the Big 4 versus Mid-tier analysis, the eighth section contains additional analysis, and the ninth concludes.

LITERATURE REVIEW AND HYPOTHESIS DEVELOPMENT

In a seminal paper, DeAngelo (1981) makes the argument that Big 4 auditors will provide higher audit quality because they have "more to lose." Big 4 auditors have a larger client portfolio than other audit firms. If the auditor is perceived to have allowed the client to manage earnings, then investors will view the earnings quality of all clients of the Big 4 auditor unfavorably. As a result, many of the auditor's clients will demand lower audit fees and/or switch to a different auditor. Therefore, a larger audit firm has more to lose by compromising its independence. In addition, there is the argument that Big 4 auditors' large size enables them to invest in more high quality training and audit technology (DeAngelo 1981; Boone et al. 2010). The early audit literature provides evidence in support of this notion. Teoh and Wong (1993) find that clients of Big 4 auditors exhibit

[4] A disadvantage of using restatements or going concern opinions for audit quality is that these are rare events. In our sample, approximately 7 percent of all firm-years involve an accounting restatement. Therefore, this audit quality proxy cannot tell us anything about how the audit quality differs among the other 93 percent of the firm-years that had no accounting restatements. It only tells us that audit quality was worse for the 7 percent of firms that did later issue a restatement.

higher earnings response coefficients, consistent with investors viewing these clients' earnings as being of higher quality. There is also evidence that Big 4 auditors report more conservatively (Francis and Krishnan 1999). Research also finds that clients of Big 4 auditors generally exhibit higher earnings quality, which is consistent with auditors tolerating less earnings management (Becker et al. 1998; Francis et al. 1999). Finally, Lennox and Pittman (2010) show that clients of Big 4 auditors are less likely to commit fraud, even when controlling for the endogenous choice of auditor.[5]

For several years, there was little research on the issue; the literature had produced a large body of evidence suggesting that Big 4 auditors provided higher quality audits than smaller auditors.[6] However, Lawrence et al. (2011) cast doubt on the superiority of Big 4 auditors. Lawrence et al. note that a firm's choice of an auditor is endogenous.[7] Firms with better performance and higher quality earnings are more likely to choose Big 4 auditors. Similarly, Big 4 auditors will prefer less risky clients with higher earnings quality. The authors control for the endogenous choice of auditor by modeling the choice of an auditor using a propensity-score matching model to match each non-Big 4 client with a Big 4 client. Using this matched sample, the authors show that clients of Big 4 auditors do not exhibit higher audit quality than clients of non-Big 4 auditors. Boone et al. (2010) is another study related to this issue. The authors examine whether clients of Big 4 auditors have higher audit quality than clients of the Mid-tier audit firms (i.e., Grant Thornton and BDO Seidman). They find that Big 4 audit firms do not perform higher quality audits than Mid-tier audit firms.

Lawrence et al. (2011) choose three proxies for the unobservable audit quality. The first proxy is an earnings quality metric, the absolute value of discretionary accruals estimated from the Jones (1991) model. The justification for using an earnings quality metric to proxy for audit quality is that higher quality auditors will tolerate less earnings management from the client. Therefore, firms with higher magnitudes of discretionary accruals are assumed to have received lower quality audits. The second proxy is the accuracy of financial analysts' forecasts (Behn, Jong-Hag, and Kang 2008). The more accurate are the forecasts, the higher is the earnings quality, which implies that the auditors must have tolerated less earnings management. The third proxy is the *ex ante* cost of equity capital. The rationale is that firms with more credible financial statements will have a lower perceived information risk, which will result in a lower cost of capital.

There are reasons to believe that the three audit quality proxies chosen by Lawrence et al. (2011) may not be capturing audit quality differences between Big 4 and non-Big 4 auditors. First, by using the magnitude of discretionary accruals to proxy for audit quality, the researcher is assuming that all discretionary accruals are equally harmful to earnings quality. This need not be the case; managers can use discretionary accruals to signal firm value (Subramanyam 1996). Second, the use of analyst forecast accuracy to proxy for audit quality relies on a very indirect link from auditor quality to earnings quality to analyst accuracy (Behn et al. 2008). It seems more reasonable to examine the relationship between the auditor and earnings quality. Finally, the cost of equity

[5] See Francis (2004) for a more complete review of the audit quality literature.

[6] Although there was little research that directly examined whether Big 4 auditors provide higher audit quality than non-Big 4 auditors, several studies of audit quality included Big 4 as a control variable. For example, Choi, Kim, Qiu, and Zang (2012) find that Big 4 auditors are associated with a lower magnitude of discretionary accruals. Asthana and Boone (2012) find that Big 4 clients are less likely to meet or beat analysts' earnings per share (EPS) forecasts by two pennies or less, and have a lower magnitude of discretionary accruals.

[7] Lawrence et al. (2011) are not the first to note that the choice of auditor is endogenous. Hogan (1997), Ireland and Lennox (2002), and Lennox and Pittman (2010) all note that the choice of auditor is endogenous, and control for this in their research designs. Early research by J. (Krish) Krishnan and J. Krishnan (1996) also controls for client risk via a two-stage approach (see also Fargher and Jiang 2008).

capital cannot speak to differences in audit quality; this variable measures investors' *perceptions* of audit quality.[8]

The purpose of this study is to re-examine whether Big 4 auditors deliver higher audit quality when the endogeneity of auditor choice is controlled via a propensity-score matching approach. Our proxy for audit quality is the likelihood of a firm issuing an accounting restatement. Consistent with recent literature, we only consider restatements attributable to a failure in the application of GAAP (Newton, Wang, and Wilkins 2013).[9] Based on prior literature, we view a restatement of audited financial statements as a strong indicator that the audit of the original financial statements was of low quality (Palmrose and Scholz 2004; Kinney, Palmrose, and Scholz 2004).[10] Plumlee and Yohn (2010) classify the explanations for accounting restatements into four main groups: internal errors by the firm (57 percent), fraud (3 percent), transaction complexity (3 percent), and application of accounting standards with judgment (37 percent). As discussed by Francis et al. (2014), the external auditor bears some responsibility for allowing a firm to issue financial statements that are materially misstated because of any one of these four reasons. Ours is not the first study to test whether clients of the Big 4 are less prone to issue restatements. Lobo and Zhao (2013, Table 5, Panel B) find evidence consistent with clients of the Big 4 being significantly less likely to issue an accounting restatement. Bentley, Omer, and Sharp (2013); Carcello, Neal, Palmrose, and Scholz (2011); and Newton et al. (2013) all find that clients of the Big 4 are no less likely to restate their earnings than are other firms.[11] However, none of the four aforementioned studies controls for the endogeneity of auditor selection. Our first hypothesis, stated in alternative form, is then:

H1: Clients of Big 4 auditors have a lower likelihood of issuing an accounting restatement than clients of non-Big 4 auditors after controlling for the client's propensity to choose a Big 4 auditor.

A separate but related question is whether the Big 4 provide higher audit quality than Mid-tier auditors. Regulators have questioned whether the Big 4 actually provide higher audit quality than Mid-tier auditors. For example, Kayla Gillan, a member of the PCAOB, suggests that audit committees "consider the so-called second-tier audit firms. I dislike using that term because it implies that the firms are secondary in quality which I strongly believe is false." (Grant Thornton 2007). We therefore test the following hypothesis.

H2: Clients of the Big 4 have a lower likelihood of issuing an accounting restatement than clients of Mid-tier auditors after controlling for the client's propensity to choose a Big 4 auditor.

We consider the two audit firms Grant Thornton and BDO Seidman to be the Mid-tier auditors, following Boone et al. (2010).

---

[8] If investors view the financial statements as being more credible due to the higher quality audit, then they will assign the firm a lower cost of equity capital. However, investors do not actually know the quality of the audit, since audit quality is unobservable. This is similar to using earnings response coefficients as a proxy for audit quality.

[9] The Audit Analytics Restatement database includes some restatements that are not due to failures in the application of GAAP. Of the 12,883 restatements available on Audit Analytics at the time of this study, 543 were not due to GAAP failure. Our results remain unchanged if we include these restatements in our analysis.

[10] Dechow, Ge, and Schrand (2010) note that "a significant benefit of using the restatement sample to identify firms with earnings quality problems is a lower Type I error rate in the identification of misstatements" (Dechow et al. 2010, 374). Restatements have been used as measures of audit quality in recent literature, such as Blankley, Hurtt, and MacGregor (2012); Francis and Michas (2013); and Francis, Michas, and Yu (2014).

[11] Many recent studies on accounting restatements restrict their sample to Big 4 clients only (e.g., Blankley et al. 2012; Francis and Michas 2013).

RESEARCH DESIGN

To test our hypotheses, we first control for the endogenous choice of auditor. We follow Lawrence et al. (2011) and Boone et al. (2010) and use a propensity-score matching approach to match each non-Big 4 client with a Big 4 client on the basis of observable firm characteristics. Note that the propensity-score matching approach does not control for unobservable differences between Big 4 and non-Big 4 clients. To control for unobservable differences, one would use the Heckman (1979) self-selection model (Lennox, Francis, and Wang 2012). We prefer the propensity-score matching approach because it does not require exclusion restrictions. In our setting, an exclusion restriction is a variable or set of variables highly correlated with *BIG4* but uncorrelated with audit quality. As noted by Lennox and Pittman (2010, 236):

> In the context of auditor choice, a researcher who wishes to use the Heckman model faces the often intractable task of identifying an independent variable that meets the following conditions: (a) it is exogenous, (b) it is a very powerful predictor of auditor choice in the first stage model, and (c) it does not affect the dependent variable in the second stage model.

If one does not have such a variable, then the results from the Heckman model can be extremely sensitive to minor changes in the specification of the model (Lennox et al. 2012). We therefore model the choice of auditor using the following logit regression, estimated separately for each year:

$$BIG4_{it} = \beta_0 + \beta_1 LNASSETS_{it} + \beta_2 ATURN_{it} + \beta_3 ROA_{it} + \beta_4 LEV_{it} + \beta_5 CURR_{it} + \sum CONTROLS_{it} + \varepsilon_{it}. \quad (1)$$

See Table 1 for variable definitions. Subscripts i and t indicate firm and year, respectively. Each independent variable is winsorized at the first and 99th percentile before estimating this regression.

We use a logit regression to estimate this model because it is the most often used approach for estimating propensity scores (Guo and Fraser 2010, 135). After obtaining the fitted values from estimating Equation (1), we match each non-Big 4 client, with replacement, to the Big 4 client with the closest fitted value in the same year and same two-digit SIC code industry, requiring a maximum distance of 0.01 between the two fitted values. Requiring this maximum distance results in a loss of observations when no good match exists but it ensures that we find a close match.[12] We then test our hypotheses on the resulting matched-pairs sample. In testing the second hypothesis, we first drop all small auditor clients and then estimate Equation (1) on the smaller sample to obtain a matched sample of Big 4 and Mid-tier clients. The key matching variable in the model is firm size, which we proxy for using the log of total assets (*LNASSETS*) and the sales of the firm (*ATURN*). Larger firms tend to select Big 4 auditors and the Big 4 auditors tend to prefer larger clients. In addition, Big 4 auditors will prefer less risky clients. Therefore, we include the current ratio (*CURR*) and firm leverage (*LEV*) to control for the financial distress of the client. We follow Lawrence et al. (2011) and include the client's return on assets (*ROA*). Finally, consistent with Lawrence et al. (2011), we include all control variables used in the respective audit-quality analysis (*CONTROLS*). The propensity-score matching procedure attempts to minimize the variation in the firm characteristics, such as size and risk. In effect, the procedure is attempting to create a "pseudo random" sample in which one group of firms (i.e., the Big 4 clients) is given the "treatment," while the other group (the non-Big 4 clients) is not given the treatment.

[12] Results in the paper are robust to imposing a maximum distance of 0.02, 0.03, 0.05, 0.10, 0.20, or 0.25 between the two fitted values. Since matching with replacement results in duplicate control firm observations, we replicate our main result after deleting these duplicate observations. Our inferences remain unchanged.

TABLE 1

Variable Definitions
(in the order they appear)

Variable		Definition
Variables Used in Restatement Analysis		
RESTATE	=	1 if the firm subsequently issues an accounting restatement, 0 otherwise.
BIG4	=	1 if the firm has a Big 4 auditor, 0 otherwise. For years before 2002, the Big 4 is actually the Big 5 as it includes Arthur Andersen.
LNASSETS	=	The natural logarithm of total assets (AT).
ATURN	=	Total sales (SALE) divided by lagged assets (AT).
ROA	=	Income before extraordinary items (IB), divided by average total assets (AT).
LEV	=	Financial leverage, defined as long-term debt plus debt in current liabilities, all scaled by total assets ([DLTT + DLC]/AT).
CURR	=	The current ratio, defined as current assets divided by current liabilities (ACT/LCT).
BM	=	The book-to-market ratio, defined as book equity scaled by market value at fiscal year-end (CEQ/[PRCC_F × CSHO]).
FIN	=	Financing raised, defined as additional cash raised from the issuance of long-term debt (DLTIS), plus cash raised from common and preferred stock (SSTK), all scaled by total assets (AT).
EPSGROW	=	1 if the firm had positive earnings (IBQ) changes for four consecutive quarters, 0 otherwise.
EP	=	The earnings-to-price ratio, defined as income before extraordinary items, scaled by market value at fiscal year-end (IB/[PRCC_F × CSHO]).
FREEC	=	Demand for external financing, defined as operating cash flows (OANCF) less capital expenditures (CAPX), all scaled by lagged assets.
AGE	=	The natural logarithm of the number of years the firm has been listed on Compustat.
QUAL	=	1 if the firm receives a qualified audit opinion, 0 otherwise. This variable equals 1 if the Compustat variable AUOP equals 2, 4, or 5.
INFLUENCE	=	The total fees (TOTAL_FEES) received from client *i* in year *t*, divided by the total fees earned by the audit office in year *t*. Audit offices are defined at the MSA level.
LNFEE	=	The natural logarithm of audit fees (AUDIT_FEES) charged to client *i* in year *t*.
Variables Used in Replication of Lawrence et al. (2011)		
|*DACC*|	=	The absolute value of discretionary accruals estimated as the residual from the Jones (1991) model, augmented with return on assets as suggested by Kothari, Leone, and Wasley (2005), estimated by industry-year, where industries are defined using two-digit SIC codes. We require at least 20 firms per industry-year to estimate this variable. We use all firms with sufficient data on Compustat to estimate this variable.
RPEG	=	The *ex ante* cost of equity capital, calculated as in Easton (2004) and Lawrence et al. (2011).
ACCY	=	Analyst forecast accuracy, calculated as −1 times the absolute value of the consensus analyst forecast error. The consensus analyst forecast error is the median consensus EPS estimate less the actual earnings, all scaled by stock price at fiscal year-end (PRCC_F).
LOGMV	=	The natural log of firm's market value (PRCC_F × CSHO).

(continued on next page)

TABLE 1 (continued)

Variable		Definition
SURP	=	Earnings surprise, calculated as change in net income (NI) scaled by the firm's market value.
LOSS	=	1 if the firm reports negative net income (NI), 0 otherwise.
ZSCORE	=	Altman's (1968) bankruptcy score, calculated as $e^X/(1 + e^X)$, where $X = -4.34 - 0.08 \times (\text{WCAP/AT}) + 0.04 \times (\text{RE/AT}) - 0.10 \times (\text{PI} + \text{XINT} - \text{IDIT})/\text{AT} - 0.22 \times (\text{PRCCF} \times \text{CSHO})/\text{LT} + 0.06 \times (\text{SALE/AT})$.
HORIZON	=	The average number of days between the forecast announcement date and the subsequent earnings announcement date.
STDROE	=	Standard deviation of net income over past 5 years.
NUMANAL	=	The number of analysts following the firm.
EL	=	Actual earnings per share during t+1.
BETA	=	Market beta estimated using daily data for the year ending on the firm's fiscal year-end.
LOGLEV	=	The natural log of the firm's leverage ([DLTT + DLC]/AT).
LOGBM	=	The natural log of the firm's book-to-market ratio (CEQ/[PRCC_F × CSHO]).
DISP	=	The standard deviation of analysts' EPS forecasts during the fiscal year-end month.
GROWTH	=	Forecasted growth, calculated as the difference between the mean analysts' two-year- and one-year-ahead earnings forecasts scaled by the one-year-ahead earning forecast.

To test the first hypothesis, we estimate the following probit model on the matched sample with standard errors clustered by firm and year (Peterson 2009):

$$\begin{aligned} RESTATE_{it} = {} & \alpha_0 + \alpha_1 BIG4_{it} + \alpha_2 LNASSETS_{it} + \alpha_3 ATURN_{it} + \alpha_4 ROA_{it} + \alpha_5 LEV_{it} \\ & + \alpha_6 CURR_{it} + \alpha_7 BM_{it} + \alpha_8 FIN_{it} + \alpha_9 EPSGROW_{it} + \alpha_{10} EP_{it} + \alpha_{11} FREEC_{it} \\ & + \alpha_{12} AGE_{it} + \alpha_{13} QUAL_{it} + \alpha_{14} INFLUENCE_{it} + \alpha_{15} LNFEE_{it} \\ & + Industry\ Fixed\ Effects + Year\ Fixed\ Effects + \varepsilon_{it}. \end{aligned} \tag{2}$$

Subscripts i and t indicate firm and year, respectively. *RESTATE* equals 1 if the client's financial statements are later restated, 0 otherwise. We choose to focus on the original period as this is the period in which the auditor overlooked misstatements, indicating lower audit quality.[13] If Big 4 auditors do perform higher quality audits, then we expect to observe a negative coefficient on *BIG4*. We control for the size of the firm (*LNASSETS* and *ATURN*), profitability (*ROA*), the financial health of the firm (*LEV* and *CURR*), the book-to-market ratio (*BM*), whether the firm raised external financing (*FIN*), capital market pressures (*EPSGROW*), the earnings-to-price ratio (*EP*), demand for external financing (*FREEC*), firm age (*AGE*), whether the firm received a qualified audit opinion (*QUAL*), and the relative importance of the client to the auditor (*INFLUENCE*). Finally, we control for the amount of audit fees charged to the client (*LNFEE*), since clients receiving higher audit fees are less likely to subsequently issue a restatement (Blankley et al. 2012; Lobo and Zhao 2013).

[13] In other words, if the client's fiscal year-end falls between the beginning of the restated period and the end of the restated period, then it is classified as a restatement year. As noted by Newton et al. (2013), focusing on the period in which the restatement is announced likely identifies client-years in which the auditor discovered misstatements, indicating higher audit quality.

TABLE 2

Sample Selection

Panel A: Big 4 versus Non-Big 4 Sample Selection

All Firm-Years on Compustat with Non-Missing CIK and SIC Code During 2000–2009:	90,634
Less: Financial Services Firms (SIC codes 6000–6999)	(14,832)
Less: Firm-years with material weaknesses	(1,599)
Less: Firm-years with auditor switches	(6,395)
Less: firm-years with missing data necessary to calculate control variables	(30,320)
Big 4 versus Non-Big 4 Full Sample:	37,488
Less: observations lost due to limited number of matches	(19,944)
Less: observations lost due to matched firm being too different	(11,594)
Big 4 versus Non-Big 4 Matched Sample:	5,950

Panel B: Big 4 versus Mid-Tier Sample Selection

All Firm-Years on Compustat with non-missing CIK and SIC code during 2000–2009:	56,966
Less: Financial Services Firms (SIC codes 6000–6999)	(8,757)
Less: Firm-years with material weaknesses	(1,368)
Less: Firm-years with auditor switches	(3,125)
Less: firm-years with missing data necessary to calculate control variables	(12,370)
Big 4 versus Mid-tier Full Sample:	31,346
Less: observations lost due to limited number of matches	(26,086)
Less: observations lost due to matched firm being too different	(2,012)
Big 4 versus Mid-tier Matched Sample:	3,248

To test whether Big 4 auditors perform higher quality audits than Mid-tier auditors (H2), we first re-estimate Equation (1) using a subsample of client-years, in which the client's auditor was either a Big 4 or a Mid-tier auditor, in order to create a matched sample. We then re-estimate Equation (2) on this sample.

SAMPLE SELECTION

We obtain financial statement data from the Compustat Fundamentals Annual file and auditor and restatement data from Audit Analytics for the period 2000–2009. The sample selection differs for each hypothesis. Table 2 outlines our sample selection procedure.

To construct our sample for H1, we begin with all client-year observations on Compustat with non-missing data on Audit Analytics and non-missing SIC code for the period 2000–2009. We then delete financial services clients (SIC codes 6000–6999), clients with material weaknesses over financial reporting, and client-years involving an auditor switch. Finally, we delete observations with insufficient data to calculate the variables used in Equation (2). This leaves us with a primary sample of 37,488 client-year observations (hereafter, the "Big 4 versus non-Big 4 full sample"). We perform the propensity-score matching regression on this sample. We lose 19,944 observations due to a limited number of matches. After imposing the requirement that the two firms' fitted values do

not differ by more than 0.01, we are left with 2,975 Big 4 clients and 2,975 non-Big 4 clients, giving us a total sample of 5,950 observations (hereafter, the "Big 4 versus non-Big 4 matched sample"). We deliberately end the sample two years before the most recent year of data available when we began the study (2011), to allow for a sufficient amount of time for a firm to subsequently restate its earnings.[14] We leave two years for clients to restate earnings because Cheffers, Whalen, and Usvyatsky (2010) show that the average time between an originally released financial statement and a restatement is approximately 700 days, which is slightly less than two years.[15]

We construct our sample for H2 in a similar manner. The only difference is that we delete all client-year observations in which the client is audited by a small auditor. This results in a primary sample of 31,346 client-year observations (hereafter, the "Big 4 versus Mid-tier full sample"). After performing the propensity-score matching regression (Equation (1)) and deleting matched pairs in which the fitted values differ by more than 0.01, we are left with 1,624 Big 4 client-years and 1,624 Mid-tier client-years, giving us a sample of 3,248 observations (hereafter, the "Big 4 versus Mid-tier matched sample"). In both samples, all continuous independent variables are winsorized at the first and 99th percentiles. We do not tabulate the results from estimating the propensity-score matching regression (Equation (1)) on each of our samples to save space.[16]

REPLICATION OF LAWRENCE ET AL. (2011)

In this section we replicate the results of Lawrence et al. (2011) using the full and matched samples to ensure that our results are not due to changes in audit quality during different sample periods.[17] Lawrence et al. (2011) test whether Big 4 auditors deliver higher audit quality, using the following three linear regression models:

$$|DACC_{it}| = \rho_0 + \rho_1 BIG4_{it} + \rho_2 LOGMV_{it} + \rho_3 ROA_{it} + \rho_4 LEV_{it} + \rho_5 CURR_{it} + Industry\ Fixed\ Effects + Year\ Fixed\ Effects + \varepsilon_{it}. \quad (3)$$

$$ACCY_{it} = \omega_0 + \omega_1 BIG4_{it} + \omega_2 LOGMV_{it} + \omega_3 SURP_{it} + \omega_4 LOSS_{it} + \omega_5 ZSCORE_{it} + \omega_6 HORIZON_{it} + \omega_7 STDROE_{it} + \omega_8 NUMANAL_{it} + \omega_9 EL_{it} + Industry\ Fixed\ Effects + Year\ Fixed\ Effects + \varepsilon_{it}. \quad (4)$$

$$RPEG_{it} = \varphi_0 + \varphi_1 BIG4_{it} + \varphi_2 BETA_{it} + \varphi_3 LOGLEV_{it} + \varphi_4 DISP_{it} + \varphi_5 LOGMV_{it} + \varphi_6 LOGBM_{it} + \varphi_7 GROWTH_{it} + Industry\ Fixed\ Effects + Year\ Fixed\ Effects + \varepsilon_{it}. \quad (5)$$

In all models, subscripts i and t denote firm and year, respectively. See Table 1 for variable definitions. Each model includes industry- and year-fixed effects. In Equation (3), audit quality is measured using the absolute value of discretionary accruals ($|DACC|$). In Equation (4), audit

[14] For example, if we used all observations up to the most recent year available on Audit Analytics (2011 at the time of this study), then we would not know whether a firm's 2010 earnings would be restated in later years, such as 2012.

[15] For the Big 4 versus non-Big 4 full sample, there are 1,791 restatements, all of which are attributable to a failure in the application of GAAP. Of these restatements, 1,530 (85.4 percent) resulted in a downward adjustment to income, 73 (4.1 percent) are attributable to financial fraud, 20 (1.1 percent) are attributable to clerical errors, and 183 (10.2 percent) are attributable to other significant issues. Certain restatements are attributable to more than one cause.

[16] We note that the R^2 of the propensity-score matching regression is 44.8 percent for the Big 4 versus non-Big 4 sample and 14.8 percent for the Big 4 versus Mid-tier sample. The lower explanatory power when using the Big 4 versus Mid-tier sample is likely due to the matching model R^2 being an increasing function of the heterogeneity of the two groups before the matching procedure (Peel and Makepeace 2012, 617).

[17] Lawrence et al.'s (2011) sample spans 1988–2006, while our sample covers 2000–2009.

quality is measured using analyst forecast accuracy (*ACCY*). In Equation (5), audit quality is measured using the firm's cost of capital, estimated as in Easton (2004).

Table 3 reports the results of estimating Equations (3), (4), and (5) on the full and matched Big 4 versus non-Big 4 samples.[18] Panel A reports the discretionary accruals results. Using the full sample, the coefficient on *BIG4* is negative and significant. Consistent with Lawrence et al. (2011), the coefficient becomes insignificant when using the propensity-score matched sample (Coeff. = 0.058, p-value = 0.552). The goodness of fit of the model is consistent with Lawrence et al. (2011); when using the full sample, the R^2 is 0.199 compared to an R^2 of 0.14 reported in Lawrence et al. (2011). Panel B reports the analyst forecast accuracy results. Using the full sample, the coefficient on *BIG4* is significantly positive, consistent with clients of the Big 4 having higher analyst forecast accuracy. However, consistent with Lawrence et al. (2011), this coefficient becomes insignificant after matching (Coeff. = 0.004, p-value = 0.606). When using the full sample, the R^2 is 0.219, compared with an R^2 of 0.28 reported by Lawrence et al. (2011). Panel C reports the cost of capital results. Here, the coefficient on *BIG4* is significantly negative when using the full sample but becomes insignificant when using the matched sample (Coeff. = 0.001, p-value = 0.837), consistent with Lawrence et al. (2011). In sum, the results of Lawrence et al. (2011) continue to hold for our sample period. This means that any divergent results we observe when using restatements as a proxy for audit quality are likely due to the choice of audit quality proxy.

There is reason to believe that the proxies used by Lawrence et al. (2011) may not be capturing audit quality. For example, using the magnitude of discretionary accruals as a proxy for audit quality assumes all discretionary accruals, regardless of whether they are positive or negative, are equally harmful to earnings quality. However, extant research provides evidence that this is not the case. Krishnan (2003) finds that the association between discretionary accruals and future profitability is stronger for Big 4 clients compared with non-Big 4 clients. Therefore, while clients of Big 4 auditors do not exhibit significantly different discretionary accruals than clients of smaller auditors, this does not imply that the two sets of firms have similar earnings quality. Clients of the Big 4 could be using their discretionary accruals to signal future profitability.[19] The use of financial analyst forecast accuracy as a proxy for audit quality relies on the assumption that auditors affect earnings quality, which affects analyst forecast accuracy. This is a very indirect measure of audit quality; it would seem that examining an earnings quality metric would be more informative about the quality of the audit. Finally, the third proxy for audit quality chosen by Lawrence et al. (2011) is the *ex ante* cost of equity capital. The argument is that firms with better audit quality will be seen as having more credible financial statements. This credibility lowers information risk, which will lower the firm's cost of capital. However, the *ex ante* cost of equity capital only speaks to investors' perceptions of audit quality. For this measure to capture audit quality, we must assume that investors can observe audit quality and can price it efficiently.

EMPIRICAL RESULTS FOR BIG 4 VERSUS NON-BIG 4

In this section, we examine whether clients of Big 4 auditors are less likely to subsequently issue accounting restatements. Table 4 contains descriptive statistics for the full and matched Big 4 versus non-Big 4 samples. Panel A reports the descriptive statistics for the full sample of Big 4 and

[18] The number of observations differs for each test, due to missing variables.

[19] Another problem with this proxy is that it assumes that the goal of management is to book the largest discretionary accruals in any given period. This is not always the case. Consider a manager of a firm with pre-discretionary earnings (i.e., earnings before discretionary accruals) of $2.00 per share and an earnings target of $1.97 per share. In this case, the manager has no incentive to record large discretionary accruals. Finally, using the absolute value of discretionary accruals to measure audit quality assumes that the discretionary accrual model accurately dissects a firm's accruals into a normal and a discretionary component.

TABLE 3

Replication of Lawrence et al. (2011)

Panel A: Discretionary Accruals Analysis

	Full Sample		Matched Sample	
Variable	**Estimate**	**p-value**	**Estimate**	**p-value**
Intercept	0.716***	0.000	1.371**	0.016
BIG4	−0.228***	0.000	0.058	0.552
LOGMV	−0.026*	0.058	−0.132*	0.090
ROA	−0.397***	0.008	−0.118	0.588
LEV	−0.050	0.620	0.200	0.562
CURR	−0.020***	0.000	−0.028**	0.039
Industry and Year FE	Yes		Yes	
n	36,153		5,907	
Adj. R^2	0.199		0.193	

Panel B: Analyst Forecast Accuracy Analysis

	Full Sample		Matched Sample	
Variable	**Estimate**	**p-value**	**Estimate**	**p-value**
Intercept	−0.047***	0.000	−0.050**	0.050
BIG4	0.005**	0.021	0.004	0.606
LOGMV	0.009***	0.000	0.020***	0.000
SURP	0.000	0.810	−0.003***	0.000
LOSS	−0.019***	0.000	−0.011*	0.077
ZSCORE	−3.047***	0.000	−4.240***	0.000
HORIZON	0.000***	0.000	−0.001***	0.000
STDROE	0.000***	0.000	0.000**	0.016
NUMANAL	0.000**	0.045	−0.003**	0.015
EL	0.005**	0.014	0.012***	0.002
Industry and Year FE	Yes		Yes	
n	19,669		2,228	
Adj. R^2	0.219		0.400	

Panel C: Cost of Capital Analysis

	Full Sample		Matched Sample	
Variable	**Estimate**	**p-value**	**Estimate**	**p-value**
Intercept	0.107***	0.000	0.168***	0.000
BIG4	−0.003*	0.064	0.001	0.837
BETA	−0.001	0.706	0.005	0.173
LOGLEV	0.002***	0.000	0.000	0.529
DISP	0.034***	0.001	0.175***	0.000
LOGMV	−0.012***	0.000	−0.021***	0.000
LOGBM	0.007***	0.000	0.006***	0.002
GROWTH	−0.005***	0.000	−0.011***	0.002

(continued on next page)

TABLE 3 (continued)

	Full Sample		Matched Sample	
Variable	**Estimate**	**p-value**	**Estimate**	**p-value**
Industry and Year FE	Yes		Yes	
n	7,975		680	
Adj. R^2	0.463		0.513	

*, **, *** Denote statistical significance at the 10, 5, and 1 percent levels, respectively, using a two-tailed test.
This table displays the results of attempting to replicate the results documented in Lawrence et al. (2011). In Panel A, the dependent variable is the absolute value of discretionary accruals (|*DACC*|). In Panel B, the dependent variable is analyst forecast accuracy (*ACCY*). In Panel C, the dependent variable is the firm's *ex ante* cost of equity capital (*RPEG*). All models include industry- and year-fixed effects. p-values are based on t-statistics, which are clustered by firm and year (Peterson 2009).

non-Big 4 clients. Surprisingly, clients of the Big 4 have a higher frequency of restatements (9.3 percent compared to 6.7 percent for the non-Big 4). However, we caution against putting too much weight on this result, as we have not controlled for other factors that influence the likelihood of issuing a restatement. Compared with clients of non-Big 4 auditors, clients of the Big 4 have significantly more assets (*LNASSETS*), a significantly higher return on assets (*ROA*), and are significantly older (*AGE*). In addition, clients of the Big 4 have a lower absolute value of discretionary accruals (|*DACC*|), a lower cost of equity (*RPEG*), and greater analyst forecast accuracy (*ACCY*). Note that these three variables were not used in the propensity-score matching regression. We only report the difference in means of these variables for comparison with prior literature (Lawrence et al. 2011). The last columns of Table 4 report a parametric t-test of the difference in means and a non-parametric Kolmogorov-Smirnov (KS) test. The KS test is a non-parametric test, which tests the difference in distributions between the two groups of firms. Differences in means for indicator variables are tested using Chi-square (χ^2) tests. The difference in means and difference in distributions of each variable are significant at the 1 percent level, highlighting the importance of controlling for endogeneity.

Panel B of Table 4 reports descriptive statistics for the Big 4 versus non-Big 4 matched sample. The important thing to note is that of the 14 control variables, only three (six) have means (distributions) that are significantly different at the 5 percent level across the two groups of auditors. Further, the magnitudes of the differences are not overly large. For example, clients of the Big 4 have an average current ratio of 3.533 compared to 3.837 for clients of non-Big 4 auditors. Finally, it is worth noting that the two groups of clients do not exhibit significantly different means or distributions of the three audit quality proxies used by Lawrence et al. (2011) (|*DACC*|, *RPEG,* and *ACCY*). This is important, since, if our matching procedure produced a sample in which Big 4 clients exhibited a lower magnitude of discretionary accruals, a lower cost of equity capital, and/or higher analyst forecast accuracy, then we could not attribute our restatement results to our choice of audit quality proxy. The results would likely be attributable to our time period and/or our matching model.[20]

Table 5 reports the results of estimating the restatement model (Equation (2)). The first set of columns displays results when using the full sample. The coefficient on *BIG4* is −0.091, significant at the 5 percent level. This implies that clients of Big 4 auditors are less likely to subsequently issue an accounting restatement. The second set of columns reports the results when using the matched

[20] We do not report correlations to save space, but we note that none of the variance inflation factors on any of the variables exceeds 5, which is well below the threshold of 10 recommended by Kennedy (1992).

TABLE 4

Descriptive Statistics for Big 4 versus Non-Big 4 Samples

Panel A: Descriptive Statistics for Full Sample

	Big 4 Clients (n = 28,716)				Non-Big 4 Clients (n = 8,772)				t-test Diff.	KS Diff.	χ^2
Variable	**Mean**	**Q1**	**Med.**	**Q3**	**Mean**	**Q1**	**Med.**	**Q3**	**p-value**	**p-value**	**p-value**
RESTATE	0.093	0.000	0.000	0.000	0.067	0.000	0.000	0.000	—	—	0.000
LNASSETS	6.230	4.747	6.157	7.632	3.396	2.225	3.299	4.474	0.000	0.000	
ATURN	1.095	0.497	0.900	1.432	1.236	0.422	1.008	1.683	0.000	0.000	
ROA	−0.037	−0.050	0.031	0.077	−0.166	−0.238	−0.015	0.063	0.000	0.000	
LEV	0.195	0.008	0.163	0.325	0.159	0.000	0.089	0.271	0.000	0.000	
CURR	3.102	1.313	2.036	3.397	3.823	1.229	2.144	4.020	0.000	0.000	
BM	0.665	0.275	0.476	0.779	0.767	0.235	0.508	0.943	0.000	0.000	
FIN	0.156	0.007	0.040	0.181	0.213	0.001	0.034	0.263	0.000	0.000	
EPSGROW	0.125	0.000	0.000	0.000	0.085	0.000	0.000	0.000	—	—	0.000
EP	−0.100	−0.052	0.032	0.061	−0.175	−0.172	−0.012	0.056	0.000	0.000	
FREEC	−0.026	−0.049	0.030	0.089	−0.175	−0.209	−0.027	0.066	0.000	0.000	
AGE	2.533	1.946	2.485	3.135	2.402	1.792	2.485	3.045	0.000	0.000	
QUAL	0.453	0.000	0.000	1.000	0.338	0.000	0.000	1.000	—	—	0.000
INFLUENCE	0.053	0.002	0.010	0.038	0.214	0.036	0.090	0.248	0.000	0.000	
LNFEE	13.377	12.367	13.349	14.274	11.776	11.002	11.704	12.464	0.000	0.000	
|DACC|	0.787	0.047	0.138	0.500	1.275	0.071	0.206	0.834	0.000	0.000	
RPEG	0.135	0.073	0.099	0.152	0.183	0.092	0.138	0.226	0.000	0.000	
ACCY	−0.018	−0.007	−0.002	−0.001	−0.041	−0.021	−0.005	−0.001	0.000	0.000	

Panel B: Descriptive Statistics for Matched Sample and Test of Covariate Balance between Matched Pairs (n = 2,975 pairs)

	Big 4 Clients (n = 2,975)				Non-Big 4 Clients (n = 2,975)				t-test Diff.	KS Diff.	χ^2
Variable	**Mean**	**Q1**	**Med.**	**Q3**	**Mean**	**Q1**	**Med.**	**Q3**	**p-value**	**p-value**	**p-value**
RESTATE	0.056	0.000	0.000	0.000	0.078	0.000	0.000	0.000	—	—	0.001
LNASSETS	4.055	2.860	3.871	5.101	4.040	2.813	3.941	5.151	0.743	0.370	
ATURN	1.117	0.459	0.919	1.493	1.083	0.484	0.945	1.430	0.159	0.157	
ROA	−0.137	−0.209	−0.010	0.063	−0.157	−0.247	−0.012	0.068	0.113	0.000	
LEV	0.129	0.000	0.033	0.230	0.136	0.000	0.063	0.230	0.037	0.109	
CURR	3.533	1.520	2.492	4.132	3.837	1.379	2.322	4.216	0.005	0.007	
BM	0.687	0.239	0.477	0.830	0.693	0.243	0.489	0.868	0.749	0.286	
FIN	0.185	0.003	0.026	0.217	0.201	0.004	0.035	0.259	0.058	0.004	
EPSGROW	0.121	0.000	0.000	0.000	0.106	0.000	0.000	0.000	—	—	0.065
EP	−0.190	−0.169	−0.006	0.049	−0.183	−0.168	−0.010	0.052	0.640	0.348	
FREEC	−0.111	−0.175	−0.015	0.073	−0.141	−0.189	−0.010	0.076	0.004	0.316	
AGE	2.400	1.946	2.398	2.890	2.417	1.946	2.485	2.996	0.368	0.000	
QUAL	0.384	0.000	0.000	1.000	0.384	0.000	0.000	1.000	—	—	1.000
INFLUENCE	0.134	0.000	0.004	0.035	0.145	0.030	0.066	0.150	0.081	0.000	
LNFEE	12.245	11.408	12.129	13.028	12.294	11.488	12.245	13.097	0.100	0.000	

(continued on next page)

TABLE 4 (continued)

	Big 4 Clients (n = 2,975)				Non-Big 4 Clients (n = 2,975)				t-test Diff.	KS Diff.	χ^2
Variable	**Mean**	**Q1**	**Med.**	**Q3**	**Mean**	**Q1**	**Med.**	**Q3**	**p-value**	**p-value**	**p-value**
|DACC|	1.335	0.086	0.274	1.160	1.269	0.088	0.275	1.124	0.352	0.492	
RPEG	0.167	0.084	0.123	0.192	0.175	0.087	0.127	0.210	0.284	0.235	
ACCY	−0.036	−0.012	−0.004	−0.001	−0.031	−0.014	−0.004	−0.001	0.209	0.386	

This table displays descriptive statistics for the Big 4 versus non-Big 4 samples. Panel A reports descriptive statistics for the full sample while Panel B reports descriptive statistics for the matched sample. In each Panel, the mean, first quartile (Q1), median, and third quartile (Q3) are displayed. For each variable, the last two columns report on the difference in the mean and the overall distribution of values for that variable among Big 4 and non-Big 4 clients. The t-test difference p-value column reports the p-value from a test of difference in means across the Big 4 and non-Big 4 clients. The KS difference p-value is the p-value from the Kolmogorov-Smirnov test of difference in distributions across the two sets of clients (Big 4 versus non-Big 4 clients). Chi-square (χ^2) tests are used to test the difference in distributions for indicator variables.
See Table 1 for variable definitions.

sample, which better controls for the endogenous choice of auditor. This is important, since it may be the case that clients with better financial reporting quality choose Big 4 auditors. The results are similar; the coefficient on *BIG4* is −0.201 and is significant at the 1 percent level. The marginal effect of *BIG4* is −2.21 percent, implying that using a Big 4 auditor results in a 2.21 percent decrease in the probability of issuing an accounting restatement.[21] At first glance, this may appear to be a rather small effect. However, it is quite large if one considers the baseline probability of issuing an accounting restatement is 5.3 percent for the matched sample and 7.5 percent for the full sample. The baseline probability is the probability that the dependent variable equals 1 if all independent variables are held at their mean values. To summarize, the results of Table 5 offer strong support for H1.[22]

EMPIRICAL RESULTS FOR BIG 4 VERSUS MID-TIER

In the previous section, we demonstrated that clients of the Big 4 are less likely to restate their earnings than are clients of other auditors, even after controlling for the choice of auditor. In this section, we examine whether this result can be explained by the inclusion of small auditors in the sample. Specifically, we examine whether clients of Mid-tier auditors are more likely to restate their earnings than are clients of the Big 4.

Panel A of Table 6 reports descriptive statistics for the full sample of Big 4 and Mid-tier clients. The results are somewhat similar to Panel A, with clients of the Big 4 generally having more assets (*LNASSETS*), greater profitability (*ROA*), less financial distress (*LEV, BM*), and being younger (*AGE*). In addition, clients of the Big 4 tend to exhibit a lower magnitude of discretionary accruals (*|DACC|*) and a lower cost of capital (*RPEG*). The difference in means and distributions for each of these variables is significant at the 10 percent level or better, with the exception of the level of financing raised (*FIN*).

[21] It is worth noting that the marginal effect of *BIG4* is much larger when using the matched sample. Given that the variance of the estimate of treatment effects is lower in matched samples (Rosenbaum and Rubin 1983, 48), we put more weight on the marginal effect of the treatment (i.e., *BIG4*) when using the matched sample.

[22] If we use matching without replacement and a 3 percent caliper distance, as in Lawrence et al. (2011), then the coefficient on *BIG4* is −0.205 with a Z-statistic of −1.14. If we use a 10 percent caliper, then the Z-statistic increases to −1.54, which is significant at the 10 percent level using a one-tailed test.

TABLE 5

Do Clients of the Big 4 Have Fewer Restatements?

	Full Sample			Matched Sample		
Variable	**Estimate**	**Marginal Effect**	**p-value**	**Estimate**	**Marginal Effect**	**p-value**
Intercept	−1.714***		0.000	−2.956***		0.000
BIG4	−0.091**	−1.35%	0.021	−0.201***	−2.21%	0.007
LNASSETS	0.067***	0.95%	0.001	0.100***	1.10%	0.003
ATURN	0.036**	0.50%	0.014	0.099***	1.20%	0.005
ROA	0.072	−0.94%	0.412	−0.153	−1.93%	0.576
LEV	−0.066	1.01%	0.410	−0.165	−1.81%	0.377
CURR	−0.005	−0.07%	0.315	−0.002	−0.03%	0.824
BM	−0.039**	−0.55%	0.038	−0.109**	−1.16%	0.025
FIN	0.040	0.55%	0.457	0.157*	1.66%	0.079
EPSGROW	−0.008	−0.12%	0.851	−0.067	−0.65%	0.584
EP	0.040	0.56%	0.362	0.052	0.60%	0.641
FREEC	0.029	0.40%	0.650	0.060	0.64%	0.632
AGE	−0.028	−0.39%	0.168	−0.029	−0.28%	0.665
QUAL	0.082***	1.17%	0.000	0.070	0.89%	0.279
INFLUENCE	−0.034	−0.48%	0.727	0.078	0.89%	0.495
LNFEE	−0.021	−0.29%	0.385	0.014	0.23%	0.714
Industry	Yes			Yes		
Year FE	Yes			Yes		
n	37,488			5,950		
Pseudo R^2	0.064			0.081		

*, **, *** Denote statistical significance at the 10, 5, and 1 percent levels, respectively, using a two-tailed test.
This table displays estimated coefficients from estimating a probit model (Equation (2)) where the dependent variable (*RESTATE*) equals 1 if the client subsequently issues an accounting restatement, 0 otherwise. The model is estimated on both the full Big 4 versus non-Big 4 sample and the matched Big 4 versus non-Big 4 sample. All independent variables are winsorized at the 1st and 99th percentiles to reduce the influence of outliers. The model includes industry- and year-fixed effects, where industries are defined using two-digit SIC codes. p-values are based on Z-statistics, which are clustered by firm and year (Peterson 2009).

Panel B of Table 6 reports descriptive statistics for the matched sample of Big 4 and Mid-tier clients. Of the 14 control variables, only three (four) exhibit significantly different means (distributions) across the two groups of auditors. One variable, which is still significantly different across the two sets of firms, is the relative importance of the client to the audit office (*INFLUENCE*). The average client of the Big 4 represents only 5.5 percent of the office's total fees, compared with 8.4 percent for the average Mid-tier client. It is worth noting that the two sets of clients do not exhibit significantly different magnitudes of discretionary accruals or significantly different costs of capital. Finally, after matching, clients of the Big 4 have significantly higher analyst forecast accuracy than clients of Mid-tier auditors (−0.022 compared to −0.029).

Panel A of Table 7 reports the main results. The first set of columns contains the results when using the full sample. Here, we find that the coefficient on *BIG4* is not significantly different from zero (p-value = 0.584). The second set of columns reports results when using the matched sample. Here, we find a significantly negative coefficient on *BIG4* (−0.266, p-value = 0.006). This suggests

TABLE 6

Descriptive Statistics for Big 4 versus Mid-Tier Samples

Panel A: Descriptive Statistics for Full Sample

	Big 4 Clients (n = 28,716)				Mid-Tier Clients (n = 2,630)				t-test Diff.	KS Diff.	χ^2
Variable	**Mean**	**Q1**	**Med.**	**Q3**	**Mean**	**Q1**	**Med.**	**Q3**	**p-value**	**p-value**	**p-value**
RESTATE	0.093	0.000	0.000	0.000	0.071	0.000	0.000	0.000	—	—	0.000
LNASSETS	6.236	4.747	6.157	7.632	4.346	3.167	4.268	5.353	0.000	0.000	
ATURN	1.091	0.497	0.900	1.432	1.304	0.673	1.138	1.754	0.000	0.000	
ROA	−0.034	−0.050	0.031	0.077	−0.072	−0.122	0.011	0.071	0.000	0.000	
LEV	0.195	0.008	0.163	0.325	0.162	0.000	0.095	0.284	0.000	0.001	
CURR	3.060	1.313	2.036	3.397	3.170	1.355	2.154	3.620	0.098	0.000	
BM	0.662	0.275	0.476	0.779	0.801	0.305	0.552	0.984	0.000	0.000	
FIN	0.154	0.007	0.040	0.181	0.171	0.003	0.025	0.200	0.002	0.157	
EPSGROW	0.125	0.000	0.000	0.000	0.102	0.000	0.000	0.000	—	—	0.001
EP	−0.099	−0.052	0.032	0.061	−0.146	−0.136	0.013	0.059	0.000	0.000	
FREEC	−0.020	−0.049	0.030	0.089	−0.041	−0.090	0.013	0.086	0.000	0.000	
AGE	2.533	1.946	2.485	3.135	2.579	2.079	2.639	3.091	0.007	0.000	
QUAL	0.453	0.000	0.000	1.000	0.388	0.000	0.000	1.000	—	—	0.000
INFLUENCE	0.051	0.002	0.010	0.038	0.104	0.022	0.051	0.123	0.000	0.000	
LNFEE	13.381	12.367	13.349	14.274	12.504	11.729	12.445	13.230	0.000	0.000	
|DACC|	0.760	0.047	0.138	0.500	0.959	0.059	0.165	0.646	0.000	0.000	
RPEG	0.135	0.073	0.099	0.152	0.169	0.088	0.129	0.205	0.000	0.000	
ACCY	−0.017	−0.007	−0.002	−0.001	−0.033	−0.016	−0.004	−0.001	0.000	0.000	

Panel B: Descriptive Statistics for Matched Sample and Test of Covariate Balance between Matched Pairs (n = 1,624 pairs)

	Big 4 Clients (n = 1,624)				Mid-Tier Clients (n = 1,624)				t-test Diff.	KS Diff.	χ^2
Variable	**Mean**	**Q1**	**Med.**	**Q3**	**Mean**	**Q1**	**Med.**	**Q3**	**p-value**	**p-value**	**p-value**
RESTATE	0.049	0.000	0.000	0.000	0.079	0.000	0.000	0.000	—	—	0.000
LNASSETS	4.506	3.300	4.322	5.574	4.653	3.475	4.604	5.693	0.010	0.000	
ATURN	1.177	0.565	0.981	1.571	1.146	0.576	1.021	1.540	0.314	0.353	
ROA	−0.088	−0.156	0.007	0.070	−0.083	−0.143	0.009	0.070	0.597	0.845	
LEV	0.148	0.000	0.072	0.263	0.141	0.000	0.067	0.242	0.213	0.400	
CURR	3.344	1.495	2.290	3.843	3.390	1.420	2.296	3.935	0.702	0.137	
BM	0.721	0.242	0.495	0.874	0.723	0.291	0.530	0.895	0.943	0.005	
FIN	0.174	0.004	0.032	0.221	0.168	0.004	0.026	0.205	0.554	0.561	
EPSGROW	0.105	0.000	0.000	0.000	0.118	0.000	0.000	0.000	—	—	0.264
EP	−0.143	−0.127	0.007	0.054	−0.148	−0.130	0.009	0.054	0.771	0.708	
FREEC	−0.054	−0.120	0.011	0.084	−0.055	−0.116	0.011	0.085	0.963	0.869	
AGE	2.469	1.946	2.398	2.917	2.497	2.079	2.485	2.996	0.271	0.002	
QUAL	0.419	0.000	0.000	1.000	0.407	0.000	0.000	1.000	—	—	0.498
INFLUENCE	0.055	0.001	0.006	0.023	0.084	0.024	0.050	0.107	0.000	0.000	
LNFEE	12.588	11.717	12.529	13.380	12.674	11.844	12.613	13.451	0.022	0.057	
|DACC|	1.147	0.075	0.204	0.931	1.073	0.074	0.219	0.906	0.368	0.586	

(continued on next page)

TABLE 6 (continued)

	Big 4 Clients (n = 1,624)				Mid-Tier Clients (n = 1,624)				t-test Diff.	KS Diff.	χ^2
Variable	**Mean**	**Q1**	**Med.**	**Q3**	**Mean**	**Q1**	**Med.**	**Q3**	**p-value**	**p-value**	**p-value**
RPEG	0.168	0.086	0.119	0.210	0.159	0.086	0.122	0.195	0.328	0.783	
ACCY	−0.022	−0.012	−0.003	−0.001	−0.029	−0.011	−0.003	−0.001	0.076	0.644	

This table displays descriptive statistics for the Big 4 versus Mid-tier samples. Panel A reports descriptive statistics for the full sample while Panel B reports descriptive statistics for the matched sample. In each Panel, the mean, first quartile (Q1), median, and third quartile (Q3) are displayed. For each variable, the last two columns report on the difference in the mean and the overall distribution of values for that variable among Big 4 and Mid-tier clients. The t-test difference p-value column reports the p-value from a test of difference in means across the Big 4 and Mid-tier clients. The KS difference p-value is the p-value from the Kolmogorov-Smirnov test of difference in distributions across the two sets of clients (Big 4 versus Mid-tier clients). Chi-square (χ^2) tests are used to test the difference in distributions for indicator variables.
See Table 1 for variable definitions.

that, after controlling for the endogenous choice of auditor, clients of the Big 4 are less likely to subsequently restate their earnings than are clients of Mid-tier auditors. Overall, Table 7 provides weak evidence that Big 4 auditors provide higher audit quality than Mid-tier auditors.

Medium and Small Auditors

In this section, we extend our analysis to small auditors (i.e., auditors other than the Big 4, Grant Thornton, and BDO Seidman). We estimate a propensity-score matching regression similar to Equation (1), with the exception that the *BIG4* variable is replaced by a Mid-tier indicator variable, which equals 1 for Mid-tier auditors, 0 otherwise. The full sample contains 8,772 firm-years, while the matched sample contains 1,406 firm-years. Panel B of Table 7 reports the results. We do not tabulate control variables in order to save space. Using either the full or matched sample, the results are similar. The coefficient on the Mid-tier indicator variable is negative, but not significantly different from zero. This suggests that clients of small auditors are no more likely to restate their earnings than are clients of Mid-tier auditors.[23]

ADDITIONAL ANALYSIS

Alternative Measures of Misstatements

A key assumption in our analysis is that restatements are an adequate proxy for audit quality. To test the robustness of our results, we repeat our analysis using the existence of Accounting and Auditing Enforcement Releases (AAERs) as a proxy for audit quality, following Lennox and Pittman (2010). We also examine the sensitivity of our results when using restatements that require a downward adjustment to net income.

Panel A of Table 8 reports the results of estimating Equation (2) after replacing the dependent variable with *AAER,* which equals 1 if the firm is subsequently targeted by the SEC in an

[23] We have also tested whether the Big 4 provide higher audit quality than small auditors. Using either the full or a matched sample, we find that clients of the Big 4 are significantly less likely to issue an accounting restatement than are clients of small auditors.

TABLE 7

Mid-Tier and Small Auditors

Panel A: Do Clients of the Big 4 Have Fewer Restatements than Clients of Mid-Tier Firms?

	Full Sample			Matched Sample		
Variable	**Estimate**	**Marginal Effect**	**p-value**	**Estimate**	**Marginal Effect**	**p-value**
Intercept	−1.951***		0.000	−5.095***		0.000
BIG4	−0.034	−0.50%	0.584	−0.266***	−2.50%	0.006
LNASSETS	0.042	0.61%	0.091	0.058	0.51%	0.236
ATURN	0.029	0.42%	0.155	0.077	0.69%	0.276
ROA	0.130	−2.62%	0.196	−0.455	−0.49%	0.232
LEV	−0.188	1.87%	0.164	−0.553*	−4.1%	0.060
CURR	−0.005	−0.07%	0.401	−0.014	−0.13%	0.555
BM	−0.020	−0.28%	0.398	0.089	0.79%	0.241
FIN	0.041	0.60%	0.556	0.313	2.80%	0.150
EPSGROW	−0.020	−0.29%	0.658	−0.327***	−2.40%	0.001
EP	0.076	1.09%	0.120	0.304**	2.72%	0.041
FREEC	0.175	2.53%	0.064	0.421**	3.77%	0.024
AGE	0.003	0.047%	0.891	0.028	0.25%	0.719
QUAL	0.090***	1.31%	0.000	0.149*	1.37%	0.087
INFLUENCE	−0.175	−2.53%	0.172	−0.287	−2.56%	0.373
LNFEE	0.019	0.27%	0.549	0.128	1.15%	0.184
Industry and Year FE	Yes			Yes		
n	31,346			3,248		
Pseudo R^2	0.071			0.119		

Panel B: Do Clients of the Mid-Tier Have Fewer Restatements than Clients of Small Auditors?

	Full Sample			Matched Sample		
Variable	**Estimate**	**Marginal Effect**	**p-value**	**Estimate**	**Marginal Effect**	**p-value**
Intercept	−1.365***		0.001	−6.055***		0.000
MIDTIER	−0.024	−0.25%	0.776	−0.127	−0.88%	0.354
CONTROLS	Yes			Yes		
Industry and Year FE	Yes			Yes		
n	8,772			1,406		
Pseudo R^2	0.067			0.133		

*, **, *** Denote statistical significance at the 10, 5, and 1 percent levels, respectively, using a two-tailed test.
This table displays coefficients from estimating a probit model (Equation (2)) where the dependent variable (*RESTATE*) equals 1 if the client subsequently issues an accounting restatement, 0 otherwise. In Panel A, the model is estimated on both the full Big 4 versus Mid-tier sample and the matched Big 4 versus Mid-tier sample. In Panel B, the model is estimated on a sample of Mid-tier and small audit firms. All independent variables are winsorized at the 1st and 99th percentiles to reduce in the influence of outliers. The model includes industry- and year-fixed effects, where industries are defined using two-digit SIC codes. p-values are based on Z-statistics, which are clustered by firm and year (Peterson 2009).

TABLE 8

Alternative Measures of Misstatements

Panel A: AAER Analysis

	Full Sample			Matched Sample		
Variable	**Estimate**	**Marginal Effect**	**p-value**	**Estimate**	**Marginal Effect**	**p-value**
Intercept	−2.595***		0.000	−11.663***		0.000
BIG4	−0.203*	−0.22%	0.092	−0.536**	−0.04%	0.034
CONTROLS	Yes			Yes		
Industry FE	Yes			Yes		
Year FE	Yes			Yes		
n	37,488			5,950		
Pseudo R^2	0.151			0.459		

Panel B: Income-Decreasing Restatements Only

	Full Sample			Matched Sample		
Variable	**Estimate**	**Marginal Effect**	**p-value**	**Estimate**	**Marginal Effect**	**p-value**
Intercept	−2.061***		0.003	−2.927***		0.000
BIG4	−0.089*	−1.16%	0.063	−0.182**	−1.70%	0.020
CONTROLS	Yes			Yes		
Industry FE	Yes			Yes		
Year FE	Yes			Yes		
n	37,488			5,950		
Pseudo R^2	0.064			0.053		

*, **, *** Denote statistical significance at the 10, 5, and 1 percent levels, respectively, using a two-tailed test.
Panel A displays coefficients from estimating Equation (2) after replacing the dependent variable with *AAER*, which equals 1 if the client is subsequently subject to an Accounting and Auditing Enforcement Release, 0 otherwise. Panel B displays coefficients from estimating Equation (2) after replacing the dependent variable with *NEGRESTATE*, which equals 1 if the firm subsequently issues an accounting restatement that results in a negative adjustment to net income, 0 otherwise. In all regressions, the control variables from Equation (2) are included. All independent variables are winsorized at the 1st and 99th percentiles to reduce in the influence of outliers. The model includes industry- and year-fixed effects, where industries are defined using two-digit SIC codes. p-values are based on Z-statistics, which are clustered by firm (Peterson 2009).

Accounting and Auditing Enforcement Release, 0 otherwise.[24] The model includes all control variables in Equation (2) but we do not report the coefficients on these controls to save space. Using either the full or the matched sample, the coefficient on *BIG4* is significantly negative, suggesting that Big 4 clients are less likely to receive an AAER.

Panel B reports the results of estimating Equation (2) after replacing the dependent variable with *NEGRESTATE,* which equals 1 if the firm subsequently issues an accounting restatement that

[24] AAER data are obtained from the Center for Financial Reporting and Management at the University of California, Berkeley. See Dechow, Ge, Larson, and Sloan (2011) for details on the dataset. The dataset contains details on AAERs up to 2011. The sample for this test is the same sample used in the main analyses.

results in a downward adjustment to net income, 0 otherwise. Using either sample, the coefficient on *BIG4* is significantly negative, suggesting that clients of Big 4 auditors are receiving higher quality audits.

To alleviate concerns that the results documented here are an artifact of the matching model, we replicate all of our analyses using the matching model of Boone et al. (2010).[25] We find that the results are qualitatively similar when using this model.

CONCLUSION

Because larger firms tend to select higher quality auditors, recent research questions whether the Big 4 provide higher audit quality than smaller auditors, once the endogenous choice of auditor has been controlled for (Lawrence et al. 2011; Boone et al. 2010). In this study, we re-examine this issue. We find that clients of Big 4 auditors are less likely to subsequently restate their earnings than are clients of non-Big 4 auditors. We also find weak evidence that clients of the Big 4 are less likely to issue a restatement than are clients of Mid-tier auditors (Grant Thornton and BDO Seidman). Taken together, the evidence is consistent with Big 4 auditors delivering higher quality audits. That we find different results when using a different audit quality proxy raises an important point. Proxies that are correlated with firm fundamentals (e.g., the absolute value of discretionary accruals) may not be capturing audit quality (see also Hribar and Nichols 2007). Similarly, readers must use caution when interpreting results from tests that use proxies that are actually measuring investors' perception of audit quality (e.g., the *ex ante* cost of equity capital).

The findings in this paper are subject to a few limitations. First, our inferences depend on the ability of accounting restatements to capture audit quality. We have presented arguments that support the use of accounting restatements as a proxy for audit quality. Nevertheless, to the extent that accounting restatements arising from materially misstated financial statements are not attributable to the auditor, our proxy may measure audit quality with error. As well, the use of accounting restatements to proxy for audit quality cannot tell us anything about audit quality differences among firms that did not issue restatements. We can only infer that firms that restated their earnings received inferior audit quality than firms that did not restate. Second, a limitation of any matching procedure is that we are unable to match firms based on pre-treatment attributes (Lawrence et al. 2011). Third, the findings in this paper are silent on the source of the Big 4 auditors' superior audit quality. The higher audit quality of Big 4 auditors may stem from these auditors' higher legal liability and/or greater economic rents to lose. On the other hand, the higher audit quality may simply be the result of better training programs (DeAngelo 1981). We leave this to future research.

REFERENCES

Asthana, S. C., and J. Boone. 2012. Abnormal audit fee and audit quality. *Auditing: A Journal of Practice & Theory* 31 (3): 1–22.

Becker, C. L., M. L. DeFond, J. Jiambalvo, and K. R. Subramanyam. 1998. The effect of audit quality on earnings management. *Contemporary Accounting Research* 15 (1): 1–24.

[25] The Boone et al. (2010) model is as follows: $BIG4 = \delta_0 + \delta_1 TA + \delta_2 lnSALES + \delta_3 LEV + \delta_4 EP + \delta_5 ISSUE + \delta_6 LOSS + \varepsilon$, where $BIG4 = 1$ if the client uses a Big 4 auditor, 0 otherwise; *TA* is the absolute value of total accruals, scaled by sales; *lnSALES* is the natural log of sales revenue; *LEV* is the debt to asset ratio; *EP* is the earnings-to-price ratio; $ISSUE = 1$ if the change in equity is greater than 10 percent, 0 otherwise; and $LOSS = 1$ if net income is negative and the absolute change in net income is greater than 10 percent, 0 otherwise.

Behn, B. K., C. Jong-Hag, and T. Kang. 2008. Audit quality and properties of analyst earnings forecasts. *The Accounting Review* 83 (2): 327–349.

Bentley, K., T. Omer, and N. Y. Sharp. 2013. Business strategy, financial reporting irregularities, and audit effort. *Contemporary Accounting Research* 30 (2): 780–817.

Blankley, A. I., D. N. Hurtt, and J. MacGregor. 2012. Abnormal audit fees and restatements. *Auditing: A Journal of Practice & Theory* 31 (1): 79–96.

Boone, J. P., I. Khurana, and K. K. Raman. 2010. Do the Big 4 and second-tier firms provide audits of similar quality? *Journal of Accounting and Public Policy* 29 (4): 330–352.

Carcello, J. V., T. L. Neal, Z.-V. Palmrose, and S. Scholz. 2011. CEO involvement in selecting board members, audit committee effectiveness, and restatements. *Contemporary Accounting Research* 28 (2): 396–430.

Cheffers, M. D., D. Whalen, and O. Usvyatsky. 2010. 2009 financial restatements: A nine year comparison. *Audit Analytics* (February). Available at: http://www.complianceweek.com/s/documents/AARestatements2010.pdf

Choi, J.-H., J.-B. Kim, A. Qiu, and Y. Zang. 2012. Geographic proximity between auditor and client: How does it impact audit quality? *Auditing: A Journal of Practice & Theory* 31 (2): 43–72.

DeAngelo, L. E. 1981. Auditor size and audit quality. *Journal of Accounting and Economics* 3 (3): 183–199.

Dechow, P., W. Ge, and C. Schrand. 2010. Understanding earnings quality: A review of the proxies, their determinants and their consequences. *Journal of Accounting and Economics* 50: 344–401.

Dechow, P., W. Ge, C. R. Larson, and R. G. Sloan. 2011. Predicting material accounting misstatements. *Contemporary Accounting Research* 28 (1): 17–82.

Easton, P. 2004. PE ratios, PEG ratios, and estimating the implied expected rate of return on equity capital. *The Accounting Review* 79 (1): 73–95.

Fargher, N. L., and L. Jiang. 2008. Changes in the audit environment and auditors' propensity to issue going-concern opinions. *Auditing: A Journal of Practice & Theory* 27 (2): 55–77.

Francis, J. R., and J. Krishnan. 1999. Accounting accruals and auditor reporting conservatism. *Contemporary Accounting Research* 16 (1): 135–165.

Francis, J. R., E. L. Maydew, and C. Sparks. 1999. The role of Big 6 auditors in the credible reporting of accruals. *Auditing: A Journal of Practice & Theory* 18 (2): 17–34.

Francis, J. R. 2004. What do we know about audit quality? *The British Accounting Review* 36: 345–368.

Francis, J. R., and P. Michas. 2013. The contagion effect of low-quality audits. *The Accounting Review* 88 (2): 521–552.

Francis, J. R., P. Michas, and M. Yu. 2014. Office size of Big 4 auditors and client restatements. *Contemporary Accounting Research* 30 (4): 1626–1661.

Grant Thornton. 2007. *After Decades Of Consolidation, the Onset of Sarbanes-Oxley and the Fall of Andersen, the Audit Environment Is Changing*. Available at: http://www.GrantThorton.com (last accessed May 30, 2007).

Guo, S., and M. Fraser. 2010. *Propensity Score Analysis: Statistical Methods and Applications*. Thousand Oaks, CA: SAGE Publications Inc.

Heckman, J. 1979. The sample selection bias as a specification error. *Econometrica* 47 (1): 153–162.

Hogan, C. 1997. Costs and benefits of audit quality in the IPO market: A self-selection analysis. *The Accounting Review* 72 (1): 67–86.

Hribar, P., and C. Nichols. 2007. The use of unsigned earnings quality measures in tests of earnings management. *Journal of Accounting Research* 45 (5): 1017–1053.

Ireland, J. C., and C. S. Lennox. 2002. The large audit firm fee premium: A case of selectivity bias? *Journal of Accounting, Auditing, and Finance* 17 (1): 73–91.

Jones, J. J. 1991. Earnings management during import relief investigations. *Journal of Accounting Research* 29 (2): 193–228.

Kennedy, P. 1992. *A Guide to Econometrics*. Cambridge, MA: MIT Press.

Kinney, W. R., Z.-V. Palmrose, and S. Scholz. 2004. Auditor independence, non-audit services, and restatements: Was the U.S. government right? *Journal of Accounting Research* 42 (3): 561–588.

Kothari, S. P., A. J. Leone, and C. Wasley. 2005. Performance matched discretionary accrual measures. *Journal of Accounting and Economics* 39 (1): 163–197.

Krishnan, G. V. 2003. Audit quality and the pricing of discretionary accruals. *Auditing: A Journal of Practice & Theory* 22 (1): 109–126.

Krishnan, J., and J. Krishnan. 1996. The role of economic trade-offs in the audit opinion decision: An empirical analysis. *Journal of Accounting, Auditing, and Finance* 11 (4): 565–586.

Lawrence, A., M. Minutti-Meza, and P. Zhang. 2011. Can Big 4 versus non-Big 4 differences in audit-quality proxies be attributed to client characteristics? *The Accounting Review* 86 (1): 259–286.

Lennox, C., and J. A. Pittman. 2010. Big Five audits and accounting fraud. *Contemporary Accounting Research* 27 (1): 209–247.

Lennox, C., J. R. Francis, and Z. Wang. 2012. Selection models in accounting research. *The Accounting Review* 87 (2): 589–616.

Lobo, G., and Y. Zhao. 2013. Relation between audit effort and financial report misstatements: Evidence from quarterly and annual restatements. *The Accounting Review* 88 (4): 1385–1412.

Newton, N. J., D. Wang, and M. S. Wilkins. 2013. Does a lack of choice lead to lower quality? Evidence from auditor competition and client restatements. *Auditing: A Journal of Practice & Theory* 32 (3): 31–67.

Palmrose, Z.-V., and S. Scholz. 2004. The circumstances and legal consequences of non-GAAP reporting: Evidence from restatements. *Contemporary Accounting Research* 21 (1): 139–180.

Peel, M. J., and G. H. Makepeace. 2012. Differential audit quality, propensity score matching, and Rosenbaum bounds for confounding variables. *Journal of Business, Finance, and Accounting* 39 (5): 606–648.

Peterson, M. 2009. Estimating standard errors in finance panel data sets: Comparing approaches. *Review of Financial Studies* 22: 435–480.

Plumlee, M., and T. L. Yohn. 2010. An analysis of the underlying causes attributed to restatements. *Accounting Horizons* 24 (1): 41–64.

Rosenbaum, P. R., and D. B. Rubin. 1983. The central role of propensity score in observational studies for causal effects. *Biometrika* 70 (1): 41–55.

Subramanyam, K. R. 1996. The pricing of discretionary accruals. *Journal of Accounting and Economics* 22 (1–3): 249–281.

Teoh, S. H., and T. J. Wong. 1993. Perceived auditor quality and the earnings response coefficient. *The Accounting Review* 68 (2): 346–366.

U.S. Chamber of Commerce. 2006. *Auditing: A Profession at Risk*. Washington, DC: U.S. Chamber of Commerce.

Social Science Research 74 (2018) 45–61

Contents lists available at ScienceDirect

Social Science Research

journal homepage: www.elsevier.com/locate/ssresearch

Occupational status benefits of studying abroad and the role of occupational specificity – A propensity score matching approach

Stine Waibel[a,*], Knut Petzold[b], Heiko Rüger[a]

[a] Federal Institute for Population Research (BiB), Friedrich-Ebert-Allee 4, 65185 Wiesbaden, Germany
[b] Ruhr-Universität Bochum, Faculty of Social Science, Universitätsstraße 150, 44801 Bochum, Germany

ARTICLE INFO

Keywords:
Study abroad
International student mobility
Labor market returns
Occupational status
Occupational specificity
Propensity score matching

ABSTRACT

Occupational status benefits of student mobility remain uncertain, despite increasing interest in the implications of international student mobility for the reproduction of societal inequality. Since mobile young people are a selective group in terms of socio-economic and achievement-oriented factors, we apply propensity score techniques to test whether German higher education graduates who did or did not study abroad differ in occupational status (based on the Socio-Economic Index of Occupational Status) three years after graduation. Analyses are based on multi-cohort representative data of the German population (Working and Learning in a Changing World). Results confirm a positively biased effect of mobility on early career occupational status driven by compositional differences. Subgroup analyses show that even when accounting for this bias, occupational status returns to mobility are positive for those graduating in occupationally unspecific fields of study. There are no returns for those graduating in occupationally specific fields of study. Findings also suggest that the effect of studying abroad is not homogeneous across the study population. Individuals less likely to study abroad are at the same time more likely to reap the occupational benefits from this experience.

1. Introduction and research question

As international exchange schemes and fellowships have gained popularity and as the international education market rapidly grew into a 'multi-billion dollar industry' (Waters, 2006, 180) different disciplines developed a sustained interest into the characteristics, determinants, and consequences of studying abroad during higher education. Up to now, most research has been driven by the question *who* studies abroad (or who doesn't) closely related to discussions about growing horizontal education-based stratification (Lörz et al., 2016; Triventi, 2013). The group of students studying abroad is highly socially selective in terms of the economic, cultural, and social capital of the students' families (Brooks and Waters, 2010; Netz and Finger, 2016; Gerhards and Hans, 2013).

Given that the share of the population reaching secondary and postsecondary levels of education has increased substantially across the educationally expanding Western world, competitions for privileged positions in society have intensified. It is likely that horizontal characteristics of education such as international mobility become more important erecting new dimensions of social stratification and new types of social inequality (Gerber and Cheung, 2008; Reimer and Pollak, 2010, 427). In fact, the benefits of international mobility are often taken for granted. Being mobile possibly strengthens graduates' skills, resources, and competitive advantage on the job market and for this reason may be valued by individuals as well as potential employers (e.g., Opper, 1991).

Based on these observations and assumptions, it is essential to figure out whether studying abroad actually yields returns in terms

* Corresponding author.
E-mail address: stine.waibel@bib.bund.de (S. Waibel).

https://doi.org/10.1016/j.ssresearch.2018.05.006
Received 25 August 2017; Received in revised form 8 May 2018; Accepted 28 May 2018
Available online 29 May 2018
0049-089X/

of earnings and occupational status and, thereby, might increase socioeconomic inequality given its social selectivity. To this end, the past decade has seen major advances in estimating the socioeconomic consequences of studying abroad, mainly in terms of income achieved a few years after graduation (e.g., Di Pietro, 2015; Kratz and Netz, 2016; Messer and Wolter, 2007; Oosterbeek and Webbink, 2006; Sorrenti, 2017; Wiers-Jenssen, 2011). However, analyzing the returns to studying abroad has confronted researchers with several methodological challenges. Chief among them is the necessity to control for systematic selection into study abroad programs on the basis of pre-study-abroad individual and family determinants that are themselves related to future career outcomes (endogeneity bias). Therefore, statistical methods estimating the effect of mobility must account for compositional differences between mobile and immobile persons in order to tackle this paper's central question: *is there a general causal relationship between studying abroad and early career occupational success*?

Although most studies fall short of methodological rigor needed to address the stated challenge, efforts have recently been made to minimize the endogeneity bias via experiments (Petzold, 2017a,b), instrumental variable estimation (Di Pietro, 2015; Messer and Wolter, 2007; Sorrenti, 2017), propensity score matching (Euler et al., 2013), regression discontinuity designs (Oosterbeek and Webbink, 2006; Rodrigues, 2013), and effect decomposition (Kratz and Netz, 2016) (for a review see Waibel et al., 2017).

Results of existing studies are ambiguous and deeply context-dependent so that the question whether studying abroad exerts some causal effect on employment outcomes is far from settled. Some studies find that income returns to studying abroad are significantly reduced (Kratz and Netz 2016 (Germany)) or even diminish (Messer and Wolter 2007 (Switzerland); Oosterbeek and Webbink 2006 (Netherlands)) once accounting for possible confounders or self-selection. Other studies identify a substantial and significant causal effect of studying abroad on the likelihood to be in employment a few years after graduation (Di Pietro 2015 (Italy)) and on starting salaries (Sorrenti 2017 (Italy)). Again others highlight that study abroad increases the probability of invitation to a job-interview, but only in international work contexts (Petzold 2017a (Germany)). Hence, discussions about if and under which conditions studying abroad brings about occupational returns are ongoing.

We propose that a more coherent idea of the professional impact of studying abroad could be established if the institutional stratification of the educational and vocational systems is taken into account. Needless to say, advancing on the job market and achieving a high occupational status is not only based on individual-level resources, but reflects the institutional relationship between the educational system and the labor market (e.g., Allmendinger, 1989; Kerckhoff, 1995; Roksa and Levey, 2010; van de Werfhorst, 2004). In this regard, fields of study in higher education differ considerably in their vocational embedding. We assume that this field-specific relation to the labor market will affect the links between studying abroad and occupational outcomes.

Solid evidence on differential returns to studying abroad contingent on the vocational embedding of the field of study is lacking so far. There is some indicative support for our claim to be outlined below that economic returns to study abroad are highest in more occupationally unspecific fields of study (Janson et al., 2009; Kratz and Netz, 2016; Opper et al., 1990). However, none of these investigations has applied methods that both control for selection and perform subgroup analysis.

Using data that is representative of the population living in Germany, the fundamental aim of this study is to provide evidence on the occupational returns of studying abroad a few years after graduating from higher education. We focus on the occupational status at the *early career* stage for two reasons. First, tertiary level graduates face a situation of heightened competition when transitioning from education to work. At this career development stage the relative value of one's accumulated educational credentials is particularly relevant for getting access to attractive labor market positions (cf., Gangl, 2002). Respondents may still be in their first job three years after graduation yet search, adaptation, and promotion processes may have already taken place during the first three years of employment. Second, we face data limitations with respect to investigating the effect of studying abroad on longer-term career outcomes. Given that we use multi-cohort data (see section 4.1.), complete employment histories are only available for a small number of respondents.

We make several contributions to the existing literature. First, we use a multi-cohort representative sample including individual-level information on residential, educational, and employment biographies that has so far not been exploited for investigating the returns to studying abroad. Second, while most research has focused on income returns to studying abroad, we examine the value of studying abroad in terms of the occupational status it brings that is of particular interest to sociological analysis. Using status as our central outcome we follow a long tradition in social mobility research characterizing the status of jobs by a socioeconomic index, representing not only economic capital but also the symbolically legitimated social organization of modern societies (Featherman et al., 1975). Third, we apply a propensity score matching approach in order to account for compositional differences between mobile and immobile graduates and to identify the causal effect of studying abroad. The matching estimates depend on the assumption that unmeasured confounders are “ignorable” once measured confounders are controlled for. Although we cannot rule out unobserved heterogeneity, we are still able to adjust effect estimates for a rich set of covariates. Moreover, we will discuss unobserved heterogeneity with respect to the question of who benefits most from studying abroad. Fourth, we perform subgroup analyses to test our hypothesis that the connection between studying abroad and occupational allocation is embedded in institutional processes specific to the field of study.

2. Theoretical considerations

2.1. Cause or selection?

In general, connections between educational attributes and labor market allocations are understood in the context of two-sided matching processes between applicants' skills and the requirements of employers’ jobs. Employers evaluate potential employees and employees evaluate potential employers in order to achieve the most optimal match to maximize both returns from educational

investments for the applicant and job performance for the employer (Kalleberg and Sørensen, 1979; Müller, 2005). Human capital and signaling approaches are the two most prominent theories that address how education in general is causally related to job assignment (cf., Bills, 2003; Kjelland, 2008).

First, economists typically model mobility and education as an investment in one's stock of human capital, which refers to the total of one's skills, knowledge, and abilities (cf. Becker, 1964; Mincer, 1958; Schultz, 1961). From a rational choice point of view, investments in human capital are the results of a cost-benefit analysis and made if they promise to yield economic returns over the long period. In these premises, students' primary consideration for becoming internationally mobile may be the acquisition of human capital (King and Findlay, 2015, 262). Study-related international mobility is an investment in foreign language proficiency, personal development, and social, intercultural, and mobility skills. Taken together, this human capital is valued on globalized labor markets that demand a high degree of global awareness, self-initiative, and personal responsibility (e.g., Gerhards and Hans, 2013; King and Findlay, 2015; Potts, 2015; Salisbury et al., 2009).

Second, according to proponents of the signaling approach educational degrees and grades do not necessarily lead to an accumulation of human capital. They mainly serve as signal or credential that conveys information about one's inherent ability and, thus, labor market productivity (cf. Spence, 1973; Stiglitz, 1975). Weiss (1995) understands the signaling approach in extension, and not in opposition, to human capital theory. In his view it is an advantage of signaling models that they conceptualize formal educational credentials not merely as representation of observable trained skills but also of unobserved individual attributes.

With respect to studying abroad it is argued that employers positively screen job applicants for taking part in study-related mobility because the former assume this to be related to unobserved characteristics such as motivation, initiative, or flexibility that also make them more productive in the workplace (e.g., Hilmer, 2002; Petzold, 2017a; Wiers-Jenssen, 2008). But regardless of whether a human capital or signaling perspective is taken – from the individual perspective it matters little what mechanisms work in practice (Kjelland, 2008, 70) – the fundamental idea is that studying abroad benefits professional career trajectories. Therefore, the juxtaposition of human capital and signaling approaches has struck observers as being overly divisive, since both can be integrated into a common argument about the labor market value of education (Rospigliosi et al., 2014).

It is well-known that educational choices interact with ability, resources, and socially generated attitudes such as risk aversion and time preferences, explaining why education and mobility strategies are more often pursued by high-achieving individuals from upper and middle class backgrounds (e.g., Boudon, 1974; Breen and Goldthorpe, 1997; Schultz, 1961). It seems that international educational mobility, too, is the result of preceding social selection processes (Lörz and Krawietz, 2011, 202; see also Di Pietro and Page, 2008; Powell and Finger, 2013; Souto-Otero, 2008), although differences across countries exist (van Mol and Timmerman, 2014). Specifically in Germany, students that have parents with an academic degree and sufficient financial resources are statistically much more likely to study abroad (Lörz et al., 2016). Hence, one cannot disregard the possibility that the expected positive occupational effect reduces to self-selection of more capable individuals into study abroad programs.

Taken together, our first hypothesis is that there is a positive association between studying abroad and occupational status that can be attributed in part to (observed) self-selection or the compositional overrepresentation of high-achievement groups and in part to causality as predicted by human capital and signaling theories. This distinction is key, yet, not the whole story. An important drawback of previous research has been the incomplete theoretical specifications about the applicability of human capital or signaling processes to different educational contexts, which may further explain inconsistencies in empirical findings.

2.2. Differential effects

Research on the transition from school to work acknowledges that national specificities and the extent to which educational programs and sectors provide occupation specific training have a substantial impact on occupational returns (e.g., Kerckhoff, 1995; Klein, 2016; Leuze, 2007; Scherer, 2005; Roksa and Levey, 2010). In Germany, tertiary education induces close links between training obtained and the occupationally segmented structure of jobs, as well, with universities providing ‘standardized educational products’ (Müller et al., 1995, 6). The state-regulated higher education system offers various occupation-specific courses – such as medicine, law, and education – that guarantee standardized access to professional and managerial positions within a few years or even months after graduation (Leuze, 2007), i.e., independent of any additional skills gained via international mobility (Kratz and Netz, 2016, 12; 18). Although additional skills such as studying abroad may not be irrelevant for practicing the profession, it is likely that these will not have a substantial effect on vertical occupational allocation.

In contrast, if study fields lack such occupation-specific orientation employers will anticipate greater training costs and will therefore screen applicants more carefully, especially given educational expansion and intensifying competition among graduates at tertiary educational level (Reimer et al., 2008; Reimer and Pollak, 2010). Career success and occupational status may depend more on means that expand and refine one's skills, and increase qualitative differences between graduates. Supporting this logic, it has been shown for the German context that completing an internship while attending university provides positive labor market returns mainly for graduates from occupationally unspecific fields of study (Saniter and Siedler, 2014).

This speaks to a growing literature seeing study abroad as ‘positional good’. A positional good can be defined as ‘a good, valuable to some people only on condition that others do not have it’ (Hollis, 1982, 236). Therefore, studying abroad may have increased in (relative) value for individuals who are increasingly alike in terms of educational attainment (Waters and Brooks, 2010, 218), particularly in unspecific fields. Our second hypothesis is that the occupational returns to studying abroad depend on the occupational specificity of the respective field of study.

3. Methodology

3.1. Causal inference

In this paper we want to identify the causal effect of studying abroad during higher education on occupational status. Causal inference is prominently conceptualized within the potential outcomes approach (Morgan and Winship, 2015). Pursuant to the vocabulary of scientific experimentation, two states – the treatment state and the control state – can be distinguished that units of observation are assigned to. In our analysis, graduates are assigned either to the study abroad group (treatment) or the non-study abroad group (control). Both states bear potential outcomes for each unit of observation and the difference between these potential outcomes for each unit *i* is the individual-level causal effect of the treatment condition on the outcome of interest. Since any unit of observation can only be subject to one of the potential states, we observe only the outcome Y_i^* that is realized under the respective treatment condition D_i. In our case, we can only observe the outcome (occupational status) either for graduates in the treatment state (study abroad) or in the control state (non-study abroad).

The individual-level causal effect, however, can be approximated. For this purpose, counterfactual values for potential outcomes are constructed using information of comparable units in the alternative state. The average treatment effect on the treated (ATT) is the mean difference between the expected outcome of the observed values of units that were subjected to the treatment condition ($D = 1$) and the expected outcome of the *counterfactual* values, which are estimated (Morgan and Winship, 2015, 55):

$$ATT = E[Y^1|D = 1] - E[Y^0|D = 1]$$

Here, the ATT is the expected what-if difference in occupational status if we could expose a randomly selected person from the study abroad group to both the study abroad and the non-study abroad condition.

In contrast to the ATT, the average treatment effect on the untreated or controls (ATC) estimates the average effect for units that did not receive treatment ($D = 0$). In other words, the ATC is the expected difference between the observed outcomes among individuals who did not study abroad and the estimated counterfactual outcomes:

$$ATC = E[Y^1|D = 0] - E[Y^0|D = 0]$$

Finally, the average of the ATT and the ATC, weighted by comparison group sizes, is the average treatment effect for the whole sample (ATE).

The comparability of average outcome values builds on the decisive assumption that units of observation being compared have to show an equal distribution of observed and unobserved characteristics (e.g., gender, age, education) except for the treatment to make sure that results are not biased by confounding variables (Morgan and Winship, 2015). As the treatment must be independent from confounding covariates X, this is referred to as the conditional independence assumption (CIA)[1]

$$Y^1; Y^0 \perp D|X$$

3.2. Propensity score matching

While in experimental studies units are randomly assigned to treatment or control conditions meeting the CIA by design, extending causal inference into observational studies is problematic since researchers cannot control the assignment of observational units to treatment and control condition (Rosenbaum, 2010, 153ff). Instead, self-selection of units into treatment and control condition takes place and a simple comparison of units in the treatment and control condition would be biased because covariates are not at balance and associated with the treatment.

A prominent procedure for balancing covariates related to the observational units in the treatment and control conditions is propensity score matching (PSM) (e.g., Morgan and Harding, 2006), which is used with increasing frequency in the literature assessing the effects of education and educational contexts (e.g., Brand and Halaby, 2006) or residential mobility (e.g., Haelermans and Witte, 2015). In this method, the first step is to reduce the multi-dimensional differences in covariates between units of treatment and control condition to one dimension, the propensity score. In the second step, the propensity score serves as distance metric for the identification of fitting counterfactual outcomes and matching them to the observed outcomes. Observed and counterfactual outcomes are compared in the third step in order to estimate causal effects. In line with the potential outcomes framework outlined above, the ATT is estimated as the average difference between observed outcomes Y_i in the treatment sample ($D=1$) and the appropriately matched observed outcomes Y_j in the control sample ($D=0$).

The propensity score is usually estimated by parametric models, such as logistic regression, and expresses the probability of experiencing the treatment given a set of observed pre-treatment covariates that determine self-selection into treatment. For the assignment of counterfactual outcomes to observed outcomes, different matching algorithms can be used that calculate weights for each comparison unit as a function of the (estimated) propensity scores reflecting the observational similarity of observations regarding the covariate space (for an overview, see for example Gangl, 2015). The more comprehensively algorithms include (weighted) sample observations to estimate the treatment effect, the more efficient matching estimators become (i.e., smaller

[1] Instead of CIA, others refer to unconfoundedness and selection on observables. An additional assumption is that the treatment effect for an individual does not depend on who else also receives the treatment, referred to as stable unit treatment value assumption (or SUTVA).

variance). For example, kernel matching estimators use all observations in the comparison group inside the common support region in constructing the estimated counterfactual outcomes. However, this efficiency comes potentially at some loss of covariate balance. We will compare estimates using different matching algorithms in order to be sensitive to the robustness of the results.[2]

PSM has major advantages over regression analyses. First, in contrast to the standard regression solution, where the goal is to simultaneously estimate the effects of the treatment and a set of covariates, X, on the outcome, Y, a matching estimate non-parametrically balances the variables in X across D (i.e., the treatment groups) in order to obtain the best possible estimate of the *causal* effect of the treatment on Y. Matching on the propensity score, thus, forces researchers to examine the distribution of covariates across units exposed to treatment and control condition and to identify sample cases that are comparable taking into account processes of self-selection that are based on previous empirical knowledge and theory (cf., Morgan and Harding, 2006). Second, matching is only justified when performed over the so called region of common support in the sample data, that is, the estimated propensity score (or the covariate data) has to overlap across the comparison groups of the analysis. If not, observations are not comparable and excluded from the analyses to avoid fundamental mismatches. If there are many non-matching cases, this will be reflected in a relatively high variance of the treatment effect estimators. Finally, in using non-parametric matching estimators, propensity score matching makes less stringent parametric assumptions and may surpass linear regression modeling when the true functional form of a relationship is nonlinear. For the same reason, propensity score matching can be applied even if outcome variables are not distributed normally.

A limitation of PSM is that since it rests on the conditional independence assumption it can only deal with confounding based on observed covariates.[3]

4. Data, sample, and variables

4.1. Data and analytic sample

Analyses are based on the forerunner study of the German National Educational Panel Study (NEPS), called Working and Learning in a Changing World (German acronym, ALWA). ALWA is a nationally representative sample of 10,177 adult German residents (regardless of their nationality) interviewed in 2007 and 2008 when they were 18–53 years old. ALWA provides comprehensive retrospective information on participants' complete educational and employment trajectories and respondent characteristics such as parental background and places of living (see Kleinert et al., 2011).

Our analytical sample is restricted to individuals who obtained a higher education degree and who were in regular gainful full or part time employment three years after graduation.[4] Graduation is defined as the point of leaving the higher education system after the latest degree has been obtained. Traineeships and practical training periods connected to educational degrees are counted as part of the training system and not as employment. We intended to keep the sample homogeneous in terms of educational trajectories and considered only educational episodes that started when a person was 30 years old or younger. As exception to this rule we considered late episodes if they corresponded to the first tertiary degree the person pursued.

Like other countries, Germany has a stratified higher education system (Leuze, 2011). However, compared to Anglo-Saxon countries, there are no pronounced differences within individual university types in terms of reputation and prestige (Müller et al., 1995, 6; Reimer and Pollak, 2010, 417). Three basic types of higher education exist: traditional universities (*Universitäten*), universities of applied sciences (*Fachhochschulen*) and universities of cooperative, administrative and economic education (*Berufsakademien*), which combine elements of firm-based training with academic education. Traditional universities cover the full range of academic disciplines and promise access to the most favorable class positions (Müller et al., 2002). Universities of applied sciences and universities of cooperative education offer a more limited, applied, business and engineering oriented set of disciplines. We chose to include individuals with degrees from all three university types because all school leavers and prospective students perceive pursuing a degree in one of these institutions as attractive educational alternatives (Trautwein et al., 2006).[5]

Our final analytic sample includes 1708 higher education graduates of which 44 percent were female and 56 percent were male. Individuals spent on average five years in higher education and graduated at mean age of 27. About 40 percent of individuals in the sample graduated between 1975 and 1990, while the remaining 60 percent graduated between 1991 and 2005.

[2] Irrespective of the specific matching algorithm, all matching estimators of the ATT can be expressed as (Gangl, 2015, 257) $ATT_M = \frac{1}{n_{D_1}} \sum_{i \in D=1} \left[y_i^1 - \sum_{j \in D=0} w_{ij} y_j^0 \right]$, where $i \in D = 1$ and $j \in D = 0$ denote individuals in the treatment and control group, respectively, n_{D_1} is the number of individuals in the treatment group (consider that the ATT is an average), and w_{ij} are the weights derived from the matching algorithms and the propensity scores. The weights express the algorithm-scaled distance between each control case and the target treatment case based on the covariate information (i.e., the similarity between observations) so that more weight is given to control group members equivalent to those in the treatment group (Morgan and Winship, 2015, 155). w_{ij} does not alter the treated cases ($w_{ij} = 1$), but only adjusts the control cases to match the treatment cases (Davidson and Sanyal, 2017, 1705). The weights for the ATC work in the opposite direction. They do not alter the control cases, but attempt to turn the treatment group into a comparison sample for the control cases with respect to the distribution of covariates (Morgan and Todd, 2008, 244).

[3] To account for unmeasured confounders, studies often use instrumental variables methods. In this approach, "instruments" have to be found that are associated with treatment choice but not with the outcome. The variation in the treatment choice affected by the instruments is then used to generate unbiased causal effect estimates. Identifying good instruments is not straightforward and was therefore not a viable strategy in our study.

[4] In the full respondent sample, 93 percent of men and 85 percent of women were in regular employment and only four percent were unemployed three years after graduation.

[5] In our analyses we excluded first-generation immigrants, as their foreign education predominantly takes place in immigrants' countries of origin and can hardly be transferred to the host-country's labor market (Chiswick and Miller, 2009; Friedberg, 2000).

4.2. Outcome: occupational status

Within the ALWA survey information on respondents' occupation is coded using the International Standard Classification of Occupations 1988 (ISCO88). We calculated the International Socio-Economic Index (ISEI; see Ganzeboom and Treiman, 1996) as our measure of occupational status using the conversion table provided by the German Microdata Lab.[6] The ISEI range is 16–85.

We chose three years after graduation as point in time for measuring occupational status assuming that we capture the labor market outcomes of graduates in their early careers, still unaffected by mobility processes occurring after labor market entry (cf. Gebel, 2009, 671). Most empirical studies in the German context show that individuals leaving the higher education system successfully enter the labor market even after initial periods of unemployment, have a low risks of involuntary job loss, and have relatively stable labor market careers (e.g., Biemann et al., 2011; Giesecke and Heisig, 2011; Lindberg, 2009; Scherer, 2005). Moreover, by choosing our time of measurement we intended to build on previous studies that use graduate tracer surveys to analyze the economic returns to study abroad sometime between labor market entry and five years after graduation (cf., Di Pietro, 2015; Kratz and Netz, 2016; Messer and Wolter, 2007; Wiers-Jenssen, 2011).

4.3. Treatment: study abroad

Our conceptualization of study abroad includes any stay abroad during educational spells, not limited to participation in a particular mobility or exchange program. Importantly, individuals pursuing an entire graduate or postgraduate degree abroad were excluded. Theoretically, we expected that these graduates, sometimes called 'degree mobiles', may face readjustment and recognition issues when entering the labor market of their origin countries similar to immigrants with foreign schooling (cf., Wiers-Jenssen, 2013). Moreover, degree-mobile students are a quite heterogeneous group so that assumptions about the eventual impact of studying abroad are difficult to make (Teichler, 2012, 44).

Given the multidimensional life-course data that is contained in ALWA it was possible to match residential and educational histories and to identify overlaps in educational and residential episodes abroad. In addition, survey participants indicated whether they spent at least one month in a foreign country during a given educational spell. In these cases, we had no information on the length of the stay abroad or the host country, but observations could none the less be integrated into our binary treatment indicator (1 = study abroad, 0 = no study abroad). Overall, we identified N = 219 individual biographies (13 percent of the analytic sample) that included at least one period abroad during higher education. For those persons in the sample for who we had information, the most common destinations of the stay abroad were Northern Europe (Great Britain and Scandinavia), North America, and France.

4.4. Covariate choice and common causes

The effect of studying abroad corresponds to the difference in mean levels of occupational status across units that differ in study abroad experience yet share a similar propensity to participate in such mobility based on a set of observed covariates. These covariates include pre-study-abroad individual characteristics that determine selection into the treatment, and are also related to the outcome. These common causes introduce bias in the effect estimates, confounding the causal effect of the treatment on the outcome.

We condition on a rich set of covariates provided by ALWA, to be presented below. These covariates figure prominently in sociological research on student mobility (e.g., Di Pietro and Page, 2008; Finger, 2011; Gerhards and Hans, 2013; Lörz et al., 2016; Schindler and Lörz, 2012; van Mol and Timmerman, 2014). An overview of operationalizations of covariates is given in Table A1 in the Appendix.

4.4.1. Parental educational background

Student mobility has repeatedly been characterized by its social selectivity. In particular, parents' education has been a stable and good predictor of variation in student mobility (cf., Lörz et al., 2016; Souto-Otero, 2008). It reflects a child's learning environment as well as the cultural and social capital of the family and has influence on educational decision making especially in stratified educational systems (Buis, 2013; Erola et al., 2016).

Accordingly, we differentiated between graduates whose parents both have education below tertiary level (ISCED 1 to 4) and graduates with at least one parent (mostly, the father) holding a tertiary educational degree (ISCED 5A and 6). We did not differentiate parental education further, as there were few cases in the sample where parents have lower secondary education (ISCED1-2). Ultimately, in our sample, differences in mobility patterns by social background emerged mainly between first-generation university students and students with at least one parent having a university degree (similarly, see Kratz and Netz, 2016; Lörz et al., 2016; Kratz and Netz, 2016; Salisbury et al., 2009).

4.4.2. Educational trajectory: Abitur, university type, student jobs and internships, vocational and public-sector training, previous international experience

Rather than having itself an effect on labor market achievement, studying abroad may simply be more likely within educational trajectories that lead to higher status jobs (Lörz et al., 2016), such as academic aspirations, extra-curricular engagement, vocational orientation of mobile students. We include several covariates capturing this process.

[6] http://www.gesis.org/missy/materials/MZ/tools/isei.

In Germany, the direct transfer from school to university is traditionally regulated by the *Abitur*, designed to certify young adults 'general academic ability' (*Studierfähigkeit*). Though educational expansion has lead more and more persons to reach this upper secondary level of education (Lörz and Schindler, 2009, 101), the *Abitur* is not a necessary requirement for non-traditional university types and a greater openness of traditional universities towards so called non-traditional students without Abitur has been formalized more recently (Brändle and Lengfeld, 2016). In general, the *Abitur* can still be seen as academic credential and indicator of higher educational aspirations associated with studying abroad.

Moreover, internationalization and Europeanization have made stronger inroads into the academically oriented traditional university system than the technical and applied system (DAAD and DZHW, 2017). Consequently, graduates from traditional university have more opportunities to study abroad than graduates from non-traditional universities. The occupational position of mobile graduates may, thus, reflect the higher returns to degrees obtained from traditional universities that are often the sole providers of the more prestigious study courses (e.g., medicine).

Students who study abroad may also be more engaged in other out-of-class experiences conforming to a 'culture of engagement' associated with intellectual and personal development helping them to aspire prestigious occupations in society (Brint et al., 2008). For example, the same persons who study abroad may also work on-campus as student assistants and collect internship experiences during their educational trajectories.

In addition, people who complete vocational training prior to tertiary education are more bound to a certain company, usually in Germany, and tend to be less mobile (Parey and Waldinger, 2011, 213). Similarly, a career in the public sector is associated with less student mobility. Public service careers are rather internally orientated with strong institution-specific labor markets and on-the-job training (Leuze, 2010, 56ff). By nature of their educational trajectory, individuals from public and vocational career tracks tend be in technical, administrative, and associate level positions corresponding to lower occupational statuses (cf., Anger et al., 2010).

Finally, the propensity to study abroad at university is also determined by previous international experiences (Brooks and Waters, 2010; Carlson, 2013; Finger, 2011; Lörz et al., 2016; Petzold and Peter, 2015), as it creates confidence in internationalized social environments, develops one's capacity to be mobile, and confirms international dispositions through interaction with similar others. That way, pre-university international experiences in primary and secondary education, as part of a family stay abroad, student exchange or voluntary service, may already endow individuals with competences that set the stage for successful careers.

4.4.3. Field of study and occupational specificity

Studying abroad is more common in some fields of studies than in others and the chosen field of study may affect the propensity to study abroad (e.g., Stroud, 2010; Salisbury et al., 2009) as well as the occupational rewards (Kratz and Netz, 2016). Higher rates of student mobility in Germany have been observed in humanities, economics, and medicine, lower rates in law, natural sciences, engineering, as well as technical fields that are more often studied at applied universities (Gerhards and Németh, 2015; DAAD and DZHW, 2017). Therefore, on the one hand, studying abroad may not by itself produce occupational gains, but may be driven by higher rates of student mobility in fields of study, like medicine, that lead directly into the most prestigious jobs. On the other hand, the effect of studying abroad on occupational status may vary by field of study.

To examine this moderation effect, we conceptualize occupational specificity of the field of study as the degree to which the field has occupational counterparts in the labor market (Roksa and Levey, 2010). Graduates from fields with high occupational specificity are predominantly employed in one occupational category related to the field of study. In contrast, graduates from occupationally unspecific fields are distributed across different occupations and not concentrated in a specific occupation that relates to their studies. Operationalization is based on the data (see Table A2 in the Appendix). We identified law, medicine, teaching (primary and secondary level), engineering, architecture, computer science, physics and related natural sciences as well as public administration as occupationally specific fields of study, whereas other educational sciences, social sciences, psychology, life science, math and statistics as well as technical and commercial subjects were identified to be occupationally unspecific fields (similarly, see Leuze, 2010, 140; Saniter and Siedler, 2014, 44).

4.4.4. Region and year of birth

Research shows that employment trajectories in the Eastern part of Germany tend to be more instable than in the West even in the higher educated segment of the population (Falk et al., 2000; Kurz and Steinhage, 2001). Significant differences between students raised in the East of Germany and students raised in the West in terms of international experience and willingness to become mobile have been demonstrated (Spieß and Brüch, 2002). Therefore, the propensity to study abroad may be related to the region of birth which may in turn be positively related to higher occupational achievements after graduation.

Finally, given the multi-cohort design of the sample, individuals experienced their education and early careers at different points in time reflecting changing labor market returns to higher qualifications. At the same time, there has been a stark increase study abroad up until the more recent years. Therefore, year of birth may be a confounder in the association between study abroad and occupational status.

4.4.5. Language skills

Although language skills (number of spoken languages and English language proficiency) may be positively related to studying abroad and occupational status, we cannot examine if they are clearly prior to studying abroad, since the information is only available in cross-sectional data records. Foreign language proficiency is likely to be influenced by studying abroad itself (cf. Kinginger, 2008). In this case, foreign language skills would correspond to human capital acquired *after* studying abroad and would mediate the association between studying abroad and occupational achievement (cf., Sorrenti, 2017). They are, thus, not eligible as

Table 1
Sample composition (full sample and by occupational specificity of study field).

	Full Sample				Graduates from specific fields				Graduates from unspecific fields			
	Mean	S.D.	Min	Max	Mean	S.D.	Min	Max	Mean	S.D.	Min	Max
ISEI 3 years after graduation	63.87	12.79	20.00	85.00	67.32	12.08	25.00	85.00	58.35	11.92	20.00	85.00
Study Abroad	.13	.33	.00	1.00	.11	.32	.00	1.00	.15	.36	.00	1.00
Higher Educated Parent	.34	.48	.00	1.00	.35	.48	.00	1.00	.33	.47	.00	1.00
Abitur	.80	.40	.00	1.00	.79	.41	.00	1.00	.81	.39	.00	1.00
Master Traditional University	.57	.50	.00	1.00	.55	.50	.00	1.00	.60	.49	.00	1.00
Applied Uni/Bachelor Trad. Uni.	.36	.48	.00	1.00	.42	.49	.00	1.00	.28	.45	.00	1.00
Uni. Coorp. Education	.07	.25	.00	1.00	.04	.18	.00	1.00	.12	.33	.00	1.00
Student Assistant	.11	.31	.00	1.00	.12	.32	.00	1.00	.09	.29	.00	1.00
Vocational Training	.37	.48	.00	1.00	.35	.48	.00	1.00	.42	.49	.00	1.00
Career in Public Sector	.05	.22	.00	1.00	.07	.26	.00	1.00	.02	.14	.00	1.00
Previous Mobility Experiences	.06	.23	.00	1.00	.05	.21	.00	1.00	.08	.27	.00	1.00
Field of Study												
Agrig./admin./techn./oth.	.14	.34	.00	1.00	.17	.37	.00	1.00	.09	.29	.00	1.00
Science, Eng., Math	.34	.48	.00	1.00	.54	.50	.00	1.00	.03	.17	.00	1.00
Medicine & Life Science	.10	.30	.00	1.00	.09	.29	.00	1.00	.11	.31	.00	1.00
Teaching	.11	.32	.00	1.00	.13	.34	.00	1.00	.09	.28	.00	1.00
Soc. Sci., Psych., Social Work	.26	.44	.00	1.00	.	.	.	.	.69	.46	.00	1.00
Law	.04	.20	.00	1.00	.07	.26	.00	1.00	.	.	.	.
Sex (Male = 1)	.56	.50	.00	1.00	.60	.49	.00	1.00	.48	.50	.00	1.00
Region of birth (West = 1)	.80	.40	.00	1.00	.82	.39	.00	1.00	.78	.41	.00	1.00
Year of birth	1965	5.96	1956	1982	1965	6.02	1956	1982	1965	5.84	1956	1981
Number of languages learned	2.42	1.03	.00	10.00	2.35	.92	.00	7.00	2.53	1.19	.00	10.00
English speaking competence	3.32	1.03	.00	5.00	3.32	.99	.00	5.00	3.32	1.08	.00	5.00
	N = 1708				*N = 1050*				*N = 658*			

pre-intervention confounder determining the propensity to study abroad. Accordingly, language skills are only included in the sensitivity analyses below.[7]

4.5. Sample composition

Descriptive statistics for the outcome variable (ISEI), the treatment variable (study abroad), and all mentioned covariates are presented in Table 1 for the full sample, the sample of graduates of unspecific fields of study, and the sample of graduates of specific fields. In addition, all zero-order correlations between studying abroad, the occupational status three years after graduation, and all covariates are shown in Table A3 in the Appendix.

The sample description confirms relatively established knowledge that graduates who did or did not study abroad differ in several respects. Mobile graduates more often come from families with high educational capital, more often complete full secondary education with the *Abitur*, and more often obtain a degree from a traditional university. They tend to show higher extra-curricular engagement with respect to working as student assistants or completing an internship. Study abroad graduates tend to have gained international experiences already before starting higher education. They less often pursue vocational training next to higher education, less often have a career in the public sector, and more often study in fields like medicine that lead directly into most prestigious jobs.

Moreover, Table A.3 shows that studying abroad is positively related to graduates' occupational status three years after finishing higher education ($r = 0.07$, $p \leq .01$). However, this bivariate association applies more strongly to graduates of occupationally unspecific fields ($r = 0.12$, $p \leq .001$) than graduates of occupationally specific fields ($r = 0.08$, $p \leq .05$) aligning with our theoretical expectations. Yet, it is still impossible to say whether studying abroad provides distinct benefits in terms of occupational status or whether systematic selection of individuals with good career prospects into study abroad programs explains this correlation. Disentangling these two processes is the main task of propensity score matching.

5. Estimation results

5.1. Main analysis

First, we estimate the propensity to study abroad including all covariates using logistic models (Table A4a-c in the Appendix). We

[7] It has also been shown that school performance determines selection into study abroad (Di Pietro and Page, 2008; Lörz et al., 2016; Kratz and Netz, 2016), which, in turn, can improve occupational success. Unfortunately, ALWA does not provide objective performance measures, but only self-reported evaluations by the respondents regarding their achievement in German and Math. We neglected these measures as they are inadequate and also uncorrelated with study abroad (see Table A3 in the Appendix). Implications will be discussed below.

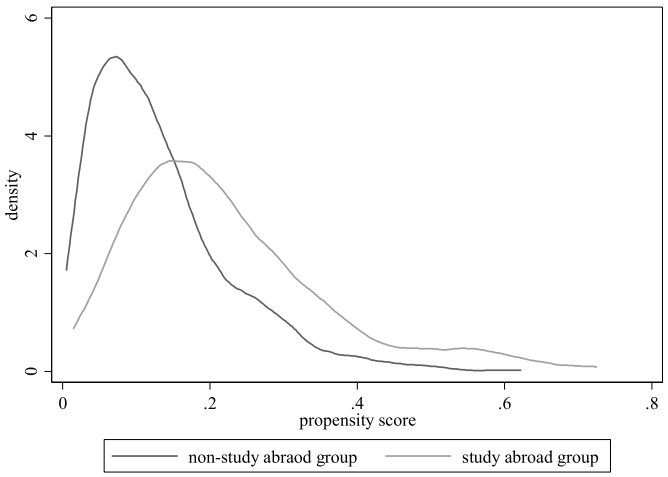


Fig. 1. Kernel density estimates of the estimated propensity score, calculated separately for study abroad (grey) and non-study abroad group (black).

then match observationally similar individuals in the study abroad group and non-study abroad group using the kernel algorithm, as it yielded best matches with respect to bias reduction. The plots of the group-specific propensity scores show considerable overlap in propensity scores between graduates who studied abroad and those that did not (Figs. 1 and 2), allowing the straightforward estimation of treatment effect with comprehensive common support.

Table 2 demonstrates that matching has worked in balancing all the variables affecting selection into studying abroad, for the full sample as well as the two subsamples indicating occupational specificity of the fields of study. The covariate distribution across treatment and control group *before matching* confirms systematic group differences, while there are no substantial *post-matching* differences in mean covariate values across the treatment and control cases, reducing the potential for selection bias. Standardized mean bias is calculated for each covariate and is defined as the difference of sample means in the treated and control subsamples divided by the square root of the average sample variances in both groups. For example, the standardized mean bias in the share of graduates from traditional universities was 43 percent before matching, reduced to four percent after matching. Overall mean bias is reduced below three percent, which is considered sufficient to balance control and treatment groups (Caliendo and Kopeinig, 2008, 48).

Table 3 presents estimates of the average effect of studying abroad for both the unmatched and matched sample. Again, we run separate analyses on the full sample and the two subsamples. Looking first at the unmatched effect and corresponding to results from pairwise correlation analysis, we find positive mean differences in occupational status between graduates who studied abroad and graduates who didn't in all samples. With 4 points (standard error = 1.30) on the ISEI scale, the difference is highest for graduates of occupationally unspecific subjects.

For the full sample, as assumed in our first hypothesis, the matching estimates for the average treatment effect on the treated (ATT) show that graduates who studied abroad seem to benefit only slightly from studying abroad after accounting for compositional

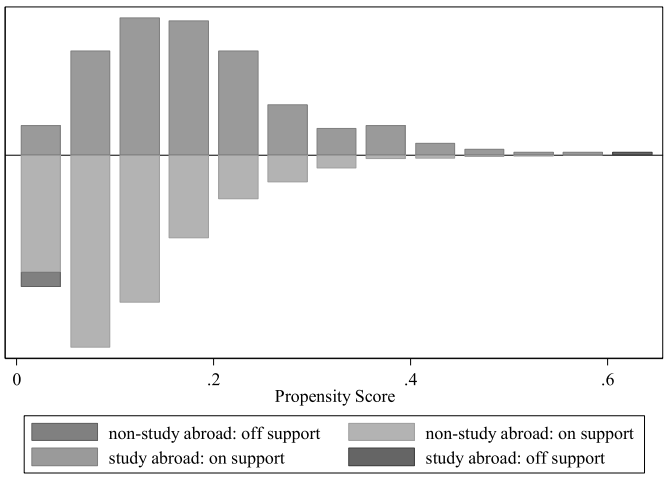


Fig. 2. Propensity score histogram by treatment status.

Table 2
Covariate balance before and after matching.

	Full sample					Occupationally specific fields					Occupationally unspecific fields				
		Unmatched control sample		Matched control sample			Unmatched control sample		Matched control sample			Unmatched control sample		Matched control sample	
	SA	*NSA*	*%bias*	*NSA*	*%bias*	*SA*	*NSA*	*%bias*	*NSA*	*%bias*	*SA*	*NSA*	*%bias*	*NSA*	*%bias*
Socio-economic background															
Parent with HE degree	.45	.33	25.70	.44	3.60	.47	.34	27.80	.48	−.30	.43	.31	24.10	.43	−1.10
Educational trajectory															
Abitur	.90	.78	32.90	.89	2.60	.89	.78	30.10	.88	1.50	.92	.79	35.80	.91	2.80
University type															
Master Trad. Uni.	.75	.54	43.10	.73	3.70	.71	.53	38.30	.70	2.60	.79	.57	47.80	.78	1.70
Degree from Uni. Corp. Ed.	.03	.08	−21.80	.03	−1.40	.01	.04	−20.00	.01	−1.70	.05	.14	−29.90	.06	−1.70
Student assistant	.18	.10	24.20	.17	1.40	.23	.10	34.20	.22	2.20	.12	.09	11.40	.11	4.20
Vocational training	.27	.39	−25.90	.28	−1.80	.26	.36	−20.50	.27	−.60	.28	.44	−35.30	.27	1.40
Public sector career	.01	.06	−27.10	.01	−1.00	.01	.08	−34.80	.01	.50	.01	.02	−10.10	.01	−2.70
Previous mobility	.12	.05	24.40	.10	3.70	.09	.04	21.50	.08	3.60	.14	.06	25.90	.12	4.70
Field of study															
Science, Eng., Math	.27	.35	−17.40	.28	−1.20	.47	.55	−14.70	.48	−1.10	.03	.03	.10	.03	2.20
Life Science & medicine	.14	.09	14.80	.13	2.70	.19	.08	34.30	.19	.60	.07	.11	−14.80	.07	2.10
Teaching	.14	.11	10.10	.15	−1.10	.20	.12	22.00	.19	3.00	.07	.09	−6.60	.07	−.80
Soc. Sci., related	.36	.25	24.30	.35	2.00	.	.	.	.	.	.80	.67	29.50	.80	−.60
Law	.04	.04	−4.00	.04	−.70	.07	.07	−1.60	.07	−.30	.	.	.	.	.
Socio-demography															
Born in West Germany	.85	.80	14.50	.85	.10	.08	3.60	.08	.87	1.40	.83	.77	13.70	.83	−2.00
Year of birth	1966	1964	34.60	1966	2.30	1966	1964	33.90	1966	6.30	1966	1964	37.00	1966	5.30
Mean bias before matching	25.00					24.90					23.00				
Mean bias after matching	1.80					1.40					2.40				

Note: *SA = Graduates with study abroad experience; NSA = graduates without study abroad experience; HE = higher education.*

Table 3
Kernel matching estimates of effect of study abroad on ISEI 3 years after graduation.

	Unmatched			Kernel matching		
	TE (s.e.)	*n1*	*n0*	TE (s.e.)	*n1*	*n0*
Full Sample						
	2.71 (.93)	*1492*	*216*			
ATT				1.59 (.90)	*1458*	*215*
ATC				2.25 (.90)		
ATE				2.16 (.78)		
Unspecific Fields						
	4.00 (1.30)	*560*	*98*			
ATT				2.91 (1.25)	539	*98*
ATC				4.58 (1.25)		
ATE				4.29 (1.22)		
Specific Fields						
	2.97 (1.78)	*932*	*118*			
ATT				−.48 (1.01)	*879*	*118*
ATC				.82 (1.27)		
ATE				.66 (1.05)		

Note: TE = treatment effect; n1 and n0 = number of treated (1) and control (0) cases within region of common support; epanechnikov kernel and a fixed bandwidth of 0.10 are used; s.e. = bootstrapped standard errors each with 100 repetitions.

differences. Their occupational status deviates by 1.59 scale points from the counterfactual average.[8] Yet, the effect estimates differ substantially across subsamples. While the ATT for graduates of unspecific fields clearly takes on a higher value of 2.91 scale points, the ATT of graduates of specific fields is −0.48 scale points and, thus, is close to zero. This effect heterogeneity across occupational specificity of study fields supports our second hypothesis.

Moreover, and unlike standard OLS regression[9] that assumes a constant coefficient for the effect of the treatment on the outcome, PSM differentiates between three average effects, the ATC, ATT, and ATE, which has been outlined above. As Table 3 shows, both ATE and the ATC consistently exceed the ATT.[10] Regarding graduates of unspecific fields of study, we find differences between the ATC and ATT of almost 2 points. If the average effect of studying abroad is larger for those who did not study abroad if they had done so (ATC) than for those who actually did study abroad (ATT), this means that individuals with a lower probability of treatment (for a distribution of propensity scores across control and treatment groups see Fig. 1) benefit more from the treatment (cf., Brand and Xie, 2010). We will discuss possible explanations for why this should be the case in the concluding section.

5.2. Sensitivity analysis

To assess the robustness of our findings regarding the effect of studying abroad on occupational status, we conducted further analyses (1) comparing alternative matching algorithms, (2) using alternative operationalizations of the outcome, (3) adding (current) language skills to the covariate space, and (4) getting separate matching estimates by gender.[11]

First, we compare kernel matching with nearest neighbor matching (with replacement) using one neighbor, nearest neighbor matching using five neighbors, and local linear matching (Table 4). Instead of weighing observations by a kernel function of the propensity score within a set bandwidth to get the counterfactual outcomes, nearest neighbor matching calculates counterfactuals for each treatment (control) case using a specified number (here, 1 and 5) of control (treatment) cases closest to it in terms of the propensity score. Local linear matching is based on the kernel estimator but it includes a linear term in the weighting function. Local linear matching is better able to deal with bias due to unequally spaced propensity scores, especially at the ends of the scale (Deaton, 1997, 197-99). As seen in Table 5, all matching algorithms estimate gaps in occupational status between study abroad and non-study abroad graduates similar to the kernel algorithm, proving the stability of our findings.

Second, as shown in Table 5, the matching estimates for graduates of occupationally specific fields across different specifications

[8] Related standard errors are commonly calculated using bootstrapping that includes the first steps of the estimation, because estimation of the variance of the treatment effect includes variance due to the estimation of the propensity score (Caliendo and Kopeinig, 2008, 53).

[9] Mathematically, kernel matching can also be seen as weighted regression of the counterfactual outcome on an intercept with weights given by the kernel weights (Caliendo and Kopeinig, 2008, 43).

[10] Remember that the ATE is the weighted average of the ATT and the ATC. Thus, it falls between the ATT and ATC. For the same reason, the ATE is always close to the ATC since almost 90 percent of the sample did not study abroad.

[11] In addition, business cycles may influence study abroad behavior as well as labor market outcomes and would possibly bias our results. As sensitivity check (not reported) and to exclude such confounding we linked the year of graduation with the economic indicator "change of GDP compared to the previous year, seasonal and calendar adjusted" (varies between −1.0 and 5.5 percent for the observation years 1975–2008) and tested whether this economic indicator is related to our treatment (study abroad) or outcome (occupational status). In both cases the correlation is virtually zero (less than .03) and not significant. There is therefore no reason to assume that our results will be distorted by the effects of the business cycle. In addition, we already control for the year of birth that is highly correlated with the year of graduation. That means we check at least for the confounding effects of trends in the economic situation that might be related to trends in study abroad participation.

Table 4
Sensitivity analysis: Alternative matching estimates of effect of study abroad on ISEI 3 years after graduation.

	Unmatched			Nearest neighbor matching (1)			Nearest neighbor matching (5)			Local Linear matching		
	TE (s.e.)	*n1*	*n0*	TE (s.e.)	*n1*	*n0*	TE (s.e.)	*n1*	*n0*	TE (s.e.)	*n1*	*n0*
Full Sample												
	2.71 (.93)	*1492*	*216*									
ATT				.53 (1.22)	*1458*	*215*	1.13 (.97)	*1458*	*215*	1.84 (.88)	*1492*	*216*
ATC				2.77 (1.24)			2.33 (.92)			2.13 (1.11)		
ATE				2.51 (1.08)			2.18 (.92)			2.09 (.87)		
Unspecific Fields												
	4.00 (1.30)	*560*	*98*									
ATT				3.33 (1.78)			2.82 (1.24)	539	*98*	2.93 (1.05)	*560*	*98*
ATC				4.87 (1.63)	*539*	*98*	5.08 (1.40)			4.94 (1.26)		
ATE				4.63 (1.39)			4.74 (1.42)			4.64 (1.16)		
Specific Fields												
	2.97 (1.78)	*932*	*118*									
ATT				−2.08 (1.48)	*879*	*118*	−1.22 (1.18)	879	*118*	−.38 (.82)	*932*	*118*
ATC				.52 (1.62)			.48 (1.27)			1.76 (1.36)		
ATE				.21 (1.31)			.29 (1.12)			1.52 (1.35)		

Note: TE = treatment effect; n1 and n0 = number of treated (1) and control (0) cases within region of common support; s.e. = bootstrapped standard errors each with 100 repetitions.

Table 5
Sensitivity analysis: Kernel matching estimates of ATT, ATC, and ATE with different model specifications.

	Full sample			Unspecific fields			Specific fields		
	ATT	ATC	ATE	ATT	ATC	ATE	ATT	ATC	ATE
(1) Different specifications of the outcome									
ISEI 3 years after graduation or ISEI of previous employment spell that lasted at least 6 months									
	1.36	1.99	1.91	2.82	3.62	3.50	−.54	.68	.53
	(.86)	(.85)	(.92)	(1.18)	(1.43)	(1.39)	(1.07)	(.91)	(1.06)
ISEI 1 year after graduation									
	2.05	3.10	2.95	2.91	3.91	3.75	.14	1.86	1.64
	(.88)	(.79)	(.86)	(1.28)	(1.38)	(1.35)	(1.04)	(1.15)	(1.01)
ISEI 5 years after graduation									
	1.32	1.92	1.84	2.31	2.85	2.76	.07	.50	.44
	(.75)	(.86)	(.90)	(1.31)	(1.40)	(1.25)	(.79)	(1.35)	(1.03)
(2) Restrictive specification of covariate structure									
Including language skills in covariate structure: number of learned foreign languages and English competences									
	.84	1.76	1.64	1.68	3.66	3.34	−.98	.28	.13
	(.88)	(1.19)	(.95)	(1.33)	(1.55)	(1.55)	(1.05)	(1.33)	(1.22)
(3) Gender-separate analyses									
Males									
	1.17	2.06	1.94	2.64	3.14	3.09	−.23	.79	.65
	(1.02)	(1.48)	(1.19)	(1.42)	(1.36)	(1.37)	(1.04)	(1.67)	(1.72)
Females									
	1.52	1.91	1.84	2.65	4.18	3.94	- .70	- .33	- .39
	(1.29)	(1.37)	(1.21)	(1.82)	(1.82)	(1.68)	(1.88)	(1.78)	(1.80)

Note: Epanechnikov kernel and a fixed bandwidth of 0.10 are used (a number of different bandwidths were tried, but estimates remained insensitive to bandwidth choice); bootstrapped standard errors with 100 repetitions in parentheses.

lend support to our finding that the effect is virtually indistinguishable from zero.[12] At the same time, matching estimates and standard errors for graduates of occupationally unspecific fields (middle columns) confirm the advantage in occupational status of those who studied abroad by around three points on the ISEI scale. The effect estimates for different outcomes specifications are close the estimates of the main analyses, however, differences between ATT and ATC are less pronounced.

Third, adding language skills as potential confounders to the set of covariates used in estimating the propensity score reduces effect sizes of the matching estimates, especially the ATT. It seems that part of the effect of studying abroad on occupational status is due to the correlation between studying abroad and language skills. Yet, with the available data we cannot be sure if graduates had good foreign language proficiency before studying abroad or if language skills were acquired after studying abroad, thus, meditating

[12] For reasons of simplicity and comprehensibility we only show effect estimates based on the kernel algorithm.

the association between studying abroad and occupational status.

Finally, we report the results of separate analyses by gender, as commonly done in labor market research. Gender-separate analyses document hardly any differences in effect estimates between men and women.

6. Discussion

Our analysis reveals that graduates who did or did not study abroad differ in several respects. Those who studied abroad combine characteristics such as high academic family background and high educational achievement that should increase their chances of reaching a higher occupational status after graduation irrespectively of studying abroad. Compositional differences, however, are not the only factors that account for the observed status returns to studying abroad. Resting upon the counterfactual inference model and the conditional independence assumption, propensity score estimates suggest slight occupational status differences between graduates who studied abroad and those who did not, even after controlling for selection of high-achievement groups into study abroad.

In addition to that, our analyses empirically validate our theoretical expectation that differentiating by level of standardization of the transition from higher education to work is crucial to quantifying the occupational benefits derived from studying abroad. We found no evidence that studying abroad increases the occupational status of graduates trained for specific occupations. Their educational degree directly translates into acknowledged specialized expertise and is almost synonymous with their economic opportunities in the labor market. Under such tight educational-professional coupling, studying abroad has limited additional value in finding a high status job and in controlling occupational access.

On the other hand, we find net occupational status benefits of about three points on the ISEI-scale for mobile graduates of unspecific fields, who have less access to professional 'labor market shelters' (Freidson, 1999). If we consider for example that the mean difference in ISEI between graduates from academic and non-academic backgrounds is around the same size in our sample, this difference does seem to be substantial. It supports our hypothesis that if educational degrees are more risky signals of practically applicable skills and competent performance, and given the incomplete information in job markets, gaining a privileged competitive position in the labor market is facilitated by the acquisition of positional goods. Of these, studying abroad is a particular example. This ties in with research that tries to show more generally how, in the course of educational expansion, tertiary education degrees that do not prepare for a specific occupation have become less reliable signals of productivity and human capital for potential employers (Klein, 2016; Reimer et al., 2008).

We are well aware that the interpretation of our findings rests on the conditional independence assumption (CIA). That is, while we were successful in matching the study abroad and the comparison groups in terms of theoretically relevant confounders we observed, we cannot rule out selection bias due to the dependence of average effects on *unobserved* differences between those who studied abroad and those who did not. Unfortunately, the data we use leaves aspects such as motivation, confidence, and general ability unmeasured. This unresolved self-selection is a drawback that our data shares with many other datasets, but nevertheless has to be kept in mind when interpreting findings, particularly if policy implications are to be drawn.

We also find systematic differences between causal effect estimates, that is, the average treatment effect on the treatment group (ATT) and the average treatment effect on the control group (ATC). Formally, the ATT differs from the ATC whenever one or more observed or unobserved variables correlate with selection into the treatment as well as the size of the individual-level treatment effect (Morgan and Todd, 2008). In our sample, the ATC exceeds the ATT indicating that status returns to studying abroad for those who have a higher propensity to study abroad based on various observed background factors (parental education, own educational achievement) are smaller in comparison to the returns achieved by low-propensity graduates (for similar findings on causal effect heterogeneity with respect to education, see Brand and Halaby, 2006; Brand and Xie, 2010; Morgan and Todd, 2008).

Whereas the true variables that generate heterogeneity in ATC and ATT are unobserved, we can nevertheless identify at least two possible explanations for why the effect of studying abroad may not be homogeneous across the study population. First, those with higher propensity to study abroad, e.g. high status individuals with more financial and social resources, will study abroad independent of expected career benefits from studying abroad, maybe following an emergent norm to study abroad (Petzold and Peter, 2015). In contrast, those with lower propensity to study abroad, e.g. low-status individuals for whom studying abroad may imply financial or emotional sacrifices, will only study abroad if they are convinced of the professional value of the experience. In other words, there would be self-selection into studying abroad based on the unobserved expected utility of studying abroad that is more meaningful for low-propensity students. Economists call this "sorting gain selection", and in this case the ATU would be biased since individuals that did not study abroad would not be comparable to those that did (cf., Tsai, 2015). Second, as argued for example by Di Pietro (2015, 2012), studying abroad may provide those with lesser propensity to study abroad with an opportunity to develop social, intercultural, and language skills that high-propensity graduates already possess. For graduates with lower propensity to study abroad, studying abroad may also play a stronger market signaling role in setting themselves apart from their peers. In this case, the ATU is unbiased and low-propensity students who currently do not study abroad would profit from the expansion of study abroad opportunities.

In any case, attending to unobserved heterogeneity *and* effect heterogeneity in future work will make an essential contribution to scientific and policy debates about the role of study abroad expansion in stratification processes. If those who are least likely to study abroad, in other words, those with the least economic, social, and cultural resources, are the ones who are most likely to benefit from studying abroad, student mobility could even reduce instead of reproduce social inequality despite the high social selectivity into studying abroad.

From a policy perspective, outreach programs drawing in more low-propensity students into study abroad as well as financial aid programs may appear as straightforward social measures. However, as long as we do not know which variables generate the

heterogeneity in returns to studying abroad the effectivity of such programs remains uncertain. Importantly, if it is true that low-propensity (as opposed to high-propensity) students are positively selected into student mobility programs based on their expected gains, low-propensity students who currently choose not to study abroad would also benefit less in a counterfactual scenario.

7. Conclusion

Studying temporarily in another country during higher education is now 'a normal option within easy reach' (Teichler, 2012, 46) and perhaps more frequently than ever before conforms to what is perceived to be a 'social norm to study abroad' (Petzold and Peter, 2015). While international education is an enriching experience in itself, there is also a common belief in its positive effect on labor market opportunities. Many speculate that studying abroad being more accessible to the affluent has turned into a factor that reproduces social inequality in society. Yet, while there is ample evidence for the social selectivity of studying abroad, its socio-economic impact in terms of income and occupational achievement is far more contested.

Our research reports new findings on this issue in the German national context. We applied propensity score techniques to investigate the causal effect of studying abroad on early career success of German higher education graduates and address concerns about selection bias. The data we used is representative of the population living in Germany and includes information on residential, education, and employment biographies.

Results confirm a positive effect of student mobility on early career occupational status. Yet, this effect is partly driven by compositional differences, i.e. graduates with better occupational prospects self-select into study abroad programs. The results also show that occupational status returns to mobility are confined to graduates from occupationally unspecific fields of study. We hypothesize that for these graduates, transition from education to work is less smooth than for graduates from specific fields and labor market allocation will depend on the relative value of their accumulated educational capital including capital acquired through studying abroad.

Furthermore, we propose that the effect of studying abroad is not homogeneous across the study population. Rather, individuals with the lowest propensity to study abroad (those with the least economic, social, and cultural resources) are the most likely to benefit from studying abroad in occupational terms. In light of our findings, it is therefore unlikely that studying abroad is substantially related to social inequality in society. Yet, the processes behind the higher returns to studying abroad experienced by low-propensity graduates remain to be explored.

To conclude, our research provides an enriched theoretical understanding of the links between studying abroad and occupational achievement and hopefully serves as avenue for future research. Our findings should be transferable to other national contexts with similar educational systems and with similar experiences of educational expansion. Future research would benefit from a closer analysis of national differences in the signaling and human capital potential of studying abroad. Another area for future research could be the investigation of how country- and period-specific conditions of national economies (e.g., business cycles) influence the labor market impact of studying abroad over the long-term. Moreover, distributional shifts in studying abroad may encourage research on whether the rapid expansion of studying abroad affects the signaling value of international mobility experiences on the labor market.

Appendix A. Supplementary data

Supplementary data related to this article can be found at http://dx.doi.org/10.1016/j.ssresearch.2018.05.006.

References

Allmendinger, Jutta, 1989. Educational systems and labor market outcomes. Eur. Socio Rev. 5 (3), 231–250. http://dx.doi.org/10.1093/oxfordjournals.esr.a036524.
Anger, Christina, Plünnecke, Axel, Schmidt, Jörg, 2010. Bildungsrenditen in Deutschland - Einflussfaktoren, politische Optionen und volkswirtschaftliche Effekte [Returns to education in Germany - political options and economic effects]. Cologne Institute for Economic Research.
Becker, Gary S., 1964. Human Capital: a Theoretical and Empirical Analysis, with Special Reference to Education. Chicago University Press, Chicago.
Biemann, Torsten, Fasang, Anette E., Grunow, Daniela, 2011. Do economic globalization and industry growth destabilize Careers? An analysis of career complexity and career patterns over time. Organ. Stud. 32 (12), 1639–1663. http://dx.doi.org/10.1177/0170840611421246.
Bills, David B., 2003. Credentials, signals, and screens: explaining the relationship between schooling and job assignment. Rev. Educ. Res. 73 (4), 441–449. http://dx.doi.org/10.3102/00346543073004441.
Boudon, Raymond, 1974. Educational growth and economic equality. Qual. Quantity 8 (1), 1–10. http://dx.doi.org/10.1007/BF00205861.
Brand, Jennie E., Halaby, Charles N., 2006. Regression and matching estimates of the effects of elite college attendance on educational and career achievement. Soc. Sci. Res. 35 (3), 749–770. http://dx.doi.org/10.1016/j.ssresearch.2005.06.006.
Brand, Jennie E., Xie, Yu, 2010. Who benefits most from college? Evidence for negative selection in heterogeneous economic returns to higher education. Am. Socio. Rev. 75 (2), 273–302.
Brändle, Tobias, Lengfeld, Holger, 2016. Erzielen Studierende ohne Abitur geringeren Studienerfolg? Befunde einer quantitativen Fallstudie [Do Students without a General Qualification for University Entrance Show Lower Academic Performance? Findings from a Quantitative Case Study]. Z. Soziol. 44 (6), 447–467. http://dx.doi.org/10.1515/zfsoz-2015-0605.
Breen, Richard, Goldthorpe, John H., 1997. Explaining educational differentials. Ration. Soc. 9 (3), 275–305. http://dx.doi.org/10.1177/104346397009003002.
Brint, Steven, Cantwell, Allison M., Hanneman, Robert A., 2008. The two cultures of undergraduate academic engagement. Res. High. Educ. 49 (5), 383–402. http://dx.doi.org/10.1007/s11162-008-9090-y.
Brooks, Rachel, Waters, Johanna, 2010. Social networks and educational mobility: the experiences of UK students. Glob. Soc. Educ. 8 (1), 143–157. http://dx.doi.org/10.1080/14767720903574132.
Buis, Maarten L., 2013. The composition of family background: the influence of the economic and cultural resources of both parents on the Offspring's educational attainment in The Netherlands between 1939 and 1991. Eur. Socio Rev. 29 (3), 593–602. http://dx.doi.org/10.1093/esr/jcs009.
Caliendo, Marco, Kopeinig, Sabine, 2008. Some practical guidance for the implementation of propensity score matching. J. Econ. Surv. 22 (1), 31–72. http://dx.doi.

org/10.1111/j.1467-6419.2007.00527.x.
Carlson, Sören, 2013. Becoming a mobile student – a processual perspective on German degree student mobility. Popul. Space Place 19 (2), 168–180. http://dx.doi.org/10.1002/psp.1749.
Chiswick, Barry R., Miller, Paul W., 2009. The international transferability of immigrants' human capital. Econ. Educ. Rev. 28 (2), 162–169. http://dx.doi.org/10.1016/j.econedurev.2008.07.002.
Davidson, Thomas, Sanyal, Paromita, 2017. Associational participation and network expansion: microcredit self-help groups and poor women's social ties in rural India. Soc. Forces 95 (4), 1695–1724. https://doi.org/10.1093/sf/sox021.
Deaton, Angus, 1997. The Analysis of Household Surveys: a Microeconometric Approach to Development Policy. World Bank Publications, Washington, D.C.
Deutscher Akademischer Austauschdienst (DAAD), and Deutsches Zentrum für Hochschul- und Wissenschaftsforschung (DZHW), 2017. Wissenschaft Weltoffen: Daten und Fakten zur Internationalität von Studium und Forschung in Deutschland [Facts and Figures on the International Nature of Studies and Research in Germany]. http://www.wissenschaftweltoffen.de/publikation/wiwe_2017_verlinkt.pdf.
Di Pietro, Giorgio, 2012. Does studying abroad cause international labor mobility? Evidence from Italy. Econ. Lett. 117 (3), 632–635. http://dx.doi.org/10.1016/j.econlet.2012.08.007.
Di Pietro, Giorgio, 2015. Do study abroad programs enhance the employability of graduates? Educ. Finance Pol. 10 (2), 223–243. http://dx.doi.org/10.1162/EDFP_a_00159.
Di Pietro, Giorgio, Page, Lionel, 2008. Who studies Abroad? Evidence from France and Italy. Eur. J. Educ. 43 (3), 389–398. http://dx.doi.org/10.1111/j.1465-3435.2008.00355.x.
Erola, Jani, Jalonen, Sanni, Lehti, Hannu, 2016. Parental education, class and income over early life course and children's achievement. Res. Soc. Stratif. Mobil. 44, 33–43. http://dx.doi.org/10.1016/j.rssm.2016.01.003.
Euler, Hanns P., Rami, Ursula, Glaser, Evelyne, Reber, Gerhard, Bacher, Johann, 2013. Lohnt sich ein Auslandsaufenthalt während des Studiums? Ergebnisse der Evaluierung eines Förderprogrammes. [Is it worth staying abroad during studies? Evaluation of a funding program]. DBW Die Betriebswirtschaft 73 (5), 425–446.
Falk, Susanne, Sackmann, Reinhold, Struck, Olaf, Weymann, Ansgar, Windzio, Michael, Wingens, Matthias, 2000. Gemeinsame Startbedingungen in Ost und West? Risiken beim Berufseinstieg und deren Folgen im weiteren Erwerbsverlauf [Common starting conditions in East and West? Risks at labor market entry and consequences for career development]. Sfb 186 -Arbeitspapier Nr. 65.
Featherman, David L., Jones, F. Lancaster, Hauser, Robert M., 1975. Assumptions of social mobility research in the U.S. The case of occupational status. Soc. Sci Res. 4 (4), 329–360. https://doi.org/10.1016/0049-089X(75)90002-2.
Finger, Claudia, 2011. The Social Selectivity of Internal Mobility Among German University Student - a Multi-level Analysis of the Impact of the Bologna Process. WZB Discussion Paper SP I 2011-503. Unpublished manuscript, last modified July 03, 2012. http://bibliothek.wzb.eu/pdf/2011/i11-503.pdf.
Freidson, Eliot, 1999. Theory of professionalism: method and substance. Int. Rev. Sociol. 9 (1), 117–129. http://dx.doi.org/10.1080/03906701.1999.9971301.
Friedberg, Rachel M., 2000. You Can't take it with You? Immigrant assimilation and the portability of human capital. J. Labor Econ. 18 (2), 221–251. http://dx.doi.org/10.1086/209957.
Gangl, Markus, 2002. Changing labour markets and early career outcomes: labour market entry in Europe over the past decade. Work. Employ. Soc. 16 (1), 67–90. http://dx.doi.org/10.1177/09500170222119254.
Gangl, Markus, 2015. Matching estimators for treatment effects. In: Best, Henning, Wolf, Christof (Eds.), The Sage Handbook of Regression Analysis and Causal Inference. SAGE Publications, Los Angeles, pp. 251–274.
Ganzeboom, Harry B., Treiman, Donald J., 1996. Internationally comparable measures of occupational status for the 1988 international standard classification of occupations. Soc. Sci. Res. 25 (3), 201–239. http://dx.doi.org/10.1006/ssre.1996.0010.
Gebel, Michael, 2009. Fixed-term contracts at labour market entry in west Germany: implications for job search and first job quality. Eur. Socio Rev. 25 (6), 661. http://dx.doi.org/10.1093/esr/jcp005.
Gerber, Theodore P., Cheung, Sin Yi, 2008. Horizontal stratification in postsecondary education: forms, explanations, and implications. Annu. Rev. Sociol. 34 (1), 299–318. https://doi.org/10.1146/annurev.soc.34.040507.134604.
Gerhards, Jürgen, Hans, Silke, 2013. Transnational human capital, education, and social inequality. Analyses of international student exchange. Z. Soziol. 42 (2), 99–117. http://dx.doi.org/10.1515/zfsoz-2013-0203.
Gerhards, Jürgen, Németh, Boróka, 2015. Ökonomisches Kapital der Eltern und Medizinstudium im Ausland. Wie Europäisierungs- und Globalisierungsprozesse die Reproduktion sozialer Ungleichheiten verändern. [Parental economic capital and medical studies abroad. How Europeanization and globalization change reproduction of social inequalities]. Berliner J. Soziol. 25 (3), 283–301. http://dx.doi.org/10.1007/s11609-015-0290-y.
Giesecke, Johannes, Heisig, Jan P., 2011. Destabilization and destandardization: for Whom? The development of west German job mobility since 1984. Schmollers Jahrb. 131 (2), 301–314. http://dx.doi.org/10.3790/schm.131.2.301.
Haelermans, Carla, de Witte, Kristof, 2015 July July. Does residential mobility improve educational outcomes? Evidence from the Netherlands. Soc. Sci. Research 52, 351–369. https://doi.org:10.1016/j.ssresearch.2015.02.008.
Hilmer, Michael J., 2002. Student migration and institution control as screening devices. Econ. Lett. 76 (1), 19–25. http://dx.doi.org/10.1016/S0165-1765(02)00021-6.
Hollis, Martin, 1982. Education as a positional good. J. Philos. Educ. 16 (2), 235–244. http://dx.doi.org/10.1111/j.1467-9752.1982.tb00615.x.
Janson, Kerstin, Schomburg, Harald, Teichler, Ulrich, 2009. The professional value of ERASMUS mobility: the impact of international experience on former students' and on teachers' careers. In: Cairns, David (Ed.), ACA Papers on International Cooperation in Education. Lemmens, Bonn.
Kalleberg, Arne L., Sørensen, Aage B., 1979. The sociology of labor markets. Annu. Rev. Sociol. 5 (1), 351–379. http://dx.doi.org/10.1146/annurev.so.05.080179.002031.
Kerckhoff, Alan C., 1995. Institutional arrangements and stratification processes in industrial societies. Annu. Rev. Sociol. 21 (1), 323–347. http://dx.doi.org/10.1146/annurev.so.21.080195.001543.
King, Russell, Findlay, Allan, 2015. 11 student migration. In: Martiniello, Marco, Rath, Jan (Eds.), An Introduction to International Migration Studies: European Perspectives. Amsterdam University Press, Amsterdam, pp. 259–280.
Kinginger, Celeste, 2008. Modern Language Journal Monograph Series. Language Learning in Study Abroad: Case Histories of Americans in France, vol. 1 Blackwell, Oxford.
Kjelland, Jim, 2008. Economic returns to higher education: signaling v. Human capital theory: an analysis of competing theories. The Park Place Economist 16 (1), 70–77.
Klein, Markus, 2016. The association between graduates' field of study and occupational attainment in West Germany, 1980–2008. J. Lab. Mark. Res. 49 (1), 43–58. http://dx.doi.org/10.1007/s12651-016-0201-5.
Kleinert, Corinna, Matthes, Britta, Antoni, Manfred, Drasch, Katrin, Ruland, Michael, Trahms, Annette, 2011. ALWA – new life course data for Germany. Schmollers Jahrb. 04, 525–634. http://dx.doi.org/10.3790/schm.131.4.625.
Kratz, Fabian, Netz, Nicolai, 2016. Which mechanisms explain monetary returns to international student mobility? Stud. High Educ. 1–26. http://dx.doi.org/10.1080/03075079.2016.1172307.
Kurz, Karin, Steinhage, Nikolei, 2001. Globaler Wettbewerb und Unsicherheiten beim Einstieg in den Arbeitsmarkt [Global competition and uncertainties at labor market entry]. Berliner J. Soziol. 11 (4), 513–531. http://dx.doi.org/10.1007/BF03204035.
Leuze, Kathrin, 2007. What makes for a good Start? Consequences of occupation-specific higher education for career mobility: Germany and great Britain compared. Int. J. Sociol. 37 (2), 29–53. http://dx.doi.org/10.2753/IJS0020-7659370202.
Leuze, Kathrin, 2010. Smooth Path or Long and Winding Road? How Institutions Shape the Transition from Higher Education to Work. Farmington Hills/Mich. Budrich UniPress, Opladen.
Leuze, Kathrin, 2011. Higher education and graduate employment: the importance of occupational specificity in Germany and Britain. In: Clasen, Jochen (Ed.), Converging Worlds of Welfare? British and German Social Policy in the 21st Century. Oxford University Press, Oxford/New York, NY, pp. 245–265. http://dx.doi.

org/10.1093/acprof:oso/9780199584499.003.0012.
Lindberg, Matti E., 2009. Student and early career mobility patterns among highly educated people in Germany, Finland, Italy, and the United Kingdom. High Educ. 58 (3), 339–358. http://dx.doi.org/10.1007/s10734-009-9198-9.
Lörz, Markus, Krawietz, Marian, 2011. Internationale Mobiltät und soziale Selektivität: ausmaß, Mechanismen und Entwicklung herkunftsspezifischer Unterschiede zwischen 1990 und 2005 [International mobility and social selectivity: extent, mechanisms, and development of origin-specific differences between 1990 and 2005]. Kölner Z. Soziol. Sozialpsychol. 63, 185–205. http://dx.doi.org/10.1007/s11577-011-0134-5.
Lörz, Markus, Netz, Nicolai, Quast, Heiko, 2016. Why do students from underprivileged families less often intend to study abroad? High Educ. 72 (2), 153–174. http://dx.doi.org/10.1007/s10734-015-9943-1.
Lörz, Markus, Schindler, Steffen, 2009. Educational expansion and effects on the transition to higher education: has the effect of social background characteristics declined or just moved to the next stage? In: Hadjar, Andreas, Becker, Rolf (Eds.), Exected and Unexpected Consequences of the Educational Expansion in Europe and the US: Theoretical Approaches and Empirical Findings in Comparative Perspective. Haupt Verlag, Berlin, Stuttgart, Vienna, pp. 97–110.
Messer, Dolores, Wolter, Stefan C., 2007. Are student exchange programs worth it? High Educ. 54 (5), 647–663. http://dx.doi.org/10.1007/s10734-006-9016-6.
Mincer, Jacob, 1958. Investment in human capital and personal income distribution. J. Polit. Econ. 66 (4), 282–302. http://dx.doi.org/10.1086/258055.
Morgan, Stephen L., Harding, David J., 2006. Matching estimators of causal effects. Socio. Meth. Res. 35 (1), 3–60. http://dx.doi.org/10.1177/0049124106289164.
Morgan, Stephen L., Todd, Jennifer J., 2008. A diagnostic routine for the detection of consequential heterogeneity of causal effects. Socio. Meth. 38 (1), 231–281. http://dx.doi.org/10.1111/j.1467-9531.2008.00204.x.
Morgan, Stephen L., Winship, Christopher, 2015. Counterfactuals and Causal Inference: Methods and Principles for Social Research. Cambridge University Press, New York, NY.
Müller, Walter, 2005. Education and youth integration into european labour markets. Int. J. Comp. Sociol. 46 (5–6), 461–485. http://dx.doi.org/10.1177/0020715205060048.
Müller, Walter, Brauns, Hildegard, Steinmann, Susanne, 2002. Expansion und Erträge tertiärer Bildung in Deutschland, Frankreich und im Vereinigten Königreich. [Expansion of tertiary education and economic returns in Germany, France and the United Kingdom]. Berliner J. Soziol. 12 (1), 37–62. http://dx.doi.org/10.1007/BF03204042.
Müller, Walter, Steinmann, Susanne, Ell, Renate, 1995. Education and Labour-market Entry in Germany. Working Papers AB I/Nr. 10.
Netz, Nicolai, Finger, Claudia, 2016. New horizontal inequalities in German higher Education? Social selectivity of studying abroad between 1991 and 2012. Sociol. Educ. http://dx.doi.org/10.1177/0038040715627196.
Oosterbeek, Hessel, Webbink, Dinand, 2006. Asessing the returns to studying abroad. CPB discussion paper 64.
Opper, Susan, Teichler, Ulrich, Carlson, Jerry, 1990. Impacts of study abroad programmes on students and graduates. In: In: Burn, Barbara B., a, u (Eds.), Higher Education Policy Series, vol. 11. Kingsley, London, pp. 2.
Opper, Susan, 1991. Study abroad: a competitive edge for women? Oxf. Rev. Educ. 17 (1), 45–64. http://dx.doi.org/10.1080/0305498910170104.
Parey, Matthias, Waldinger, Fabian, 2011. Studying abroad and the effect on international labour market mobility: evidence from the introduction of ERASMUS. Econ. J. 121 (551), 194–222. http://dx.doi.org/10.1111/j.1468-0297.2010.02369.x.
Petzold, Knut, 2017a. Studying abroad as a sorting criterion in the recruitment process. A field experiment among German employers. J. Stud. Int. Educ. 21 (5), 412–430. http://dx.doi.org/10.1177/1028315317697543.
Petzold, Knut, 2017b. The role of international student mobility in hiring decisions. A vignette experiment among German employers. J. Educ. Work 30 (8), 893–911. https://doi.org/10.1080/13639080.2017.1386775.
Petzold, Knut, Peter, Tamara, 2015. The social norm to study abroad: determinants and effects. High Educ. 69 (6), 885–900. http://dx.doi.org/10.1007/s10734-014-9811-4.
Potts, Davina, 2015. Understanding the early career benefits of learning abroad programs. J. Stud. Int. Educ. 19 (5), 441–459. http://dx.doi.org/10.1177/1028315315579241.
Powell, Justin J.W., Finger, Claudia, 2013. The Bologna Process's model of mobility in Europe: the relationship of its spatial and social dimensions. Eur. Educ. Res. J. 12 (2), 270–285. http://dx.doi.org/10.2304/eerj.2013.12.2.270.
Reimer, David, Noelke, Clemens, Kucel, Aleksander, 2008. Labor market effects of field of study in comparative perspective. Int. J. Comp. Sociol. 49 (4–5), 233–256. http://dx.doi.org/10.1177/0020715208093076.
Reimer, David, Pollak, Reinhard, 2010. Educational expansion and its consequences for vertical and horizontal inequalities in access to higher education in west Germany. Eur. Socio Rev. 26 (4), 415–430. http://dx.doi.org/10.1093/esr/jcp029.
Rodrigues, Margarida, 2013. Does Student Mobility during Higher Education Pay? Evidence from 16 European Countries. Publications Office of the European Union, Luxembourg.
Roksa, Josipa, Levey, Tania, 2010. What can you do with that Degree? College major and occupational status of college graduates over time. Soc. Forces 89 (2), 389–415. http://dx.doi.org/10.1353/sof.2010.0085.
Rosenbaum, Paul R., 2010. Design of Observational Studies. Springer New York, New York, NY.
Rospigliosi, Asher P., Greener, Sue, Bourner, Tom, Sheehan, Maura, 2014. Human capital or signalling, unpacking the graduate premium. Int. J. Soc. Econ. 41 (5), 420–432. http://dx.doi.org/10.1108/IJSE-03-2013-0056.
Salisbury, Mark H., Umbach, Paul D., Paulsen, Michael B., Pascarella, Ernest T., 2009. Going global: understanding the choice process of the intend to study abroad. Res. High. Educ. 50, 119–143. http://dx.doi.org/10.1007/s11162-008-9111-x.
Saniter, Nils, Siedler, Thomas, 2014. Door Opener or Waste of Time? the Effects of Student Internships on Labor Market Outcomes. IZA Discussion Papers 8141.
Scherer, Stefani, 2005. Patterns of labour market entry – long wait or career Instability? An empirical comparison of Italy, great Britain and west Germany. Eur. Socio Rev. 21 (5), 427–440. http://dx.doi.org/10.1093/esr/jci029.
Schindler, Steffen, Lörz, Markus, 2012. Mechanisms of social inequality development: primary and secondary effects in the transition to tertiary education between 1976 and 2005. Eur. Socio Rev. 28 (5), 647–660. http://dx.doi.org/10.1093/esr/jcr032.
Schultz, Theodore W., 1961. Investment in human capital. Am. Econ. Rev. 51 (1), 1–17.
Sorrenti, Giuseppe, 2017 (October). The Spanish or the German Apartment? Study abroad and the acquisition of permanent skills. Econ. Educ. Rev. 60, 142–158. http://dx.doi.org/10.1016/j.econedurev.2017.07.001.
Souto-Otero, Manuel, 2008. The socio-economic background of erasmus students: a trend towards wider inclusion? Int. Rev. Educ. 54 (2), 135–154. http://dx.doi.org/10.1007/s11159-007-9081-9.
Spence, Michael, 1973. Job market signaling. Q. J. Econ. 87 (3), 355–374. http://dx.doi.org/10.2307/1882010.
Spieß, Erika, Brüch, Andreas, 2002. Auswirkungen von interkulturellen Erfahrungen für die Motivation beruflicher Auslandsaufenthalte ost-und westdeutscher Studierender [The impact of intercultural experiences on the motivation of East and West German students for job related stays abroad]. Z. für Sozialpsychol. 33 (4), 219–228. http://dx.doi.org/10.1024//0044-3514.33.4.219.
Stiglitz, Joseph E., 1975. The theory of 'screening', education, and the distribution of income. Am. Econ. Rev. 65 (3), 283–300.
Stroud, April H., 2010. Who plans (not) to study abroad? An examination of U.S. student intent. J. Stud. Int. Educ. 14 (5), 491–507. http://dx.doi.org/10.1177/1028315309357942.
Teichler, Ulrich, 2012. International student mobility and the Bologna process. Res. Comp. Int. Educ. 7 (1), 34–49. http://dx.doi.org/10.2304/rcie.2012.7.1.34.
Trautwein, Ulrich, Maaz, Kai, Lüdtke, Oliver, Nagy, Gabriel, Husemann, Nicole, Watermann, Rainer, Köller, Olaf, 2006. Studieren an der Berufsakademie oder an der Universität, Fachhochschule oder Pädagogischen Hochschule? [Studying at the Berufsakademie or at the University, Fachhochschule or Pädagogische Hochschule?]. Z. für Erziehungswiss. (ZfE) 9 (3), 393–412. http://dx.doi.org/10.1007/s11618-006-0057-5.
Triventi, Moris, 2013. Stratification in higher education and its relationship with social inequality: a comparative study of 11 european countries. Eur. Socio Rev. 29 (3), 489–502. http://dx.doi.org/10.1093/esr/jcr092.
Tsai, Shu-Ling, 2015. Revisiting selection in heterogeneous returns to college education. 人文及社會科學集刊 27 (2), 323–360.

van de Werfhorst, Herman G., 2004. Systems of Educational Specialization and Labor Market Outcomes in Norway, Australia, and The Netherlands. Int. J. Comp. Sociol. 45 (5), 315–335. https://doi.org/10.1177/0020715204054154.
van Mol, Christof, Timmerman, Christian, 2014. Should I stay or should I Go? An analysis of the determinants of intra-european student mobility. Popul. Space Place 20 (5), 465–479. http://dx.doi.org/10.1002/psp.1833.
Waibel, Stine, Rüger, Heiko, Ette, Andreas, Sauer, Lenore, 2017. Career consequences of transnational educational mobility: a systematic literature review. Educ. Res. Rev. 20, 81–98. http://dx.doi.org/10.1016/j.edurev.2016.12.001.
Waters, Johanna, Brooks, Rachel, 2010. Accidental Achievers? International higher education, class reproductionand privilege in the experiences of UK students overseas. Brit. J. Sociol. Educ. 31 (2), 217–228. http://dx.doi.org/10.1080/01425690903539164.
Waters, Johanna L., 2006. Geographies of cultural capital: education, international migration and family strategies between Hong Kong and Canada. Trans. Inst. Br. Geogr. 31 (2), 179–192. http://dx.doi.org/10.1111/j.1475-5661.2006.00202.x.
Weiss, Andrew, 1995. Human capital vs. Signalling explanations of wages. J. Econ. Perspect. 9 (4), 133–154. http://dx.doi.org/10.1257/jep.9.4.133.
Wiers-Jenssen, Jannecke, 2008. Does higher education attained abroad lead to international jobs? J. Stud. Int. Educ. 12 (2), 101–130. http://dx.doi.org/10.1177/1028315307307656.
Wiers-Jenssen, Jannecke, 2011. Background and employability of mobile vs. Non-Mobile students. Tert. Educ. Manag. 17 (2), 79–100. http://dx.doi.org/10.1080/13583883.2011.562524.
Wiers-Jenssen, Jannecke, 2013. Degree mobility from the nordic countries: background and employability. J. Stud. Int. Educ. 17 (4), 471–491. http://dx.doi.org/10.1177/1028315312463824.

Appendix

Table A.1: Operationalization of Covariates

Variable name	Operationalization
Socio-economic background	
Parental education	Coded 0/1; 1= Father or mother completed higher education degree (ISCED 5A or 6)
Educational trajectory	
Abitur	Coded 0/1; 1= full secondary education completed with Abitur
University type	Categorical variable; (1) Master from traditional university/higher public service career; (2) Bachelor from traditional university, degree from applied university, higher intermediate public service career; (3) Degree from university of cooperative education
Student assistant	Coded 0/1; 1=worked as student assistant during studies or completed an internship
Vocational training	Coded 0/1; 1=completed vocational training prior to or after higher education
Public sector career	Coded 0/1; 1=pursued a career in the public service (higher or higher intermediate level)
Previous mobility experiences	Coded 0/1; 1=previous international experience during primary or secondary school or during voluntary service
Field of Study	Categorical variable; (1) Agrig./Admin./Techn./Oth.; (2) Science, Engineering, Math; (3) Life Science & medicine; (4) Teaching; (5) Social Science & related, Psychology, Social Work; (6) Law *[in case of multiple degrees, the study field with established higher occupational returns was coded]*
Socio-demography	
Region of Birth	Coded 0/1; 1=Born in West Germany
Year of birth	Continuous variable; range is 1965 to 1982
Language skills	
Self-report of number of foreign languages learned (current situation)	Numbers range from 0 to 10
Self-report of competence in English language (current situation)	Answers range from 0 to 5 (0=no competence, 1=very bad; 5= very good)

Table A.2: Fields of study and degree of occupational specificity

Field of Study	Occupational specificity	Primary Occupational Category (three years after degree completion)
Law	High	Lawyers (63%)
Medicine	High	Medical doctors (94%)
Teaching (primary level)	High	Teachers (86%)
Teaching (secondary level)	High	Teachers (65%)
Engineering, architecture	High	Engineers, architects (71%)
Computer science	High	IT professional (75%)
Physics, Chemistry, & related	High	Physicist, chemists, & related scientists (68%)
(Public) Administration incl. Police	High	Public administration professionals and police commissioners (76%)
Other educational studies	Moderate	Teachers (39%), government associate professionals (16%)
Social science & related, Psychology, Social work	Moderate	Social scientist & information professionals (18%), associate professionals (14%)
Life Science	Moderate	Life science professionals (33%), medical assistants (11%)
Math, Statistics	Moderate	Mathematicians, statisticians (45%); teachers (15%)
Technical and Commercial Professionalization	Moderate	Technical professionals (14%), clerks (11%)

Note: based on ALWA data; calculations based on procedure by Roksa & Levey (2010); see also Leuze (2010, 140) and Saniter & Siedler (2014, 44)

Table A.3: Bivariate Correlations (r) between outcome, treatment, and covariates by specificity of field of study

	Full sample		**Graduates from specific fields**		**Graduates from unspecific fields**	
	ISEI 3 YAG	StudyAbroad	ISEI 3 YAG	StudyAbroad	ISEI 3 YAG	StudyAbroad
ISEI 3 years after graduation	1.00		1.00		1.00	
Study Abroad	.07**	1.00	.08*	1.00	.12***	1.00
Higher Educated Parent	.10***	.09***	.13***	.09**	.03	.09*
Abitur	.16***	.10***	.18***	.09**	.17***	.11**
Master Traditional University	.25***	.14***	.34***	.12***	.21***	.16***
Applied Uni./Bachelor Trad. Uni.	-.16***	-.10***	-.28***	-.10**	-.12**	-.10**
Uni. Coorp. Education	-.19***	-.06**	-.15***	-.05^{+}	-.15***	-.09*
Student Assistant	.09***	.09***	.07*	.12***	.12**	.04
Vocational Training	-.19***	-.08***	-.16***	-.06*	-.19***	-.12**
Career in Public Sector	-.06**	-.07**	-.15***	-.09	.01	-.03
Previous mobility experiences	.03	.10***	.05	.08**	.07^{+}	.11**
Field of Study						
Agrig./admin./techn./oth.	-.26***	-.10**	-.44***	-.10***	-.06	-.09*
Science, Eng., Math	.25***	-.06*	.08**	-.05	.14***	.00
Medicine & Life Science	.26***	0.05*	.43***	.13***	.07^{+}	-.05
Teaching	-.05*	.04	-.13***	.08*	.00	-.02
Soc. sci., Psych., Social Work	-.28***	.08***	.	.	-.07^{+}	.10*
Law	.18***	-.01	.16***	-.01	.	.
Sex (Male=1)	.10***	-.02	.07*	-.01	.04	-.01
Region of birth (West=1)	.06*	.05^{+}	.06	.05	.03	.05
Year of birth	.04^{+}	.12***	.05	.11***	.00	.13***
Number of languages learned	.08***	.22***	.14***	.17***	.09*	.27***
English speaking competence	.14***	.25***	.15***	.20***	.15***	.30***
	N=1,708		*N=1,050*		*N=658*	

*Note: $^{+}$p<=.1, *p<=.05, **p<=.01, ***p<=0.001; YAG = years after graduation*

Table A.4a: Logistic regression predicting participation in study abroad (full sample)

	Coef.	Std. Err.	[95% Conf. Interval]	
Socio-economic background				
Parent with HE degree	.33*	.16	.02	.64
Educational trajectory				
Abitur	.20	.28	-.35	.74
Degree type				
Reference BA Trad. Uni. or Applied Uni.				
MA Trad. Uni.	0.36^{+}	.213	-.055	.778
Corp. Uni.	-1.05*	.461	-1.950	-.142
Vocational training	-.15	.19	-.50	.24
Student assistant	0.47*	.21	.06	.88
Public sector career	-1.18	.74	-2.63	.28
Previous mobility experiences	.53*	.26	.01	1.05
Field of Study				
Reference Agrig./Admin./Techn./Oth.				
Science, Eng., Math	.44	.37	-.29	1.17
Life Science & Medicine	.83*	.42	.01	1.65
Teaching	.75^{+}	.42	-.07	1.62
Soc. Sci., Psych., Social Work	1.15**	.37	.42	1.88
Law	.22	.52	-.80	1.25
Socio-demography				
Born in West Germany	.44	.37	-.29	1.17
Year of birth	.83*	.42	.01	1.65

N	1,708
LR chi2(15)	111.34
Prob > chi2	.000
Pseudo R2	.0859

*Note: $^{+}$p<=.1, *p<=.05, **p<=.01, ***p<=0.001; HE = higher education*

Table A.4b: Logistic regression predicting participation in study abroad (only graduates from occupationally unspecific fields)

	Coef.	Std. Err.	[95% Conf.	Interval]
Socio-economic background				
Parent with HE degree	.34	.24	-.13	.81
Educational trajectory				
Abitur	.20	.44	-.66	1.07
Degree type				
Reference BA Trad. Uni. or Applied Uni.				
MA Trad. Uni.	1.49**	.53	.46	2.53
Corp. Uni.	.78	.55	-.30	1.86
Student assistant	-.01	.37	-.74	.72
Vocational training	-.38	.27	-.91	.16
Public sector career	-.01	1.08	-2.13	2.12
Previous mobility experiences	.55	.37	-.18	1.28
Field of Study				
Reference Agrig./Admin./Techn./Oth.				
Science, Eng., Math	.38	.90	-1.39	2.16
Life Science & Medicine	.01	.75	-1.45	1.47
Teaching	.28	.75	-1.19	1.75
Soc. Sci., Psych., Social Work	.98	.63	-.25	2.21
Law	.	.	.	.
Socio-demography				
Born in West Germany	.22	.31	-.39	.83
Year of birth	.07***	.02	.03	.11

N	658
LR chi2(15)	53.20
Prob > chi2	.000
Pseudo R2	.0961

*Note: $^{+}p<=.1$, *p<=.05, **p<=.01, ***p<=0.001; HE = higher education*

Table A.4c: Logistic regression predicting participation in study abroad (only graduates from specific fields)

	Coef.	Std. Err.	[95% Conf. Interval]	
Socio-economic background				
Parent with HE degree	.28	.22	-.15	.71
Educational trajectory				
Abitur	.31	.37	-.42	1.05
Degree type				
Reference BA Trad. Uni or Applied Uni.				
MA Trad. Uni.	-.01	.29	-.58	.56
Corp. Uni.	-1.23	1.05	-3.28	.83
Student assistant	.79***	.26	.28	1.30
Vocational training	.15	.27	-.38	.68
Public sector career	-1.68	1.04	-3.71	.35
Previous mobility experiences	.58	.38	-.17	1.33
Field of Study				
Reference Agrig./Admin./Techn./Oth.				
Science, Eng., Math	.59	.45	-.29	1.47
Life Science & Medicine	1.49**	.53	.45	2.54
Teaching	1.15*	.52	.13	2.17
Soc. Sci., Psych., Social Work	.	.	.	.
Law	.53	.59	-.61	1.68
Socio-demography				
Born in West Germany	.60*	.31	.00	1.21
Year of birth	.05**	.02	.02	.09
N	1,050			
LR chi2(15)	68.36			
Prob > chi2	.000			
Pseudo R2	.0926			

Note: $^{+}$p<=.1, *p<=.05, **p<=.01, ***p<=0.001*; HE = higher education*