UNIVERZA V LJUBLJANI
FAKULTETA ZA DRUŽBENE VEDE

Rok Platinovšek

Statistično modeliranje neodgovora na anketno vprašanje in prekinitve anketiranja

Statistical Modeling of Item Nonresponse and Respondent Breakoff

Doktorska disertacija

Ljubljana, 2013

UNIVERZA V LJUBLJANI
FAKULTETA ZA DRUŽBENE VEDE

Rok Platinovšek

Mentorica: doc. dr. Katja Lozar Manfreda
Somentor: prof. dr. Joop J. Hox

Statistično modeliranje neodgovora na anketno vprašanje in prekinitve anketiranja

Statistical Modeling of Item Nonresponse and Respondent Breakoff

Doktorska disertacija

Ljubljana, 2013

Univerza
v Ljubljani
Fakulteta
za družbene vede

Kardeljeva ploščad 5
1000 Ljubljana, Slovenija
telefon 01 58 05 120

I Z J A V A O A V T O R S T V U
doktorske disertacije

Podpisani/-a Rok Platinovšek, z vpisno številko 74080809, sem avtor/-ica doktorske disertacije z naslovom: Statistično modeliranje neodgovora na anketno vprašanje in prekinitve anketiranja (Statistical Modeling of Item Nonresponse and Respondent Breakoff).

S svojim podpisom zagotavljam, da:

- je predložena doktorska disertacija izključno rezultat mojega lastnega raziskovalnega dela;
- sem poskrbel/-a, da so dela in mnenja drugih avtorjev oz. avtoric, ki jih uporabljam v predloženem delu, navedena oz. citirana v skladu s fakultetnimi navodili;
- sem poskrbel/-a, da so vsa dela in mnenja drugih avtorjev oz. avtoric navedena v seznamu virov, ki je sestavni element predloženega dela in je zapisan v skladu s fakultetnimi navodili;
- sem pridobil/-a vsa dovoljenja za uporabo avtorskih del, ki so v celoti prenesena v predloženo delo in sem to tudi jasno zapisal/-a v predloženem delu;
- se zavedam, da je plagiatorstvo – predstavljanje tujih del, bodisi v obliki citata bodisi v obliki skoraj dobesednega parafraziranja bodisi v grafični obliki, s katerim so tuje misli oz. ideje predstavljene kot moje lastne – kaznivo po zakonu (Zakon o avtorski in sorodnih pravicah (UL RS, št. 16/07-UPB3, 68/08, 85/10 Skl.US: U-I-191/09-7, Up-916/09-16)), prekršek pa podleže tudi ukrepom Fakultete za družbene vede v skladu z njenimi pravili;
- se zavedam posledic, ki jih dokazano plagiatorstvo lahko predstavlja za predloženo delo in za moj status na Fakulteti za družbene vede;
- je elektronska oblika identična s tiskano obliko doktorske disertacije ter soglašam z objavo doktorske disertacije v zbirki »Dela FDV«.

V Ljubljani, dne 11.11.2013 Podpis avtorja/-ice: _____

Povzetek

Neodgovor predstavlja na področju anketne metodologije izrazit problem, saj manjkajoče vrednosti, ki so posledica neodgovora, zmanjšujejo zaupanje v anketne ocene. Manjkajoče vrednosti se lahko pojavijo na različne načine. *Neodgovor enote* se zgodi, kadar ne uspemo pridobiti meritev za celotno vzorčeno enoto (vzorčena oseba zavrne sodelovanje ali je ne uspemo kontaktirati). O *neodgovoru na anketno vprašanje* govorimo, ko je vzorčena oseba pripravljena sodelovati v anketi, vendar pa podatki za določena anketna vprašanja niso na voljo. O *prekinitvi anketiranja* pa govorimo, ko anketiranec začne izpolnjevati anketo, vendar preneha, še preden jo dokonča. V doktorski disertaciji se osredotočamo na obliki neodgovora, ki sta bili doslej v literaturi deležni manj pozornosti: neodgovor na anketno vprašanje ter prekinitev anketiranja.

V empiričnem delu disertacije smo neodgovor na anketno vprašanje ter prekinitev neodgovora proučevali na primeru podatkov, pridobljenih s pilotsko anketo *Generations and Gender Programme* (v nadaljevanju GGP). Isti vprašalnik je bil izveden v treh načinih anketiranja: osebno, telefonsko in spletno. Da bi pri danih sredstvih maksimizirali velikost vzorca, smo se odločili podatke zbirati v dveh fazah. V prvi fazi so bili anketiranci člani spletnega panela podjetja Valicon, v drugi fazi pa so bili vzorčeni iz Slovenske populacije.

Čeprav je izpolnjevanje vprašalnika GGP tipičnemu anketirancu vzelo približno eno uro, pa v prvih dveh fazah zbiranja podatkov skoraj ni bilo prekinitev anketiranja; presenetljivo jih ni bilo niti v spletnem načinu. Takšni podatki nam ne omogočajo analizirati, kako so lastnosti anketiranca in anketnih vprašanj povezane z prekinitvijo anketiranja, zato smo se odločili za dodatno zbiranje podatkov. V tej tretji fazi zbiranja podatkov smo anketirance na anketo vabili z oglasi na spletni strani Facebook. Za razliko od anketirancev v prvih dveh fazah, ki so bili o trajanju ankete obveščeni v vabilu k sodelovanju, anketirancem v tretji fazi zbiranja podatkov nismo vnaprej povedali, kako dolgo bo trajalo izpolnjevanje vprašalnika. Prekinitev anketiranja je bila v tej tretji fazi zbiranja podatkov res mnogo pogostejša, saj je anketiranje prekinila več kot polovica anketirancev.

Za analizo neodgovora na anketno vprašanje smo uporabili *posplošene linearne mešane modele*. Raba teh modelov na področju *teorije odgovora na postavko* je analogna naši aplikaciji na problem neodgovora na anketno vprašanje. Za teorijo odgovora na postavko je značilna predpostavka, da nemerljive lastnosti oseb na eni strani ter postavk po drugi strani določajo izid. Bolj konkretno: statistični model predpostavlja, da razlika med *sposobnostjo* osebe in *zahtevnostjo* testnega vprašanja (pri testu znanja) določa verjetnost, da bo dana oseba na zadano vprašanje odgovorila pravilno. Naša aplikacija je analogna: predpostavljamo, da je verjetnost neodgovora na anketno vprašanje določena z razliko med anketirančevo motivacijo ter bremenom zastavljenega anketnega vprašanja. Za razliko od modelov v teoriji odgovora na postavko, kjer je najpogosteje cilj zgolj *opisna meritev* lastnosti oseb in postavk, pa v naših modelih nastopajo tudi pojasnjevalne spremenljivke, saj nas zanima predvsem, kako je

verjetnost neodgovora na anketno vprašanje povezana z lastnostmi anketnih vprašanj, anketirancev ter anketarjev.

Za analizo prekinitve anketiranja smo uporabili metode analize preživetja, predvsem Coxov model sorazmernih ogroženosti. Metode analize preživetja so namenjene analizi časa, do katerega se zgodi določen *dogodek*, ter omogočajo upoštevanje krnjenih enot v analizi. V naši aplikaciji kot dogodek definiramo prekinitev anketiranja ter vse anketirance, ki so anketo dokončali, obravnavamo kot krnjene. Razširjeni Coxov model nam omogoča, da poleg časovno neodvisnih učinkov (lastnosti anketiranca) kot pojasnjevalne spremenljivke v model vključimo tudi časovno odvisne spremenljivke (lastnosti anketnih vprašanj). Kadar se je izkazalo, da je predpostavka sorazmernih ogroženosti za določeno pojasnjevalno spremenljivko kršena, smo postopali tako, da smo z grafično metodo določili prelomno točko in s tem celotno časovno obdobje razdelili na dva intervala, na katerih je bilo omenjeni predpostavki zadoščeno.

Rezultati analiz potrjujejo, da sta tako neodgovor na anketno vprašanje kot prekinitev anketiranja bolj pogosta v spletni verziji vprašalnika kot pri osebnem in telefonskem anketiranju. Tak rezultat je bil pričakovan in je v skladu s teoretično podlago, ki predpostavlja, da je zadrževalni prag Galesic (2006) za obe obliki neodgovora višji pri načinih anketiranja, kjer je prisoten anketar.

Raziskave neodgovora na anketno vprašanje ter prekinitve anketiranja navadno kot pojasnjevalne spremenljivke vključujejo tudi anketirančeve demografske lastnosti. V literaturi je pogosta uporaba anketirančeve starosti in izobrazbe kot *proxy* mer anketirančeve kognitivne sposobnosti. Anketiranci z višjimi kognitivnimi sposobnosti naj bi bili tako manj obremenjeni, ko odgovarjajo na anketna vprašanja, in zaradi tega pri njih pričakujemo manj prekinitev ter neodgovora na anketno vprašanje. Rezultati naših analiz so v skladu z opisano logiko, kar se tiče izobrazbe anketiranca: pri bolj izobraženih anketirancih opažamo manj neodgovora na anketno vprašanje ter nižje tveganje za prekinitev anketiranja. Kar se tiče anketirančeve starosti, pa se rezultati za neodgovor in prekinitev razlikujejo. Višja starost je res povezana s pogostejšimi neodgovori na anketno vprašanje. V nasprotju s pričakovanji pa je višja starost povezana z *nižjim* tveganjem za prekinitev. Podrobnejša analiza pokaže, da so mladi anketiranci anketo prekinjali že kmalu po začetku anketiranja. Po približno sto anketnih vprašanjih pa anketirančeva starost nima več učinka na tveganje za prekinitev.

V vprašalnik smo dodali tri vprašanja, s katerimi smo merili anketirančevo splošno naravnanost do anket (npr. strinjanje z izjavo, da so ankete pomembne za znanost, politiko in gospodarstvo). Pričakovali smo, da bodo anketiranci, ki so bolj pozitivno naravnani do anket, bolj skrbno izvajali vsako izmed faz procesa odgovarjanja na anketna vprašanja (glej Stocke 2006). Rezultati naših analiz potrjujejo, da je anketirančeva pozitivna naravnanost do anket povezana z manj neodgovora na anketno vprašanje ter nižjim tveganjem za prekinitev anketiranja.

Trije neodvisni eksperti so vsako anketno vprašanje v vprašalniku GGP kodirali na treh merah: vsiljivost teme vprašanja, nevarnost razkritja (občutljivih informacij) ter potencial za pretirano pozitivno predstavitev. Rezultati analiz kažejo, da so anketna vprašanja, ki predstavljajo nevarnost razkritja, povezana z višjo mero neodgovora, vendar pa je omenjeni učinek statistično značilen samo pri spletnem anketiranju. Ne-

varnost razkritja občutljivih informacij je povezana tudi z višjim tveganjem za prekinitev anketiranja, vendar pa ta učinek ni bil značilen na začetku (približno prvih sto vprašanj) vprašalnika GGP.

Če anketiranec preskoči, zavrne odgovor, ali prekine anketo pri vprašanju, ki omogoča pretirano pozitivno predstavitev (npr. pomoč prijateljem pri skrbi za otroke), je to moč razumeti, kot da anketiranec implicitno priznava, da se ni vedel na družbeno zaželen način (Bradburn et al. 1978). Rezultati analiz so v skladu z opisano logiko: potencial za pretirano pozitivno predstavitev je povezan z manj neodgovora na anketno vprašanje in nižjim tveganjem za prekinitev anketiranja. V osebnem in telefonskem načinu anketiranja je opisani učinek še močnejši (interakcija z načinom anketiranja je statistično značilna): pri anketnih vprašanjih, ki omogočajo pretirano pozitivno predstavitev, je manj neodgovora, če je prisoten anketar (v primerjavi s spletnim samo-anketiranjem).

Rezultati pričakovano kažejo, da je vsiljivost teme anketnega vprašanja povezana z višjo verjetnostjo neodgovora. Vpliv vsiljivosti teme na tveganje za prekinitev pa se izkaže za bolj kompleksnega kot smo pričakovali. Rezultati kažejo, da bolj vsiljivo anketno vprašanje zniža tveganje za prekinitev, vendar obenem poviša tveganje za prekinitev dve anketni vprašanji naprej. Možna razlaga za omenjeni rezultat je, da anketiranci ne želijo razkriti informacije, da jim je določena tema res vsiljiva in zato pri anketnem vprašanju, ki zadeva takšno temo, ne prekinejo anketiranja. Prekinitve pri vsiljivih vprašanjih se vzdržijo in se za prekinitev odločijo raje kmalu po tem, ko jim je bilo postavljeno vsiljivo vprašanje.

V statističnih modelih smo kot pojasnjevalne spremenljivke vključili tudi objektivne lastnosti anketnih vprašanj kot je število ponujenih odgovorov in dolžina vprašanja (število besed). Rezultati kažejo, da ima dolžina vprašanja statistično značilen vpliv na neodgovor zgolj pri spletnem anketiranju: anketiranci so na spletu zagrešili več neodgovora pri dolgih vprašanjih. Vpliv dolžine vprašanja v Coxovem modelu za prekinitev anketiranja ni bil statistično značilen. Rezultati kažejo, da so odprta vprašanja ter vprašanja z velikim številom ponujenih odgovorov povezana z več neodgovora in višjim tveganjem za prekinitev anketiranja. V nasprotju s pričakovanji pa je ugotovitev, da je ta učinek šibkejši pri spletnem anketiranju: anketiranci na spletu so pri takšnih vprašanjih zagrešili *manj* neodgovora v primerjavi z osebami, ki so bile anketirane osebno ali po telefonu (statistično značilna interakcija z načinom anketiranja).

V Coxov model za prekinitev odgovora smo kot pojasnjevalni spremenljivki vključili še indikatorja za to, 1) ali je bil odgovor na dano vprašanje *obvezen* (program za anketiranje ni dovolil, da se vprašanje preskoči) ter 2) ali je dano anketno vprašanje vpeljalo novo temo. Skladno s pričakovanji se izkaže, da je tveganje za prekinitev v obeh primerih višje. V model za prekinitev smo kot pojasnjevalno spremenljivko vključili tudi mero pogostosti neodgovora na nedavna anketna vprašanja. V skladu s pričakovanji se tudi ta učinek izkaže za pozitiven in statistično značilen: kadar anketiranec pogosto izpušča odgovor na anketna vprašanja, je tveganje za prekinitev višje. S tem smo replicirali rezultate Galesic (2006).

Pričujoča doktorska disertacija predstavlja naslednje izvirne prispevke k razvoju področja anketne metodologije. Predmet študije sta dve obliki neodgovora, ki doslej ni-

sta bili raziskovani tako obsežno kot neodgovor enote, zato empirična študija razširja obstoječe znanje o faktorjih vpliva na neodgovor na anketno vprašanje ter prekinitev anketiranja. Razumevanje faktorjev vpliva, ki izhaja iz rezultatov študije, se lahko uporabi 1) za preprečevanje neodgovora na anketno vprašanje in prekinitve anketiranja (npr. s prilagoditvijo vprašalnika) ali 2) za izboljšanje postopkov, ki omogočajo analizo podatkov v prisotnosti manjkajočih vrednosti (npr. večkratno vstavljanje manjkajočih vrednosti).

Kolikor nam je znano, ni pred našo nobena raziskava hkrati obravnavala vpliva lastnosti anketnega vprašanja, anketiranca in anketarja na neodgovor na anketno vprašanje v treh različnih načinih anketiranja. Prav tako ne poznamo drugih raziskav, ki bi za Coxov model za prekinitev anketiranja preverile predpostavko sorazmernih ogroženosti in v primeru kršitev prilagodile model. Naša raziskava je prva, ki je v modelu za prekinitev anketiranja vključila mero predhodnih neodgovorov na anketna vprašanja ter pokazala statistično značilen vpliv le-teh.

Ključne besede: neodgovor na anketno vprašanje, prekinitev anketiranja, Generations and Gender Programme, teorija odgovora na postavko, posplošeni linearni mešani modeli, analiza preživetja, Coxov model

Abstract

Nonresponse is a prominent problem in survey methodology, as missing values caused by nonresponse reduce trust in survey estimates. Such missing values occur in various patterns. *Unit nonresponse* occurs when measurements cannot be obtained for the entire sampled unit (the sample person refuses to cooperate or cannot be contacted). *Item nonresponse* occurs when the sample person agrees to take the survey, but the data for certain items are unavailable. Finally, there is *breakoff* when the respondent starts the survey but stops prior to completing it. The present doctoral dissertation focuses on the two types of nonresponse that have received less attention in the literature: item nonresponse and breakoff.

In the empirical part of the dissertation we analyze item nonresponse and breakoff using data collected in the *Generations and Gender Programme* (GGP) pilot survey. The same questionnaire was administered in three modes of administration: face-to-face, telephone, and web. The data were collected in two rounds in order to maximize the total sample size within the constraints of the budget. The respondents in the first round were members of a commercial panel put together by the market research company Valicon, while for the second round respondents were sampled from the Slovenian population.

Even though filling out the GGP questionnaire took the typical respondent about an hour, almost no breakoff occurred in the first two rounds of data collection; surprisingly, not even in web mode. Because such data do not allow us to analyze how respondent and item characteristics are associated with breakoff, we decided to launch another round of data collection. The respondents in this third round of data collection were recruited via advertisements on Facebook. Unlike respondents in the first two rounds of data collection, who were told of the survey's duration in the advance letter, respondents in the third round of data collection were not told upfront how long it would take to fill out the questionnaire. This, indeed, led to a higher breakoff rate, with more than half the respondents breaking off before reaching the end of the survey.

Item nonresponse was analyzed by fitting generalized linear mixed models to the collected data. Our application of these models to the problem of item nonresponse is analogous to how they are applied in item response theory. The assumption made in item response theory is that unmeasurable characteristics of persons and items determine the outcome. Put more concretely, the statistical model assumes that the difference between a particular respondent's ability and an item's difficulty determines the probability of a correct answer in an achievement test. Our application is analogous: we assume that the probability of item nonresponse is determined by the difference between the respondent's motivation and the administered item's burden. However, unlike item response models whose aim is most often limited to descriptive measurement of persons and items on theoretical constructs, our models involve explanatory variables, for we are first and foremost interested in how item nonresponse is connected to characteristics of items, respondents, and interviewers.

Survival analysis methods, especially the Cox proportional hazards model, were used to analyze breakoff. Survival analysis methods have been developed to analyze the time until a specified *event* occurs and allow the inclusion of censored units in the analysis. In our application, we define respondent breakoff as the event of interest and regard all respondents who completed the survey as censored. The extended Cox model allows time-dependent predictors (item facets) to be included in the model in addition to the time-independent predictors (respondent characteristics) that can be used in the unextended form of the model. When the proportional hazards assumption was found to be suspect for a particular predictor, we used a graphical method to determine a cut-point. The proportional hazards assumption for the suspect predictor was thereby rendered tenable on each of the time intervals thus obtained.

Our findings corroborate that both item nonresponse and breakoff are more common with web administration than with face-to-face and telephone interviewing. This result was expected and agrees with the theory that posits that the inhibitory threshold (Galesic 2006) for both types of nonresponse is higher in interviewer-administered modes than in web administration.

Research on item nonresponse and breakoff commonly includes the respondent's demographic characteristics as explanatory variables. Using the respondent's age and education as proxy measures of their cognitive ability is common in the literature. Respondents higher in cognitive ability are thus theorized to experience less burden when answering to questionnaire items which is why we expect them to produce less item nonresponse and to break off later in the survey, if at all. Our findings are congruent with this rationale as far as the respondent's education is concerned: a higher level of respondent education is connected to less item nonresponse and a lower risk of breakoff. The results for item nonresponse and breakoff differ, however, with respect to the respondent's age. Higher age is, indeed, connected to more item nonresponse. Contrary to our expectations, however, higher age is associated with a *lower* risk of breakoff. Upon further investigation, younger respondents were found to break off shortly into the questionnaire. After about one hundred items have been administered, however, respondent age no longer has a bearing on the risk of breakoff.

Three questions were added to the questionnaire with the aim of measuring the respondents' attitude toward surveys in general (e.g. agreement with the statement that surveys are important for science, politics, and the economy). We hypothesized that respondents with a more positive attitude toward surveys would more carefully execute each phase of the question-answer process (see Stocke 2006). Our findings corroborate that the respondents' positive attitude toward surveys is associated with less item nonresponse and a lower risk of breakoff.

Three independent experts rated each item of the GGP questionnaire on three measures: the intrusiveness of the item's topic, the threat of disclosure, and the potential for overclaiming. Our findings imply that items high in the threat of disclosure are associated with more item nonresponse, but this effect was only found to be statistically significant for web administration. The threat of disclosure is also connected to a greater risk of breakoff, but this effect was found not to be significant in the first part (approximately the first one hundred items) of the GGP questionnaire.

A respondent's act of skipping, refusing to answer to, or breaking off at an item that allows them to present themselves in a more favorable light (e.g. the item on helping friends with childcare) may be understood as implicitly admitting to not having acted in socially condoned ways (Bradburn et al. 1978). The results of our analyses are in accord with this rationale: the potential for overclaiming is associated with less item nonresponse and a lower risk of breakoff. This effect is even stronger in interviewer-administered modes (we find a statistically significant interaction with the mode of administration): we find less item nonresponse at items that allow overclaiming when the interviewer is present in comparison to web based self-administration.

As expected, we find the item's intrusiveness to be connected to more item nonresponse. The effect of intrusiveness on breakoff, however, was found to be more complex than hypothesized. The results convey that a more intrusive item lowers the risk of breakoff, while simultaneously increasing the risk of breakoff two items further into the questionnaire. One possible explanation for this finding is that respondents do not want to disclose the information that they regard a particular topic highly intrusive by breaking off at the particular item that concerns such a topic. They actually refrain from breakoff at such items and rather break off shortly after the intrusive item was administered.

The statistical models included objective characteristics of questionnaire items like the number of response alternatives and the length of the item's wording. We only find the wording length to have a statistically significant effect on item nonresponse with web administration: web respondents produced more item nonresponse at longer items. The effect of the length of the item's wording was insignificant in the Cox model for breakoff. Our findings indicate that open-ended items and items with many answer alternatives are associated with more item nonresponse and a higher risk of breakoff. Contrary to our expectations, however, we find this effect to be weaker with web administration: web respondents produced *less* item nonresponse than respondents in interviewer-administered modes (we find a significant interaction with the mode of administration).

The Cox model for breakoff included as predictors the indicators for 1) whether the item was *required* (the software did not allow the item to be skipped) and 2) whether the item introduced a new section of the questionnaire. As expected, the risk of breakoff was found to be higher in both cases. We also included in the model for breakoff a measure of recent item nonresponse. In accordance with our expectations, we find this effect to be positive and significant: when the respondent has been producing item nonresponse at a higher rate, the risk of breakoff is higher. We have thus replicated Galesic's (2006) results.

The present dissertation makes the following contributions to the field of survey methodology: the objects of the study are two types of nonresponse that have not been studied as extensively as unit nonresponse, and the empirical study thus enriches extant knowledge regarding factors affecting item nonresponse and breakoff in surveys and has the potential to inform procedures for prevention (e.g. by adapting the questionnaire design) and treatment (e.g. by multiple imputation of missing values) of item nonresponse and breakoff.

To the best of our knowledge, no previous study has simultaneously considered the effect of item, respondent, and interviewer on item nonresponse across three different modes of administration. We also know of no other study to have checked the proportional hazards assumption in the Cox model for breakoff and adjusted for violations. Our study is also the first to have included a measure of recent item nonresponse in a model for breakoff and to have demonstrated a significant effect.

Keywords: item nonresponse, breakoff, Generations and Gender Programme, item response theory, generalized linear mixed models, survival analysis, Cox model

Contents

Table of abbreviations

Abbreviation	**Meaning**
AIC	Akaike information criterion
BIC	Bayesian information criterion
CATI	computer assisted telephone interviewing
F2F	face-to-face
GGP	generations and gender programme
GLMM	generalized linear mixed model
HR	hazard ratio
INR	item nonresponse
IRT	item response theory
KM	Kaplan-Meier
LMM	linear mixed model
MI	multiple imputation
ML	maximum likelihood
PH	proportional hazards
f2f.pnl	face-to-face mode; first round of administration (commercial panel)
f2f.smp	face-to-face mode; second round of administration (random sample)
cati.pnl	telephone mode; first round of administration (commercial panel)
cati.smp	telephone mode; second round of administration (random sample)
web.pnl	web mode; first round of administration (commercial panel)
web.smp	web mode; second round of administration (random sample)
web.fb	web mode; additional third round of administration (Facebook)

Table of predictors used in statistical models

Predictor[a]	Meaning
Mode of administration and round of data collection	
(mode) cati	mode of administration: computer assisted telephone interviewing
(mode) web	mode of administration: web based self-administration
panel	first round of data collection (Valicon's panel)
Interviewer	
male	interviewer sex
age10	interviewer age divided by 10
education	interviewer education measured on a scale from 1 to 9
log(I. experience)	logarithm of the number of months of experience with interviewing
Respondent	
male	respondent sex
age10	respondent age divided by 10
education	respondent education measured on a scale from 1 to 9
attitude toward surveys	respondent's attitude toward surveys; scale from 1 (negative attitude) to 5 (positive attitude)
Item	
Expert ratings	
intrusiveness	item topic is inappropriate in everyday conversation
disclosure	response alternatives ask respondent to admit to counternormative behavior/opinions
overclaiming	response alternatives allow the respondent to portray themselves in a more favorable light
Objective measures	
log(n. words)	logarithm of the number of words in the item's wording
log(n. alternatives)	logarithm of the number of offered response alternatives
input numeric	open-ended item that requires numeric input
input string	open-ended item that requires string input
radio yes	item appears in a battery with radio buttons for yes/no responses
section intro	item introduces a new section of the questionnaire
required	the item is required; cannot be skipped by respondent
Respondent self-assessment of cognitive state[b]	
partner activity	I am familiar with details concerning my partner's job/activity (recoded)
personal information	I sometimes have problems recalling information like relatives' birth-days
rel. quality	I rarely reflect on my relationships with other people
HH finances	I am thoroughly familiar with my household's financial situation and transactions (recoded)
Respondent self-assessment of topic sensitivity[c]	
networks	my relationships with people and the help and support we provide to each other
rel. partner	my relationship with my partner
rel. children	my relationship with my children
rel. parents	my relationship with my parents
having children	having (more) children; my and my partner's fertility
income	my household's income and possessions
values	my attitude toward issues like marriage, relations between genders, inter-generational relations
Previous item nonresponse	
log(cum. INR)	logarithm of the cumulative number of item nonresponses from start until the current item
serial INR	indicator that captures the respondent's tendency to produce item nonresponses in series
sqrt(INR last 10)	square root of the number of item nonresponses during the last 10 administered items

[a] Some predictors may have additional information given in square brackets next to their names. The abbreviations stand for: cn – centered; dch – dichotomized; MI – multiply imputed.

[b] Respondents were asked to choose a response on the agreement scale with the following response options: "strongly agree," "agree," "neither agree nor disagree," "disagree," and "strongly disagree."

[c] Respondents were asked to assess how sensitive they regarded certain topics that appeared in the GGP survey. The response scale spanned from 1 "very sensitive" to 7 "not sensitive at all" with unlabeled intermediary points.

1 Introduction

Nonresponse is a prominent problem in survey methodology that has received a lot of attention in the literature (e.g. Groves et al. 2002; Groves and Couper 1998; Koch and Porst 1998). Missing values caused by nonresponse reduce trust in survey estimates. Drawing unbiased inferences from probability samples is dependent on the collection of data from *all* sample persons. In other words, a response rate of less than 100 percent introduces the possibility of bias (Peytchev 2013). Sample persons' tendency to cooperate as reflected by response rates has been in decline in the past few decades (Curtin et al. 2005; de Leeuw and de Heer 2002; de Heer 1999; Hox and Leeuw 1994). This has led to increased interest in understanding the causes of nonresponse and techniques for preventing and adjusting for nonresponse.

Missing values caused by nonresponse occur in different patterns. *Unit nonresponse* (also referred to as *survey nonresponse*) occurs when measurements cannot be obtained for the entire sampled unit (Dillman et al. 2002), which usually happens, e.g., when a sample person refuses to cooperate in the survey altogether or when the sample person cannot be contacted. The second type of nonresponse is referred to as *item nonresponse*, and occurs when the sample person agrees to take the survey, but the data for certain items are unavailable (de Leeuw et al. 2003). Finally, there is *breakoff* (also referred to as *dropout* and *premature termination* in the literature), when the respondent starts the survey but stops prior to completing it (Peytchev 2009).

While unit nonresponse has received a great deal of attention in the literature, there has been much less research on item nonresponse. The research on breakoff is sparser still and is most often limited to web surveys. The present dissertation aims to address this lack of attention by focusing on the two less commonly studied types of nonresponse, namely item nonresponse and breakoff. We argue that these two types of nonresponse are connected, as they are affected by the same underlying factors. While an empirical association between item nonresponse and breakoff has already been found and described in the literature(Galesic 2006), studies that jointly consider both types of nonresponse are rare.

The present dissertation aims to make a contribution to the field of survey methodology by providing both theoretical insight into and an innovative application of the statistical modeling of nonresponse and breakoff. By enriching extant knowledge regarding factors affecting item nonresponse and breakoff in surveys, we aim to provide

an understanding of why the resulting missing values occur which is the prerequisite for *any* statistical treatment of missing values (de Leeuw et al. 2003).

The extant research has identified four fundamental levels of factors influencing item nonresponse and breakoff: respondent characteristics, interviewer characteristics, questionnaire item facets, and survey design characteristics. Kveder (2005) has demonstrated that omitting any of these three levels from analysis can lead to misleading conclusions about the effect of the predictors on item nonresponse. A great deal of attention in this dissertation is devoted to presenting and applying statistical models that can accommodate predictors at all three aforementioned levels. Our study aims to enrich extant knowledge regarding factors affecting item nonresponse and breakoff in surveys, thus providing a basis for more informed procedures for preventing item nonresponse and breakoff and addressing the resulting missing values when they do occur.

The dissertation is divided into three parts; The first part includes the Chapters 2 and 3, which review the relevant literature on survey nonresponse and introduce the statistical models later used to analyze item nonresponse and breakoff. The second part, comprising Chapter 4, describes the dataset that is analyzed in the empirical part. The third part consists of Chapters 5, 6, and 7, wherein the models for item nonresponse and breakoff are applied and results are interpreted and discussed.

Chapter 2 introduces background literature with an account of the question-answer process and Krosnick (1991)'s concept of satisficing when answering surveys. Having laid out this theoretical basis, we proceed to discuss item nonresponse and breakoff in turn. We describe Beatty and Herrmann's (2002) response decision model as the theoretical framework for item nonresponse and proceed to review the correlates of item nonresponse that previous research has discovered. We proceed in a similar fashion with breakoff, first reviewing the conceptual frameworks that authors have referenced in explaining breakoff, before giving an account of the more concrete findings in the extant research on breakoff. We continue with a discussion and synthesis of the presented material, and conclude by stating the hypotheses are to be tested in the empirical part.

Chapter 3 gives an account of the statistical models that are applied in the empirical part: generalized linear mixed models for item nonresponse and the Cox proportional hazards model for breakoff. The chapter starts with a section on multilevel modeling, which serves as an informal introduction to multilevel techniques, upon which later sections expand.

Chapter 4 is somewhat heterogeneous in its content, first describing the dataset that

that is analyzed in the empirical part of our study and the data collection procedures that were followed. It then describes how a number of additional predictors to be used in the analyses of item nonresponse and breakoff were obtained by coding each questionnaire item and by administering the respondents additional self-assessment items. Some of the predictors themselves have missing values, and we accordingly give an account of how multiple imputation was used to address this issue. The chapter concludes by operationalizing the hypotheses put forward at the end of Chapter 2.

Chapters 5 and 6 constitute the main empirical part of the present dissertation, where statistical models are applied to item nonresponse and breakoff, respectively. We start by performing preliminary analyses and proceed to applying statistical models of increasing complexity to the data. Our hypotheses are evaluated and discussed on the basis of the results. Chapter 7 concludes with a review of the material presented in the dissertation and a joint discussion of findings for item nonresponse and breakoff.

2 Survey nonresponse theory: item nonresponse and breakoff

This chapter will provide the conceptual framework for understanding item nonresponse and breakoff as two types of survey nonresponse. As these two survey phenomena take place within the wider context of the survey process, we will begin by giving an account of the question-answer process and satisficing. Having laid out this theoretical basis, we proceed to discuss item nonresponse and breakoff in turn. We continue with a discussion and synthesis of the theoretical frameworks presented and conclude by forming the hypotheses that will be tested in the empirical study.

2.1 The question-answer process

Survey statisticians have long been aware that that the question-answer process can be a source of response effects that contribute to non-random measurement error in survey statistics (Groves 1989). Starting in the second half of the 1970s, the survey interview was conceptualized through concepts from social and cognitive psychology (Bradburn 2004). Some influential works from this field are Sudman and Bradburn (1974), Sudman, Bradburn and Schwarz (1996), Schwarz and Sudman (1996), and Tourangeau, Rips and Rasinski (2000).

The survey interview is seen as a structured interaction between two people who play the distinctive roles of interviewer and respondent. The interview takes place in a social context that is governed by socially shared expectations and norms like mutual respect for individuals (in particular respect for the privacy of respondents), truthfulness, and confidentiality (Bradburn 2004).

Responding to a particular survey item involves considerable cognitive work on the part of the respondent. A number of models have been put forward in the literature (Cannell et al. 1981; Strack and Martin 1987; Sudman et al. 1996; Tourangeau and Rasinski 1988), which, though differing in details, generally agree on a series of processes that the respondent must go through in responding to a questionnaire item. These processes are *comprehension*, *retrieval*, *judgment*, and *formatting*. While conceptually viewed as a linear sequence, it is recognized that these processes take place within a conversation, and that different processes may go on in parallel or by cycling back and forth (Bradburn 2004). Next, we provide a short description of the processes,

commenting on where difficulties can arise.

In order to reply to a questionnaire item, the respondent must first *comprehend* what they are being asked. The researcher who developed the questionnaire aims for the respondent to understand the item in the same way as the researcher does. If the respondent is unsure whether they understand the item at hand the way the researcher intended, the respondent may respond with "don't know" or skip the item entirely (de Leeuw et al. 2003).

Having comprehended the questionnaire item, the respondent must *retrieve* from memory the necessary information to provide a response. Depending on the topic, this might be a cognitive task of some difficulty. For an opinion item, for example, it might mean remembering the requested information or generating it on the spot. If the item concerns a particular behavior, the respondent must recall or reconstruct relevant instances of this behavior and determine whether it occurred in the period mentioned in the item wording (de Leeuw et al. 2003). The next process concerns integrating the requested information into a summary *judgment*, i.e. formulating a concrete answer. Problems can arise in both of these phases. If the respondent does not have the required knowledge, has difficulty retrieving the requested information, or is simply not willing to spend the effort, this can, again, result in a "don't know" response or a skipped item. Alternatively, the respondent may reduce their cognitive effort by choosing the first reasonable response or choosing a response at random[1].

Having formulated a response, the respondent must *formatformatting* the response to fit one of the response formats offered by the interviewer. Since today almost all questionnaires depend on closed or pre-coded questions, this means choosing an appropriate response category. If none of the appearing response alternatives quite fit their imagined answer, the respondent can either select the next most appropriate response category, or decide to answer "don't know" or refuse to respond. Once a response alternative has been chosen, the respondent might still decide to edit their response, since they might be concerned with self-presentation. Research shows that responses to sensitive questions might be seriously distorted by respondents' unwillingness to admit to behavior that would put them in a bad light in the interviewer's eyes or by their respondents' desires to exaggerate socially desirable behavior (Bradburn 2004). The respondent can avoid a potentially embarrassing situation by telling a white lie or by outright refusing to answer (de Leeuw et al. 2003).

Krosnick (1991) makes the assertion that the question-answer process outlined above

[1]Two response behaviors that (Krosnick 1991) refers to as *satisficing*. We discuss satisficing in the remainder of this section.

is merely an ideal that is rarely achieved in practice. He refers to performing each of the four processes carefully and comprehensively as *optimizing*. Some respondents are, indeed, motivated to expend the substantial amount of cognitive effort required to optimize because of desires for self-expression, interpersonal response, intellectual challenge, self-understanding, or altruism. Krosnick argues, however, that respondents are likely to satisfy whatever desires motivated them to participate soon after beginning the interview and "become increasingly fatigued, disinterested, impatient, and distracted as the interview progresses" (Krosnick 1991).

As this happens, respondents are likely to shift their response strategy. Krosnick hypothesizes that at first this change in response strategy is reflected merely in being less thorough in comprehension, retrieval, judgment, and formatting. Respondents still go through all four steps, but less diligently and comprehensively as before when they were optimizing. "Instead of attempting to generate an optimal answer, respondents settle for generating merely satisfactory answers" (Krosnick 1991). Krosnick uses the term *weak satisficing* to refer to this response strategy.

Krosnick goes on to assert that after this strategy is used for a while, the respondent's fatigue continues to increase and the question and answer process becomes even more taxing. At some point the respondent may simplify their response strategy even further by omitting the retrieval and judgment steps altogether. The respondent thus interprets each questionnaire item only superficially, selecting what they believe "will appear to be a reasonable answer" (Krosnick 1991). Krosnick calls this response strategy *strong satisficing*.

Krosnick argues that a satisficing response strategy is often the reason behind the measurement effects that have been identified by survey methodologists. He mentions a number of forms of satisficing; A satisficing respondent may: 1) select the first response alternative that seems reasonable, 2) choose a response alternative at random, 3) avoid the cognitive effort of optimizing by choosing "don't know," 4) agree with whatever assertion is being made without really considering it, 5) endorse the status quo (e.g. in a political survey) simply because this appears to be a reasonable answer, 6) or choose the same answer to all items that appear in the same battery.

The likelihood that a given respondent will satisfice when responding to a particular item, Krosnick posits, is a function of three factors: the difficulty of the task, the respondent's ability to perform the task, and their motivation for doing so. Each of these factors, in turn, is argued to depend on the characteristics of the item or the respondent.

Krosnick's work on satisficing is relevant to the topic of this dissertation as satisfic-

ing respondents are more likely to generate item nonresponse by skipping over items, answering "don't know", or refusing to respond to items. The concept of satisficing is much broader, however, and, as mentioned, encompasses other forms of response behavior like choosing the first reasonable answer alternative or choosing one at random (Krosnick 1991).

2.2 Item nonresponse

This section will narrow the focus from the more general discussion of the question-answer process to item nonresponse in particular. De Leeuw et al. (2003) define item nonresponse as the unavailability of data on particular items. The term *unavailable* is used to stress that whether or not the value on an item is regarded as missing depends on the goal of the analysis. A "don't know" to a question on voting preference can, for instance, be regarded as a meaningful answer, while a "don't know" on an item inquiring about the respondent's income has no informational value. Item nonresponse also occurs when a respondent refuses to provide an answer to specific items, or when an item is skipped over (intentionally or not; by the respondent or the interviewer).

We will first describe Beatty and Herrmann's (2002) response decision model as a theoretical framework for item nonresponse, highlighting how their model relates to the theory on the question-answer process and satisficing presented above. We go on to give an account of the research that has identified a number of factors that affect item nonresponse.

2.2.1 Beatty and Herrmann's response decision model

While Krosnick regards item nonresponse as only one of several forms of satisficing, Beatty and Herrmann (2002) provide a framework for item nonresponse in particular. They, too, base their model on the question-answer process literature, but strongly emphasize that responses to questionnaire items are based on different degrees of knowledge. When asked if they had had a medical examination in the past 12 months, for example, one respondent may respond based on their memories of particular times and places where medical exams took place, while another respondent more or less guesses based on vague or incomplete memories (Beatty and Herrmann 2002). The authors argue that, in addition to what they call *errors of omission* (item nonresponse), respondents might also commit *errors of commission*, i.e., provide a substantive response where item nonresponse would have been more accurate.

Beatty and Herrmann postulate that three factors drive the respondent's decision on whether to respond to a particular questionnaire item or not:

1. the cognitive state: the availability of the requested information;
2. adequacy judgments: the respondent's perception of the level of accuracy required by the questioner; and
3. communicative intent: the respondent's decision on what to report (Beatty and Herrmann 2002, 72).

We will briefly describe the role of each of these three factors, starting with the respondent's *cognitive state*. Drawing upon previous psychological research, Beatty and Herrmann argue that the respondent's knowledge is a matter of degree rather than a dichotomy. The authors' own previous research (Beatty et al. 1998) prompts them to believe that four "cognitive states" can be identified on this continuum from knowing to not knowing:

1. available: the requested information can be retrieved with minimal effort;
2. accessible: the requested information can be retrieved with effort or prompts;
3. generatable: the requested information is not known exactly, but may be estimated using other information in memory; and
4. inestimable: the requested information is not known and has virtually no basis for estimation (Beatty and Herrmann 2002, 73).

The authors assert that the cognitive state is the most obvious determinant of whether or not the respondent provides a response to an item, but warn that the relationship between cognitive state and item nonresponse is not straightforward. Even in the extreme cognitive states (available or inestimable), the respondent's communicative intent may come into play. In the intermediate states, on the other hand, inference plays an important role in coming up with an answer (see Bradburn et al. 1987).

When a potential response contains a degree of uncertainty, estimation, or guessing, the respondent needs to *judge the adequacy* of the potential response. In the intermediate cognitive states, the respondent is able to provide a response, but may be uncertain as to whether this information meets the requirements of the question. Beatty and Herrmann suggest that the decision on whether to report the potential substantive answer or to resort to item nonresponse is based on the respondent's judgment of the level of precision called for by the survey. They conduct an experiment with three forms of the same questionnaire, each form differing only in the instructions to the respondent regarding uncertainty about the precision of the response. The form of the questionnaire that encouraged the use of "don't know" when uncertain indeed resulted in the highest rates of "don't know" responses, providing support for Beatty

and Herrmann's thesis (2002).

The last factor, *communicative intent*, refers to the respondent's decision on what to report. The respondent may have come up with a reasonable response at this point, but chooses not to report it. The most common reason for this is that the respondent believes their behaviors or attitudes are socially undesirable, or even that providing a truthful response might put them at risk of legal consequences (e.g. replying to an item on drug use). Beatty and Herrmann stress that the converse may also occur: unwilling to admit ignorance, the respondent may choose a substantive answer over "don't know." Respondents' tendency to avoid admitting ignorance has been demonstrated, e.g., by experiments in which respondents were asked to provide their opinions on fictitious issues (e.g. non-existent pieces of legislation), where some of the respondents provided their opinions, rather than admitting to not being familiar with the issue at hand (Bishop et al. 1986).

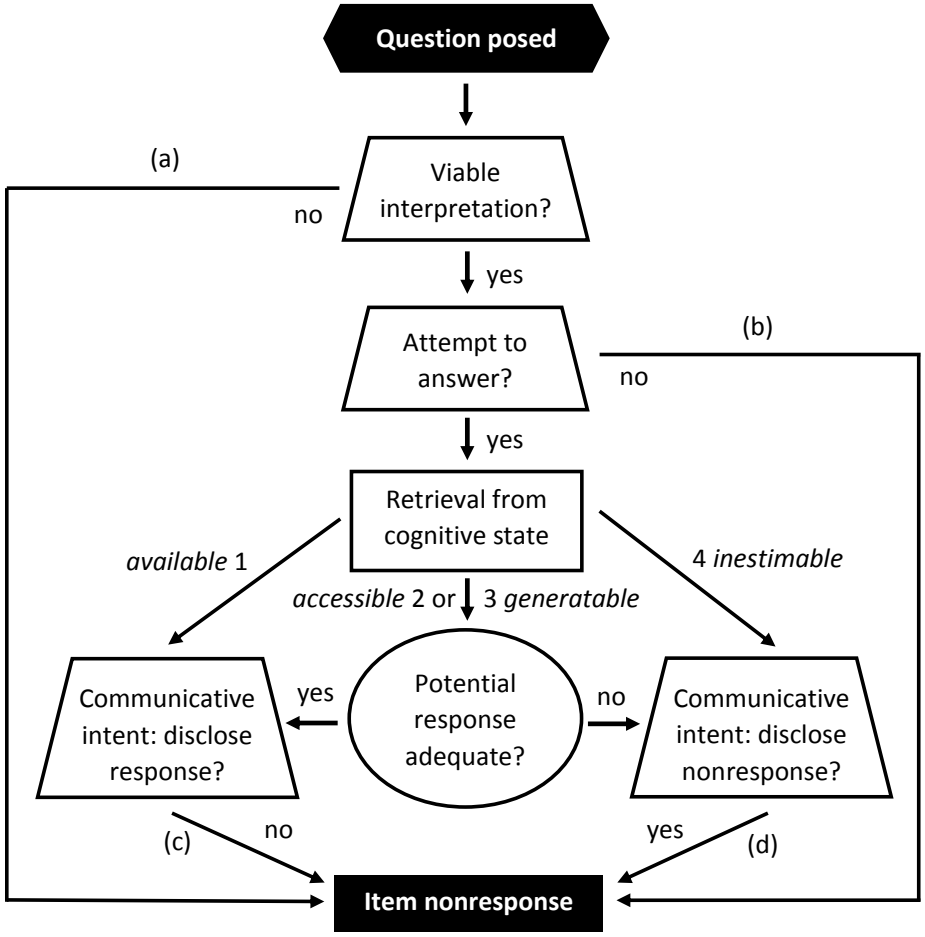
Beatty and Herrmann's response decision process is elucidated by Figure 2.1. The process starts when the respondent is administered the questionnaire item and must come up with an interpretation of what is being asked. The first opportunity for item nonresponse occurs here if the respondent does not understand the task at hand. This step corresponds to the *comprehension* phase in the question-answer process (see Section 2.1)

If the respondent believes they have a viable interpretation of the request, the process continues. The respondent must then decide whether or not to make the effort required to answer. If their motivation is low, the respondent may opt out of the response process (cf. Krosnick's concept of satisficing, 1991). This decision may be influenced by the length of the interview and the complexity of the item (Beatty and Herrmann 2002).

The respondent who continues must then retrieve the information that is relevant to the item and available. If the requested information is *available* or *inestimable*, communicative intent is the only remaining determinant of item nonresponse (Beatty and Herrmann 2002). In the available cognitive state the respondent knows the answer, so the only explanation for item nonresponse is a decision not to report the answer. In the inestimable cognitive state, conversely, the respondent does not know the answer, making "don't know" the most accurate response. As their communicative intent comes into play, however, the respondent may decide to avoid admitting ignorance and provide a substantive answer.

If the requested information falls into one of the intermediate cognitive states of (*accessible* or *generatable*, see Figure 2.1), the respondent's adequacy judgments will come

Figure 2.1: Beatty and Herrmann's (2002, 77) response decision model; italics added



into play as they first evaluate the quality of the potential response. Following this judgment, the respondent continues with communicative intent decisions, ultimately leading to a substantive response or item nonresponse (Beatty and Herrmann 2002).

2.2.2 Correlates of item nonresponse

This section will provide a short overview of the extant research into the factors that correlate with item nonresponse. These factors can be broadly categorized into *survey design features*, *item facets*, *respondent characteristics*, and *interviewer characteristics*. Although these issues are discussed separately in the literature, it is likely that the factors interact to influence item nonresponse, and are as such difficult to separate in operational settings (Wolfe et al. 2008). This is likely the reason why research on the correlates of item nonresponse sometimes produces conflicting results: the underlying studies are conducted on surveys concerning different topics, administered in different modes, to different populations.

Survey design features

The mode of administration in which the survey is conducted has been found to have a profound effect on the item nonresponse rate. Web administration was found to produce higher rates of item nonresponse than face-to-face (Heerwegh and Loosveldt 2008) and telephone interviewing (Roster et al. 2004; Smyth et al. 2008). This effect is attributed to the lack of interviewer probing and interviewer-provided extrinsic motivation in self-administered modes (de Leeuw 1992).

Comparisons of web and mail surveys are more common but report conflicting results. The majority of studies found more item nonresponse in mail surveys (Boyer et al. 2002; Klassen and Jacobs 2001; Kwak and Radler 2002; Lorenc 2010; Shin et al. 2012; Truell et al. 2002; Stanton 1998), but others have reported the opposite result (Bates 2001; Lozar Manfreda and Vehovar 2002; Denniston et al. 2010). The authors attribute the better performance in web mode to the benefits of computerized questionnaires, such as real-time validation of responses and automatic routing, which dramatically reduces the number of erroneously skipped items.

There has been extensive research into the effect of incentives (small sums of money given to the respondent) in the survey methodology literature (for a review see Singer and Ye 2013; Cantor et al. 2008; Singer 2002). Authors investigating the effect of incentives regard (the lack of) item nonresponse an indicator of response quality[2]. Two competing alternatives have been put forward with respect to the impact of incentives on data quality; according to the first hypothesis, incentives induce a response from sample persons who would otherwise have refused, thus leading to a decline in response quality. The alternative hypothesis, on the other hand, predicts that incentives, by rewarding the participants, will lead to a better quality of responses (Singer and Ye 2013). Empirical studies have found incentives to lead to less item nonresponse (James and Bolstein 1990; Mack et al. 1998; Singer et al. 2000), or have found no relationship between incentives and item nonresponse (Berk et al. 1987; Davern et al. 2003; Goyder 1994; Shettle and Mooney 1999; Singer et al. 1999; Teisl et al. 2006; Tzamourani and Lynn 1999; Willimack et al. 1995; Dirmaier et al. 2007; Petrolia and Bhattacharjee 2009; Curtin et al. 2007; Cantor et al. 2008). The only study reporting a significant *increase* in item nonresponse when using incentives is by Jäckle and Lynn (2008).

[2]Another readily available measure that is often used is the length of answers to open-ended items. Other response quality indicators like the reliability and validity of responses, which could be argued to be more important, have not been considered in the studies regarding incentives (Singer and Ye 2013).

Item facets

Items most obviously differ with regard to topic. Items dealing with sensitive topics have been found to exhibit more nonresponse, with typical examples being the question on income (Pickery and Loosveldt 2001, 2004; Gruskin et al. 2001) and items dealing with sexual behavior (Catania et al. 1996; Tu and Liao 2007; Gruskin et al. 2001; Kupek 1998). The effect of sensitive items, furthermore, has been found to vary with regard to the mode of administration. In a study that examined the reporting of sensitive behavior across different modes, Kreuter et al. (2008) found that more sensitive information was reported in web-administered surveys as compared to ones administered over the telephone. Increasing the distance between respondents and interviewers by switching the mode of administration or reducing the presence of the interviewer has been recommended as a general strategy for reducing social desirability effects (Bradburn 2004).

Though the majority of items in contemporary questionnaires are closed-type (offer the respondent a limited choice of response alternatives), researchers also make use of open-ended items. Open-ended items require that the respondent formulate their own answer without reference to predetermined categories, and are, as such, used 1) to avoid biasing results by suggesting responses, and 2) to allow the possibility of discovering responses given spontaneously and which the researcher might not have thought to include among the response categories (Reja et al. 2003). Open-ended items have been consistently found to induce more item nonresponse, especially in self-administered modes (Denscombe 2009; Börkan 2010; Reja et al. 2003; Aoki and Elasmar 2000). This finding has been attributed to the additional effort that open-ended items require of the respondent. The absence of the interviewer who could administer additional probes at such items exacerbates this effect in self-administered modes (Reja et al. 2003).

Klein et al. (2011) included the item's position in the questionnaire as a predictor in their statistical model for item nonresponse and found that items that appeared later in the questionnaire were subject to more item nonresponse. Wolfe et al. (2008), on the other hand, found little evidence of a relationship between item position and the probability of item nonresponse in their analysis.

The likelihood of item nonresponse rises with the increase of the cognitive burden the item puts on the respondent. In a study on the elderly, Knäuper et al. (1997) argued that this effect is due to an interaction between the item's difficulty and the respondent's cognitive ability. The respondents were administered a cognitive ability test involving recall from memory and measures of knowledge, language, and orientation.

The questionnaire items, on the other hand, were coded according to nine indicators of difficulty: question length and complexity; the presence of additional instructions, introductory phrases, and ambiguous terms; whether the item asked for retrospective reports, frequency reports, or quantitative reports; and whether a response scale was used. The authors found that of the nine considered measures of difficulty, five exhibited a significant interaction effect (in the expected direction), with the respondent's cognitive ability in line with the authors' main hypothesis (Knäuper et al. 1997).

Respondent characteristics

Studies that attempt to explicitly measure respondents' cognitive ability are rare (e.g. Knäuper et al. 1997). More often the assumption is made that the respondent's age and education can serve as *proxies* for their "cognitive sophistication", which is seen as a causal factor (Peytchev 2009; Krosnick 1991). Indeed, respondents' age and education have been found to consistently correlate with item nonresponse (de Leeuw et al. 2003). A statistically significant effect indicating that older respondents produce more item nonresponse was found by Bell (1984); Elliott et al. (2005); Hox et al. (1991); Klein et al. (2011); Pickery and Loosveldt (1998); Shin et al. (2012); Singer et al. (2000); Gruskin et al. (2001). Pickery and Loosveldt found that older respondents were more likely to produce item nonresponse on political statement items, while the opposite effect was found for items on income (Pickery and Loosveldt 2001, 2004). Respondents' education was found to have a significant effect on item nonresponse by Bell (1984); Klein et al. (2011); Pickery and Loosveldt (1998, 2001, 2004); Shin et al. (2012); Singer et al. (2000): more educated respondents were found to produce less item nonresponse.

Survey methodologists have attempted to measure respondents' attitude toward surveys by asking them how much they agree with statements like "surveys are valuable for society at large" or "surveys are a waste of people's time." Respondents with more positive attitudes toward surveys have been found to produce less item nonresponse (Singer et al. 1998; Stocke 2006). Stocke argues that respondents with a positive attitude toward surveys frame their particular survey interview in such a way that they regard supporting this survey as an important goal. In other words, respondents with a positive attitude toward surveys are more motivated to optimize, and this increased motivation counteracts the high burden imposed by certain questionnaire items (Stocke 2006).

When a statistically significant effect of respondent sex is found in studies, the direction of the effect is usually that women produce more item nonresponse (Bell 1984; Elliott et al. 2005; Klein et al. 2011; Pickery and Loosveldt 1998, 2001, 2004; Singer et al. 2000). An exception that we encountered when studying the extant research was in

Shin et al. (2012), who found more item nonresponse for men. Another respondent characteristic that has been considered in studies of on item nonresponse is income: wealthier respondents have been found to produce less item nonresponse, (Singer et al. 2000; Shin et al. 2012).

Interviewer characteristics

Rather than being seen merely as a cognitive process involving the respondent, the survey interview is viewed also as a communicative process (Schwarz and Sudman 1996) in which an important role is played by the interviewer. The literature on the effect of the interviewer as a source of measurement error is extensive (see e.g. Groves 1989, 357-406 for a review) and recognizes the interviewer as an active participant, able to influence the answers obtained, rather than as simply a neutral collector of data. The standardized interviewing approach to minimizing the interviewer's influence stresses that all respondents should receive exactly the same stimulus (wording) without additional comments or omissions (Fowler and Mangione 1990). This approach has, on the other hand, been criticized by proponents of conversational interviewing, who claim that the standardized approach can prevent interviewers from resolving misunderstandings thereby increasing measurement error (Suchman and Jordan 1990).

Studies on the interviewers' effect on item nonresponse have shown that interviewers differ substantially with regard to how much item nonresponse their respondents produce (Catania et al. 1996; Hox et al. 1991; Pickery and Loosveldt 1998, 2001, 2004; Singer et al. 1983; Tu and Liao 2007). Researchers have attempted to explain this variability by using interviewer characteristics as predictors in multilevel models, but have found that characteristics such as interviewer demographics have little or no explanatory power for item nonresponse (Pickery and Loosveldt 1998, 2001, 2004).

The study by Hox et al. (1991) reported that introverted interviewers and interviewers who, when asked, expressed a preference for face-to-face mode produced interviews with a higher proportion of item nonresponse. In a telephone study on sexual behavior, Catania et al. (1996) found that when the respondents were able to choose the sex of their interviewer, the probability of breakoff was lower and the quality of the gathered data higher. Singer et al. (1983) report that more educated interviewers obtained interviews with less item nonresponse.

When the survey is conducted by an interviewer (as opposed to being self-administered), the interviewer can be instructed to evaluate the respondent's willingness to answer the questions once the interview has been completed. Such interviewer ratings of respondent cooperation have been found as significant predictors of item nonre-

sponse in statistical models: more cooperative respondents were found to produce less item nonresponse (Hox et al. 1991; Tu and Liao 2007).

2.3 Breakoff

This section will provide a theoretical background for breakoff. Because it came under more intense study only with the advent of web surveying in the 1990s, the literature on breakoff is less extensive than the literature on item nonresponse. No conceptual model has been put forward that would focus specifically on breakoff. Researchers interested in studying breakoff must therefore borrow from theoretical frameworks developed for other survey phenomena like item nonresponse and unit nonresponse (Peytchev 2009). We will first give an account of the more theoretical approaches to breakoff and then move on to discuss specific correlates of breakoff that have been identified in the literature.

2.3.1 Breakoff and the question-answer process

Breakoff occurs when a respondent starts the survey, but stops answering prior to completing it. The incidence of breakoff is strongly influenced by the mode of administration. Metacontent analyses of web surveys found breakoff rates ranging from 1% to 87% with an average of 34% (Musch and Reips 2000) and ranging from 0% to 73% with an average of 16% (Lozar Manfreda and Vehovar 2002). The proportion of breakoff in face to face and telephone surveys is, by comparison, usually lower than 5% (Galesic 2006). This dramatic difference is attributed to the absence of the interviewer in the web mode. This allows respondents to reevaluate their decision to participate throughout the web survey, compared to an interviewer-administered survey, where the respondent makes a more permanent decision to participate at the time of the survey request (Peytchev 2007).

Administration reports for face to face and telephone surveys usually treat breakoff as unit nonresponse (if it appears early in the interview) or item nonresponse (if it appears toward the end of the interview) (de Leeuw et al. 2003). Because breakoff became an issue of concern to survey methodologists only with the dawn of web surveying, most of the research aimed specifically at explaining breakoff uses data from web surveys. However, even with the high reported breakoff rates in web surveys, breakoff has received scant attention in the research literature (Peytchev 2009). Peytchev attributes this deficiency in scholarly attention to the "lack of a unified framework that places breakoff in relation to other forms of nonresponse, a lack of theories specifying causal mechanisms, and empirical difficulties" specific to breakoff (2009). We find that a

number of authors have, nevertheless, contributed ideas that could be valuable in explaining breakoff: in addition to Peytchev (2009), we will consider Galesic (2006), Bosnjak and Tuten (2001), and Yan and Curtin (2010).

In calling for a theoretical framework, Peytchev (2009) argues that such a framework would place breakoff in the context of the extant theories on unit nonresponse in household surveys (Groves and Couper 1998), item nonresponse (Beatty and Herrmann 2002), and the question-answer process (see Section 2.1). The framework for unit nonresponse focuses on the respondent-interviewer interaction and the initial decision to participate, classifying the factors affecting participation into environment, respondent, survey, and interviewer (Groves and Couper 1998; Peytchev 2009). Breakoff is conditional on the initial decision to participate, but is thereafter affected by questionnaire characteristics that are not seen by unit nonresponders. Because breakoff occurs on the item level, models like Beatty and Herrmann's (2002) response decision model are relevant after the initial decision to cooperate. Peytchev sees the respondent as continuously re-evaluating their participation in the (web) survey and suggests that breakoff can be regarded as an alternative to item nonresponse (2009). He identifies three sets of factors affecting breakoff that are worthy of study: respondent characteristics, survey design, and item facets.

Attempting to provide a theoretical basis for the respondent's continuous decision whether to continue or not, Galesic (2006) resorts to decision field theory (Busemeyer and Townsend 1993). A central concept in her application of this approach to survey behavior is the *inhibitory threshold*: the point that determines when the difference in the preference for one action is large enough to trigger some sort of behavior. "As long as the preference for one action is larger than that for the other, but not enough to cross the threshold, a person shows only an inclination towards the preferred activity but is not actually performing it" (Galesic 2006, 314). The author asserts that, at the beginning of the survey, the factors that influenced the initial decision are still influential, but, as the study continues, negative aspects of participation like fatigue and boredom become stronger, as does the preference to stop participating. The decision to break off, however, will not be made until this change exceeds the inhibitory threshold. In her empirical study, Galesic finds that item nonresponse increases immediately prior to breakoff, and attributes this to the intermediate period when the respondent would already prefer to stop responding, but this preference is still not strong enough to provoke breakoff (2006).

Bosnjak and Tuten (2001) classify response behavior in web surveys based on two dimensions: 1) the number of *displayed* items and 2) the number of *answered* items. On the basis of these two dimensions, they define seven segments of response behaviors, e.g.

complete respondents with the maximum value on both dimensions (all items displayed and answered); unit nonresponders (zero on both dimensions), *answering drop-outs* (intermediate but equal value on both dimensions), *lurkers* (maximum value on first, zero on second dimension) etc. (for details see Bosnjak and Tuten 2001). The authors suggest that the response behavior results from three factors (motivation, opportunity, and ability). We regard their classification scheme as inherently descriptive and do not consider it further (see also Peytchev 2009 for a critique).

We will mention one final contribution that might be useful in explaining breakoff. Survey methodologists have long speculated about and reported on collecting data of lower quality from reluctant respondents and respondents who initially refuse to cooperate in the survey (see Yan and Curtin 2010, and references therein). Yan and Curtin explicate these speculations in what they term the *response continuum perspective*. They posit that a continuous factor—the sample person's *response propensity*—determines both unit and item nonresponse. A person with a very low response propensity thereby becomes a unit nonresponder when approached with a survey request, a respondent with an intermediate response propensity takes part in the survey but produces some item nonresponse, while a respondent with a high response propensity answers all items. The authors examine 20 years of data from the Survey on Consumers and consider as evidence in favor of their thesis the fact that during the last five examined years of the series the unit response rate increased while the item nonresponse rate simultaneously sharply decreased (Yan and Curtin 2010). Even though Yan and Curtin do not explicitly mention breakoff, we argue it would be entirely reasonable to put breakoff on the same response continuum, somewhere between unit and item nonresponse. We therefore interpret Yan and Curtin's work as suggesting a common underlying cause for the various forms of nonresponse.

2.3.2 Correlates of breakoff

This section gives an overview of breakoff correlates that has been identified in the survey literature. The factors studied can be broadly classified into *survey design features*, *item facets*, and *respondent characteristics*. Given the previously mentioned lower prevalence of breakoff in interviewer-administered surveys, most studies on breakoff have used web-administered questionnaires.

Survey design features

A metacontent analysis of 74 web surveys found that the breakoff rate was lower when the survey was conducted on special populations (as opposed to the general

population), as well as when incentives were used (Lozar Manfreda and Vehovar 2002).

Crawford et al. (2001) found that more sample persons started filling out the questionnaire when the announced survey length was lower (8 to 10 minutes, as opposed to 20 minutes). The study found, however, that once respondents started the survey, those in the 20-minute group had a lower breakoff rate.

A great deal of research has focused on certain features of the web questionnaire which can be used to lower the breakoff rate. *Progress bars* (visual aids that display the respondent's current position in the web questionnaire) have received particularly extensive treatment (see e.g. Couper et al. 2001; Crawford et al. 2001; Conrad et al. 2003; Heerwegh and Loosveldt 2006; Peytchev 2009; Matzat et al. 2009). The researchers initially assumed that the information conveyed by progress bars would motivate the respondent to persevere to the end of the survey and were surprised to find that the breakoff rate was higher when the progress bar was used (Crawford et al. 2001). This result was replicated by subsequent studies, some of which further investigated the effect of different types of progress indicators, such as a progress bar that showed (deceitfully) fast progress at the beginning of the survey (fast-then-slow), which was compared to the converse type (slow-then-fast) and the linear progress indicator (Conrad et al. 2003; Matzat et al. 2009). Conrad et al. (2003) found that when respondents first received encouraging information from the fast-then-slow progress bar, the breakoff rate was lower than in other settings, the respondents skipped fewer items (produced less item nonresponse) and evaluated the study as having been more interesting. Matzat et al. (2009), on the other hand, found the effect of any type of progress indicator to be either negative or nonexistent. Both studies agree that if the questionnaire is long (exceeding 20 minutes), any kind of progress indicator is likely to increase the breakoff rate.

Respondent characteristics

Studying the effect of respondent characteristics and item facets on breakoff involves a number of empirical difficulties. If we want to investigate, e.g., the effect of gender on breakoff, then we need information on the gender of respondents who broke off. Information on respondent characteristics will not generally be available unless the respondent provides it *prior to* breaking off. When studying the effect of item facets, there must be sufficient variation in the facet under study (e.g. intrusiveness of item topic) across items in the questionnaire. Finally, to be able to separate the effect of item facets from the location of items in the questionnaire, the respondents should receive different versions of the questionnaire, otherwise the effect of item facets is completely confounded with the order of the items in the questionnaire (Peytchev

2007, 60).

Studies have also investigated the effect of respondent characteristics on breakoff. When demographic characteristics are studied, they are not seen as causing breakoff, but are instead theorized to be proxy measures for causes that cannot be measured directly (Peytchev 2009). Age and education have been, e.g., used as proxy measures of respondents' "cognitive sophistication" (see e.g. Krosnick 1991), with higher levels of education and lower age associated with higher cognitive sophistication. Studies on breakoff (Peytchev 2009; Galesic 2006) have indeed found that educated respondents are less inclined to break off.

The researchers have not, however, found a consistent effect of age; Matzat et al. (2009) found the risk of breakoff was higher for older respondents, while (Peytchev 2009) and (Galesic 2006) found the opposite effect, contrary to the "cognitive sophistication" hypothesis. The effect of the respondent's gender was mostly found to be insignificant (Peytchev 2009; Galesic 2006), though Peytchev (2011), in a study on the student population found that significantly more men broke off. Matzat et al. (2009) found respondents with more experience taking surveys to have a lower risk of breakoff.

Item facets

Peytchev (2009) studied the effect of item facets on breakoff and found that the risk thereof was increased when items wordings were longer and when items required manual input (as opposed to selecting a category). The metacontent analysis by Lozar Manfreda and Vehovar (2002) found that the survey breakoff rate was positively correlated to the proportion of open-ended items in the questionnaire.

The introduction of a new section in the questionnaire foreshadows additional material to come, and thus provides a natural breaking point in the conversation (Peytchev 2007, 62). Breakoffs were found to occur more frequently at section introductions in telephone surveys (Groves and Kahn 1979). This result was replicated by Peytchev (2009), who in a web-survey found that pages introducing a new section had more than twice the relative risk of inducing breakoff.

Galesic (2006) studied the effect of professed interest and experienced burden on breakoff. The survey consisted of blocks of questionnaire items, each block followed by a request to rate how interesting and burdensome the respondent found the items in that block. Galesic found that lower interest and higher burden were associated with a higher risk of breakoff.

Galesic speculates that the effect of burden might be *cumulative*, "whereby the bur-

den experienced at each question is a function of both specific characteristics of that question and burden experienced while answering the preceding questions" (Galesic 2006, 315). Peytchev (2009) tests this hypothesis by operationalizing the cumulative burden as the cumulative number of items administered to a respondent up to the point considered in the questionnaire. The corresponding effect is, however, found not to be statistically significant.

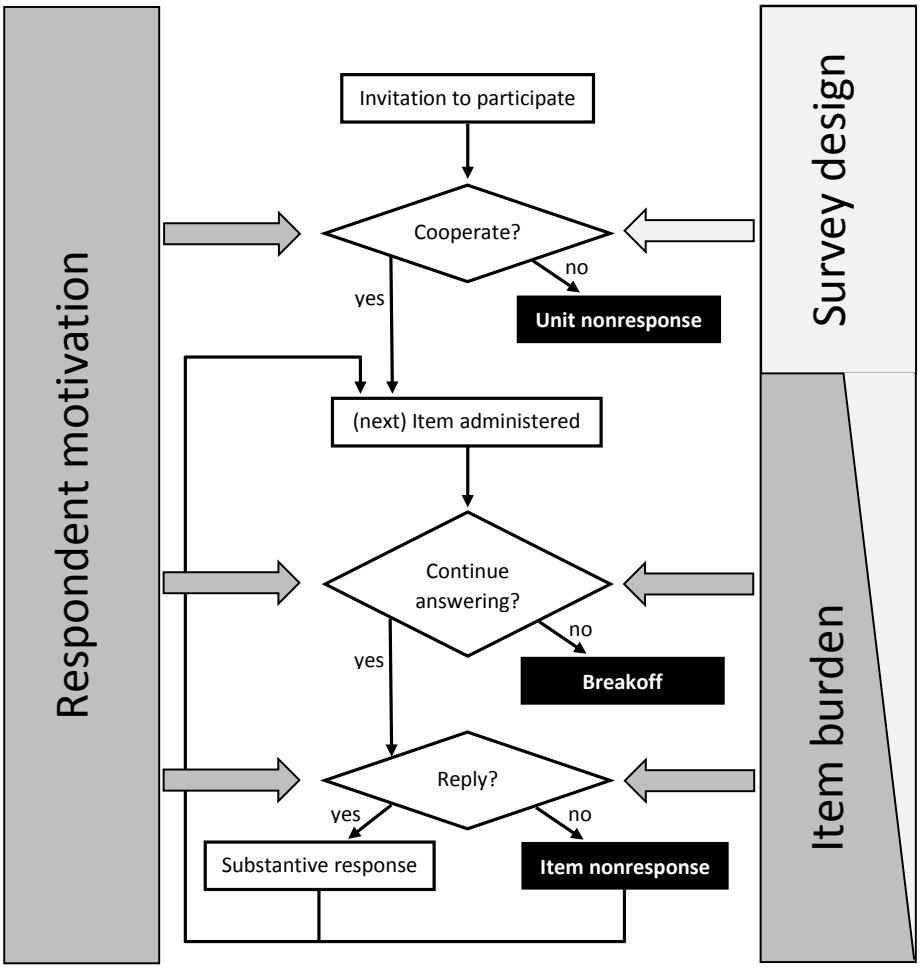
2.4 Synthesis and discussion

In this section we discuss the cited literature and attempt to find aspects where the different frameworks can complement each other. First of all, we would like to point out that (Galesic 2006) makes an argument very similar to Krosnick's (1991), i.e. that as motives that motivated the respondent to take part in the survey lose their influence, the respondent's attitude toward the survey task changes. The respondent will now expend less effort in executing the question-answer processes, resulting in lower-quality data—the response strategy that Krosnick (1991) refers to as satisficing. The difference is that Galesic is commenting on web respondents while Krosnick (1991) made his arguments before the advent of web surveys. It is reasonable to assume that the inhibitory threshold for breakoff is much lower in web surveys than in face to face or telephone surveys, where ending the interview involves terminating interaction with the interviewer. The satisficing period is therefore much shorter in web surveys as compared to interviewer-administered surveys, where the end of the questionnaire is likely to be reached before the inhibitory threshold is exceeded.

Having agreed to partake in the survey, the respondent's motivation to optimize—perform each phase of the question-answer process carefully and comprehensively (Krosnick 1991)—is affected by item characteristics. The respondent's motivation is unaffected (perhaps even increased) if the respondent can understand the item, can retrieve the relevant information with ease, and has no concerns about self-presentation or threats involved in reporting the constructed response (the phases of Beatty and Herrmann's (2002) response decision model). If, however, the questionnaire item is difficult to understand, requires substantial cognitive effort in retrieving the information, or concerns a topic perceived as sensitive or threatening, the respondent's motivation to optimize will decrease. If the respondent's motivation drops below the inhibitory threshold, they will break off. This interplay of item burden and the respondent's motivation is the rationale behind the *cumulative effect of item burden on breakoff* hypothesized by Peytchev (2009) and Galesic (2006).

Our view of the survey process and the various forms of nonresponse that can occur is

Figure 2.2: Decision model for unit nonresponse, breakoff, and item nonresponse



illustrated in figure 2.2. The respondent's initial decision to participate is influenced by the respondent's motivation on the one hand, and survey design features on the other. A negative decision at this point results in unit nonresponse—a form of nonresponse that is not considered further in this dissertation. If the respondent agrees to participate, on the other hand, questionnaire items will begin to be administered. Now that the respondent has been exposed to the content of the questionnaire, item burden plays an important role in addition to survey design features in influencing the respondent's decisions. If the respondent's motivation is not sufficiently high to counteract a particular item's burden, the respondent may decide to break off. Even if the respondent does not break off, they may still decide to omit the response to the item in question. This decision is, again, influenced in part by the respondent's motivation and in part by item burden and survey design features.

The respondent's motivation is, furthermore, influenced[3] by respondent characteristics, including their attitude toward surveys and the interviewer-provided extrinsic

[3]The factors influencing the respondent's motivation and item burden are not shown in Figure 2.2 as to avoid further cluttering the diagram.

motivation. The item burden, on the other hand, is influenced by the item's topic and item facets like response format. Survey design features include the mode of administration, incentives, questionnaire design, the use of progress bars (in web surveys) etc. The interplay of the same underlying factors—the respondent's motivation on the one hand and item burden and survey design features on the other—influence both the respondent's decision as to whether or not to break off and the respondent's decision about replying to the currently administered item. We argue that this is reflected in empirical findings like the increase in item nonresponse immediately prior to breakoff (Galesic 2006). It also leads us to expect the empirical analysis to show respondent-level and item-level predictors to have similar effects on both item nonresponse and breakoff.

The motivation to optimize that the respondent has at the beginning of the interview can vary widely across respondents. Some respondents may be genuinely interested in the survey, while others are "reluctant respondents", meaning that "they may feel pressured to participate in the study because of follow-up procedures, because they do not like to refuse a strong request from another person or for some other reason" (Bradburn 2004). In this respect, the *motivation to optimize* is similar to the factor that (Yan and Curtin 2010) posit as the determinant of both unit and item nonresponse—the response propensity. We prefer the term motivation to optimize, however, as the concept of response propensity is not grounded in the question-answer theory. The implication of extending the motivation to optimize to nonresponders is that there is a tradeoff inherent in converting refusing and reluctant sample persons. If convinced to participate, such respondents will likely care more about getting the interview finished than taking the time to carefully proceed through the question-answer processes (Bradburn 2004). This result is data of lower quality, a higher rate of item nonresponse and greater chance of breakoff, especially in web mode.

2.5 Hypotheses

In this section we make hypotheses based on the substantive theory discussed above. Each hypothesis will be operationalized at the end of Chapter 4, after we have introduced the statistical models that we use to model item nonresponse and breakoff and have given a description of the data. Because we see breakoff as a more extreme alternative to item nonresponse, occurring when the respondent's inhibitory threshold is exceeded, we will make essentially the same hypotheses for both breakoff and item nonresponse.

Hypothesis 1: Item nonresponse and breakoff will be most common in web mode and least common in face-to-face mode.

Refusing to reply to an item or terminating the interview are actions that violate behavioral expectations and norms in the survey interview. Answering "don't know" similarly requires the respondent to admit ignorance about a certain issue which might raise concerns about the respondent's self-presentation in the social situation of the interview. The inhibitory threshold for any of these actions will be highest when the respondent is in the same room with the interviewer; somewhat lower when the communication between the respondent and interviewer is conducted over the phone, as this increases the distance; and lowest when the questionnaire is self-administered.

Hypothesis 2: Item nonresponse and breakoff will be more common for cognitively less sophisticated respondents. This effect will be even more pronounced in web mode.

Respondents with lower cognitive sophistication will need to expend more effort to perform the stages of the question-answer process. The motives that motivated the respondent to take part in the survey will therefore lose their influence sooner at which point the respondent's response strategy will change to satisficing and perhaps even breaking off if the inhibitory threshold is reached. We hypothesize this effect to be present in all modes, but most pronounced when the interview is self-administered and the respondent cannot receive any help or elucidation from the interviewer in case they have trouble understanding the task at hand.

Hypothesis 3: Item nonresponse and breakoff will be less common among respondents with a more positive attitude toward surveys in general.

The rationale behind this hypothesis is simply that respondents with a more positive attitude toward surveys will have a higher motivation to perform each phase of the question-answer process carefully and comprehensively. They will thus be more willing to respond to items that are sensitive or seen as threatening and more willing to expend effort, despite the fact that the requested information is not in the available cognitive state.

Hypothesis 4: Items that are sensitive or present a threat of disclosure will induce more item nonresponse and breakoff. This effect will be less pronounced in web mode.

If the item topic is perceived as sensitive or threatening, the respondent's communicative intent will be lower, increasing the likelihood of nonresponse. If the respondent's

preference for discontinuing the interview is already high, such an item can induce breakoff. The effect of sensitive and threatening items will be lower when the questionnaire is self-administered, because the respondent's concerns for self-presentation will be lower in the absence of an interviewer.

Hypothesis 5: Items that are complex or deal with topics that the respondent is less familiar with will induce more item nonresponse and breakoff. This effect will be even more pronounced in web mode.

Complex items require more cognitive effort on the part of the respondent. If lacking motivation, the respondent faced with a difficult task might decide to produce a "merely satisfactory" (Krosnick 1991) response like a "don't know", or skip an item. If the respondent's preference for discontinuing the interview is already high, such an item can induce breakoff. We expect the effect of difficult items to be even more pronounced in web mode, where the absence of an interviewer means the respondent cannot receive any help with items they find difficult.

3 Modeling item nonresponse and breakoff: a review of relevant statistical models

This chapter will introduce the statistical models that will be used to analyze item nonresponse and breakoff in the empirical part of the present dissertation. The information on whether the respondent provided a substantive response to a particular item or not is recorded in a *binary* variable. We want to regress item nonresponse on available predictor variables in order to examine the relationship between the occurrence of item nonresponse, and respondent characteristics and item facets. Because each measurement occasion is associated with (nested in) a particular item on the one hand, and with a particular respondent on the other, the model for item nonresponse must be able to accommodate multilevel structures. Models for binary response variables containing both random and fixed effects are referred to as *generalized linear mixed models*, and are introduced in Section 3.2.

Unlike item nonresponse, breakoff is a *terminal event* in the survey interview. A respondent can produce several item nonresponses during the course of the interview, but can only break off once at the most. The information on whether the respondent broke off or completed the interview is also recorded in a binary variable. Unlike item nonresponse, however, our primary interest in breakoff lies in *when*—how far into the questionnaire—the breakoff occurred (if at all). Survival analysis methods offer a way of modeling such data and will be introduced in Section 3.3.

We start with a section on multilevel modeling with the purpose of providing an informal introduction to multilevel techniques. Section 3.2 builds on this and provides a more thorough and technical discussion of models that will be applied to item nonresponse.

3.1 Multilevel modeling

Many kinds of data have a hierarchical, nested, or clustered structure. A typical example that many multilevel modeling textbooks use to illustrate this is that of students who are grouped into schools. This is an example of a hierarchy in which *units* are grouped at different *levels* (Goldstein 2011, 3). Let us imagine that we have data from an educational study and are interested in modeling students' grades on a standardized test in each school. Variables that can serve as predictors are available

at the student level (e.g. sociodemographics and previous grades) and school level (e.g. the socioeconomic status of the school's neighborhood, and whether the school is public or private). We will briefly describe three approaches to modeling such data, which Gelman and Hill (2006) refer to as *complete pooling*, *no pooling* and *partial pooling*.

Complete pooling would, in this case, correspond to disaggregating the school-level predictors to the student-level (ascribing the school characteristic to each student) and performing a regression of the students' grades on the student- and school-level predictors. Such an analysis does not satisfy the assumption required in classical regression that the units of analysis be independent. In fact, students' grades within the same school are usually found to be more similar than the grades of students from different schools. One of the consequences of violating the necessary assumption is that the precision of the regression coefficients is overestimated in such a model, especially for school-level predictors (Snijders and Bosker 1999).

No pooling, on the other hand, would mean fitting a regression model for each school separately. The disadvantages of such an approach are that the school-level predictors cannot be included (because the value of such predictors is constant for a given school) and that the precision of the estimates is decreased because the sample size for each school is only a fraction of the combined sample size.

Partial pooling is implemented by fitting a multilevel model and can be thought of as a compromise between the two extremes of complete pooling and no pooling. Whereas complete pooling ignores variation between schools, the no-pooling analysis overstates it. In other words, the no pooling analysis overfits the data within each school (Gelman and Hill 2006, 253). The following set of equations defines a simple multilevel model:

$$\begin{aligned} y_i &= \alpha_{j[i]} + \beta x_i + \epsilon_i, \quad \text{for students } i = 1, \ldots, n \\ \alpha_j &= a + b u_j + \eta_j, \quad \text{for schools } j = 1, \ldots, J. \end{aligned} \tag{3.1}$$

The first equation describes the student-level regression. A student's grade y_i is regressed on a single student-level predictor x_i, (e.g. the student's average grade in the previous year). β is the regression coefficient corresponding to x_i. The term ϵ_i is the residual for student i; the residuals are assumed to be distributed normally: $\epsilon_i \sim N(0, \sigma_\epsilon^2)$.

$\alpha_{j[i]}$ is the *intercept* in the student-level equation. Unlike the intercept in a classical regression, this intercept is allowed to differ across schools: each school is allowed to have a unique intercept. This is reflected in the intercept's index: $j[i]$ refers to that

school j to which the ith student belongs. The benefits of this manner of indexing will be discussed shortly.

The intercept itself is, furthermore, modeled as the second equation shows. α_j is regressed on the school-level predictor u_j (e.g. an indicator of whether the school is a private institution). b is the corresponding regression coefficient, while a is the overall intercept. The school-level regression, too, includes a residual η_j which is assumed to be distributed normally: $\eta_j \sim N(0, \sigma_\eta^2)$.

Unlike the complete pooling regression, where a single residual accounts for deviations from the model prediction, the multilevel model (3.1) involves two residuals. η_j can be thought of as accounting for the deviations in the schools' average grade from the overall school mean. Because the model includes predictors, η_j actually accounts for deviations from the overall mean unaccounted for by the predictors. ϵ_i, on the other hand, accounts for students' deviations from their corresponding school-level mean (again, unaccounted for by the predictors).

Model (3.1) is called a *random intercept model* because the school's influence is modeled by means of the normal distribution. Alternatively, the school's influence on the grade could be modeled with a *fixed effect* by including dummy indicators for schools in a classical (one-level) regression. Note that in this case the school level predictors could not be included in the model because any such predictor would be collinear with the dummy indicators (see Gelman and Hill 2006, 68).

The school effect can therefore be modeled as either fixed or random. The literature recommends using random effects when the grouping factor has a large number of levels that are typically not *repeatable* in the sense that "if the experiment were to be repeated, it would be with different levels of the grouping factor" (Doran et al. 2007). Frequently the levels of such a factor are a sample from a population. In the present example this means that if the study were to be repeated, this would involve drawing a different sample of schools. Other authors suggest using fixed effects when the groups in the data represent all possible groups, and random effects when the population includes groups not in the data (see Gelman and Hill 2006, 246).

In addition to modeling the intercept as random, the regression coefficient corresponding to x_i in Equation (3.1) could be allowed to vary across schools resulting in a so-called *random slope model*. The student level regression would then be expressed as $y_i = \alpha_{j[i]} + \beta_{j[i]} x_i + \epsilon_i$, while the random slope β_j would be further modeled in a similar manner as the random intercept. The inclusion of a random slope further complicates the model somewhat because the values of α_j and β_j could in principle be correlated. In order to model this appropriately, the pairs (α_j, β_j) should be modeled jointly and

assumed to follow a bivariate normal distribution (we will not further consider random slope models, for details see e.g. Snijders and Bosker 1999; Hox 2002; Goldstein 2011).

3.1.1 Practical considerations in multilevel modeling

The indexing used in the example (of the form $j[i]$) is referred to as *nested indexing* and allows for great flexibility in the identification of units, especially when the levels are crossed. This would be the case, e.g., if students were simultaneously classified into schools on the one hand, and home districts on the other. We could then use index $j[i]$ (as previously) to identify the school j to which the ith student belongs, and $k[i]$ to identify the home district k to which the ith student belongs. The subscript i is attached to the lowest level units and uniquely identifies every measurement and random effect (Goldstein 2011).

Another practical issue that worthy of mention is that of *centering*. The intercept in the model is interpreted as the mean value of the response variable when the values of all the predictors in the model are zero. The value of the intercept is thus meaningless if there are variables among the model predictors that never have values close to zero. If the students' IQ was used as a predictor of the grade in the above example, then the value of the intercept would refer to the average grade when the value of the student's IQ is zero. Since there are no such students, the value of the intercept is essentially meaningless.

The interpretation can be simplified if we *center* the predictor by subtracting the mean IQ from each student's IQ before entering it into the model. The value of the intercept then has the interpretation of the mean value of the response variable (the grade) at the average student IQ. Subtracting the mean merely shifts the values of the predictor but preserves the scale. The centering transformation therefore does not alter the value of the regression coefficient for student IQ, nor any other model parameters except the intercept (Gelman and Hill 2006). In models where the intercept can vary across groups, particular interest is typically paid to the intercept estimate and its variance. Centering the predictors is therefore especially recommended in multilevel models to ease the interpretation of those values.

3.1.2 Explained variation in multilevel models

An important statistic in regression models is the squared multiple correlation coefficient R^2 which is interpreted as the proportion of the variation of the response variable explained by the model predictors. The issue of explained variation is more complex

in multilevel models because we have residual variation to contend with at several levels. In models with random slopes the concept of explained variation has no unique definition (Hox 2002, 62).

In random intercepts models, however, calculating the explained variation for a particular level is quite straightforward, but requires an additional baseline model to be fit. Such a baseline model will contain all the random intercepts, while omitting the fixed effects of predictors. Upon including predictors in the model, the variation at each level will typically decrease. The proportion of explained variation can be calculated as:

$$R^2 = \frac{\sigma^2_{\epsilon|b} - \sigma^2_{\epsilon|m}}{\sigma^2_{\epsilon|b}}, \tag{3.2}$$

where $\sigma^2_{\epsilon|b}$ is the variance component at a particular level in the baseline model and $\sigma^2_{\epsilon|m}$ is the residual variation at the same level in the model including predictors.

The proportion of explained variation as defined above corresponds to the *unadjusted* R^2 in regression analysis, and therefore does not penalize the inclusion of additional predictors. We also note that including a predictor at a particular level can *increase* the residual variation at another level, which is unusual from the perspective of classical regression, where adding a predictor can only decrease the residual variation (see Gelman and Hill 2006, 480-481 for an example and an explanation).

We thus conclude our informal overview of multilevel modeling. The next section will build on the material presented here and will provide a more thorough treatment of models that include random intercepts.

3.2 Item response models

Item response theory (IRT) is a general framework for specifying mathematical functions that describe the interactions of persons and items. The family of models developed in IRT has been demonstrated to be useful in the design and evaluation of educational and psychological tests. Traditionally, the aim of item response modeling has been to *measure* individuals and items on hypothesized underlying constructs. The difference between a particular respondent's *ability* and an item's *difficulty* is posited to determine the probability of a correct answer in the IRT model, and so test results can be used to assign numeric values to persons and items on the dimensions of ability and difficulty, respectively.

We intend to apply models that have been developed in the IRT tradition to analyze item nonresponse. Like in the typical IRT setting, our application also pertains to the interplay of persons and items. The response variable, however, does not contain the information on whether or not the *correct* answer was obtained, but whether a *substantive* response was obtained (with item nonresponse as the other alternative). Because we are interested in *explaining* item nonresponse through respondent characteristics and item facets (rather than measuring respondents and items on underlying dimensions), the statistical models for item nonresponse will need to involve fixed effects of predictors. We will thus apply *explanatory* IRT models (De Boeck and Wilson 2004b) to the problem of item nonresponse.

This section will introduce models that were developed in the IRT tradition. The discussion follows the line of reasoning and notation introduced by De Boeck and Wilson (2004b). We will first describe a linear mixed model for a continuous response variable, then show how such a model can be generalized to accommodate a dichotomous response. This framework is the basis for a range of different item response models. We continue by describing a number of various item response models and finish with an account of how residual dependencies can be modeled, along with the issues this involves.

3.2.1 A linear mixed model for continuous data

In this section we will specify a linear mixed model (LMM) that can be used to model a continuous response variable. Because the object of this dissertation is to model dichotomous data, the next section will show how the LMM can be generalized to accommodate such a response variable.

We do not wish to give a general account of linear mixed models, but rather to illustrate how this class of models is used in item response theory. We will therefore use a rather simple LMM, with only a random intercept on the person side and fixed predictors on the item side:

$$Y_{pi} = \sum_{k=0}^{K} \beta_k X_{ik} + \theta_p + \epsilon_{pi}. \tag{3.3}$$

Indices p ($p = 1, \ldots, P$) and i ($i = 1, \ldots, I$) are used to denote persons and items respectively. The left-hand side of (3.3) therefore denotes the measurement on the pth person and ith item.

Index k ($k = 1, \ldots, K$) denotes the model's fixed item-level predictors. X_{ik} is the value of predictor k for item i. In the simplest case, items are represented by dummy indicators, in which case $X_{ik} = 1$ if $i = k$, and $X_{ik} = 0$ if $i \neq k$. β_k is the fixed regression coefficient corresponding to X_{ik}.

There are two random components in the model: θ_p denotes the random intercept corresponding to person p, while ϵ_{pi} is the error term for person p and item i. Both are assumed to be distributed normally, i.e., $\theta_p \sim N(0, \sigma_\theta^2)$. The error terms are, furthermore, assumed to be independently and identically distributed, i.e., $\epsilon_{pi} \sim N(\mathbf{0}, \mathbf{\Omega})$, where $\mathbf{0}$ is a vector of zeroes and $\mathbf{\Omega}$ is a diagonal matrix with the same value, σ_ϵ^2, on all diagonal elements.

The indexing in (3.3) assumes that the responses for the same set of items are available for each person. For now, we will continue to work under this assumption, as it keeps the notation simple, but note that we need to switch to nested indexing (see Section 3.1.1) if some items are not administered to certain persons.

3.2.2 Application to dichotomous data

We specified the model (3.3) for a continuous response variable. This section will present a heuristic argument that illustrates how to extend the linear mixed model to dichotomous data (see Lord and Novick 1968 or Thissen and Orlando 2001 for a full discussion). In achievement testing, where item response theory has been traditionally applied, the probability of a *correct* response is modeled. Because the same models can be applied in other settings, the term *1-response* is used the following paragraphs to allow for more generality.

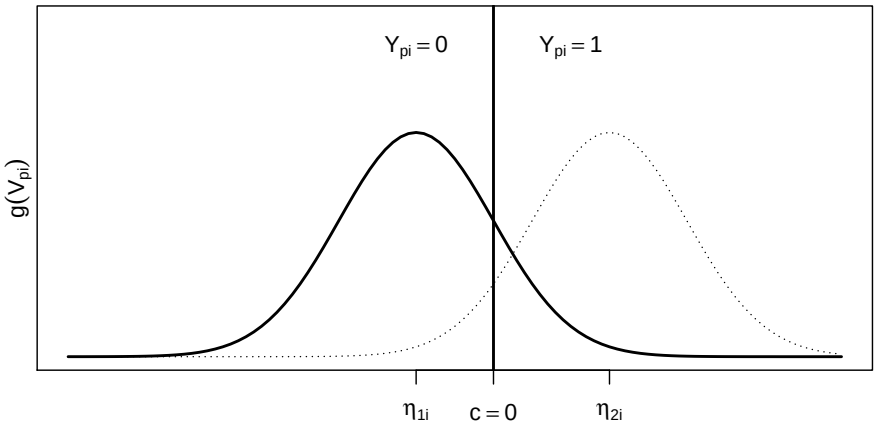
Let us suppose the binary data Y_{pi} stemmed from the dichotomization of an underlying continuous variable which we will denote as V_{pi}. The model (3.3) could then be applied

to this underlying V_{pi}. Such a model implies that the probability of a 1-response ($Y_{pi} = 1$) can be derived from the distribution of V_{pi} in some way. We will use π_{pi} to denote the probability that $Y_{pi} = 1$.

Assuming that V_{pi} follows (3.3) implies the assumption that V_{pi} is distributed normally (De Boeck and Wilson 2004c, 28). Let us denote with η_{pi} the expected value of V_{pi}. The variance σ_ϵ^2 is the same for all pairs of (p, i). Dichotomization is implemented through the use of a cut-off value c, which is defined so that $Y_{pi} = 1$ if $V_{pi} > c$ and $Y_{pi} = 0$ otherwise. The probability that $Y_{pi} = 1$ therefore equals the probability that V_{pi} exceeds c.

In order to determine the probability that $Y_{pi} = 1$, the values of σ_ϵ^2 and c need to be specified. However, the choice of unit and origin on which the V_{pi} vary has no consequence: the value of π_{pi} is invariant under linear transformations of the V-scale (De Boeck and Wilson 2004c, 28). We can therefore choose $\sigma_\epsilon^2 = 1$ and $c = 0$ for simplicity. Figure 3.1 is a graphical representation of the dichotomization for item i paired with two persons. The area under the curve to the right of the horizontal line corresponds to the probability that $Y_{pi} = 1$. This probability is higher for person 2, whose distribution is shifted toward the right (denoted by the dotted line) in comparison to person 1.

Figure 3.1: Distribution of V_{pi}; the vertical line represents the cut-off point used for dichotomization.



Because we chose $\sigma_\epsilon^2 = 1$, the cut-off value $c = 0$ corresponds to a value of $-\eta_{pi}$ under the standard normal distribution (i.e. $(0 - \eta_{pi})/1$). Under this distribution, the cumulative probability of $-\eta_{pi}$ is $(1 - \pi_{pi})$, so that under the same distribution the cumulative probability of η_{pi} is π_{pi} (De Boeck and Wilson 2004c, 28). In other words, the function that maps η_{pi} onto π_{pi} is the cumulative normal distribution function, also referred to as the *normal-give* function. By applying the inverse function, we can map π_{pi} onto η_{pi}, thus obtaining the expected value of the hypothetical underlying V_{pi}.

This inverse function is referred to as the probit function $\eta_{pi} = f_{\text{probit}}(\pi_{pi})$.

It follows from the described dichotomization procedure and the independence of the error terms σ_ϵ^2 that Y_{pi} has an independent Bernoulli distribution with mean π_{pi} and a variance of $\pi_{pi}(1 - \pi_{pi})$. The equation (3.3) without the error term is a model for η_{pi}, and thus for $f_{\text{probit}}(\pi_{pi})$:

$$\eta_{pi} = \sum_{k=0}^{K} \beta_k X_{ik} + \theta_p, \tag{3.4}$$

where $\theta_p \sim N(0, \sigma_\theta^2)$. Thus, $\pi_{pi} = f_{\text{probit}}^{-1}(\sum_{k=0}^{K} \beta_k X_{ik} + \theta_p)$. π_{pi} and η_{pi} also depend on the value of θ_p, but this dependence is not explicitly shown due to convention (De Boeck and Wilson 2004c, 30). We have thus far described the *normal-ogive random-intercepts model*, which belongs to the family of *generalized* linear mixed models (GLMMs). The model has three components:

1. The *distributional* or *random component* connects Y_{pi} to π_{pi}. Formally, $Y_{pi} \sim$ Bernoulli(π_{pi}), and all Y_{pi}s are independent. To reiterate, π_{pi} is the probability of $Y_{pi} = 1$ given θ_p (the value of the random intercept for the person in question).

2. The component that links the expected value of the binary observations to the underlying continuous variable, η_{pi}, is the *link function*, in this case the probit function*probit function.*

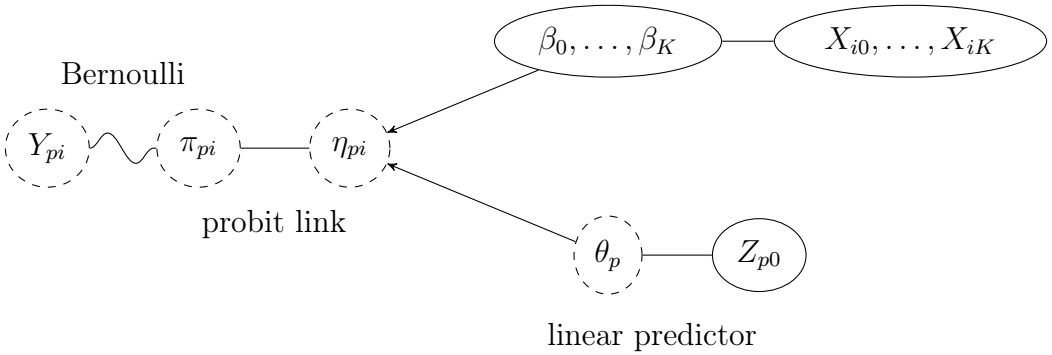
3. Finally, the component that links the expected value of the underlying continuous variable η_{pi} to the predictors is the *systematic component*, given by Equation 3.4. The function value η_{pi} is often referred to as the *linear predictor*.

The model is graphically represented in Figure 3.2. The arrows represent the linear effects of the predictors. Random quantities appear in dashed ellipses while solid ellipses are used for fixed quantities in the model. The ellipse with Z_{p0} in the figure represents the value of 1 that is, technically speaking, multiplied with the random intercept of person p. The squiggly line represents the Bernoulli distribution.

GLMMs as item response models

An alternative to the probit link is the *logit link*: $\eta_{pi} = f_{\text{logit}}(\pi_{pi})$. In fact, the logit link is used more often in item response models. The logit function is defined as the natural logarithm of the probability of a 1-response divided by the probability of a

Figure 3.2: Graphical representation of the normal-ogive random-intercepts model



Source: De Boeck and Wilson (2004c, 30)

0-response:

$$\eta_{pi} = \log(\pi_{pi}/(1 - \pi_{pi})). \tag{3.5}$$

The logistic model is no longer based on an underlying *normally* distributed error term ϵ_{pi}, but rather on a *logistic error term*. Its distribution has a larger variance and somewhat heavier tails than the standard normal distribution (De Boeck and Wilson 2004c, 32). It turns out, however, that the logit function multiplied by 1.7 is a good approximation of the probit link: $1.7 f_{\text{logit}}(\pi_{pi}) \approx f_{\text{probit}}(\pi_{pi})$. This means that models employing the probit and logit link produce very similar estimates and predictions, the only difference being that parameter estimates under a logistic model are 1.7 times higher than in a corresponding probit model. The popularity of the logit link stems from its simple mathematical form and the fact that it is the canonical link for the binomial distribution (see e.g. Olsson 2002, 40-42 for a definition of the canonical link and examples).

The normal-ogive and logistic random intercepts models presented above are two well known item response models. When the item predictors are dummy variables identifying the items ($X_{ik} = 1$ if $i = k$, and $X_{ik} = 0$ if $i \neq k$), then the fixed effects ($\sum_{k=0}^{K} \beta_k X_{ik}$ in Equation (3.4)) correspond to the *item parameter*, usually denoted simply as β_i, and interpreted as *item difficulty*. The person parameter θ_p, on the other hand, is interpreted as the *ability* of person p. The convention in IRT literature is to use a negative sign for β_i:

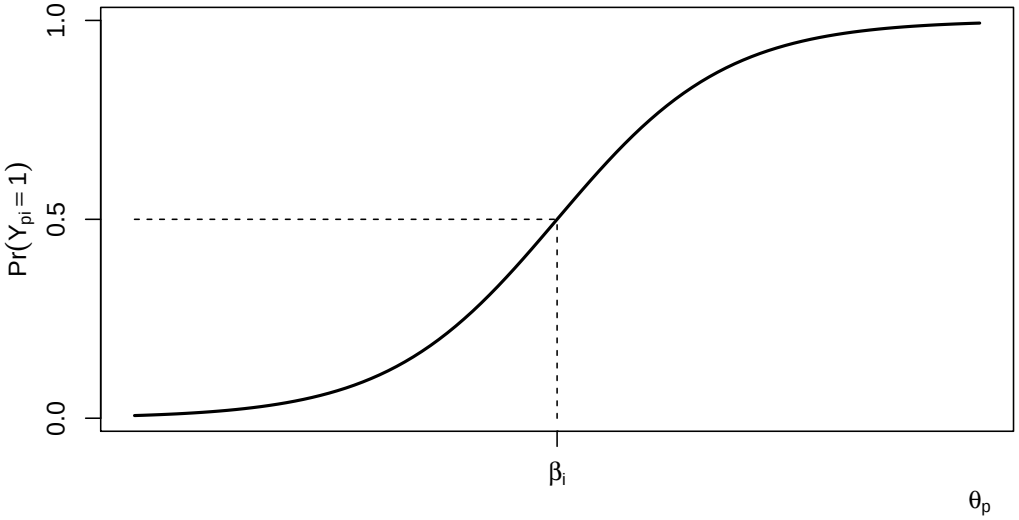
$$\eta_{pi} = \theta_p - \beta_i. \tag{3.6}$$

The probability of obtaining a 1-response, π_{pi}, thus depends on the difference between the person's ability θ_p and the item's difficulty β_i. In the case when $K = I$ item dummies are used to identify the items, the constant predictor must be omitted so

that the model is identifiable.

Figure 3.3 illustrates how the interplay of θ_p and β_i determines the probability of a 1-response, π_{pi}. The curve in the figure corresponds to an item with difficulty β_i. If the person's ability, θ_p, is equal to the item's difficulty, β_i, then, according to (3.6), the value of η_{pi} is zero. This value is then transformed by the inverse logit function to obtain $\pi_{pi} = f_{\text{logit}}^{-1}(0) = 0.5$. If the person's ability is higher than the item's difficulty, the probability of a 1-response is higher than 0.5. The function that maps θ_p onto π_{pi} given β_i called the *item characteristic curve* or *item response function*. The value of β_i locates the curve (De Boeck and Wilson 2004c, 34).

Figure 3.3: Item response function for the logistic random-intercept model



Source: De Boeck and Wilson (2004c, 34)

Interpretational issues with GLMMs

In a linear mixed model, the fixed effects can also be interpreted marginally, meaning that they apply to the population average and each individual separately (for a discussion on marginal, random-effects, and conditional models see Molenberghs and Verbeke 2004). This important and useful property of LMMs does not, however, carry over to *generalized* linear mixed models. Nonetheless, even though the fixed effects in a GLMM have no marginal interpretation, they do show a strong correlation to their marginal counterparts (see Molenberghs and Verbeke 2004, 126 for a more thorough discussion).

Another noteworthy issue pertains to the scaling of normal-ogive and logistic models. The fixed effects and the variance of the random effects in such models are expressed

relative to σ_ϵ, the standard deviation of the error term for the underlying variable. Due to the hypothetical nature of this variable, the value of σ_ϵ is fixed to a value determined by convention. The consequence of this relative way of expressing effects is that the scale is reduced when a source of variation is not included in the model. Non-included effects become part of the error term and since the value of σ_ϵ is fixed by convention, other model effects will be expressed through reduced values. This happens, e.g., to fixed effects if random intercepts are omitted from the model (De Boeck and Wilson 2004c, 33).

3.2.3 Descriptive and explanatory item response models

According to De Boeck and Wilson (2004b), two philosophical orientations regarding measurement and statistical modeling converge in item response modeling. They label the first philosophical orientation *explanatory analysis*. Its principal aim involves explaining the dependent variable in terms of the predictors under consideration.

The other philosophical orientation is *descriptive measurement*, whose aim is measuring individuals (and, consequently, items) on one or more constructs, which may or may not be theoretically derived. The purpose of using these measures is often descriptive and aims to assign numeric values to persons (and items). Explanation of these values is only considered in the second step, if at all (De Boeck and Wilson 2004c, 37).

Under the perspective of explanatory analysis, the analyst might prefer to ignore the individual differences that are the target of descriptive measurement, and historically this has been common. In the context of descriptive measurement, however, measuring individual differences is the prime objective, without a necessary interest in systematic effects to explain the observations. The authors claim that, although the two philosophical orientations seem to be in conflict, they can in fact be combined in what they term *explanatory measurement*. A common core of statistical models (GLMMs) can be used under either of these perspectives, as well as under their combination (De Boeck and Wilson 2004c, 38).

We will now describe a number of item response models, slowly proceeding from simpler models to more complex ones. For each model, we will provide a formula and a graphical representation. We will describe the newly introduced features of each model and note which features of the model are considered explanatory or descriptive.

The Rasch model

The Rasch model (Rasch 1960) has actually already been introduced in Equation (3.6). Taking into account that $\eta_{pi} = \log(\pi_{ip}/(1-\pi_{ip}))$, we can obtain the *exponential form* of the Rasch model by exponentiating both sides of (3.6):

$$\frac{\pi_{pi}}{1-\pi_{pi}} = \frac{\exp(\theta_p)}{\exp(\beta_i)}. \tag{3.7}$$

This form of the model lends itself more readily to interpretation. The intuition reflected in Equation (3.7) is that ability (θ_p) facilitates success, whereas difficulty (β_i) causes one to fail. The ratio of these (exponentiated) quantities determines the odds of success (De Boeck and Wilson 2004a). De Boeck and Wilson invoke the metaphor of the hurdler (the person) and a series of hurdles (the items) to be overcome. The hurdler is seen as having the ability (indicated by θ_p) to leap over hurdles of a certain height, while the series of hurdles has heights indicated by the corresponding item difficulties ($\beta_1, \ldots, \beta_I$). When the hurdler's ability is equal to the height of the hurdle, the probability that the leap will be successful is 0.5. When the hurdler's ability is higher than the hurdle's height, the probability of a successful leap will be greater than 0.5 and vice versa.

The person parameter varies *randomly* over persons: the persons in the sample are regarded as exchangeable. The person parameter provides a measure of a latent variable such as ability, achievement level, skill, cognitive process, cognitive strategy, developmental stage, motivation, attitude, personality trait, state, emotional state, or inclination (De Boeck and Wilson 2004a, 43).

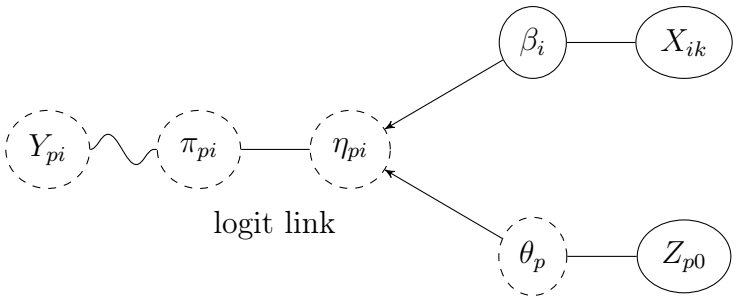
We have mentioned above that the person parameters are assumed to be distributed normally around zero: $\theta_p \sim N(0, \sigma_\theta^2)$. If the mean of this distribution were not constrained to zero, the presented model equations would have an identification problem. If we add the same constant to both the person and the item parameters in (3.6), the probability of a 1-response does not change. In other words, the person and item parameters are not identified separately, but only one in relation to the other. Alternative solutions exist to solve the identification problem (such as fixing a particular β_i or the mean of the β_is to zero), but the most common solution when the person parameter is modeled as random is to fix the mean of the person parameters to zero (De Boeck and Wilson 2004a, 53).

An important feature of the Rasch model and other models still to be presented (except for models for residual dependencies of Section 3.2.4) is the co-called *local in-*

dependence assumption. This assumption implies that, for a given response vector $\mathbf{y_p} = (y_{p1}, \ldots, y_{pI})'$, the conditional probability of the whole vector is the product of the conditional probabilities for each response. Under this assumption, θ_p is the only source of dependence between items. Therefore, for a given value of θ_p the observations are assumed to be independent (De Boeck and Wilson 2004a, 52).

Figure 3.4 is a graphical representation of the Rasch model. The item side of the linear predictor is slightly simpler in comparison to Figure 3.2, as the item effect β_i is multiplied only with the corresponding dummy indicator X_{ik}. The constant predictor that is multiplied with the random person effect is again shown as Z_{p0}. We note that the link function in the Rasch model is the logit link and that models employing the probit link are not referred to as a Rasch models in item response theory literature.

Figure 3.4: Graphical representation of the Rasch model



Source: De Boeck and Wilson (2004a, 52)

The Rasch model is characterized by De Boeck and Wilson (2004a) as "doubly descriptive", meaning that upon fitting the model, the parameter estimates (the θ_ps and β_is) allow only for descriptive measurement on the person as well as the item side.

The latent regression Rasch model

This is not so for the latent regression Rasch model, which includes an explanatory component on the person side. The equation for this model differs from the Rasch model (3.6), in that θ_p is replaced with a linear regression equation:

$$\theta_{pi} = \sum_{j=1}^{J} \vartheta_j Z_{pj} + \epsilon_p, \tag{3.8}$$

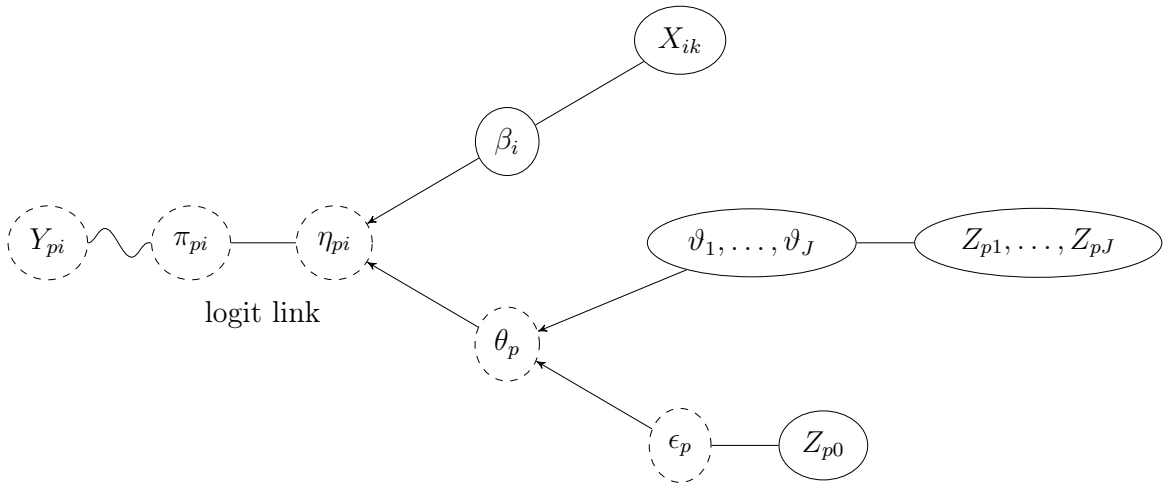
yielding:

$$\eta_{pi} = \sum_{j=1}^{J} \vartheta_j Z_{pj} + \epsilon_p - \beta_i. \tag{3.9}$$

Index j ($j = 1, \ldots, J$) in the above equation is used to refer to measured person characteristics and Z_{pj} is the value of person p on the jth characteristic. ϑ_j is the fixed regression coefficient for person characteristic j.

Figure 3.5 is a graphical representation of the latent regression Rasch model. The difference between it and Figure 3.4 is that the person parameter θ_p is (partially) explained in terms of the external person characteristics (denoted Z_{pj}) and their effects (ϑ_js). The unexplained part or the error term is the random effect of the constant predictor (De Boeck and Wilson 2004a, 59).

Figure 3.5: Graphical representation of the latent regression Rasch model



Source: De Boeck and Wilson (2004a, 59)

Model 3.9 is referred to as the latent regression Rasch model, as the latent person variable θ_p can be thought of as being regressed on external person characteristics (De Boeck and Wilson 2004a, 58). The person characteristics are considered as variables with fixed values, meaning that the fact that they may include measurement error is ignored (like in classical regression models).

The linear logistic test model

In the linear logistic test model (LLTM), item facets are used to explain the differences between items in terms of the effect they have on η_{ip}. The model is therefore considered explanatory on the item side and descriptive on the person side (De Boeck and Wilson 2004a, 62). We obtain the LLTM equation by substituting the term β_i with a linear function:

$$\beta_i' = \sum_{k=0}^{K} \beta_k X_{ik} \tag{3.10}$$

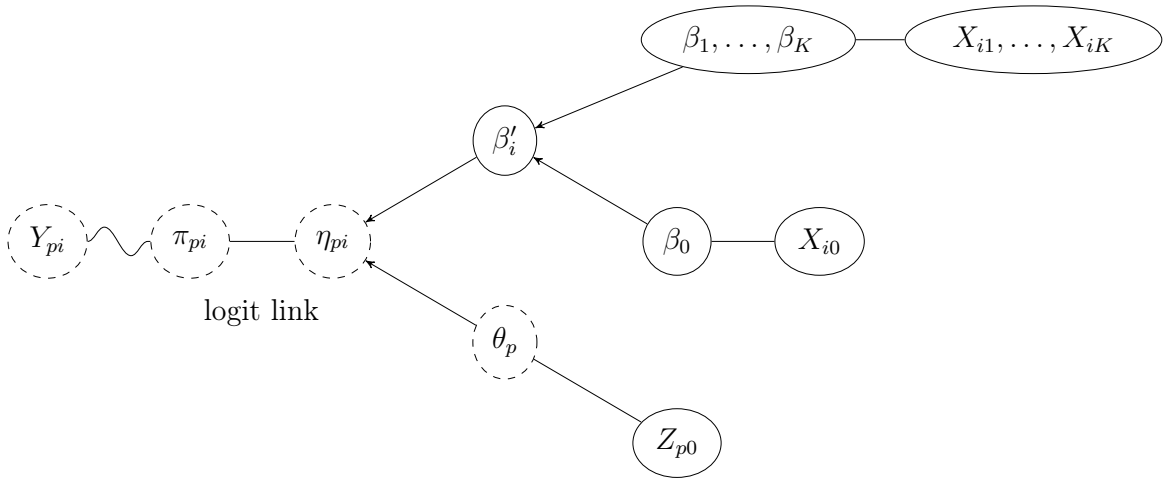
in the Rasch model equation (3.6), yielding:

$$\eta_{pi} = \theta_p - \sum_{k=0}^{K} \beta_k X_{ik}. \tag{3.11}$$

As there is no error term in Equation (3.11), this substitution implies that the item effects can be explained perfectly by the item facets: that β_i from the Rasch model equals β_i'. This is a strong assumption that makes the model highly restrictive (De Boeck and Wilson 2004a, 63).

Because the mean of the person effects is fixed to zero, we need to include the overall intercept by including an item facet with value 1 for all items (X_{i0}). This is accentuated in Figure 3.6, where β_0 (the overall intercept) and X_{i0} (the constant predictor) appear as a separate branch. It is less apparent in Equation (3.11), where we wish to stress that the summation starts at $k = 0$ (thus including the constant predictor) and not $k = 1$.

Figure 3.6: Graphical representation of the latent trait test model



Source: De Boeck and Wilson (2004a, 63)

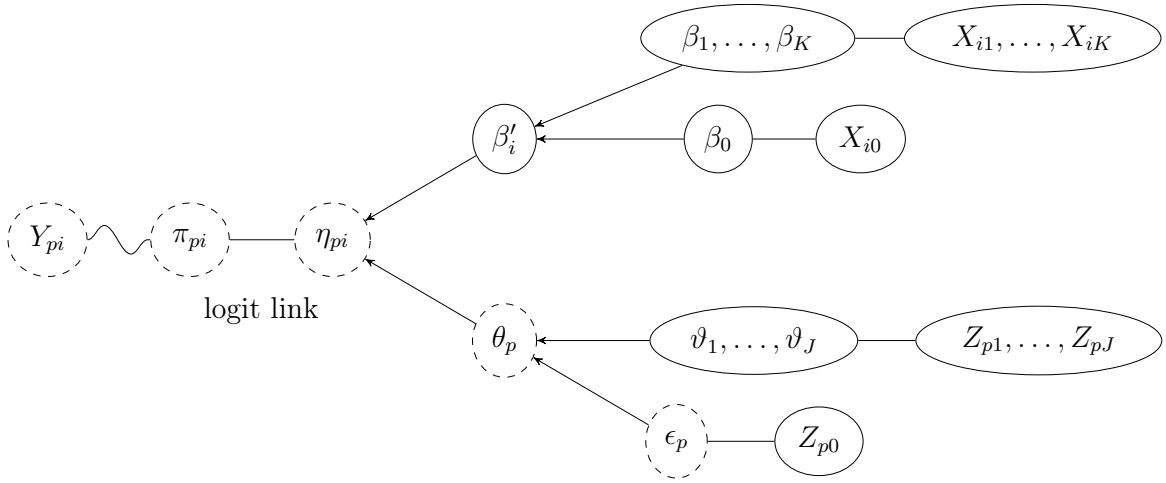
The latent regression LLTM

Both described extensions can be simultaneously applied to the Rasch model to obtain the *latent regression LLTM*, a model considered "doubly explanatory" by (De Boeck and Wilson 2004a, 66), as it involves an explanatory component both for the persons and the item side:

$$\eta_{pi} = \sum_{j=1}^{J} \vartheta_j Z_{pj} + \epsilon_p - \sum_{k=0}^{K} \beta_k X_{ik}. \tag{3.12}$$

Model 3.12 is a generalized linear mixed model with person predictors as well as item predictors, each having a fixed effect, and a random intercept, which is the error term for the person contribution (De Boeck and Wilson 2004a, 66). Like the LTMM, the latent regression LTMM does not include an error term on the item side and thus assumes that the item effects can be wholly explained by the included item-level predictors. The model structure is graphically depicted in Figure 3.7.

Figure 3.7: Graphical representation of the latent regression LLTM



Source: De Boeck and Wilson (2004a, 67)

The LLTM with an error term for items

Because the linear logistic test model assumes that item difficulty is equal to a linear combination of item-level predictors (see Equation (3.10)), this model's goodness of fit is almost invariably worse than that of a corresponding Rasch model (Janssen et al. 2004). We will now relax this assumption by adding an error term on the item side:

$$\begin{aligned} \beta_i &= \sum_{k=0}^{K} \beta_k X_{ik} + \epsilon_i, \\ &= \beta_i' + \epsilon_i, \end{aligned} \tag{3.13}$$

where we assume the error terms to follow the normal distribution $\epsilon_i \sim N(0, \sigma_\epsilon^2)$.

Modeling item effects as random is relatively uncommon in item response modeling. Equation (3.13) breaks down the item effect into a *structural* component, described

by the linear combination of item predictors, and an *item-specific deviation* part, described by the ϵ_i. Analogously to linear regression, the variance σ_ϵ^2 refers to the residual variance in the regression of the β_i on the item predictors X_{ik}. A higher σ_ϵ^2 in comparison to the total variance of the β_i means that the explanatory power of the item predictors is higher (Janssen et al. 2004, 191).

An alternative interpretation is that the item parameters β_i are randomly sampled. Under this interpretation, items that share the same values on their item predictors belong to the same population. Given these values, the items are considered exchangeable. Since the items are seen as a random sample from the population, the individual item's difficulty β_i may differ from β_i', which is the expected difficulty under the model (Janssen et al. 2004, 191).

Modeling both item and person parameters as random leads to a so-called *crossed random effects model*. Substituting Equation (3.13) into the Rasch model equation yields:

$$\eta_{pi} = \theta_p - \sum_{k=0}^{K} \beta_k X_{ik} - \epsilon_i, \tag{3.14}$$

where both the random person intercept and the item residual are assumed to follow the normal distribution: $\theta_p \sim N(0, \sigma_\theta^2)$ and $\epsilon_i \sim N(0, \sigma_\epsilon^2)$. We note that in the above model the linear predictor η_{pi} is conditional on both θ_p and ϵ_i.

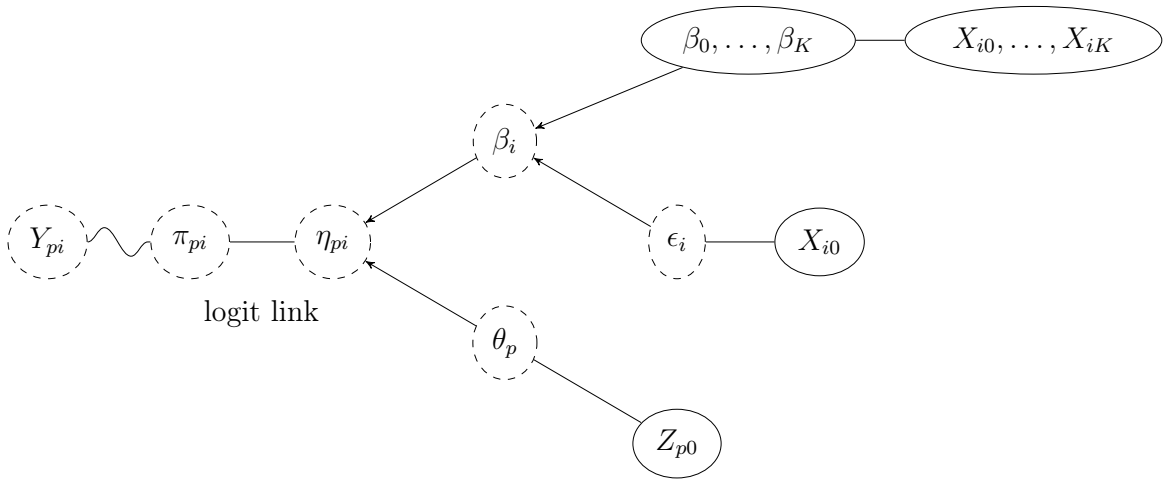
Figure 3.8 is a graphical representation of the model. The model includes an overall intercept, but in the figure, this fact is not stressed by putting the intercept on a separate branch. Rather, the overall intercept appears as β_0 in the linear combination of the item predictors and their corresponding regression coefficients.

Model with person-by-item predictors

The models we have considered so far have all separated the effect of items from the effect of persons, never considering a combined effect of the two. Put more technically, the models have not included person-by-item predictors, thereby assuming that the effect of a particular item is the same for all persons (De Boeck and Wilson 2004a, 46).

A person-by-item predictor is derived as the product of an item indicator and a person predictor, indicating group membership. In the traditional setting of achievement testing such an indicator is usually included in a model in order to investigate *differential item functioning*. Such studies are usually concerned with the question whether a particular item is "fair" for members of a particular focal group, compared to members

Figure 3.8: Graphical representation of the model with item properties, random item effects, and random person effects



Source: Janssen et al. (2004, 195)

of a reference group (Meulders and Xie 2004). A related concept is that of differential facet functioning, which concerns differential effects of item properties.

We will index the person-by-item predictors with index h ($h = 1, \ldots, H$). The notation W_{pih} denotes the value on the hth such predictor, pertaining to person p and item i. The corresponding regression coefficients are denoted as δ_h in the model:

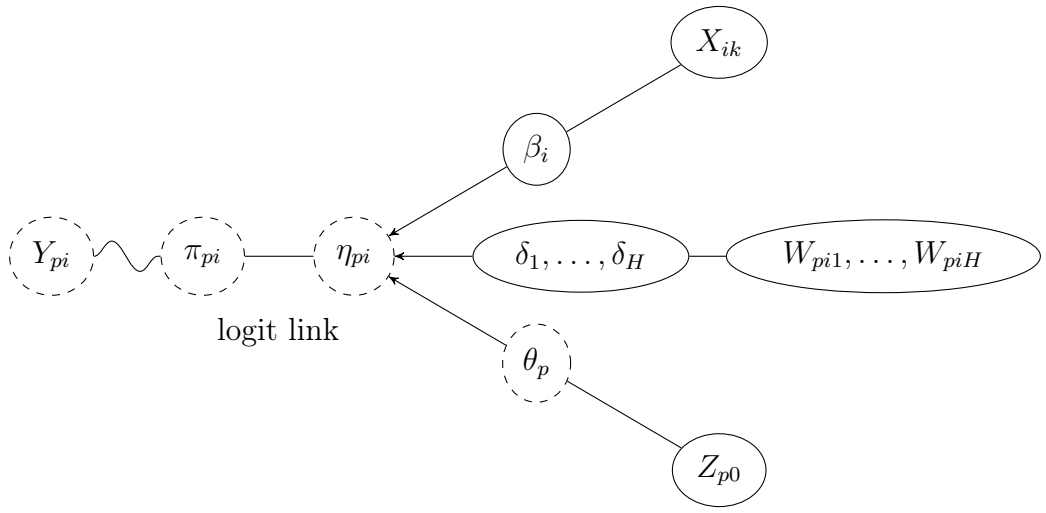
$$\eta_{pi} = \theta_p - \beta_i + \sum_{h=1}^{H} \delta_h W_{pih}. \tag{3.15}$$

Equation (3.15) shows how person-by-item predictors can be added to the Rasch model. Figure 3.9 is the corresponding graphical representation. We will discuss the construction of the person-by-item predictors in Section 4.3 by applying this notation to the concrete data that we wish to model in the empirical part of this dissertation.

3.2.4 Modeling residual dependencies

The models we have described so far require that, conditional on the random effects in the model, the responses to the different items be independent. We have already mentioned that this requirement is called the conditional (or local) independence assumption. In many applications, however, not all dependence between the responses can be explained by the random effects assumed to underlie the responses. Examples of this are abundant in the achievement test domain: if a test consists of several paragraphs of text followed by several questions, the data from each such item bundle may show more dependencies than can be accounted for by a single reading

Figure 3.9: Graphical representation of the Rasch model, extended to include person-by-item predictors



Source: Meulders and Xie (2004, 215)

ability dimension. Another example when the conditional assumption is violated is when outcome-dependent learning takes place while taking the test (Tuerlinckx and De Boeck 2004, 289).

Residual dependencies therefore often occur as a result if an additional organizing principle is present in the data that the model does not account for. We have already mentioned the example of a bundle of items that are all similar in some way. Another organizing principle can be that the items be ordered along the *time* dimension. We will limit our discussion to this example from now on.

When an *a priori* ordering of items exists, the probability of a 1-response to a particular item can in principle depend on the responses to the preceding items. Such models, where no feedback loops are allowed (a response to one item cannot affect the response to a later item) are called *recursive* models (Tuerlinckx and De Boeck 2004, 295). In recursive models, the response to a particular item is modeled by including a certain function of the responses to preceding items as a predictor.

This leads to certain interpretational difficulties, though, as we now illustrate. Let us assume that the probability of a response to a certain item is influenced only by the response to the item immediately preceding the current item. The probability of person p for a response y_{p1} (y_{p1} can assume the value of 0 or 1) on the first administered item is, of course, unaffected by this and can be calculated, in the simplest case, according

to the Rasch model:

$$\Pr(Y_{p1} = y_{p1}) = \frac{\exp(y_{p1}(\theta_p - \beta_1))}{1 + \exp(y_{p1}(\theta_p - \beta_1))}. \tag{3.16}$$

The probability for the response y_{p2} on the second item given the response on the first item then equals:

$$\Pr(Y_{p2} = y_{p2}|y_{p1}) = \frac{\exp(y_{p2}(\theta_p - \beta_2 + y_{p1}\delta))}{1 + \exp(y_{p2}(\theta_p - \beta_2 + y_{p1}\delta))}, \tag{3.17}$$

where the parameter δ captures the dependence on the previous item. If $\delta > 1$, the probability of a 1-response to the second item is higher if the response to the first item was also a 1-response. In the achievement test domain the parameter δ is therefore interpreted as the *learning parameter*. If $\delta = 0$, there is no learning and the model simplifies to the Rasch model (Tuerlinckx and De Boeck 2004, 298).

The interpretational difficulty that arises with the inclusion of such a parameter concerns the item parameters. In the Rasch model, the functional form of a particular item's characteristic curve is a logistic function and the parameter β_i has the natural and simple interpretation of the difficulty of the item. This follows because β_i marks the point where the probability of a 1-response is 0.5 for a person with $\theta_p = \beta_i$. With the inclusion of the δ parameter in recursive models this property no longer holds (except for the first item). The probability of a 1-response on the second item that we obtain by substituting (3.16) into (3.17) is (Tuerlinckx and De Boeck 2004, 299):

$$\begin{aligned}
\Pr(Y_{p2} = 1) &= \sum_{m=0}^{1} \Pr(Y_{p1} = m, Y_{p2} = 1) \\
&= \sum_{m=0}^{1} \Pr(Y_{p2} = 1|Y_{p1} = m) \Pr(Y_{p1} = m) \\
&= \frac{\exp(\theta_p - \beta_2)}{(1 + \exp(\theta_p - \beta_2))} \frac{1}{(1 + \exp(\theta_p - \beta_1))} + \\
&\quad \frac{\exp(\theta_p - \beta_2 + \delta)}{(1 + \exp(\theta_p - \beta_2 + \delta))} \frac{\theta_p - \beta_1}{(1 + \exp(\theta_p - \beta_1))}
\end{aligned} \tag{3.18}$$

Two noteworthy observations can be made from Equation (3.18). First, the marginal probability does not have a logistic form and its shape depends on the dependence structure of the model, therefore on parameters δ and β_1 rather than just on β_2. Second, the parameter β_2 no longer has the interpretation of marking the point on the latent scale where the probability of a 1-response is 0.5. Different values of δ lead to different locations of the scale where $\Pr(Y_{p2} = 1) = 0.5$. When $\delta \neq 0$, the item

parameters cannot be interpreted as item difficulties (Tuerlinckx and De Boeck 2004, 299).

This concludes our account of models developed in the IRT framework. As mentioned, a model akin to those described in this section can be used to model item nonresponse in the survey data. We will discuss this application in the last section of this chapter. We now turn our discussion to survival analysis methods, which will be used to analyze survey breakoff.

3.3 Survival analysis

Survival analysis is a collection of statistical methods for which the response variable is the time until a specified *event* occurs. The term survival analysis comes from biomedical research, where the interest lies in studying mortality or patients' survival times from the diagnosis of a particular disease to death. Classical examples of *events* (sometimes also referred to as *failures*) are death, heart attack, divorce, birth of first child, etc. Survival analysis methods bear a different name in other disciplines: *event history analysis* in sociology, *duration* or *transition analysis* in economics, and *failure-time analysis* in engineering (Guo 2010).

Survey breakoff is a terminal event in the context of the survey interview. Like with patients in a clinical study, we are interested in how long a respondent perseveres ("survives") in the interview, and which respondent characteristics are related to the risk of breakoff. Because the research problem, when expressed this way, resembles the study of patients' survival times, survey methodologists have applied survival analysis methods to study survey breakoff (Galesic 2006; Peytchev 2009; Matzat et al. 2009).

Due to the fact that survival data are typically *censored*—a problem we turn to in the the first section—special methods are required to analyze such data. We proceed slowly in our treatment of survival analysis methods, first giving an account of basic concepts and descriptive methods. We then proceed to describe the Cox proportional hazards model and its characteristics. In order to demonstrate how survival analysis is typically applied, we use a survival dataset from a clinical study of leukemia patients to illustrate the methods. We describe how the Cox model will be applied to the survey breakoff data in Section 3.4.2.

3.3.1 Censoring

Superficially, one might think that time-to-event data are merely measurements on a scale, and thus expect to be able to employ the multitude of well developed statistical methods for analyzing continuous data. This will not work, however, due to *censoring*, which is the fundamental problem that survival analysis methods were developed to address. The point is that, when studying an event of interest, we have to actually wait for the events to take place. As our study ends, we will almost invariably find that the event in question has occurred for some individuals in the study but not for others (Aalen et al. 2008). Some people will not have died (or had a heart attack, divorced, had their first child, etc.) during our period of observation.

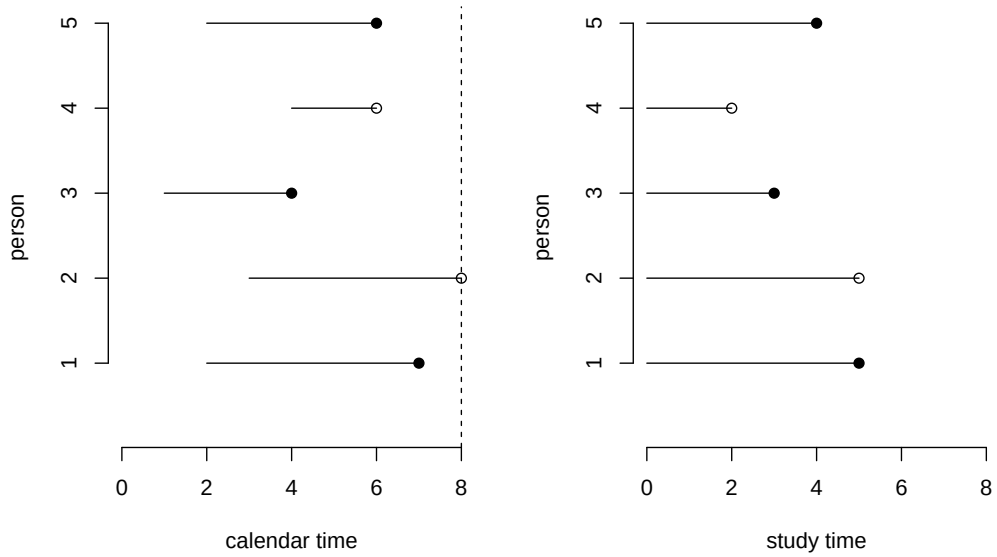
Censoring, broadly speaking, occurs when some lifetimes are known to have occurred only within certain intervals. There are various categories of censoring, but we will only mention two types. For a fuller discussion see e.g. Klein and Moeschberger (2003, 63-78).

Right-censoring occurs when the ending point for the interval is not known. This usually happens when 1) a person does not experience the event before the study ends, 2) the person is lost to follow-up during the study period, 3) or the person withdraws from the study for some other reason (a competing risk). Right-censoring is the most common type of censoring with survival data (Kleinbaum and Klein 2005, 6-7).

Left-censoring occurs when the origin of the interval is not known. Klein and Moeschberger (2003) give the example of a study that aims to determine the age at which a child learns to accomplish a particular task. "Often, some children can already perform the task when they start in the study. Such event times are considered left censored" (Klein and Moeschberger 2003, 71).

We provide a simple fictitious example of right-censoring in Figure 3.10. We denote the start of the study period with calendar time zero (see left panel). Person 1 enters the study at time 2 and experiences the event at time 7. Person 2 enters the study at time 3 does not experience the event before the study period ends at time 8. Persons 3 and 5 experience the event during the study period, while person 4 is lost to follow-up at time 6.

Figure 3.10: Simple fictitious example of calendar time and study time; filled circle indicates failure; open circle indicates right-censoring.



As indicated by the right-hand panel of Figure 3.10, the study time-scale used in

survival analysis will usually not be calendar time. The appropriate time scale is chosen pragmatically. Time zero is set to the initiating event, which could be the time of diagnosis, time of entry into the study, time of remission, etc. (Aalen et al. 2008, 4).

Note that even though the lines for persons 1 and 2 in the right-hand panel are of the same length, the survival time for person 1 is *exactly* 5, while the survival time for person 2 is *at least* 5. Using the + sign to indicate censoring, the data from the right-hand panel are:

$$5,\ 5+,\ 3,\ 2+,\ 4.$$

This example illustrates why censored survival times cannot be handled by means of conventional statistical methods. Even the simple mean of the survival times cannot be calculated due to the censored observations. If we cannot calculate the mean, then we cannot find the standard deviation, perform a t-test, fit a regression model, or perform almost any other conventional statistical analysis (Aalen et al. 2008, 5).

Though the Cox PH model that we describe in the following sections can handle censored data, it requires the key assumption that censoring is *non-informative*. This assumption implies that the censoring mechanism is under the researcher's control and out of the study subject's control. While this is clearly the case with persons who are censored because they have not experienced the event by the end of the study period (participating persons have no control over when the study will end), the assumption can be violated in case persons drop out of the study.

Guo 2010 gives an example of a study on the relapsing of alcoholics, in which several subjects dropped out. If a subject dropped out of the study because they, moved to another city, then they are considered to be a case of noninformative censoring. If the subject, however, dropped out because they started drinking again and stopped notifying the researchers of their whereabouts, then this is considered to be a case of informative censoring, since the censoring mechanism is under the study subject's control.

3.3.2 Survival function and hazard rate

Even though conventional statistical methods cannot handle censored data, tackling such data is quite straightforward provided with the right concepts. Two basic concepts that will be defined in this section pervade the whole theory of survival analysis (Aalen et al. 2008), namely the *survival function* and the *hazard rate*.

We begin by describing how the response variable is usually denoted. In order to take censoring into account, the response variable in survival analysis typically consists of two pieces of information, which distinguishes it from response variables used in conventional statistical analyses (Guo 2010, 6). The first piece is a variable recording the time, and the second is a dichotomous variable indicating whether the person experienced the event or was censored.

We will denote with capital T the random variable for a person's survival time. We will use lower-case t to denote a specific value for T. δ will be used to denote the event indicator; $\delta = 1$ if the person experienced the event and $\delta = 0$ if the person was censored.

The survival function $S(t)$ gives the expected proportion of persons for which the event has not happened by time t (Aalen et al. 2008). If the event is death, then $S(t)$ gives the probability that the person survives longer than some specified time t (Kleinbaum and Klein 2005):

$$S(t) = P(T > t). \tag{3.19}$$

The survival function is a central concept in survival analysis because obtaining survival probabilities for different values of t provides crucial summary information for survival data (Kleinbaum and Klein 2005). Theoretically, the survival function can be traced as a smooth curve as t ranges from zero to infinity. The shape of the concrete survival function depends on the data, but all survival functions have the following characteristics:

- They are nonincreasing; they have a downward tendency because, as time passes, more and more individuals experience the event.
- At time $t = 0$, $S(t) = 1$. Since no one had experienced the event at the start of the study, the probability of surviving past time $t = 0$ is one.
- Theoretically, if the study period could be increased ad infinitum, we would observe the event for all persons[4] (nobody would survive), therefore $S(\infty) = 0$ (Kleinbaum and Klein 2005, 9).

The survival function (3.19) gives the unconditional probability that a person survives past time t. The *hazard rate*, on the other hand, is defined by means of a conditional

[4]If we, however, apply survival analysis to events that do not necessarily occur to all individuals, like divorce or testicular cancer, the survival function will decrease toward a positive value as t approaches infinity (Aalen et al. 2008).

probability. Assuming that T is continuous, we look at the individuals who have not experienced the event by time t and consider the probability of experiencing the event in the small time interval $[t, t + dt)$. This probability then equals $h(t)dt$. The hazard rate $h(t)$ is defined as a limit (Aalen et al. 2008):

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{3.20}$$

It is important to note that the hazard rate is not a probability. The concrete value of the hazard rate depends on the unit of time used—the hazard will be different if we measure time e.g. in weeks as opposed to days—and may even assume values larger than one (Kleinbaum and Klein 2005). Like the probability, the hazard rate will always be nonnegative, as both the numerator as the denominator in (3.20) are nonnegative.

The hazard rate (3.20) gives "the instantaneous potential per unit time for the event to occur, given that the individual has survived up to time t" (Kleinbaum and Klein 2005, 10). While the survival function focuses on the event *not* occurring, the hazard focuses on the event's occurrence. The two concepts can therefore somehow be considered to give the opposite information. There is, in fact, a clearly defined mathematical relation between the two:

$$S(t) = exp\left[-\int_0^t h(u)du\right] \tag{3.21}$$

and

$$h(t) = -\left[\frac{dS(t)/dt}{S(t)}\right]. \tag{3.22}$$

In the following section, we illustrate the Kaplan-Meier method for estimating the survival function from censored survival data.

3.3.3 The Kaplan-Meier method

In order to make our account of survival analysis methods more concrete, we will illustrate the use of the methods for a typical survival dataset, which we will refer to as the remission data (Freireich et al. 1963). The dataset includes the results of a clinical trial of a drug (6-mercaptopurine) vs. a placebo in 42 leukemia patients. The selected patients' cancer went into remission, and the subjects were followed until their

leukemia returned (the event of interest), or until the end of the study. In addition to treatment status (treatment or placebo), we have the information on each patient's sex (0 - female, 1 - male) and the logarithm of the white blood cell count, which is a known indicator of survival for leukemia patients. The data are given in Table 3.1.

Table 3.1: The remission data (Freireich et al. 1963)

placebo group				treatment group			
time	δ	sex	logWBC	time	δ	sex	logWBC
1	1	1	2.80	6	0	0	3.20
1	1	1	5.00	6	1	0	2.31
2	1	1	4.91	6	1	1	4.06
2	1	1	4.48	6	1	0	3.28
3	1	1	4.01	7	1	0	4.43
4	1	1	4.36	9	0	0	2.80
4	1	1	2.42	10	0	0	2.70
5	1	1	3.49	10	1	0	2.96
5	1	0	3.97	11	0	0	2.60
8	1	0	3.52	13	1	0	2.88
8	1	0	3.05	16	1	1	3.60
8	1	0	2.32	17	0	0	2.16
8	1	1	3.26	19	0	0	2.05
11	1	0	3.49	20	0	1	2.01
11	1	0	2.12	22	1	1	2.32
12	1	0	1.50	23	1	1	2.57
12	1	0	3.06	25	0	1	1.78
15	1	0	2.30	32	0	1	2.20
17	1	0	2.95	32	0	1	2.53
22	1	0	2.73	34	0	1	1.47
23	1	1	1.97	35	0	1	1.45

We wish to estimate the survival function for each group separately and plot the corresponding survival curves. In order to apply the Kaplan-Meier method of estimating the survival function, the remission data in reorganized, as shown in Table 3.2, so that each row corresponds to one distinct survival time. The first column of Table 3.2 gives the ordered distinct survival times and the second column (denoted n_j) gives the size of the *risk set*: the number of patients who have not failed or been censored up to time $t_{(j)}$. In other words, n_j gives the number of patients at risk for failing instantaneously prior to time $t_{(j)}$ (Kleinbaum and Klein 2005). The third column (m_j) gives the counts of failures at each distinct failure time. Finally, the fourth column (q_j) gives the number of persons censored in the time interval starting with $t_{(j)}$ up to (but not including) the next failure time $t_{(j+1)}$.

The figures in the rightmost column denoted $\hat{S}(t_{(j)})$ are the KM estimates of the

Table 3.2: Computation of KM survival probabilities for the remission data

placebo group					treatment group				
$t_{(j)}$	n_j	m_j	q_j	$\hat{S}(t_{(j)})$	$t_{(j)}$	n_j	m_j	q_j	$\hat{S}(t_{(j)})$
0	21	0	0	1.00	0	21	0	0	1.00
1	21	2	0	0.90	6	21	3	1	0.86
2	19	2	0	0.81	7	17	1	1	0.81
3	17	1	0	0.76	10	15	1	2	0.75
4	16	2	0	0.67	13	12	1	0	0.69
5	14	2	0	0.57	16	11	1	3	0.63
8	12	4	0	0.38	22	7	1	0	0.54
11	8	2	0	0.29	23	6	1	5	0.45
12	6	2	0	0.19					
15	4	1	0	0.14					
17	3	1	0	0.10					
22	2	1	0	0.05					
23	1	1	0	0.00					

survival function. Their computation is straightforward for the placebo group as there was no censoring (the column q_j contains only zeroes). In the first row there is nothing to compute as $\hat{S}(t_{(j)}) = 1.00$ by definition. The number in the second row is $19/21 = 0.90$, because two people failed in the same week, so that 19 out of 21 remain at risk past one week. The remaining survival probabilities are calculated in the same manner: count the number of subjects surviving past the considered time and divide with the number of subjects at the start of the study.

When some units are censored, as in the treatment group, the calculation is not as simple, as we must use the product of conditional probabilities to calculate the KM estimate of the survival function. The probability of surviving past the 23rd week in the treatment group is calculated as:

$$\hat{S}(4) = 1 \cdot \frac{18}{21} \cdot \frac{16}{17} \cdot \frac{14}{15} \cdot \frac{11}{12} \cdot \frac{10}{11} \cdot \frac{6}{7} \cdot \frac{5}{6} = 0.45$$

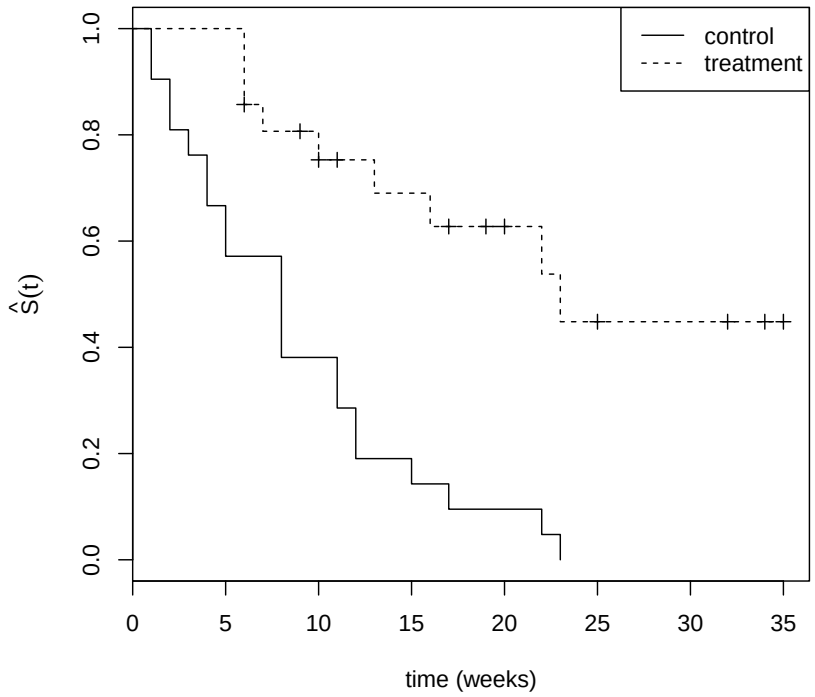
Because the formula for survival probability is limited to product terms up to the survival time being specified, the Kaplan-Meier formula is often referred to as the *product-limit* formula (Kleinbaum and Klein 2005). The general expression for the product limit formula is:

$$\hat{S}(t_{(j)}) = \prod_{i=1}^{j} \hat{Pr}\left(T > t_{(i)} | T \geq t_{(i)}\right) \tag{3.23}$$

$$= \hat{S}(t_{(j-1)}) \cdot \hat{Pr}\left(T > t_{(i)} | T \geq t_{(i)}\right). \tag{3.24}$$

The estimated survival functions for the placebo and treatment groups are plotted in Figure 3.11. Because the probabilities were calculated at discrete intervals (at each distinct survival time), the plot for a particular group is a decreasing broken line instead of a smooth curve. The line starts at probability 1 and drops at each subsequent survival time. Note that, if we had more data, the steps would be smaller and closer, and the plot would more closely resemble a smooth curve. The plus signs ("+") superimposed on the treatment curve indicate censored units.

Figure 3.11: Kaplan-Meier survival curve for the remission data



For the concrete case of the remission data, the most important finding is that the KM curve for the treatment group is consistently higher than the KM curve for the control group. This indicates that the survival prognosis for the treatment group is better than for the control group. Moreover, the two curves diverge as time increases, suggesting that the effect of treatment over the placebo is greater, the longer a patient stays in remission (Kleinbaum and Klein 2005).

3.3.4 The Cox proportional hazards model

The Cox PH model (Cox 1972) is the most commonly used regression model for censored survival data (Aalen et al. 2008, 8). The model assumes that the hazard rate at time t for an individual with the vector of predictors $\mathbf{X}$ takes the form:

$$h(t, \mathbf{X}) = h_0(t) \cdot \exp\left(\sum_{i=1}^{p} \beta_i X_i\right). \tag{3.25}$$

An important feature of (3.25) is that it breaks down the hazard into the product of two quantities. The first, $h_0(t)$, is called the baseline hazard. It is a function of time, but does not involve the predictor variables $\mathbf{X}$. If all the predictors have the value of zero (or if there are no predictors in the model), the Cox PH formula reduces to $h_0(t)$. Accordingly, this quantity is called the *baseline* hazard.

The second part of (3.25) is called the relative risk function and corresponds to the linear predictor in the classical regression model. The hazard rate on the left-hand side of (3.25) should never be negative, which is why the the linear sum of the predictors and their corresponding coefficients, $\sum_{i=1}^{p} \beta_i X_i$, cannot appear simply as an additive term on the right-hand side. Instead, the baseline hazard is multiplied by the *exponential* of the linear sum, ensuring that the product is positive.

In the simplest case, the relative risk function is independent of time, in which case the predictors are aptly called time-independent. In this case the model satisfies the *proportional hazards assumption*, which will be discussed in the following section. It is, nevertheless, also possible to consider time-dependent predictors. If such predictors are considered, the Cox model can still be used, but such a model no longer satisfies the PH assumption and is referred to as the *extended* Cox model (Kleinbaum and Klein 2005).

The beta coefficients in (3.25) can be estimated with a number of statistical software tools, most often employing the maximum likelihood method of estimation. Once the ML estimates are obtained, statistical inferences usually center on *hazard ratios*, which can be defined in terms of these estimates. A hazard ratio (HR) is defined as the hazard for one individual divided by the hazard for a different individual (Kleinbaum and Klein 2005). The individuals being compared are distinguished only in terms of their values on the predictors.

Let us denote the set of predictor values for the first individual as $\mathbf{X}^*$ and the set of values for the second individual as $\mathbf{X}$. We can obtain the hazard ratio for comparing these two individuals by substituting the Cox model expression (3.25) into both the numerator and the denominator of the hazard ratio:

$$\begin{aligned} \hat{HR} &= \frac{\hat{h}(t, \mathbf{X}^*)}{\hat{h}(t, \mathbf{X})} = \frac{\hat{h}_0(t) \cdot \exp\left(\sum_{i=1}^{p} \hat{\beta}_i X_i^*\right)}{\hat{h}_0(t) \cdot \exp\left(\sum_{i=1}^{p} \hat{\beta}_i X_i\right)} \\ &= \exp\left[\sum_{i=1}^{p} \beta_i (X_i^* - X_i)\right]. \end{aligned} \tag{3.26}$$

We will illustrate how the estimates from a fitted Cox model can be interpreted for the remission data. Table 3.3 shows the estimated beta coefficients in the first column. The predictors in the model are the treatment status of the patient (denoted Rx; 0 for the placebo group and 1 for treatment), the logarithm of the white blood cell count (logWBC), and sex (1 for male, 0 for female).

Table 3.3: Model estimates for the Cox PH model fit to the remission data with treatment (Rx), the logarithm of the white blood cell count (logWBC), and sex as predictors

	est	se	z	Pr(>|z|)		exp(est)	[95% conf.	interval]
Rx	-1.39	0.46	-3.05	0.00	**	0.25	0.10	0.61
logWBC	1.59	0.33	4.83	0.00	**	4.92	2.58	9.40
sex	0.26	0.45	0.59	0.56		1.30	0.54	3.14

The central research question in the remission study is whether or not the treatment is successful in delaying leukemia relapse. The purpose of including additional predictors in the model is to control for possible confounding effects. Assuming logWBC and sex are the same for two compared individuals, the two terms corresponding to logWBC and sex in (3.26) are equal to zero. The hazard ratio for comparing a treated patient to an untreated one is therefore $\exp\left[\beta_1(X_1^* - X_1)\right] = \exp\left[\beta_1(1-0)\right] = \exp(\beta_1)$. Controlling for logWBC and sex, the hazard ratio of being in the treatment group as compared to the placebo group is $exp(-1.39) = 0.25$. The results therefore show that for the considered sample of individuals the effect of treatment is beneficial in lowering the hazard for relapse.

Software packages used to estimate the model parameters also provide the ML estimates of the standard error. This allows the standard Wald test to be applied (see e.g. Harrell 2001). Assuming that the sample was obtained from a large population by means of simple random sampling, we are interested if the population value of the coefficient is different from zero. The test statistic (the coefficient estimate divided by the corresponding estimated standard error) is compared to the quantile of the standardized normal distribution for a predetermined significance level (usually alpha=0.05). For the example presented in Table 3.3, treatment and logWBC have a highly significant effect, while the effect of sex is not significant at the .05 level. The last three columns of the table give the exponentiated estimate (the hazard ratio) and its corresponding 95% confidence interval.

Alternatively to the Wald test, the likelihood ratio test can be performed by comparing the model in Table 3.3 to the same model excluding the predictor of interest (treatment). The two tests usually lead to the same conclusions, but, where they differ, the likelihood ratio test should be preferred (Collett 1994).

A key feature of the Cox model is that it does not assume a distribution for the time-to-event. In comparison to a parametric model, where the baseline hazard has a specific parametric form (e.g. exponential or Weibull, see Klein and Moeschberger 2003, ch. 12), the baseline hazard in the Cox model is left unspecified, which makes it a "robust" model in the sense that it will in general closely approximate the results for the correct parametric model (Kleinbaum and Klein 2005, 96). Because the Cox model contains a nonparametric part (baseline hazard) and a parametric part (the relative risk function), it is said to be *semiparametric* (Aalen et al. 2008, 133).

As a consequence of the baseline hazard being unspecified, the full likelihood for the Cox model cannot be formulated. Unlike the likelihood for a parametric model based on the distribution of the outcome variable (the time-to-event), the Cox likelihood is based on the *order* of events rather than the joint distribution of events (Kleinbaum and Klein 2005, 111). This means that any transformation that preserves the order of events (a monotonic transformation) can be applied to the outcome variable and the model estimates will not be changed.

3.3.5 The proportional hazards assumption

A critical assumption that the Cox PH model makes is that the hazard ratio comparing any two sets of predictors is constant over time. This equivalently means that the hazard for a certain individual is proportional to the hazard for any other individual, and that the proportionality constant is independent of time (Kleinbaum and Klein 2005). This assumption is obviously violated if, e.g., the hazard for individual A is higher than the hazard for individual B at the start of the observation period, while the hazard for individual A is lower than that of individual B after some time has passed. But even if the violation is not as gross as in the mentioned cross-hazard situation, the PH assumption may still be questionable.

A number of approaches have been developed to investigate whether the PH assumption is tenable. One graphical approach involves plotting estimated -log(-log) survival curves over different combinations of predictors being investigated. Parallel curves, e.g. for treatment groups, indicate that the PH assumption is supported (see Kleinbaum and Klein 2005, ch. 4 for details and other graphical methods).

A more quantitative approach involves calculating the so-called Schoenfeld residuals and produces a test statistic for each predictor in the model. The Schoenfeld residual (Schoenfeld 1982) is defined only for individuals who experience the event. Unlike residuals in a simple regression model, a different residual is defined for each predictor in the model. For an individual who fails at time t_j, the Schoenfeld residual for a

particular predictor is defined as that individual's observed value on the said predictor minus its expected value. The expected value is that predictor's weighted average for the other subjects still at risk at time t_j, the weights being each individual's hazard.

The statistical test for the PH assumption is based on the fact that if the PH assumption holds for a particular predictor, then the Schoenfeld residuals for that predictor will not be related to survival time (Kleinbaum and Klein 2005, 151). Rejecting this null hypothesis leads to the conclusion that the PH assumption is violated for the predictor in question. The test that we use in the analyses actually uses the *scaled* Schoenfeld residuals (Grambsch and Therneau 1994), but the purpose of the test is the same, and the tests typically yield similar results (Kleinbaum and Klein 2005, 152). The results of the test for the Cox PH model with three predictors are given in Table 3.4.

Table 3.4: The test of the proportional hazards assumption for the Cox model with treatment (Rx), the logarithm of the white blood cell count (logWBC), and sex as predictors.

	rho	chisq	p
Rx	0.12	0.41	0.52
logWBC	0.07	0.19	0.66
sex	-0.37	3.84	0.05

The p-values in Table 3.4 indicate that there is not enough evidence to reject the null hypothesis for the treatment and logWBC predictors. The test for the predictor sex is, however, significant at the alpha=0.05 level, indicating that the PH assumption is violated for sex.

The described test has the same disadvantage as any goodness of fit test: the null hypothesis can never be proven, and we can only determine that there is not enough evidence to reject it. The p-value, of course, depends on sample size, which means that a gross violation of the PH assumption may not be detected if the sample is too small, and, conversely, that a minor violation may result in a significant p-value if the sample is large.

If the PH assumption is found to be violated, several approaches can be taken. We can fit a Cox model *stratified* with regard to covariates that do not satisfy the PH assumption. This approach circumvents the problem by allowing a different baseline hazard function for each stratum (in the above example, a different $h_0(t)$ for each sex group). The PH assumption is still required to hold within each stratum (see Kleinbaum and Klein 2005, ch. 5). The disadvantage of stratification is that the covariates used to define the strata are excluded from the relative risk function, meaning that the coefficients (and hazard ratios) for such predictors are not estimated. The model

controls for the effect of covariates that define the strata, but cannot be conveniently quantified like it can be for predictors that appear in the relative risk function.

If the PH assumption for a certain predictor is suspect, another approach that can be taken is including as a predictor in the model the interaction of the suspect predictor with some function of time. The *extended* Cox model with one such interaction assumes the following form:

$$h(t, \mathbf{X}) = h_0(t) \cdot \exp\left(\sum_{i=1}^{p} \beta_i X_i + \delta X^* \times g(t)\right), \tag{3.27}$$

where X^* is the predictor under suspicion and $g(t)$ is a certain function of time (e.g. the logarithm). If the coefficient for δ is found to be significant in a Wald test (or alternatively the LR test for comparing the model with the interaction to a model without the interaction is significant), this is proof that the PH assumption has been violated.

3.3.6 The extended Cox model

The extended Cox model, as mentioned, does not satisfy the PH assumption because time enters into the model's relative risk function. The hazard ratio is thereby a function of time (see Kleinbaum and Klein 2005). In the previous section, we mentioned that interactions of time-independent covariates (like sex) with a function of time can be added to the model to account for deviations from the PH assumption.

Some predictors, however, are *inherently* time-dependent. In the remission data, the logarithm of the white blood cell count is regarded as time-independent, because it was only measured as each patient entered the study and was assumed to remain constant thereafter. If, however, the white blood cell count were to be measured at several time points for each patient, fluctuations of logWBC over time would be captured and could be used as additional information to model the time-to-relapse. The extended Cox model with time-dependent predictors assumes the form:

$$h(t, \mathbf{X}) = h_0(t) \cdot \exp\left(\sum_{i=1}^{p_1} \beta_i X_i + \sum_{j=1}^{p_2} \delta_j X_j(t)\right), \tag{3.28}$$

where the second sum inside the exponential refers to time-dependent predictors $X_j(t)$.

When the extended Cox model is used, each individual is represented not by one but

by several rows of data. This also means that the dependent variable is no longer defined by two pieces of information[5] (survival or censoring time plus event indicator). An additional assumption that the extended Cox model makes is that the effect of a time-dependent predictor $X_j(t)$ on the survival probability at time t depends on the value of this predictor at that same time t, and not on the value at an earlier or later time (Kleinbaum and Klein 2005, 220).

3.3.7 Global measures of model performance

The predictive performance of Cox models is commonly evaluated through measures of discrimination, which refers to the model's ability to distinguish between high-risk and low-risk persons. Two such measures are commonly reported with Cox PH models: the *c-index* and *pseudo* R^2.

The c-index is defined as the proportion of all usable person pairs in which model predictions and actual outcomes are concordant (Harrell et al. 1996). Computing the index boils down to evaluating all pairs of persons where at least one of the persons failed. If the model prediction (e.g. person A survives longer than person B) is concordant with the actual outcome, such a pair is assigned the value of 1. If the model predicts the opposite of the actual outcome, such a pair is assigned the value of 0. The c-index is the average of such ones and zeroes across all usable pairs. A c-index value of 0.5 therefore indicates no predictive discrimination, and a value of 1.0 indicates perfect separation of persons with different outcomes (Harrell et al. 1996).

Measures of explained variation can be defined in various ways in survival analysis (see Schemper and Stare 1996). The most common quasi R^2 reported with the Cox PH model is defined with regard to model likelihood. There has been some debate over how exactly this quantity should be computed[6]. As O'Quigley, Xu and Stare (2005) argue and demonstrate with a simulation study, the correct estimate for the proportion of explained variation in a proportional hazards model is:

$$R^2 = 1 - \exp\left\{\frac{-2(LL_b - LL)}{k}\right\}, \tag{3.29}$$

where LL_b is the loglikelihood of the baseline model with no predictors, LL is the

[5]The survival package (Therneau 2013) for R (R Core Team 2013) that is used for survival analyses in the empirical part requires us to specify the dependent variable with three pieces of information in such cases: the *start* and *stop* of the time period to which the information in the current row of data pertains, as well as the event indicator (as before) (see Therneau 1999).

[6]We note that the value of R^2 reported by the survival package in R is incorrect. It is calculated with the *number of rows in the datafile* as the denominator in (3.29). The R^2 estimates reported in the survival package are therefore too low, especially when several rows per respondent are used to estimate the extended Cox model.

loglikelihood of the model with predictors, and k is the number of events.

This concludes our review of relevant statistical models. The next section will describe how this theoretical framework will be applied to the problems of item nonresponse and breakoff in the survey data.

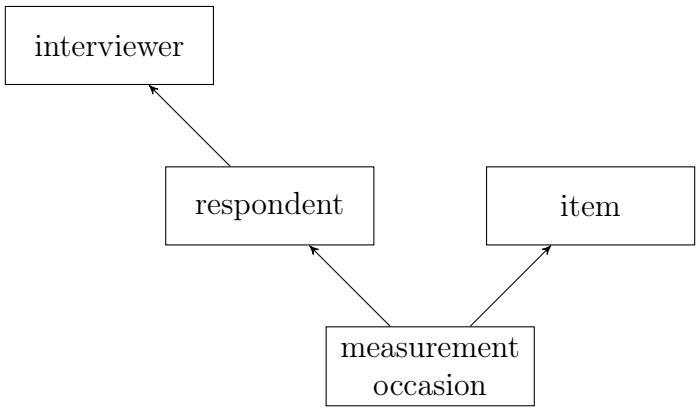
3.4 Application of the models to survey data

This section will outline how the statistical models that have been introduced will be applied to the survey data. For now, we wish to provide only a broad overview of our application and comment on it using the concepts that have been introduced in this chapter, while leaving technical details to later chapters.

3.4.1 Generalized linear mixed model for item nonresponse

The model for item nonresponse that we will fit to the data will combine many of the features introduced in Section 3.2. The data are clustered in a quite complex way, as shown in Figure 3.12 by a *classification diagram* (Browne et al. 2001). We will refer to the lowest level as the *measurement occasion*: a particular item being administered to a particular respondent. Each measurement occasion is nested within an *item* and cross-classified with a *respondent*. For telephone and face-to-face modes, the respondent is, further, nested within the *interviewer*.

Figure 3.12: Classification diagram for the survey data



The outcome of each measurement occasion is either a substantive response (0-response) or an item nonresponse (1-response). Because we chose to code item nonresponses with 1 (and not zero), a positive estimate of a fixed effect in the model will mean that high values of the corresponding predictor increase the probability of item nonresponse.

The GGP questionnaire asked very detailed questions, a great number of which was not applicable to all respondents. The respondent was, e.g., asked "Were you born in Slovenia?" If they answered "yes," the following item inquired into the *municipality* of birth, otherwise another item was administered asking about the *country* of birth. The point we wish to make with this example is that, due to the questionnaire routing, the respondents were not all administered the same set of items. This issue is quite typical for any kind of survey.

The indexing with the two indices p and i (as in η_{pi} in Equation (3.6)) that was used in Section 3.2 assumes that each item was administered to all respondents and that the data can therefore be neatly organized into a rectangular scheme, with index i referring to the rows and index j to the columns (or vice versa). When the model is applied to survey data, each respondent is (typically) not administered all items in the questionnaire as a result of questionnaire routing. The aforementioned way of indexing with two indices must therefore be abandoned in favor of *nested indexing* that was introduced in Section 3.1.1. In order to keep our notation comparable to the notation used in Section 3.2, we will introduce index m to refer to the measurement occasion. The subscript $p[m]$ will then refer to the person (respondent) p to whom the mth measurement occasion belongs, and $i[m]$ will identify the item i to which the mth measurement occasion belongs. Another index, n, is introduced to identify interviewers. Subscript $n[p]$ refers to interviewer n, to whom the pth respondent belongs.

Because of the many levels we will refrain from specifying the model in a single equation, and rather write one equation for each level. The following four equations fully specify the model:

$$\eta_m = \sum_{h=1}^{H} \delta_h W_{mh} + \theta_{p[m]} + \beta_{i[m]} \tag{3.30}$$

$$\beta_i = \sum_{k=0}^{K} \beta_k X_{ik} + \epsilon_i \tag{3.31}$$

$$\theta_p = \sum_{j=1}^{J} \vartheta_j Z_{pj} + \gamma_{n[p]} + \epsilon_p \tag{3.32}$$

$$\gamma_n = \sum_{l=1}^{L} \gamma_l V_{nl} + \epsilon_n. \tag{3.33}$$

Note that, contrary to IRT convention, we use the *positive sign* for the random item intercept $\beta_{i[m]}$ in (3.30). This way, the direction of the fixed effects at the item-level will have the same meaning as for fixed effects at other levels: an effect's positive value of an effect means that high values of the corresponding predictor increase the probability of item nonresponse. The predictors at the measurement occasion level (the respondent-by-item predictors described in Section 4.3) are indexed with index h ($h = 1, \ldots, H$). The hth predictor for the mth measurement occasion is denoted W_{mh}; the fixed effect of the hth predictor is δ_h.

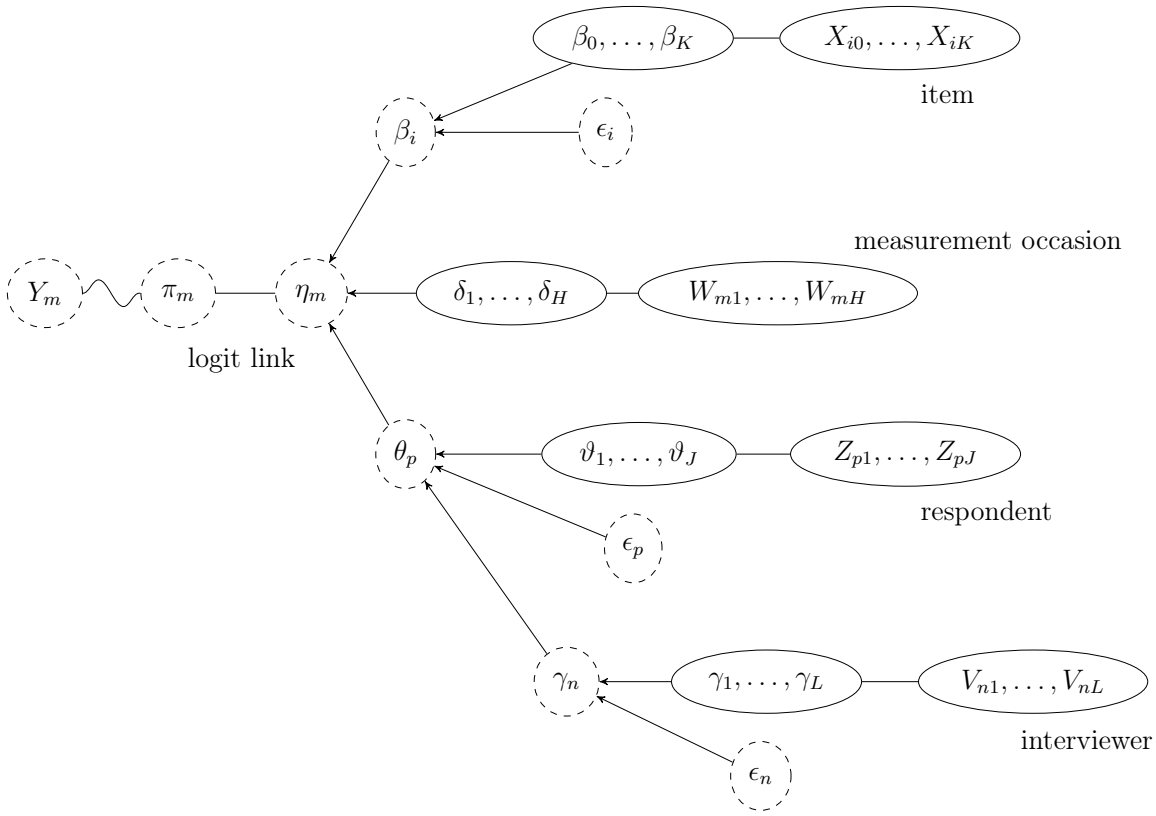
The notation in the other equations follows the same logic; there are K fixed effects at the item-level, J at the respondent level, and L at the interviewer level. The kth predictor for the ith item is denoted X_{ik}, the jth predictor for the pth respondent is

denoted Z_{pj}, and the lth predictor for the nth interviewer is denoted V_{nl}. Residuals appear at each of the higher levels; ϵ_i captures the ith item's deviation from the model's prediction, ϵ_p captures the pth respondent's deviation from the model's prediction, and ϵ_n captures the nth interviewer's deviation from the model's prediction.

$\gamma_{n[p]}$ in the respondent-level equation is the effect of the nth interviewer. This is further regressed on interviewer level predictors in Equation (3.33). We note that the overall intercept is included at the item level: the summation starts at $k = 0$ with X_{i0}, corresponding to a vector of ones that is, technically speaking, multiplied by the overall intercept β_0.

The model, defined in Equations (3.30) through (3.33) is graphically represented in Figure 3.13 in a manner following De Boeck and Wilson (2004b) and figures in Section 3.2. The levels are labeled at the right hand side of the figure to help with the interpretation of the diagram.

Figure 3.13: Graphical representation of the logistic model for item nonresponse



3.4.2 Cox PH model for breakoff

We will model breakoff using the Cox model. The *number* of items occurring from the start of the interview will play the role of the time variable. Alternatively, we could consider the time in minutes since the beginning of the interview as the measure of

time, but we argue that the burden experienced by the respondent is better measured in the number of items than in elapsed minutes.

In this context, we will consider *breakoff* as the terminal *event*, whereas respondents who *complete* the interview will be considered *censored*. We believe that the assumption of non-informative censoring is tenable in this case, as the respondent had no control over how many items would still be administered before completion. It is true that some respondents could have learned how to answer certain items in order to avoid additional questions (e.g. by naming no previous partners, a respondent could avoid all subsequent items inquiring into previous partnerships), but we believe this scenario not to be very likely.

The number of breakoffs was very low in Rounds 1 and 2 (see Table 4.8), which is why an additional round of data was collected via recrutation through advertisements on the social network website Facebook. There were enough breakoffs in this additional round to employ survival analysis methods. Unfortunately, as a result we will not be able to compare modes of administration with regard to breakoff, apart from the simple comparison of the *proportion* of breakoff across modes.

When analyzing breakoff, we will therefore not need to consider interviewers as a source of clustering. The respondent characteristics, of course, do not change during the course of the interview and will thus be considered time-*independent*. Item predictors and respondent-by-item interactions, on the other hand, do change with each administered item. Because we are interested in examining the effect of the item facets on the risk of breakoff, we will use the *extended* Cox model with time-dependent predictors (see Section 3.3.6) to model breakoff.

4 Methodology of the empirical research

The first section of the present chapter provides a detailed description of the data collected in the Generations and Gender Programme survey, and the data collection procedures that were followed. The next two sections describe the construction of predictors that will be used in the statistical models. Section 4.2 describes how item-level predictors were obtained by coding each questionnaire item by measures of topic sensitivity. We explain how respondents' self-assessments were combined with questionnaire codes to form item-by-respondent interactions in Section 4.3. Section 4.4 then briefly introduces the rationale behind multiple imputation and describes how this procedure was applied to address the issue of missing values in the predictors. The final section of this chapter operationalizes the hypotheses that were put forward in Section 2.5.

4.1 The Generations and Gender Programme survey

The data that is analyzed in the present dissertation was collected in a pilot study funded by the Generations and Gender Programme (GGP). The GGP is a pan-European research infrastructure that has received funding from the 7th framework programme. It has the aim of supplying a data source for academic research and population-related policy formulation. The questionnaire's central topics are fertility, partnership, transition to adulthood, economic activity, and inter-generational and gender relations. As of the writing of the present dissertation, the GGP survey has been implemented in 19 countries (see Generations and Gender Programme 2013).

In 2011, a new version of the GGP questionnaire was developed and needed to be field-tested in a pilot study. Several questionnaire items were modified and would be tested against their original counterparts in split-ballot experiments (see, e.g. Schuman and Presser 1981). The pilot's second aim was to determine whether face-to-face interviewing could be supplemented by other, cheaper modes without jeopardizing data quality.

The questionnaire of the 2011 pilot was divided into eleven modules. The modules contained items inquiring into the following topics:

1. **Personal information**: the respondent's sex, date and place of birth, education; information on and satisfaction with dwelling unit.

2. **Partnerships and children**: the respondent's current partner's basic demographics; legal status of current partnership; satisfaction with relationship; frequency and topic of disagreements with partner; children's basic demographics; history of previous partnerships; intentions of union formation; information on grandchildren.
3. **Household composition**: information on other household members, satisfaction with the relationship with each member; household and childcare organization (who takes care of what); decision making, income organization.
4. **Parents and parental home**: basic demographics and living arrangement for each of the respondent's parents; frequency of meeting each parent; satisfaction with the relationship with each parent; intentions to move out of parental home or move back in; information on the parental home during the respondent's childhood.
5. **Networks and support**: information on the respondent's emotional support network; childcare provision network (incoming and outgoing), information on professional childcare providers; incoming and outgoing networks for practical help, personal care, and financial support; basic demographics for each alter named in all aforementioned name generators.
6. **Fertility**: information on current pregnancy; the respondent's and partner's ability to have more children; information on infertility treatment or birth control use; intentions of having children.
7. **Health and well-being**: self-assessment of health; inventories on the respondent's personality traits, sense of control and well-being.
8. **Respondent's activity**: satisfaction with current activity; information on current occupation and workplace; intentions to retire; income.
9. **Partner's activity**: information on partner's current occupation and workplace; partner's income.
10. **HH Possessions and income**: inventories on household possessions and ability to maintain the household; total household income.
11. **Values and attitudes**: respondent's religious denomination, attendance of religious ceremonies, generalized trust, attitudes toward inter-generational and gender relations.

Because one of the aims of the pilot study was to evaluate the feasibility of conducting an interview of this length and complexity in cheaper modes of administration, the new questionnaire was implemented in three modes: 1) face-to-face, as in all previous GGP surveys; 2) computer assisted telephone interviewing (CATI); and 3) over the internet in a self-administered questionnaire (web mode). Assessing the feasibility of transition to web and CATI modes in the pilot study consisted of two tasks:

1. Estimating the *measurement* effects of CATI and web modes as compared to F2F interviewing. Studying the measurement effect of the mode of administration requires strict control over other factors that can cause differences across modes in the obtained responses. Particular attention needs to be paid to possible *coverage* and *selection* effects that can occur when sample persons cannot be interviewed in a particular mode, or prefer one mode over others. If older respondents, e.g., have a low preference for web mode, they will be underrepresented in the web sample as compared to other samples. If a certain questionnaire item is then found to have a different average in web mode as compared to other modes, it is not clear whether this difference can be attributed to measurement effects, or coverage and selection effects.
2. Identifying the *most efficient mixed-mode design* in terms of costs and response rates. One way to overcome coverage bias is to employ mixed-mode methodology, in which two or more modes are used to collect data for a single data set (Dillman and Tamai 1988; Dillman 2000; Diment and Garret-Jones 2007). An example of such a strategy is to start with the cheapest mode of data collection (web); then invite non-responding sample persons to participate in a phone interview; and finally send an interviewer for a face-to-face interview if previous attempts at eliciting cooperation have failed. In the 2011 pilot, several sequences of data collection modes were considered and compared (CATI→web→F2F vs. F2F→web→CATI vs. web→CATI→F2F).

In an attempt to meet both goals of the pilot, it was decided that the data collection should proceed in *two rounds*. The evaluation of measurement effects would be performed not by sampling from the population but by resorting to a panel of respondents that the selected fielding agency (Valicon) had already recruited for their own purposes. The fielding agency would select those panel members that had *provided all three types of contact* (address, telephone number, and e-mail address) and randomly assign them to the mode of administration. The rationale behind this decision was to save costs on recruitment, since these persons had already agreed to be interviewed and had given all required contact information—a result that would be very costly to replicate on the general population because of an additional screening phase. The fact that no random sampling would be performed in recruiting the respondents (before random assignment to mode) was seen as unsubstantial for the study of measurement effects. The fielding agency, furthermore, assured that the panel structure was comparable to the Slovenian population with regard to key demographic variables, meaning that findings from the measurement effects study could be generalized onto the general population.

The second task of identifying the most efficient mixed-mode design could not, of

course, avoid sampling from the general population, as the purpose of this phase was to find the best way to approach a sampled person. Because of refusal to cooperate (unit nonresponse), however, the funds allocated for the second phase of interviewing would result in relatively fewer successful interviews as compared to the funds used for interviewing the fielding agency's panel members.

4.1.1 Sampling procedures

The first round of interviewing was conducted in September and October 2011, while the second round was conducted in November and December 2011. To make the headings in the tables clearer we will refer to first round of data collection as "the panel" (abbreviation "pnl" in tables; because the commercial panel was used) and the second as "the sample" (abbreviation "smp" in tables, as individuals were sampled from the population). For the purposes of analyzing item nonresponse and breakoff, this two-round design adds an undesired level of complexity that we will have to take into account in the analyses found in the empirical part of this dissertation. We will now describe the data collection procedures for each round of data collection in turn.

Round 1 - the panel

The first round of data collection was performed without random sampling from the population. The majority of respondents were members of a web panel maintained by the fielding agency Valicon for its own interviewing purposes. The agency reported to have recruited the panel members through various methods (via telephone, F2F, using banners on different portals and websites) in an attempt to obtain a demographic structure similar to that of the overall Slovenian population. Upon completing one of the regular commercial surveys, existing panel members were asked if they would be willing to participate in the GGP pilot. They were specifically told that the GGP was a multimode experiment survey. Only those panel members who agreed to give all three required pieces of contact information were eligible to be included in the study (see Table 4.1).

Table 4.1: Two-step recruitment of existing web panel members into the GGP pilot study

			n	%
Step 1	Willing to participate?	Yes	1050	75.4
		No	342	24.6
		Total	1392	100.0
Step 2	Please give contact info	Obtained all three contacts	743	53.4

The goal was set of realizing 200 respondents in each of the three modes of administration. Several months before the interviewing period began, the fielding agency was already preparing for the possibility that the panel members who were willing to participate and gave the required contact information would not be sufficient for the goal of 200 units per mode. Accordingly, the fielding agency's recruiters started recruiting additional members in their areas of residence, solely for the purpose of the GGP pilot. . Each recruiter briefly presented the GGP survey to potential sample persons and asked them for the required contact information. Of the 178 persons enlisted in this way, a random subsample of 104 was later actually used.

Panel members were randomly assigned to modes: 248 to F2F and web modes and 247 to CATI. Because the goal of obtaining 200 interviews per mode was not obtained (see Table 4.2), sample persons from the additional step of recruitment were added to the face-to-face (32 additional persons) and CATI (72 additional persons) modes toward the end of the interviewing period.

Table 4.2 examines the final disposition codes for each mode. In survey research, final disposition codes (completed survey, refusal, breakoff etc.) are examined with the goal of calculating the *response rate*: the proportion of completed questionnaires out of all eligible units in the sampling frame (see, e.g., AAPOR 2009). No sampling frame was used in the first round of data collection, thus the response rate cannot be calculated. We can, however, calculate the *completion rate*: the proportion of completed surveys among those persons who agreed to participate and were approached with an interview request. The completion rate can serve as a cursory indicator of the sample persons' willingness to cooperate, but is in this respect admittedly a poor substitute for the response rate.

As Table 4.2 shows, about three quarters of respondents who agreed to participate and were approached with an interview request actually completed the survey. The completion rate for the panel members was higher than the completion rate for the additionally recruited persons (74% vs. 69% for F2F mode; and 65% vs. 47% for CATI). The high completion rate (87%) for the web mode can be attributed to the fact that the panel members had already completed a number of Valicon's surveys on the web prior to the GGP questionnaire. Web interviewing was therefore, in a sense, their "native" mode, which also involved no need to coordinate with the interviewer for an interview appointment.

Table 4.2: Disposition codes and completion rates for Round 1

	F2F	CATI	web		Total
Commercial panel	248	247	248	initial sample	743
	184	161	216	completed	567
		14	12	break off	26
	18	4		refused	22
	74%	65%	87%	completion rate	76%

	F2F	CATI	web		Total
Additional	32	72		initial sample	104
	22	34		completed	56
				break off	
		1		refused	1
	69%	47%		completion rate	54%

	F2F	CATI	web		Total
Total	280	319	248	initial sample	847
	206	195	216	completed	623
		14	12	break off	26
	18	5		refused	23
	74%	61%	87%	completion rate	74%

Round 2 - the sample

The purpose of the mode-systems study was to estimate which type of mixed-mode combination would be most efficient from the perspective of response rate and costs. Three systems of mixed mode data collection were compared against each other:

S1: CATI→web→F2F

S2: F2F→web→CATI

S3: web→CATI→F2F

There were two subtypes for S3: one with incentives and one without. These are referred to as S3+ and S3-, respectively. The central population register of the Statistical office of Slovenia was used to draw a sample and each sample person was assigned to the mode system randomly.

As random sampling was used in the second round, we are able to calculate the response rate for each mode system. This requires us to determine the final disposition codes, a task complicated by the fact that sample persons who refused to cooperate in one mode were attempted to be re-contacted in the next mode. We use the S1 mode system as an example of how the final disposition codes were determined and how the response rate was calculated. Tables 4.3, 4.4, and 4.5 show the relevant figures. For the other mode systems we omit this information and only provide the response rates.

The disposition codes used in Tables 4.3, 4.4, and 4.5 have the following meaning:

Soft refusal: the sample person did not wish be interviewed, but gave justifications like "no time at the moment", and seemed likely to participate if approached at a later time and/or by different mode. Soft refusals were countered with an offer of a different interview mode.

Hard refusal: the sample person gave a clear indication that they do not wish to participate at all. Respondents giving hard refusals were not approached again.

No attempt (web mode)**:** the sample person never accessed the questionnaire web page.

Breakoff, refusal: the respondent started filling in the questionnaire, but gave up before reaching the beginning of Module 2.

Breakoff, partial: the respondent started filling in the questionnaire but broke off mid-survey. This happened after the beginning of Module 2.

Note that two types of breakoff are distinguished above. If the breakoff took place very early in the interview (before reaching the question on whether the respondent

currently has a partner in Module 2), this was regarded as a refusal and an attempt was made to re-contact the sample-person in the next mode. If, however, the respondent filled out at least the initial module of the questionnaire and then broke off, this was considered a partially completed interview and the respondent was encouraged to complete the interview in the same mode.

We will now describe the mode transitions for the example of the first mode system. The starting sample size was 108 (Table 4.3). Of these units, 14 complete interviews were obtained as well as one partial interview. There were 24 hard refusals and 7 ineligible units, all of which were not approached again. The fielding agency was not able to contact 51 sample persons at the listed telephone numbers. An additional 11 gave a soft refusal. These 62 units were sent an invitation to fill in the web questionnaire.

Table 4.3: Disposition codes for S1, CATI

	CATI	→web
completed	14	
breakoff, partial	1	
hard refusal	24	
ineligible	7	
soft refusal	11	11
no contact	51	51
Total	108	62

Of these 62 sample persons, the majority (51) did not begin filling in the web questionnaire. All of these were moved to the next mode (F2F), as was an additional sample person who started filling in the questionnaire, but quit before reaching Module 2.

Table 4.4: Disposition codes for S1, web mode

	web	→F2F
completed	10	
breakoff, refusal	1	1
no attempt	51	51
Total	62	52

Of the 52 sample persons who were first approached over the telephone, then asked to fill in the web questionnaire, and finally approached in person, 32 gave a hard refusal. One person was ineligible, and 11 were not contacted at their listed address. Eight interviews were completed.

Table 4.6 draws on Tables 4.3, 4.4, and 4.5 and gives the final disposition codes for the first mixed mode system. Two types of response rate are calculated on the basis

Table 4.5: Disposition codes for S1, F2F

	F2F
completed	8
hard refusal	32
ineligible	1
no contact	11
Total	52

of the figures in Table 4.6, following the guidelines by AAPOR (2009). The first one excludes breakoffs in the numerator, while the second response rate includes them:

$$RR_1 = \frac{\text{completed}}{\text{total} - \text{ineligible}} = \frac{32}{108 - 8} = 32\% \tag{4.1}$$

$$RR_2 = \frac{\text{completed} + \text{breakoff}}{\text{total} - \text{ineligible}} = \frac{32 + 1}{108 - 8} = 33\% \tag{4.2}$$

Table 4.6: Final disposition codes for S1: CATI→web→F2F

	CATI	web	F2F	Total
completed	14	10	8	32
breakoff	1			1
refusal	24		32	56
no contact in last mode			11	11
ineligible	7		1	8
Total	46	10	52	108

The same procedure was followed to determine the response rate for the other mode systems. Table 4.7 gives the response rates for all mode systems. We note again that S3+ refers to the third mode system *with* incentives, and S3- refers to the same system *without* incentives.

Table 4.7: Response rates by mixed mode system

	RR_1	RR_2
S1 (CATI→web→F2F)	32%	33%
S2 (F2F→web→CATI)	43%	44%
S3+ (web→CATI→F2F)	35%	36%
S3- (web→CATI→F2F)	27%	28%

Overall, the response rate is quite low: none of the mode systems exceeds 50%. The highest response rate was achieved in S2 where the respondent was initially approached

face-to-face. Mode system 3 achieved a noticeably higher response rate when incentives were used than the setting with no incentives.

The mode system's design in the second round adds further undesired complexity to the sample, as far as our goal of analyzing item nonresponse and breakoff is concerned. In order to simplify matters, we will only consider the *final* mode of administration in the analyses of the second round data and disregard the information on the mode system. We will thereby not distinguish between a sample person who was contacted face-to-face and cooperated, and a sample person who first refused to be interviewed over the web and telephone and decided to cooperate only after being approached face-to-face.

4.1.2 Breakoff rates and the third round of data collection

Because the purpose of this dissertation is to analyze item nonresponse and breakoff, our concern before data collection started was whether the rates of item nonresponse and breakoff would be high enough to allow to fit complex models that were described in the previous chapter. Our assumption was that this particular questionnaire was sufficiently long and contained many difficult and sensitive items to get "good" data for our analyses (high rates of item nonresponse and breakoff would, of course, not be considered beneficial if we were interested in the substance of the items).

After preliminary analyses, we were somewhat surprised to find that the overall item nonresponse rate was very low (see Table 5.6) and breakoff was almost non-existent (Table 4.8). We theorize that the absence of breakoff, despite the length of the interview, was caused by a selection effect: the sample persons (in both rounds) were sent an advance letter that informed them of the purpose of the study and also mentioned the approximate length of the interview. Those sample persons who were unwilling to sacrifice an hour of their time for the purpose of the interview probably refused outright (resulting in unit nonresponse), while those willing persevered until the end of the interview.

Not wanting to give up on studying breakoff, we initiated another round of data collection with the same questionnaire (with our own funding). The respondents were recruited via advertisements displayed on the social network site Facebook inviting them to take part in a demographic survey. The duration of the interview was not mentioned and this, indeed, resulted in a substantial breakoff rate of 60.3 %. This additional round of data collection was conducted in November and December 2012 and is labeled "web.fb" in the tables. Advertisements on the Facebook page were only displayed to Slovenian nationals aged eighteen or more.

Table 4.8: Frequency and proportion of breakoff by sample

	n	n breakoff	% breakoff
f2f.pnl	206	0	0.0
f2f.smp	107	1	0.9
cati.pnl	209	14	6.7
cati.smp	59	6	10.2
web.pnl	228	12	5.3
web.smp	45	5	11.1
web.fb	262	158	60.3
Total	1116	196	17.6

A number of questionnaire items were added to the GGP questionnaire for the purposes of the pilot study. These additional items inquired into the respondent's subjective experience of the interview: whether it was too long, how clear the questions were, etc. We have excluded these additional items for the purposes of our analyses: we consider the interview to have started with the first "regular" GGP item, and finished with the last "regular" GGP item. If the respondent did not reach this final item, their interview is considered to have ended in breakoff.

4.1.3 Questionnaire routing and interview length

Different respondents were administered different questions during the course of the interview. One reason is the inclusion of the aforementioned split-ballot items, whereby one respondent would be administered the old version of a questionnaire item while another respondent would be administered the new version. This assignment was randomized.

Another reason is questionnaire routing; a substantial portion of the questionnaire consisted of questions on the respondent's relation to their children, previous partners and cohabiting household members. A respondent with many children and previous partners would need to answer substantially more questions than a respondent with no children and no previous partners.

This means that the questionnaire items differ widely with regard to their sample size: certain core items were administered to all respondents, while other items were only visited by respondents in a specific life situation (e.g. those expecting a child). The figures in the first column of Table 4.8 refer to the number of respondents who *started* the interview. The first two questionnaire items which inquired into the respondent's

sex and birth date have this sample size. Subsequent items have a lower (or equal) sample size due to questionnaire routing and breakoff.

Table 4.9 gives the average number of items that completing respondents were administered during the course of the interview, as well as the average interview duration in minutes for each sample. Respondents who *broke off* have been *removed* to make figures comparable across samples (otherwise the mean number of items and the mean duration for web.fb would be substantially lower due to breakoff).

Table 4.9: Interview duration by sample for completing respondents (breakoffs excluded)

		number of items				duration (minutes)			
	n resp.	mean	std.dev	min	max	mean	std.dev	min	max
f2f.pnl	206	314.7	38.7	215	390	46.5	11.3	20.5	92.4
f2f.smp	106	292.8	42.4	209	385	37.9	10.3	14.5	67.8
cati.pnl	195	318.6	38.0	218	412	54.4	11.2	21.0	104.7
cati.smp	53	296.8	40.2	212	374	54.7	12.6	25.2	93.5
web.pnl	216	327.3	39.0	220	455	45.0	13.1	20.4	96.7
web.smp	40	303.3	42.4	206	428	51.9	15.5	26.4	92.6
web.fb	104	315.7	42.0	228	402	41.1	12.9	20.4	80.1
Total	920	314.5	41.0	206	455	46.9	13.3	14.5	104.7

On average, completing respondents were administered more than 300 items and took about 45 minutes to complete the interview. The duration of the GGP questionnaire for completing respondents was longest when interviewing was conducted over the telephone (54.4 minutes in the first round and 54.7 in the second) and the shortest when the interview was conducted in person (46.5 and 37.9 minutes respectively).

4.1.4 Demographic structure of the sample

In this section we examine the demographic structure for each of the seven aforementioned samples that are defined by mode and round of data collection. This is of interest, because substantial differences in demographic found across samples could indicate a sign of selection effects. The tables in this section give the demographic structure according to gender, age, and education for each sample. Table 4.10 gives the *absolute* frequencies for each demographic category by sample and Table 4.11 gives the *relative* frequencies (percentages). The rightmost column in the tables refers to the population of Slovenian residents *aged eighteen or more* in the year 2011.

The additional round of data collection with recruitment through Facebook clearly

stands out. The second round used random sampling and the structure of the web panel used in the first round was adjusted to more closely match the population with regard to core demographics. No such procedures could be used in the additional Facebook sample.

An important issue with the third round is *self-selection*. Couper (2000) gives examples of self-selected web surveys and comments on the example of a National Geographic Society survey that the self-selected nature of the survey, coupled with its placement on the National Geographic Society's web site, was likely to yield respondents who are more interested in cultural events and differ from the general population in other aspects. The same could be argued for the case of the additional Facebook sample: those Facebook users who clicked on the advertisement are likely to be more interested in demographic issues or taking part in surveys than those who did not. An additional issue with self-selected surveys is that the response rate cannot be calculated because the denominator of the ratio (see Equations (4.1) and (4.2)) is unknowable: there is no way to determine the number of eligible persons who were invited to partake in the survey (Couper 2000).

It is no surprise, then, that the structure of the additional sample does not reflect the overall Slovenian population: young people are overrepresented, with more than half the sample members under the age of 26, as compared to 11.6% in the actual population (see last column of Table 4.11). Women make up three quarters of the sample as compared to 50.9% in the population of 2011.

Table 4.10: Demographic structure by sample (frequencies)

	f2f.pnl	f2f.smp	cati.pnl	cati.smp	web.pnl	web.smp	web.fb	popul.(2011)
sex								
female	109	56	117	36	118	27	194	865267
male	97	51	92	23	110	18	68	834226
age								
18-25	26	14	26	2	33	12	136	197628
26-35	47	18	46	6	63	10	50	307005
36-45	52	16	52	14	60	5	36	302771
46-55	40	23	43	11	33	8	28	309823
56+	41	36	42	26	39	10	12	582266
education								
low	3	25	6	3	6	6	21	453448
middle	129	61	113	36	121	16	164	938376
high	74	21	90	20	101	23	77	307669
Total	206	107	209	59	228	45	262	1699493

Source for population statistics: Statistical office of the republic of Slovenia (2013)

Women are somewhat overrepresented in other samples too, but the percentage of women exceeds 60% only for the cati.smp and web.smp samples. The age structure in face-to-face mode and in cati.pnl roughly reflects that of the actual population. In web mode (and cati.smpl) respondents aged 56+ are underrepresented and respondents younger than 26 are overrepresented. We attribute this to younger persons' familiarity with information technology and preference for web interviewing over other modes. Highly educated respondents are somewhat overrepresented and low-education respondents underrepresented in all samples in comparison to the overall Slovenian population of 2011.

Table 4.11: Demographic structure by sample (percentages)

	f2f.pnl	f2f.smp	cati.pnl	cati.smp	web.pnl	web.smp	web.fb	popul.(2011)
sex								
female	53	52	56	61	52	60	74	50.9
male	47	48	44	39	48	40	26	49.1
age								
18-25	13	13	12	3	14	27	52	11.6
26-35	23	17	22	10	28	22	19	18.1
36-45	25	15	25	24	26	11	14	17.8
46-55	19	21	21	19	14	18	11	18.2
56+	20	34	20	44	17	22	5	34.3
education								
low	1	23	3	5	3	13	8	26.7
middle	63	57	54	61	53	36	63	55.2
high	36	20	43	34	44	51	29	18.1
Total	100	100	100	100	100	100	100	100.0

Source for population statistics: Statistical office of the republic of Slovenia (2013)

The data collection procedures that have been described so far do not allow us to *generalize* the results of statistical analyses to the overall population. Generalization in this strict statistical sense would require that a population be defined and that a random sample be drawn from it with a non-zero selection probability for each unit in the population. In the first and third round, no random sampling procedures were employed[7] while in the second round the sample design was complicated by the mixed mode design. The web.smp sample, for example, contains respondents from all three mode systems (CATI→web→F2F, F2F→web→CATI, and web→CATI→F2F). Since sample persons had the possibility to refuse an interview in one mode and later accept in another mode, the possibility of self-selection effects cannot be excluded.

[7]Strictly speaking, the first round was also subject to self-selection, but this is less apparent as the self-selection had already taken place at the panel recruitment stage. The panel member decided to join the panel of their own volition, perhaps after an invitation by a friend or acquaintance (option or access panel, Couper 2000).

Even without the possibility of generalizing the results, we see the analyses presented in the empirical part of this dissertation as relevant for survey methodology, in that they demonstrate the feasibility of our approach to analyzing item nonresponse and breakoff. Many survey methodology studies are based on samples from special populations, like university students. The sample members, in our case, do not all belong to such a special group, but do roughly represent the Slovenian (18+) population. For example, low-education persons, even though they are underrepresented, constitute a part of each sample. The additional third round of data collection, though, is clearly not comparable to the others, thus special consideration will need to be taken when interpreting the results from the Facebook sample.

This concludes our description of the GGP survey data. When analyzing item nonresponse and breakoff, both respondent-level and item-level predictors will be used in the statistical models. Some respondent characteristics that will be used as predictors have already been mentioned in this section, e.g., respondent education and age. The item-level predictors, on the other hand, need to be constructed, e.g., by expert judgment. The next section describes the details.

4.2 Expert judgment of item characteristics

As we mentioned in previous chapters, we hypothesize that the sensitivity of the item topic will be associated with item nonresponse and breakoff. In order to construct item-level predictors that will be used in statistical models, each item needs to be rated on how sensitive it is. The concept of questionnaire item sensitivity, however, has been argued in the literature to have several meanings, hence we decided to attempt to capture item sensitivity with several measures rather than just one. This section provides a short review of how the concepts of item sensitivity and social desirability have been treated in the survey literature. We then present the instructions that were developed and given to raters, and finally examine the inter-rater agreement for each measure.

Tourangeau, Rips and Rasinski (2000) argue that *three* distinct meanings of the concept of sensitivity appear in the survey literature:

1. A survey question might be sensitive if it is perceived as *intrusive*. Such a question might touch on taboo topics that are inappropriate in everyday conversations or are off-limits for the government to ask. Questions of this sort are seen as an invasion of privacy regardless of the correct answer, and thus risk offending all respondents.

2. The second meaning of sensitivity involves the *threat of disclosure*. The respondent might be concerned about the possible consequences of giving a truthful answer. Common examples in the literature include questions on drug use, criminal behavior, etc. It is important to note that only certain answers (e.g. admitting to having used drugs) are considered sensitive, while others are not. This constitutes a clear distinction between threat of disclosure and intrusiveness.
3. The last meaning of question sensitivity is related to *social desirability* and pertains to the extent to which a question elicits answers that are socially unacceptable or socially undesirable. This conception of sensitivity presupposes the existence of clear social norms regarding a given behavior or attitude. A question is sensitive in this sense if it asks for a socially undesirable answer. Social desirability concerns can be seen as a special case of threat of disclosure, involving social disapproval as the potential consequence of a truthful answer (Tourangeau and Yan 2007).

In order to measure different aspects of item sensitivity, we decided to measure the *converse* of the third meaning of sensitivity mentioned above. The concept of *intrusiveness* therefore captures the inappropriateness of the item's topic, but does not depend on any particular answer alternative. The *threat of disclosure*, on the other hand, pertains to admitting to having done something counter-normative or hold such attitudes: it is not the topic of the item that is necessarily threatening, but the particular answer alternative, if it is chosen by the respondent. A particular item can therefore be intrusive even when none of the answer alternatives are threatening, or vice versa.

The third concept we attempted to measure was that of *overclaiming*. This pertains to highly desirable or socially condoned behavior, like voting, for which *not* reporting the behavior in question is considered contra-normative (Bradburn et al. 1978). We expect items that concern socially condoned behavior to induce *less* item nonresponse and breakoff, as skipping over or refusing to answer such an item could be seen as admitting to a "sin of omission" (not having acted in socially desirable ways, Bradburn et al. 1978).

The raters were presented with the following instructions and asked to rate each item in the GGP questionnaire.

Intrusiveness

Does the question inquire into topics that are inappropriate in everyday conversation, e.g. when talking to a stranger in a waiting room?

1. no, the topic is very casual
2. 3. 4.
5. yes, the topic would be extremely inappropriate in such a situation

Threat of disclosure

Does (at least) one of the possible answers ask the respondent to admit to holding opinions or acting in ways that are not in accordance with generally accepted norms?

1. no
2. yes, weak/moderate norm
3. yes, strong norm

Sub-rating for threat of disclosure

(rated only if a (2) or (3) was given on the threat of disclosure)

Was the positive rating (2 or 3) of the threat of disclosure based on answer alternatives that all pertain to very small proportions of the population?

0. no, at least one of the threatening answers pertains to a substantial proportion
1. yes, all threatening answers pertain to small proportions of the population

Potential for overclaiming

(portraying oneself in an overly positive manner)

Does (at least) one of the possible answers allow the respondent to portray themselves in a more favorable light by claiming to hold opinions or act in ways that are generally considered desirable?

1. no
2. yes, it allows the respondent to portray him/herself in a *somewhat* more favorable light
3. yes, it allows the respondent to portray him/herself in a *very* favorable light

Each item in the GGP questionnaire was independently rated by three expert raters with substantial experience in survey methodology, employed at the Faculty of Social Sciences in Ljubljana. The final rating for the threat of disclosure was obtained by

combining the threat of disclosure rating and the sub-rating. The value of the rating was taken as the final value unless the sub-rating value was (1), indicating that threatening answer categories pertain to small proportions of the population. In this case the value for the final combined rating of threat of disclosure was recoded to (1). The combined rating therefore assumes values (2) or (3) only if the threatening answer alternative pertains to a substantial proportion of the population.

Even though the raters' background was very similar, the inter-rater agreement as reflected by the value of Krippendorff's alpha was quite low. The procedure for calculating Krippendorff's alpha for *ordinal* ratings was employed (see Krippendorff 2004; Hayes and Krippendorff 2007 for details). The lowest inter-rater agreement was found for the measure of threat of disclosure, namely 0.27. This value is low enough to warrant concern for the reliability of the threat of disclosure data. Krippendorff's alpha values for intrusiveness and potential for overclaiming were 0.65 and 0.51 respectively. We will consider these values of alpha high enough to make tentative conclusions on the basis of the ratings.

The mean value of each rating across the three raters was calculated. The resulting variables will be used in the statistical models as item-level predictors. Research has shown that low values of inter-rater agreement are common when expert raters attempt to identify problematic questionnaire items (Olson 2010; DeMaio and Landreth 2003; Presser and Blair 1994). Olson (2010) has demonstrated that, despite the lack of reliability, the average expert ratings successfully identify questionnaire items with a higher item nonresponse rate. We, too, will thus include the threat of disclosure as a predictor in the models despite its low inter-rater agreement, but will interpret its effect with particular caution.

4.3 Respondents' self-assessments of sensitivity and difficulty

According to Beatty and Herrmann, the respondent is faced with two decisions when administered a survey question: whether they *can* respond and whether they *will* respond. Item nonresponse results when the respondent decides negatively in either case (Beatty and Herrmann 2002). Stocke (2006) takes this notion further and speaks of the *cognitive costs* that an item poses, and which are associated with the former decision and *psychological and social costs* that, he argues, are connected with the latter decision.

In order to test hypotheses related to item difficulty and sensitivity, the corresponding

item facets will be included as predictors in the statistical models for nonresponse and breakoff. Relying merely on item-level predictors, however, assumes that a particular item is perceived in the same way by all respondents and has the same effect on them. While we consider the inclusion of such general item-level effects reasonable, we want to consider *respondent-specific* sensitivity and difficulty effects in our analysis. This requires us to include a number of additional items in the questionnaire, inquiring about the particular respondent's perception of items' sensitivity and difficulty.

Asking respondents about the sensitivity and difficulty of *each* item would be extremely impractical, to say the least. To make this task manageable, we need to ask the respondents about the sensitivity and difficulty of certain *groups* of items. We argue that the most reasonable approach is to define the groups according to items *topic*. We therefore asked the respondents to assess how sensitive items concerning certain topics would be for them. In order to assess a given respondent's degree of knowledge about items concerning certain topics, we formulated a number of statements. By measuring the respondent's agreement with these statements we hope to measure this respondent's cognitive state (see Section 2.2.1), which is an important factor in how difficult this item will be for them.

Because the GGP questionnaire was already very long (a typical respondent took about an hour to complete it, see Table 4.9), we were limited in the number of topics that we could inquire about, since each additional question would further prolong the interview. Upon carefully reviewing the items in the GGP questionnaire, we thus decided to include four additional items to measure the respondents' cognitive state about certain item topics, and seven additional items to inquire about the sensitivity of certain item topics.

In an attempt to measure the respondent's cognitive state for certain item topics, the respondent was asked to choose a response on the agreement scale ranging from "strongly agree" to "strongly disagree" for the following statements:

1. *I am familiar with details concerning my partner's job/activity.* The GGP questionnaire contained a module dedicated to recording detailed information on the respondent's partner's occupation and activity. If a respondent disagreed with the aforementioned statement, it is likely that the requested information for this respondent is in a higher cognitive state (generatable or inestimable), and that this respondent will require more effort to construct an answer to items concerning this topic.
2. *I sometimes have problems recalling information like relatives' birth-days.* The GGP questionnaire asked for detailed information on the respondent's partner,

children, other household members, parents, previous partners, and social network alters. The items inquired into birth dates, the date of achieving the respondent's partner's current level of education, the date of marriage and divorce (for previous partners), etc. If the respondent agreed with the aforementioned statement, it is likely that items of this type will require more cognitive effort on their part in comparison to a respondent who readily remembers people's personal information.

3. *I rarely reflect on my relationships with other people.* A number of items throughout the GGP questionnaire asked the respondent to evaluate the satisfaction with the relationship with various people in their life. We argue that a respondent who rarely reflects on relationships will need to generate the answer to such questionnaire items, whereas a respondent who often reflects on their relationships will have this information available.
4. *I am thoroughly familiar with my household's financial situation and transactions.* The questionnaire also inquired into the household's combined income and financial transactions. Again, the argument is that a respondent agreeing with the above statement will require less cognitive effort to respond to such items.

The second battery of items asked the respondent to assess how sensitive they would find answering questions concerning the following topics for the purpose of the survey:

1. my relationships with people and the help and support we provide to each other;
2. my relationship with my partner;
3. my relationship with my children;
4. my relationship with my parents;
5. having (more) children; my and my partner's fertility;
6. my household's income and possessions;
7. my attitude toward issues like marriage, relations between genders, inter-generational relations.

It is clear from how these items are worded that the corresponding variables should not be entered into models as respondent-level predictors. This would mean that, e.g., the respondent's self-assessment of familiarity with their partner's activity would be used as a predictor (of item nonresponse and breakoff) for *all* items, including those items that pertain topics other than the partner's activity. These self-assessments should, rather, be added as respondent-by-item predictors, which were mentioned toward the end of Section 3.2.3.

In order to accomplish this, each item in the GGP questionnaire was coded on whether

or not it concerned each of the eleven topics mentioned in the self-assessment items administered to the respondents. A GGP questionnaire item could concern several topics simultaneously. For example, a particular GGP item asked the respondent to evaluate the agreement with the statement "My parents think that I should have a/another child." This GGP item could be sensitive both because it concerns having children and because it concerns the respondent's relationship with their parents (corresponding to items 4 and 5 in the second self-assessment battery, see Section 2.5). Each GGP item was given code 1 on those topics it concerned and code 0 on those topics it did not concern.

A particular respondent-by-item predictor was constructed by multiplying the respondent's self-assessment with the corresponding item topic code. Before entering this product, the respondents' self-assessments were *recoded*, so that high values corresponded to high cognitive states (for the first battery) and high sensitivity (for the second battery). They were recoded so that the lowest possible cognitive state and the lowest sensitivity self-assessment corresponded to the value of zero. After multiplication:

$$W_{pih} = \text{self-assessment}_{ph} \times \text{topic code}_{ih}, \tag{4.3}$$

the respondent-by-item predictor assumes the value of zero if either 1) the item's topic does not pertain to the self-assessment in question, or 2) the respondent finds this topic "not sensitive at all" or assesses that they have the relevant information in the *available* cognitive state. Indices in (4.3) pertain to person p, item i, and respondent-by-item predictor h. We constructed $H = 11$ such predictors.

4.4 Multiple imputation

This section will describe where missing values appear in the GGP dataset. A general framework for analyzing data in the presence of missing values called *multiple imputation* will be briefly outlined. We will then describe how this procedure was applied in two concrete cases.

When analyzing item nonresponse, missing values do not appear in the response variable: even if the respondent failed to provide an answer to a particular item, this is coded as a 1-response and therefore does not constitute a missing value for our purposes. A similar point holds in the case of breakoff analysis.

We are only faced with the problem of missing data when missingness due to item nonresponse or breakoff occurs for variables that we want to use as *predictors* in our

models. This is the case for items that we added to the GGP questionnaire for the purpose of explaining item nonresponse and breakoff: the respondents' self-assessment of the cognitive state and sensitivity that have been mentioned, and three items on the respondents' attitudes toward surveys.

Another case of missingness occurs because we were not able to obtain information on all interviewers who participated in the GGP. We successfully obtained the information on each interviewer's sex, age, education, and experience with interviewing (measured in months) for 31 out of 36 interviewers, but did not receive this information for the remaining five.

Ad-hoc procedures for dealing with incomplete data like listwise deletion and mean imputation have been shown to be very problematic, as they only produce unbiased estimates with unbiased standard errors if very strict assumptions apply (see Schafer 1997). The motivating idea for multiple imputation is to provide a statistically sound alternative to such ad-hoc procedures. Rather than imputing a missing value once, each missing value is imputed m times (where m is typically a low number like 5 or 10). Multiple imputation is a model-based procedure: a statistical model is specified so that the imputations for a particular incomplete variable are informed by covariates in the model. The MI procedure results in a set of m plausible imputations for each missing value. The variation across the values of the imputations reflects the amount of uncertainty in the missing value under the specified model (see Little and Rubin 2002 for a comprehensive treatment of multiple imputation).

Each set of imputations is used to create a complete dataset. The m completed datasets are then analyzed as if the data were complete to yield *completed data statistics*, which typically encompass estimates with their corresponding standard errors and p-values. The estimates and standard errors are pooled according to so-called Rubin's rules (Little and Rubin 2002). The pooled estimate is simply the average of the m values, while the associated standard error is calculated to take into account both variance within imputations and across imputations.

Multiple imputation can also be employed when data are missing for several variables. Two general approaches for imputing such multivariate incomplete data have emerged: joint modeling and fully conditional specification. Under the first, specifying the imputation model involves specifying the multivariate distribution of the variables in the dataset to be imputed, e.g., multivariate normal, or log-linear (Schafer 1997). This is an attractive method if the multivariate distribution is a reasonable description of the data.

Under the fully conditional specification, on the other hand, the MI model is specified

on a variable-by-variable basis by a set of conditional densities, one for each incomplete variable (van Buuren and Groothuis-Oudshoorn 2011). Rather than specifying the joint multivariate distribution of all variables, this means specifying a (regression) model for each variable to be imputed. This method has been found to work well in a variety of simulation studies (see van Buuren 2012 and references therein).

We will use the fully conditional approach as implemented in the mice package (Multiple Imputation by Chained Equations, van Buuren and Groothuis-Oudshoorn 2011) for R (R Core Team 2013). The semi-parametric *predictive mean matching* (PMM) method will be used because it has the desirable characteristics that the imputations are restricted to the observed values and that it can preserve non-linear relations, even when the structural part of the imputation is wrong (van Buuren and Groothuis-Oudshoorn 2011, for details on PPM, see Little 1988). We first describe the imputation procedure for respondent-level missingness due to item nonresponse and breakoff, and then proceed to do the same for missingness in interviewer-level variables.

4.4.1 Respondent-level predictors

As mentioned, a number of items was added to the GGP questionnaire to serve as predictors in the statistical models. The items' wordings are repeated here, and the scale for each battery is specified. In the figures that follow, the variables that correspond to the items will be referred to by their item names, given in bold before each wording.

The first battery of items was aimed at assessing the respondent's cognitive state for information concerning certain topics. There were five response options: "strongly agree", "agree", "neither agree nor disagree", "disagree", and "strongly disagree".

> To what extent do you agree or disagree with each of the following statements?
>
> **a1203.1** I am familiar with details concerning my partner's job/activity.
> **a1203.2** I sometimes have problems recalling information like relatives' birthdays.
> **a1203.3** I rarely reflect on my relationships with other people.
> **a1203.4** I am thoroughly familiar with my household's financial situation and transactions.

The second battery asked the respondent to assess how sensitive they regarded certain topics that appeared in the GGP survey. The response scale spanned from 1 "very sensitive" to 7 "not sensitive at all" with unlabeled intermediary points.

Even though confidentiality is guaranteed, some survey questions can be more sensitive than others. Please tell me how sensitive you found answering questions concerning each of the following topics for the purpose of this survey.

a1204.1 my relationships with people and the help and support we provide to each other
a1204.2 my relationship with my partner
a1204.3 my relationship with my children
a1204.4 my relationship with my parents
a1204.5 having (more) children; my and my partner's fertility
a1204.6 my household's income and possessions
a1204.7 my attitude toward issues like marriage, relations between genders, intergenerational relations

The last battery of questions concerned the respondent's attitude toward surveys. The wordings were taken from a longer, 16-item instrument for measuring respondents' attitudes toward surveys compiled by Stocke and Langfeldt (2004). Due to concerns that the questionnaire would become excessively long, only three items were retained. The response scale was the same as in the case of the first battery and spanned from "strongly agree"to "strongly disagree".

I'm going to read out three statements that pertain to surveys in general and not merely to the survey you have just participated in. Please tell me to what extent you agree or disagree with these statements, choosing your answer from this card.

a1205.1 Surveys are important for science, politics, and the economy.
a1205.2 Surveys only keep me from doing more important things.
a1205.3 In surveys I have the opportunity to articulate my own opinion.

The author of the present dissertation opted for the three batteries of questions to be positioned *before* the other items of the GGP questionnaire. This way, these variables would be available for respondents who broke off and could be used as predictors of breakoff. The GGP management, however, was concerned that adding such items to the beginning of the questionnaire could introduce adverse context effects, and only agreed to the items' being added *after* the GGP items. As the fielding period was completed, it turned out that not much information was lost because of this, as there was hardly any breakoff in the first two rounds (see Table 4.8).

In the additional third round of data collection, however, the three aforementioned batteries of items were positioned at the very beginning of the questionnaire[8]. If they had not been, this information would not be available for respondents who broke off, thus defeating the purpose of serving as predictors of breakoff.

Even though the item wordings were modified only slightly, the effect of the *position* of these items could be considerable, as the respondent was asked about the cognitive costs and sensitivity of *hypothetical* topics when these items appeared before the GGP items. These same items, when administered after the GGP items, did not have the same hypothetical character, as the respondent could actually remember how sensitive the items concerning a particular topic felt when they were asked to answer them. The respondent's attitude toward surveys was most likely also much more affected by the current survey when they were asked after the fact, as opposed to being asked before the GGP items.

For the reasons outlined above, the correlation structure between the variables might be noticeably different for the Facebook sample, which is why we performed the multiple imputation *separately* for the Facebook sample. The multiple imputation was conducted jointly for all other samples, using the information on mode of administration and round of data collection as covariates in the imputation model.

A number of additional covariates were added to inform the multiple imputation; the respondent's sex, age, education, the (logged) duration of the interview, as well as several items on the respondent's impression of the GGP questionnaire were added to the imputation model. The first questionnaire evaluation item asked about the general pleasantness of the interview, while the others inquired into various aspects of the questionnaire. The scale for the latter battery spanned five points from "definitely not" to "definitely yes".

a1201.1 Overall how did you feel about completing this questionnaire?

5-point scale spanning from "very enjoyable" to "very unpleasant" with unlabeled intermediary points

I'm going to read out five questions about the survey you've just participated in.

a1202.1 Was it difficult to answer the questions?
a1202.2 Were the questions clear?

[8]The item wordings were also modified accordingly e.g. "please tell me how sensitive you *found*..." was changed into "please tell me how sensitive you *would find*..."

a1202.3 Did the questions made you think?
a1202.4 Was the topic interesting?
a1202.5 Was the questionnaire too long?

Some of the self-assessments values are invalid, not because of item nonresponse or breakoff, but because the respondents' life situations were such that the items *did not apply* to them. Items 1203.1 and 1204.2 were only applicable if the respondent had a partner, item 1204.3 was only applicable if the respondent had children, and item 1204.4 was only applicable if the respondent's parents were alive at the time of the survey. The aforementioned items had the explicit answer category "not applicable" available in the Facebook survey. In other modes, the items were administered *after* the GGP items and the information on whether the respondent had a partner, children, and parents was already available. These items were therefore only administered to respondents in the appropriate life situation and skipped for others.

Tables 4.12 and 4.13 show the frequencies of each reason for the missingness of the MI variables by sample. The rows labeled "N.A." in 4.13 refer to "not applicable" in the case of the Facebook sample, and to "not administered" in the case of rounds 1 and 2[9].

Even though these values were missing by design or because the item was not applicable, they were treated as other missing values and imputed. This resulted in values that can only be interpreted as counterfactual scenarios, e.g. how sensitive a respondent would find answering survey questions about their children if they, in fact, had children at the time of interviewing. If such imputed variables were used as predictors in statistical models, this could lead to erroneous inferences.

The self-assessments of cognitive state and sensitivity, however, were never used as respondent-level predictors. They were combined with item-level codes to form respondent-by-item interactions, as we explain in Section 4.3. This means that, e.g., for a respondent with no children, a nonsensical value was imputed for variable a1204.3. This respondent, however, was not administered any questions concerning their children due to questionnaire routing. This means that that the nonsensical imputation was always multiplied by zero, resulting in a vector of zeroes as the respondent-by-item predictor for this particular respondent.

[9]The number of breakoffs in Table 4.13 can exceed the number of breakoffs listed in Table 4.8. The additional breakoffs occurred *after* the last item of the GGP questionnaire was administered. Such a respondent is regarded as completing the questionnaire without breakoff for the purposes of breakoff analysis.

Table 4.12: Missing values for variables a1201.1 and a1202 by sample and reason for missingness

	f2f.pnl	f2f.smp	cati.pnl	cati.smp	web.pnl	web.smp	web.fb
a1201.1							
breakoff	0	1	14	6	12	5	158
INR	0	0	0	0	0	0	1
a1202.1							
breakoff	0	1	14	6	12	5	158
INR	2	0	0	0	1	0	1
a1202.2							
breakoff	0	1	14	6	12	5	158
INR	0	0	0	0	1	1	2
a1202.3							
breakoff	0	1	14	6	12	5	158
INR	0	0	0	0	1	1	2
a1202.4							
breakoff	0	1	14	6	12	5	158
INR	0	0	0	0	1	1	2
a1202.5							
breakoff	0	1	14	6	12	5	158
INR	0	0	0	1	3	1	2

An alternative approach to circumvent nonsensical imputations altogether would have been to model the "not applicable"/"not administered" answer alternative as a distinct answer category. In this case, the scale of the variables could no longer be regarded as numeric, but rather as nominal. A different multiple imputation method would therefore need to be employed (linear discriminant analysis, see van Buuren and Groothuis-Oudshoorn 2011). Explicitly including the "not applicable"/"not administered" category would thus come at the cost of disregarding the information on the order of the categories in the scale. Because the nonsensical imputed values are all multiplied by zero when forming the final predictor variable, it was therefore decided to regard the scales of the variables as numeric.

The mice 2.14 package for R was used to implement the imputation. The full predictor matrix was used, i.e., each variable was used as a predictor for all other variables. Predictive mean matching was utilized as the imputation method, as all the variables to be imputed have a limited range (most often from 1 to 5). We thus avoided imputed values that would fall out of the range or between the points of the scale that would have been produced by other methods for imputing numerical variables.

We executed 20 iterations, which proved sufficient for convergence. The traceplots have been deferred to Appendix A. They show good mixing for all variables in the imputation model, with no apparent trends for the last ten iterations. We imputed each missing value *five times*. A higher number of imputations would, of course, have

Table 4.13: Missing values for variables a1203, a1204, and a1205 by sample and reason for missingness

	f2f.pnl	f2f.smp	cati.pnl	cati.smp	web.pnl	web.smp	web.fb
a1203.1							
breakoff	0	1	14	6	12	5	0
INR	0	0	0	0	0	1	6
N.A.	54	35	52	12	59	8	32
a1203.2							
breakoff	0	1	14	7	12	5	0
INR	1	0	0	0	0	1	8
a1203.3							
breakoff	0	1	14	7	12	5	0
INR	0	0	0	0	1	1	8
a1203.4							
breakoff	0	1	14	7	12	5	0
INR	0	0	0	0	1	1	8
a1204.1							
breakoff	0	1	14	7	12	5	0
INR	0	0	1	0	1	0	4
a1204.2							
breakoff	0	1	14	7	12	5	0
INR	0	0	0	0	0	0	6
N.A.	54	35	52	11	60	8	22
a1204.3							
breakoff	0	1	14	7	12	5	0
INR	0	0	0	0	0	0	11
N.A.	91	32	73	7	94	19	62
a1204.4							
breakoff	0	1	14	7	12	5	0
INR	0	0	0	0	0	0	6
N.A.	41	42	33	24	32	6	5
a1204.5							
breakoff	0	1	14	7	13	5	0
INR	0	0	0	0	0	1	7
a1204.6							
breakoff	0	1	14	7	13	5	0
INR	1	4	0	0	0	1	6
a1204.7							
breakoff	0	1	14	7	13	5	0
INR	0	0	1	0	0	1	6
a1205.1							
breakoff	0	1	14	7	13	5	0
INR	0	0	0	1	2	0	3
a1205.2							
breakoff	0	1	14	7	15	5	0
INR	0	0	0	1	0	1	5
a1205.3							
breakoff	0	1	14	7	15	6	0
INR	0	0	0	0	0	0	6

been desirable to lower the simulation component of the total MI variance. This rather low number of multiple imputations was chosen because each model involving imputed predictors would have to be fit $m = 5$ times. This would considerably increase the time for item nonresponse analyses.

Figure 4.1 shows the smoothed distribution for each of the imputed variables in the MI model[10]. The thick blue line corresponds to the estimated (kernel smoothed) density distribution of the observed data points (with missing values removed), while each of the five superimposed red lines corresponds to the density of the imputed data for a particular MI dataset. Figure 4.1 is the diagnostic plot for the imputation model used for all samples other than web.fb, while the analogous plot for the Facebook sample is deferred to Appendix A (Figure A.1).

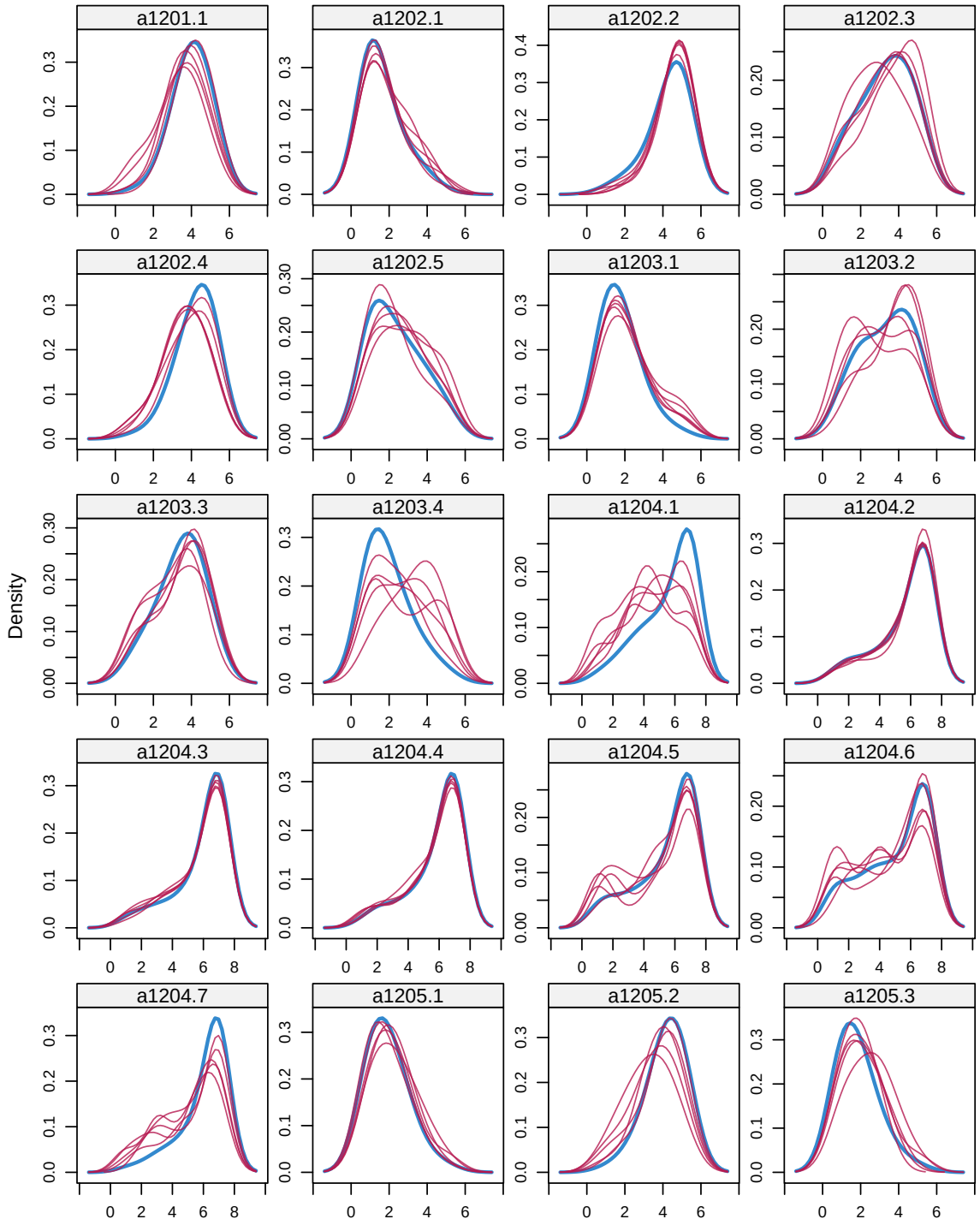
Differences in the densities between the observed and imputed values may suggest a problem with the imputation model that needs to be further checked (van Buuren and Groothuis-Oudshoorn 2011). In the present example, the marginal density for the imputed values quite closely follows the density for the observed data thus revealing no potential problems.

Items a1205.1 through a1205.3 will be combined into a scale that we will refer to as the respondent's *attitude toward surveys*. The scale variable is formed by calculating the mean of the three items with the second item inverted (so that the value 5 is recoded into 1 etc.). Cronbach's alpha for the three-item scale is quite low: 0.61[11]. However, as the three items were a subset of a 16-item instrument with a Chronbach's alpha of 0.73 (Stocke and Langfeldt 2004), we will assume that the value of alpha for the three-item scale is sufficiently high for tentative conclusions on the basis of the scale.

[10]Demographic variables, as well as mode and round of data collection, do not appear in the figure because they have no missing values

[11]This is the average value across the $m = 5$ complete-data statistics for the combined sample of the first and second rounds (excluding the Facebook sample).

Figure 4.1: Kernel density plot for marginal distributions of observed data (blue) and the five densities per variable calculated from imputed data (red); all samples except web.fb



4.4.2 Interviewer-level predictors

As mentioned, we were not able to obtain the demographics and experience information for 5 out of 36 interviewers. The data show that no interviewer worked in both modes of administration in the GGP survey: some interviewers only performed face-to-face interviews, while others only conducted interviews over the phone. All missing information pertains to interviewers in CATI mode. Tables 4.14 and 4.15 show the interviewer-level data.

Table 4.14: Interviewer item nonresponse percentage, demographics, and workload for F2F interviewing

	% INR	n respondents	male	age	education	experience
	1.57	1	1	20.5	4	1
	1.04	2	0	25.2	5	24
	1.63	2	0	19.8	5	24
	2.09	3	0	22.3	5	12
	1.36	7	0	28.2	5	87
	0.33	7	1	33.8	8	6
	0.65	7	1	60.7	5	6
	0.96	9	0	30.8	5	72
	0.43	11	1	24.5	5	1
	0.41	18	0	47.8	4	19
	2.78	30	0	34.7	5	36
	1.83	31	0	21.1	4	12
	0.68	31	0	29.6	9	24
	0.57	34	0	35.1	4	17
	0.34	45	1	32.2	7	29
	0.95	75	1	25.2	4	16
mean	1.10	19.6	0.38	30.7		24.1

The first column of Table 4.14 shows the percentage of item nonresponse, calculated at interviewer level. This is lower than 1% for most interviewers, but for two interviewers the percent of item nonresponse exceeds 2%. The table's rows have been sorted according to the interviewer workload, shown in the second column, which varies from 1 to 75 with a mean of 19.6 interviews per interviewer. Most of the face-to-face interviewers were women (62%). The interviewers' average age was 30.7 years, on average they had two years of interviewing experience (24.1 months). All interviewers had at least a middle education[12].

The data for telephone interviewers contains missing data. The values in the bottom

[12]The answer alternatives for education were: 1) no education, incomplete primary; 2) primary; 3) lower or middle vocational; 4) middle technical; 5) middle general; 6) higher; 7) higher technical; 8) university; 9) msc, phd.

Table 4.15: Interviewer item nonresponse percentage, demographics, and workload for CATI

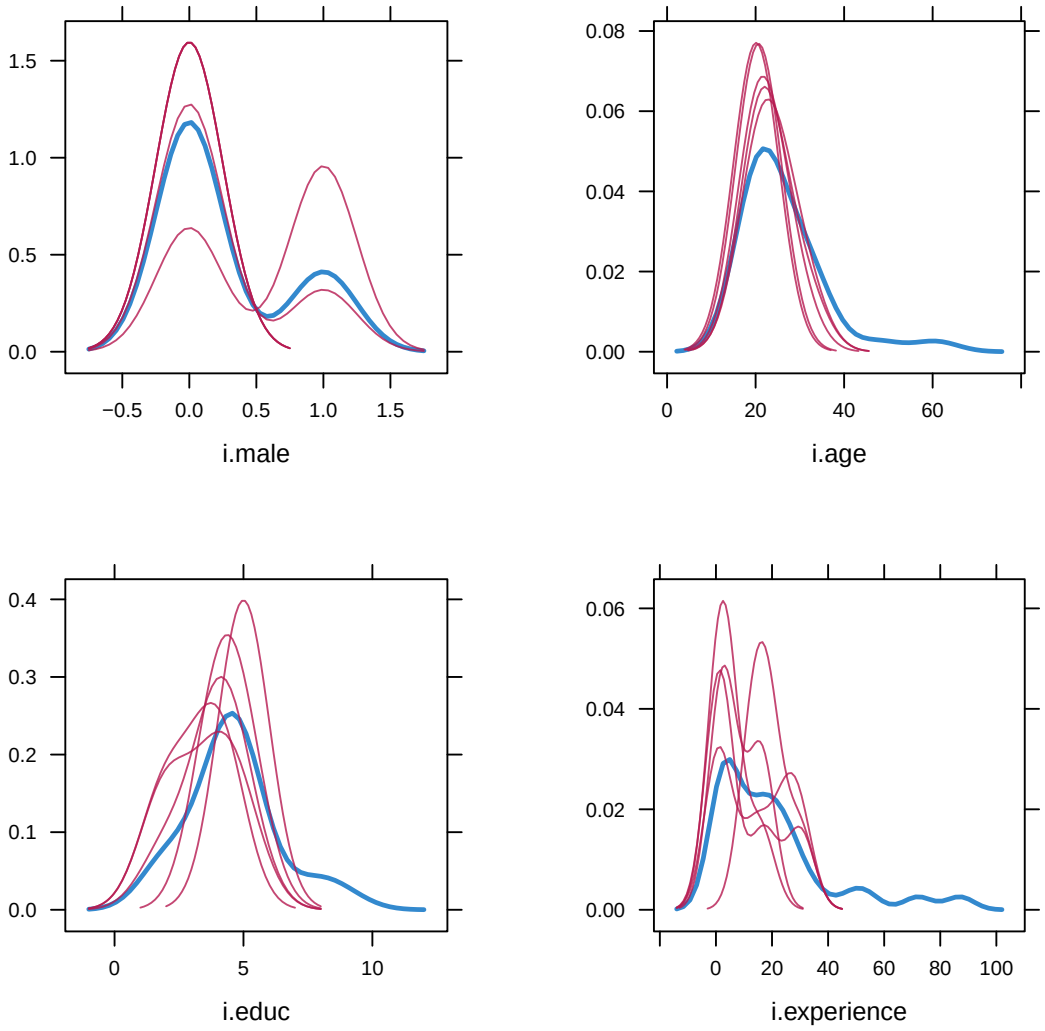
	% INR	n respondents	male	age	education	experience
	0.36	1				
	0.40	1				
	1.42	1				
	0.34	2	1	28.2	8	54
	0.33	5	0	19.7	5	1
	0.74	5	0	22.7	5	48
	0.59	5	0	30.6	4	24
	0.96	7	1	20.5	2	19
	0.78	8	0	21.5	4	2
	1.89	8	0	18.9	4	30
	1.62	9	0	21.9	5	6
	0.50	11	0	17.5	2	3
	0.94	13				
	0.98	15				
	0.50	16	0	23.2	5	14
	1.42	21	0	17.2	2	2
	1.16	21	0	20.8	4	6
	0.81	28	0	17.3	2	5
	1.48	40	0	19.2	2	1
	1.74	51	0	22.0	4	17
mean	0.95	13.4	0.13	21.4		15.5

row have been calculated by excluding the missing values. The percentage of item nonresponse for CATI interviewers does not vary as widely as for F2F interviewers. The interviewer-level percentage of item nonresponse is positively correlated to the workload[13]: those interviewers who interviewed more respondents showed, on average, more item nonresponse. The prevalence of women was even stronger among CATI interviewers (87% women). The CATI interviewers were on average younger (with mean age of 21.4 years) and less experienced (mean experience of 15.5 months) than F2F interviewers, and their education was also at a lower level.

We imputed the missing interviewer-level information by means of multiple imputation, using chained equations with predictive mean matching as our method. We used the full predictor matrix, so each variable shown in Tables 4.14 and 4.15 was used to predict all other variables (the mode of administration was also used as a predictor). We produced $m = 5$ imputed datasets. After 20 iterations, the traceplots of the mean and standard deviation for the imputed values were examined (see Figure A.6 in Appendix A). They show good mixing for all variables in the model and no apparent

[13]The correlation between the interviewer-level percentage of item nonresponse and the number of interviewed respondents is .51 for the CATI interviewers as compared to -0.12 for F2F interviewers.

Figure 4.2: Kernel density plot for marginal distributions of observed data (blue) and the five densities per variable calculated from imputed data (red); interviewer sex, age, education, and experience in months



trends.

Figure 4.2 shows the smoothed distribution for each of the imputed variables in the MI model for interviewer-level variables. The five superimposed red lines, which correspond to the density of the imputed data, approximately follow the contours of the blue line, which corresponds to the density distribution of the observed data points. The discrepancies are, however, larger than for the imputation of respondent-level data (see Figure 4.1).

This was expected, as for five interviewers *all* information is missing except for workload and average item nonresponse. If this were not the case, the information on an interviewer's age (assuming it were available) could, e.g., be used to narrow the range

of reasonable values for this interviewer's education and experience. Because of the particular pattern of missingness, however, the imputed values vary widely, as not much is known about a particular interviewer whose information we failed to obtain.

The MI procedure, however, insures that the imputed values are in the right range. The fact that, e.g., the imputed age for a particular interviewer varies widely across the $m = 5$ imputations simply reflects the high uncertainty in this value under the imputation model. This variability is reflected in a high between-imputation variance and is, after pooling the results of the analysis performed on each completed dataset, finally expressed through increased standard errors of estimates.

What we wish to stress here is that, even though the MI model cannot predict the missing values well, the uncertainty about missing values is finally reflected in the estimates' increased standard errors. We argue that this is a much sounder approach to analyzing the data in the presence of missing values than mere ad-hoc procedures like mean imputation or listwise deletion. In the case of mean imputation (or any other kind of single imputation), the imputed values would be treated as known, and the standard error of the completed data analysis estimates would be underestimated as a consequence. If we were to delete the five interviewers with missing information (listwise deletion), on the other hand, we would also need to delete all 31 respondents that were interviewed by these five interviewers: a substantial loss of sample size (see Schafer 1997 for a longer discussion of the disadvantages of ad-hoc procedures).

4.5 Operational hypotheses

Having provided a thorough description of the data, we proceed to operationalize the hypotheses put forward in Section 2.5. As mentioned, survival analysis methods will only be applied to the data additionally gathered in the third round of data collection, with the result that we will be able to test more detailed operational hypotheses about the mode of administration for item nonresponse, but will be limited in testing such hypotheses regarding breakoff.

We will first repeat each hypothesis put forward in Section 2.5, and then describe how we intend to test it in the empirical part. An account will be given of the results that are expected under each hypothesis, as well as what will be considered as evidence in favor of the hypothesis in question. We will consider a p-value lower than 0.05 to indicate that the sign of the effect is stable and not just an artifact of small sample size. Operational hypotheses will be made separately for item nonresponse and breakoff. The suffix "a" will be used for operationalized hypothesis concerning item

nonresponse, and suffix "b" will be used for breakoff.

Hypothesis 1: Item nonresponse and breakoff will be most common in web mode and least common in face-to-face mode.

The generalized linear mixed model for item nonresponse will include mode of administration as a predictor. We will regard face-to-face as the baseline for comparison, because this mode of administration was used in nearly all previous implementations of the GGP survey. One of the purposes of the pilot study, which provided the dataset we will use in our analyses, was to investigate whether alternative modes could be used. We expect that the estimated coefficient for web mode will be positive and statistically significant (in comparison to F2F mode). We also expect the estimated coefficient for telephone mode to be positive, but lower than the coefficient for web mode.

Hypothesis 1a.1: a positive and significant effect of web mode in the model for item nonresponse.

Hypothesis 1a.2: a positive effect of CATI mode, but lower than that of web mode in the model for item nonresponse.

Because breakoff was so uncommon in the first and second rounds of data collection, we will not attempt to apply survival analysis methods to the full dataset including all modes. We will rather use logistic regression to model the respondent's probability of breakoff and include the mode of administration as a predictor. Our operational hypotheses for breakoff mirror those for item nonresponse.

Hypothesis 1b.1: a positive and significant effect of web mode in the logistic regression model.

Hypothesis 1b.2: a positive effect of CATI mode in the logistic regression model, but lower than the effect of web mode.

Hypothesis 2: Item nonresponse and breakoff will be more common for cognitively less sophisticated respondents. This effect will be even more pronounced in web mode.

We follow the rationale that age and education can serve as proxy measures for cognitive sophistication. Krosnick defines cognitive sophistication as "the ensemble of abilities needed to retrieve information from memory and integrate that information

into verbally expressed summary judgments" (Krosnick 1991). The reliance on proxies can be justified, as it has been shown that educational attainment strongly correlates to more direct measures of cognitive sophistication (Krosnick and Alwin 1987) and that cognitive ability diminishes with old age (see Knäuper et al. 1997 and references therein).

The GLMM for nonresponse will include person's characteristics as predictors. We expect to find a positive effect of age and a negative effect of education. If the effects are statistically significant at the standard alpha=.05 level, we will consider this as evidence in support of our hypothesis. We will also include among the model predictors the interaction of the web mode indicator with the respondent's age and education. We expect that the interaction effect for age will be positive, and the interaction effect for education will be negative, indicating that these respondent characteristics exert an even stronger effect in web mode.

Hypothesis 2a.1: a positive and significant effect of respondent age in the model for item nonresponse.

Hypothesis 2a.2: a negative and significant effect of respondent education in the model for item nonresponse.

Hypothesis 2a.3: a positive and significant effect of the interaction of respondent age and web mode in the model for item nonresponse.

Hypothesis 2a.4: a negative and significant effect of the interaction of respondent education and web mode in the model for item nonresponse.

As we will only apply the Cox PH model to data gathered in web mode, the operational hypotheses for respondent age and education do not involve an interaction with mode.

Hypothesis 2b.1: a positive and significant effect of respondent age in the Cox proportional hazards model for breakoff.

Hypothesis 2b.2: a negative and significant effect of respondent education in the Cox PH model for breakoff.

Hypothesis 3: Item nonresponse and breakoff will be less common among respondents with a more positive attitude toward surveys in general.

The attitude toward surveys will be included as a respondent-level predictor in the GLMM for item nonresponse, and as a time independent predictor in the Cox PH

model for breakoff. We expect a negative effect, indicating that a positive attitude toward surveys lowers the probability of item nonresponse and risk of breakoff.

Hypothesis 3a: a negative and significant effect of the attitude toward surveys in the model for item nonresponse.

Hypothesis 3b: a negative and significant effect of the attitude toward surveys in the model for breakoff.

Hypothesis 4: Items that are sensitive or present a threat of disclosure will induce more item nonresponse and breakoff. This effect will be less pronounced in web mode.

The models for item nonresponse and breakoff will include as predictors both expert ratings of the sensitivity of each item, as well as respondent-specific measures of the sensitivity of certain topics. We expect both types of predictors to have positive effects in the models, indicating that sensitive and threatening items increase the probability of item nonresponse and risk of breakoff. The GLMM for item nonresponse will also include interactions of these predictors with web mode. We expect the interaction effects to be negative.

Hypothesis 4a.1: a positive and significant effect of both item-level and respondent-specific measures of item sensitivity in the model for item nonresponse.

Hypothesis 4a.2: a negative and significant effect of the interaction between item-level and respondent-specific measures of item sensitivity with the web-mode indicator in the model for item nonresponse.

Hypothesis 4b: a positive and significant effect of both item-level and respondent-specific measures of item sensitivity in the Cox PH model for breakoff.

Hypothesis 5: Items that are complex or deal with topics that the respondent is less familiar with will induce more item nonresponse and breakoff. This effect will be even more pronounced in web mode.

The statistical models will include as predictors two types of measures aimed at capturing how difficult an item is for the respondent: objective measures like the number of words in the item's wording and the number of answer alternatives, as well as the respondents' self-assessments of their cognitive state for certain item topics. Both types of predictors are expected to have positive effects in the models, indicating

that complex items and items dealing with topics that the respondent is less familiar with increase the probability of item nonresponse and risk of breakoff. We will, again, include interactions of these predictors with web mode in the model for item nonresponse, and this time expect to find positive interaction effects.

Hypothesis 5a.1: a positive and significant effect of both item-level and respondent-specific measures of item difficulty in the model for item nonresponse.

Hypothesis 5a.2: a positive and significant effect of the interaction between item-level and respondent-specific measures of item difficulty with the web-mode indicator in the model for item nonresponse.

Hypothesis 5b: a positive and significant effect of both item-level and respondent-specific measures of item difficulty in the Cox PH model for breakoff.

5 Item nonresponse analysis

Having described how the data were collected and adjusted for missing values, we proceed to analyze the GGP dataset. The present chapter focuses on item nonresponse analyses, while the subsequent one analyzes breakoff. We first investigate the relevant descriptive statistics in Section 5.1 and also examine charts of bivariate relations of item nonresponse to explanatory variables that will be used as predictors in the statistical models. Section 5.2 presents the results of fitting a separate model for each dataset. The main part of the chapter consists of Sections 5.3, 5.4, and 5.5, which describe and interpret the results of statistical models applied to item nonresponse. We discuss the results and evaluate the hypotheses in the final section.

5.1 Preliminary analyses

Before fitting statistical models for item nonresponse, we will examine the relevant descriptive statistics. Tables 5.1 and 5.2 give the absolute and relative frequencies for various response categories separately for each sample. The first two columns refer to the substantive responses that were obtained. These are given separately for required and non-required items. Particular attention is given to *required* items, as they will be *excluded* from our analysis of item nonresponse.

Table 5.1: Response composition by sample; absolute frequencies

	substantive response		item nonresponse				
	not req.	required	skipped	refusal	DK	Total	n resp.
f2f.pnl	49007	15275	391	61	91	64825	206
f2f.smp	23122	7604	292	33	100	31151	107
cati.pnl	47896	14878	582	18	110	63484	209
cati.smp	11966	4111	203	15	18	16313	59
web.pnl	54489	16646	1103	109	273	72620	228
web.smp	9047	2913	370	50	57	12437	45
web.fb	37476	13686	1346	91	298	52897	262

Because the GGP questionnaire had quite an elaborate routing scheme, it was necessary for the respondent to provide an answer at least to certain *filter* questions before proceeding with the interview. To repeat the previously given example, each respondent was asked "Were you born in Slovenia?" If they answered affirmatively,

the following item inquired into the municipality of birth, otherwise another item was administered asking about the country of birth. The first item serves as a filter that determines which item will be administered next. Given that the routing depends on the answers to filter items, an error message was displayed if the respondent tried to proceed to the next item without responding to a filter item. Such "hard controls" were imposed *only* for filter items, i.e., the respondent could skip all non-filter items. The same procedure was used in interviewer-administered modes: the interview could not continue before an answer to a required item was given.

Table 5.2: Response composition by sample; weighted percentages

	substantive response		item nonresponse				
	not req.	required	skipped	refusal	DK	Total	n resp.
f2f.pnl	75.60	23.56	0.60	0.09	0.14	100.00	206
f2f.smp	74.23	24.41	0.94	0.11	0.32	100.00	107
cati.pnl	75.45	23.44	0.92	0.03	0.17	100.00	209
cati.smp	73.35	25.20	1.24	0.09	0.11	100.00	59
web.pnl	75.03	22.92	1.52	0.15	0.38	100.00	228
web.smp	72.74	23.42	2.97	0.40	0.46	100.00	45
web.fb	70.85	25.87	2.54	0.17	0.56	100.00	262

We use another term in the context of item nonresponse analysis—*required* item—which is not quite synonymous with *filter* item. What we mean by "required" is that a substantive response was mandatory before proceeding. While a filter item cannot be *skipped* by the respondent, item nonresponse to a filter item is still possible if the respondent is allowed to answer "don't know" or refuse to answer. We thus consider as *required* filter items with no explicit refusal or "don't know" response alternatives. As the second column of Table 5.2 shows, about a quarter of the administered items were required. The proportion of required items varies somewhat across samples, because certain filter items were nested within others, making the proportion of required items dependent on the concrete responses given by respondents in a particular sample.

Before continuing, we would like to elaborate on two alternative ways of calculating proportions (of required items, item nonresponse, etc.). The absolute frequencies in Table 5.1 (except for the rightmost column) refer to the number of *measurement occasions*. In the first round, e.g., a total of 64825 items were administered in face-to-face mode (see first row labeled f2f.pnl). Of these, 49007 were substantive responses to non-required items, 15275 were required items etc. The percentages in the first row of Table 5.2 refer to the proportion of measurement occasions within the total of 64825. An alternative way of calculating these proportions would be to first determine the

proportion for each respondent and then average this figure across respondents.

The two aforementioned ways of calculating proportions do not, in general, yield the same results. The reason for this is that the number of items administered to a respondent varies across respondents due to routing and breakoff. If we first calculate the proportion for each respondent and then average across respondents, we are ascribing the same weight to each respondent. A respondent who broke off after ten items will have the same weight as a respondent completing the survey. If we, however, calculate the proportion of particular kinds of measurement occasions within all measurement occasions, we are effectively weighting the respondents with regard to the number of items they were administered:

$$\bar{Y} = \frac{\sum_{p=1}^{P} W_p \bar{Y}_p}{\sum_{p=1}^{P} W_p}. \tag{5.1}$$

In Equation (5.1) $\bar{Y}$ is the overall proportion, $\bar{Y}_p$ is the proportion for respondent p, P is the total number of respondents, and W_p is the weight corresponding to respondent p. Calculating the proportion for each respondent and then averaging across respondents would mean setting the same weight for all respondents (e.g. $W_p = 1$ for all p). Calculating the proportion of measurement occasions of a particular kind within all measurement occasions, on the other hand, corresponds to setting W_p equal to the number of items administered to respondent p. We will refer to the former type of proportion as *unweighted* and to the latter as *weighted*. We will prefer weighted proportions as they take into account the number of administered items.

Continuing with the description of Tables 5.1 and 5.2, columns three though five refer to different categories of item nonresponse. As mentioned, the way item nonresponse is defined depends on the goal of the particular substantive analysis. A "don't know" to a question on voting preference can be regarded as a meaningful answer, while the same reply on an item inquiring into income has no informational value (de Leeuw et al. 2003). For the purposes of our analysis *we will consider "don't know" answers, as well as refusals and skipped items, as item nonresponse*. We argue this definition would suit the majority of substantive analyses that could be performed on the GGP data, although alternative definitions for specific items certainly could be considered.

Table 5.3 gives the composition of item nonresponse under this definition. The great majority, about three quarters, of item nonresponse stems from skipped items. This proportion is even higher under telephone administration where it exceeds 80%. The remainder of item nonresponse is accounted for by "don't know" answers and refusals.

Table 5.3: Item nonresponse composition by sample; weighted percentages

	skipped	refusal	don't know	Total	n INR	n resp.
f2f.pnl	72.01	11.23	16.76	100.00	543	206
f2f.smp	68.71	7.76	23.53	100.00	425	107
cati.pnl	81.97	2.54	15.49	100.00	710	209
cati.smp	86.02	6.36	7.63	100.00	236	59
web.pnl	74.28	7.34	18.38	100.00	1485	228
web.smp	77.57	10.48	11.95	100.00	477	45
web.fb	77.58	5.24	17.18	100.00	1735	262

These answer alternatives were explicitly given only for those items where refusal was considered likely (e.g. income), or "don't know" was considered a meaningful answer (e.g. the question on the father's occupation when the respondent was 15 years old). A lower proportion of item nonresponse is accounted for by refusals, because fewer items had this answer alternative available as compared to "don't know."

In all subsequent analyses, required items have been excluded, because they provide no information on the respondents' tendency to produce item nonresponse. Required items do not have any item nonresponse precisely due to the fact that the respondent was *required* to provide a substantive answer. When we speak in this section of, e.g., the proportion of item nonresponse, this pertains to the share of skipped items, refusals, and "don't know" answers *within non-required items.*

Table 5.4 shows how much item nonresponse can be attributed to extreme respondents. In the case of face-to-face administration in the first round, 5% of respondents with the highest item nonresponse rate account for 32.9% of the total item nonresponse; 10% of respondents account for 42.4% of item nonresponse, etc. The figures vary by sample, but in general convey the message that a small number of respondents account for a rather large proportion of item nonresponse. In other words, item nonresponse is not uniformly distributed across respondents, but is quite concentrated in a number of extreme individuals.

We proceed by examining bivariate relations between item nonresponse and other variables: we are interested in how the proportion of item nonresponse varies with the levels of each variable that we later intend to use as a predictor in the statistical models. We will report the proportion of item nonresponse separately for each sample, as it is possible that the relation between a certain predictor variable and item nonresponse might differ across samples. The following variables will be used as predictors in the models for item nonresponse:

Table 5.4: Weighted percentage of item nonresponse accounted for by proportion of respondents given in the first column (by sample)

respondents	f2f.pnl	f2f.smp	cati.pnl	cati.smp	web.pnl	web.smp	web.fb
5%	32.9	17.7	15.3	13.4	23.4	34.2	24.9
10%	42.4	24.8	24.9	22.3	30.5	22.6	36.8
20%	57.6	41.8	42.4	36.4	47.9	31.7	52.7
30%	65.4	54.7	55.7	44.1	59.9	35.6	63.3
50%	77.5	71.7	74.7	58.3	75.4	44.2	77.4

- **Interviewer characteristics**[14]**:** sex, age, education, and experience.
- **Respondent characteristics:** sex, age, education, attitude toward surveys.
- **Item facets:**
  - **Measures of item sensitivity:** intrusiveness, threat of disclosure, potential for overclaiming (see Section 4.2).
  - **Measures of item complexity:** number of words in the item's wording, number of answer alternatives.
  - **Item format:** numeric input, string input, radio button for yes/no.
- **Item-by-respondent interactions:** self-assessments of cognitive state and sensitivity to items concerning certain topics (see Section 4.3).

We will first examine the relation of item nonresponse with regard to basic respondent demographics: sex, age and education. For the purposes of presentation in the tables, respondent age was categorized into five categories. Education was categorized into low, middle, and high[15]. Tables 5.5 and 5.6 give the unweighted and weighted percentages respectively. The total unweighted percentages (bottom row) are higher than corresponding weighted percentages, indicating that respondents who were administered a lower number of items have on average a higher percentage of item nonresponse. The inner cells of the tables give the percentage for certain categories of respondents.

The trends in the proportion of item nonresponse are better discerned when plotted as in Figure 5.1. Each line in the plots denotes a particular sample. The same scheme is used in all subsequent plots: red lines denote face-to-face mode, green CATI mode, and blue web mode; thick solid lines correspond to the first round while thin dashed lines correspond to the second round. The Facebook sample of the third round is denoted by a thick black line.

[14]We will not analyze the bivariate relationship of item nonresponse to interviewer characteristics because of very low sample size (36 interviewers).

[15]The answer alternatives for education were: 1) no education, incomplete primary; 2) primary; 3) lower or middle vocational; 4) middle technical; 5) middle general; 6) higher; 7) higher technical; 8) university; 9) msc, phd. Categories 1 and 2 constitute *low* education, categories 3 through 5 constitute *middle* education, and categories 6 through 9 constitute *high* education.

Table 5.5: Unweighted percentage of item nonresponse by sample and basic demographics

	f2f.pnl	f2f.smp	cati.pnl	cati.smp	web.pnl	web.smp	web.fb
sex							
female	1.21	2.06	2.19	2.55	2.50	6.00	5.80
male	1.05	1.69	1.71	2.41	3.16	11.44	6.54
age							
18-25	0.64	1.13	1.07	0.43	1.96	11.71	7.03
26-35	0.96	1.28	1.23	2.62	2.68	2.83	5.22
36-45	0.93	2.14	1.68	1.73	2.09	1.49	5.35
46-55	0.99	2.09	2.55	3.21	3.17	2.69	3.97
56+	2.02	2.24	3.17	2.73	4.60	17.02	4.11
education							
low	0.32	2.33	2.35	0.51	3.65	21.81	7.40
middle	1.18	1.78	1.88	2.84	2.82	7.26	6.37
high	1.08	1.67	2.09	2.17	2.77	5.26	4.81
Total	1.13	1.89	1.98	2.50	2.82	8.18	5.99

Respondent sex is plotted in a similar way to continuous variables. This is somewhat inappropriate, as the lines suggest that there are intermediary values between the categories of male and female. We nonetheless provide such plots for more convenient interpretation. The horizontal lines suggest that there are no differences in the proportion of item nonresponse across respondent sex. The lines increase slightly from left to right for web.pnl and the Facebook sample, indicating that there is more item nonresponse among males in those samples.

There is a slight positive trend apparent for respondent age, except for the Facebook sample where the correlation of age and proportion of item nonresponse seems to be negative. There are no consistent trends of respondent education apparent in the lower left panel of Figure 5.1.

The lower right panel of the figure plots the percentage of item nonresponse against the respondents' attitude toward surveys. Higher values denote a more positive attitude toward surveys. This scale was categorized into four categories as denoted by the x-axis labels. The items that make up the scale had some missingness that was addressed by multiple imputation as described in Section 4.4. In the presence of missing values, the x-axis value was determined by averaging the imputed values across the five imputed datasets. The proportion of item nonresponse does decrease with increasingly positive attitude toward surveys, as expected. This is not consistent for all samples, though, as f2f.smp and the Facebook sample seem to exhibit an increasing trend.

Table 5.6: Weighted percentage of item nonresponse by sample and basic demographics

	f2f.pnl	f2f.smp	cati.pnl	cati.smp	web.pnl	web.smp	web.fb
sex							
female	1.14	1.89	1.47	1.93	2.26	5.00	4.32
male	1.05	1.71	1.45	1.94	3.08	5.03	4.74
age							
18-25	0.66	1.19	1.08	0.43	1.74	3.41	4.73
26-35	0.96	1.28	1.18	1.61	2.39	2.80	4.16
36-45	0.88	2.12	1.39	1.05	1.95	1.82	5.10
46-55	0.98	2.07	1.77	2.98	3.10	2.62	3.54
56+	1.98	2.02	1.81	2.24	4.76	14.01	3.56
education							
low	0.31	2.12	1.42	0.51	3.24	3.70	4.81
middle	1.16	1.73	1.47	2.06	2.75	6.76	4.43
high	1.02	1.70	1.45	1.99	2.51	4.10	4.36
Total	1.10	1.80	1.46	1.93	2.65	5.01	4.43

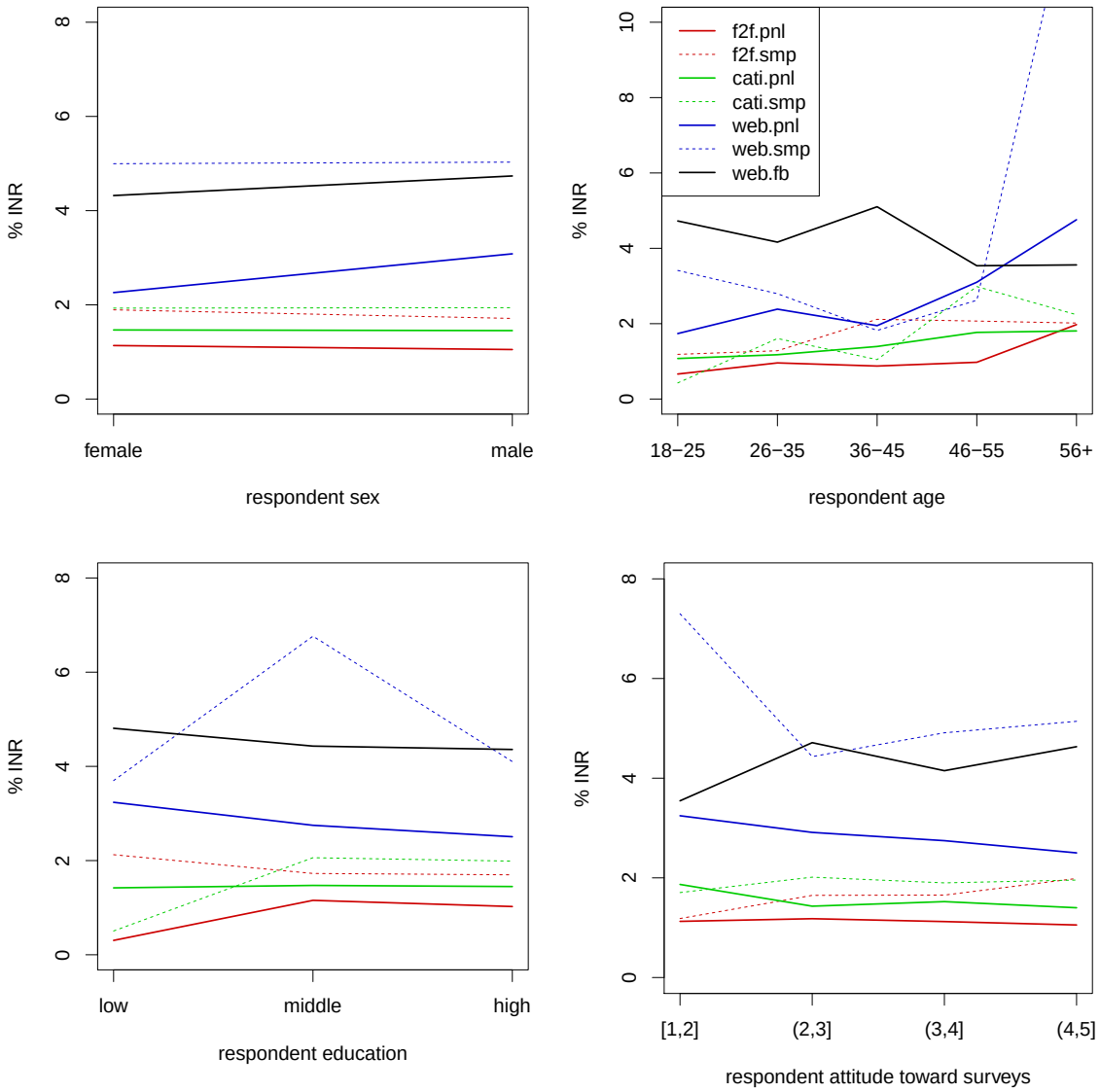
Figure 5.2 shows the proportion of item nonresponse plotted against measures of item sensitivity. The mean value across the three raters was computed and categorized as denoted by the categories marked on the x-axis. The trends are consistent across the samples and have the expected direction: item nonresponse increases with item intrusiveness and threat of disclosure, and decreases with potential for overclaiming.

One feature of the influence of intrusiveness plotted in the top left panel of Figure 5.2 is particularly striking. The proportion of item nonresponse is constant for low and intermediate values of intrusiveness and dramatically increases at values exceeding 4 (intrusiveness was rated on a scale from 1 to 5). A similar pattern is discernible in the plot for the item's threat of disclosure: the item nonresponse percentage is constant for low values of the threat of disclosure, and spikes at values exceeding 2 (the threat of disclosure was rated on a scale from 1 to 3).

This is valuable information to be used when modeling item nonresponse; rather than including intrusiveness as a continuous predictor in the model, this variable should be dichotomized (0 - low values; 1 - values exceeding 4.0) before being entered into the model for item nonresponse. The threat of disclosure will also be dichotomized before being entered into the model as a predictor (0 - low values; 1 - values exceeding 2.0)

Figure 5.3 depicts the bivariate relations of measures of item complexity and the weighted percentage of item nonresponse. The number of words in the item wording has been categorized into intervals of unequal width, with narrower categories for

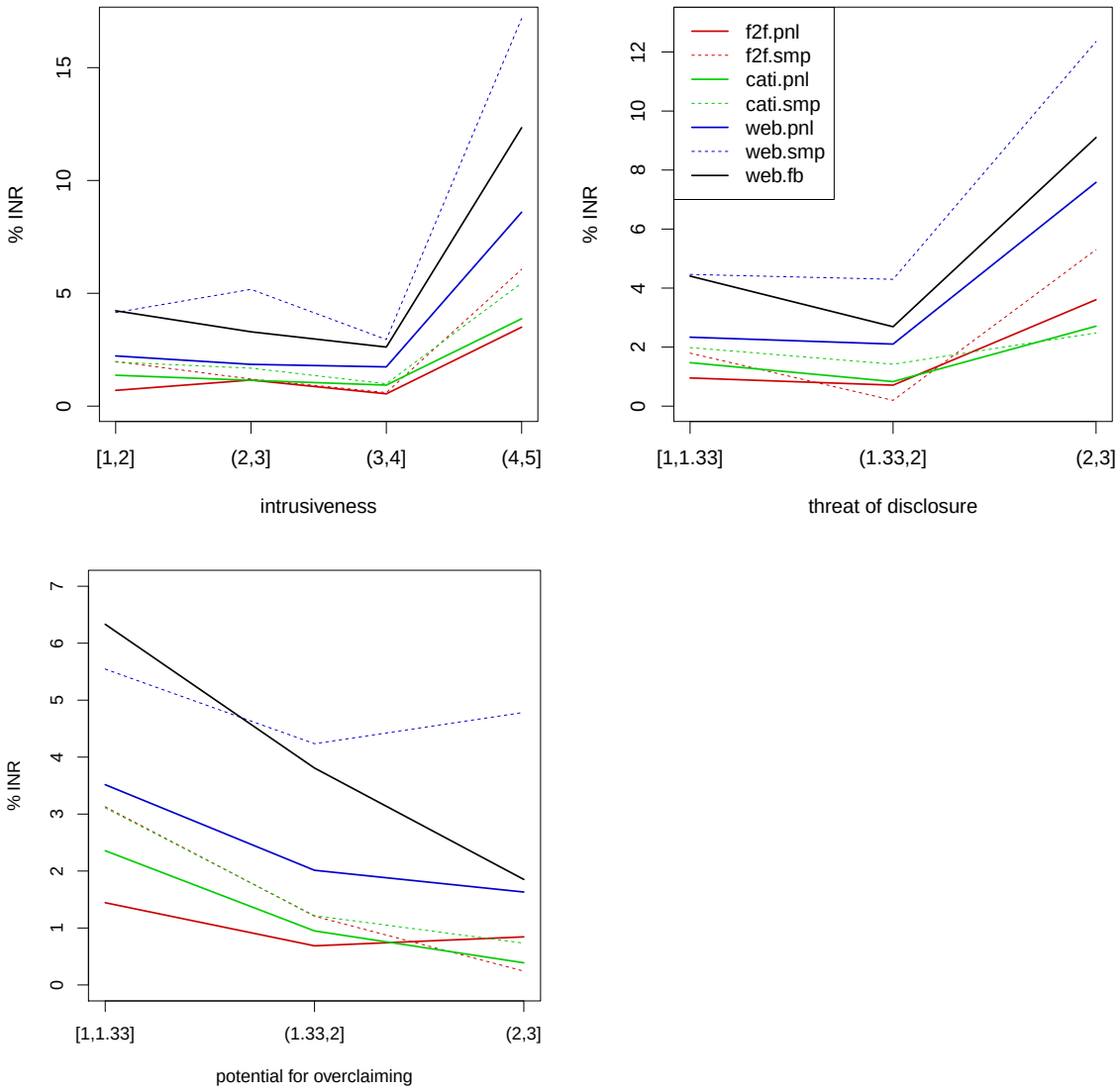
Figure 5.1: Weighted percentage of item nonresponse by respondent characteristics and sample; the mean value across imputations was used in the case of respondents' attitudes toward surveys.



short item wordings (the first three intervals are 10 units wide) and wider intervals for longer wordings (the last three intervals are 20, 30, and 50 units wide respectively). The percentage of item nonresponse in most samples (with the exception of web.smp) seems to decrease as the item wording increases in length.

The right panel of Figure 5.3 shows the relation between the item nonresponse rate and the number of answer alternatives. There is an obvious spike in item nonresponse for items with ten answer alternatives. The reason is that the question on the respondent's (and the respondent's partner's) income had exactly ten answer categories: eight income brackets plus unsubstantive answers of refusal and "don't know." We do not consider this spike to be a problem for the purposes of statistical modeling. We

Figure 5.2: Weighted percentage of item nonresponse by ratings of item sensitivity and sample



expect the nonlinear relationship between the number of answer alternatives and item nonresponse to be accounted for by other predictors that will be included in the model, most notably item intrusiveness, which was coded as very high for all income-related items.

Before entering them into the model as predictors, measures of item complexity will be *log-transformed*. This is because these measures are *count* variables (expressing the count of words or alternatives) and since we are not interested in fine discrimination at high values. For example, we do not consider the difference between 90 vs. 100 words as significant as the difference between an item wording of ten words vs. twenty words.

Figure 5.3: Weighted percentage of item nonresponse by measures of item complexity and sample

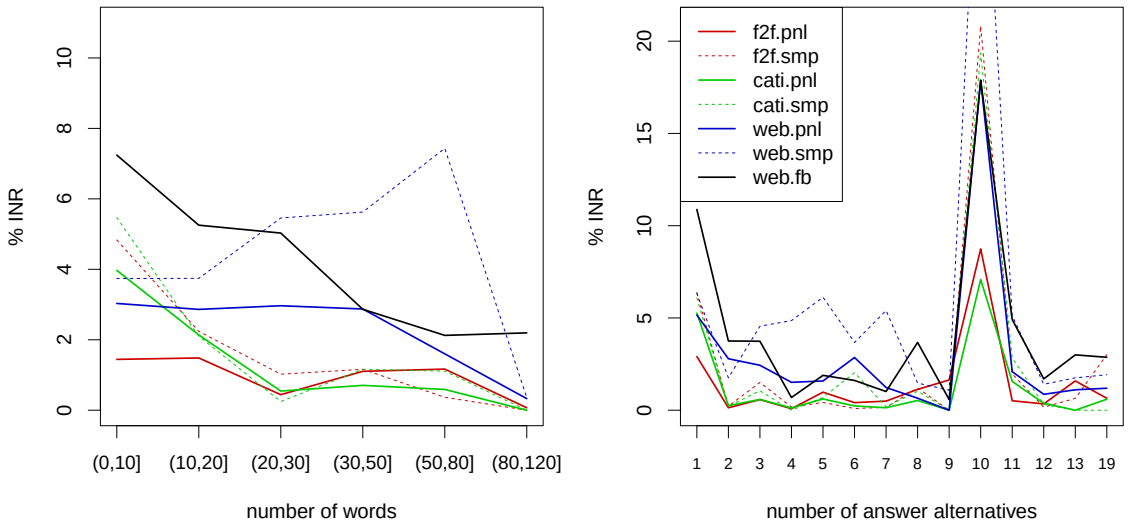
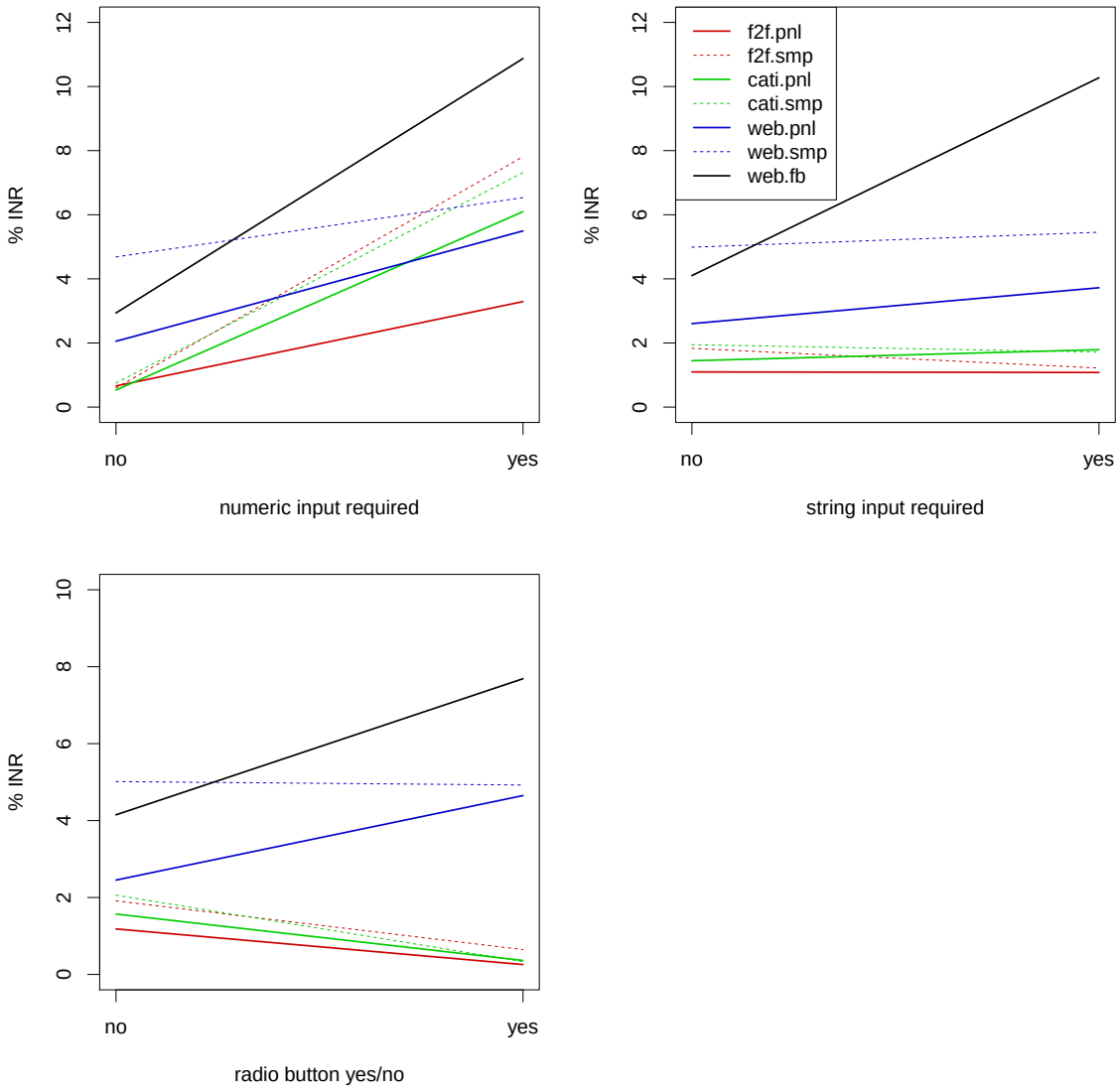


Figure 5.4 shows the relation of item nonresponse to indicators of item format. We differentiate between open-ended items that require *numeric* input (e.g. "How many rooms are there in the dwelling where you live?") and *string* input (e.g. "What was your father's occupation when you were 15?"). Because open-ended items require more effort to fill in, we expect more item nonresponse for such items. The upper panels of Figure 5.4 indicate that this is, indeed, the case for the GGP data.

Upon visually inspecting the data, many item nonresponses were found to have occurred because web respondents did not mark "no" on a particular type of item battery. This occurred for survey questions that asked the respondent to use radio buttons to choose between "yes" and "no" on a long list of items, e.g., types of birth control that were used in the last 12 months. Instead of choosing "yes" and "no," some web respondents only marked those "yes" items as applicable, leaving the others unmarked, resulting in item nonresponse. We will include in the models the indicator named "radio yes" to control for this effect. As the lower left panel of Figure 5.4 shows, items with radio buttons for yes/no responses have substantially more item nonresponse in the web.fb and web.pnl samples, and somewhat less item nonresponse in other samples.

The remainder of the figures in this section pertain to respondents' self-assessments of cognitive state and the sensitivity of certain item topics. Low values on the x-scale indicate low cognitive state (easily accessible information) and low topic sensitivity. Only the subset of items applying to a particular topic was used to calculate the percentage of item nonresponse. For example, the upper left panel of Figure 5.6 shows the item nonresponse rate by the respondents' self-assessed cognitive state for items

Figure 5.4: Weighted percentage of item nonresponse by indicators of item format

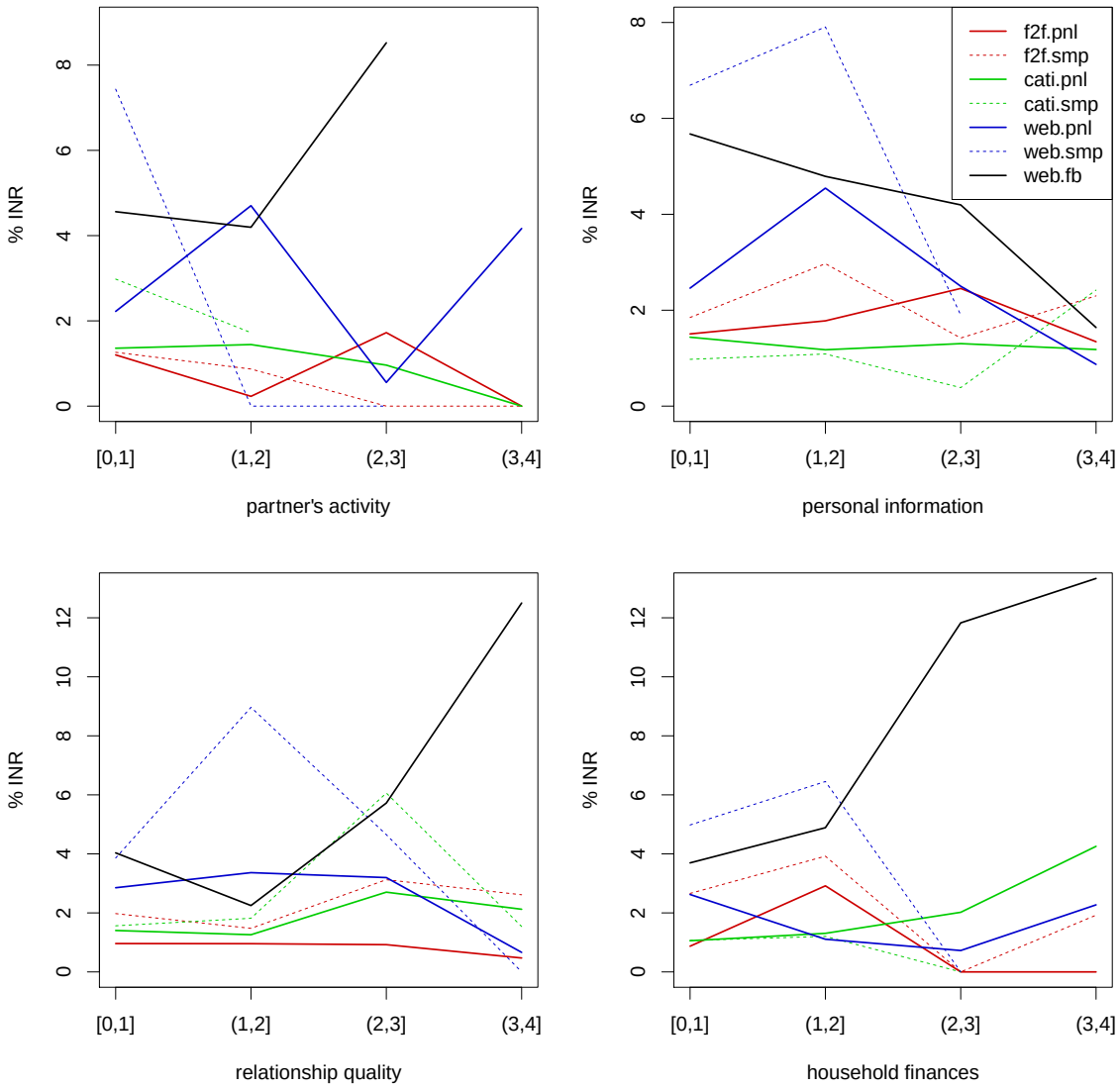


pertaining to the partner's activity. Only the items pertaining to the partner's activity were used to calculate the proportion of item nonresponse, while all other items were excluded. In the presence of missing values, the value on the x-axis was determined by averaging the imputed values across the five imputed datasets.

According to the hypotheses put forward in Section 4.5, we expect an increase in item nonresponse with the values of the self-assessments[16], i.e. we expect the broken lines in the figures to increase from left to right. Because only a subset of items was used for each figure, however, the sample size can be quite low which is why we will not interpret each figure separately. We will interpret the effects of cognitive state and respondent-specific sensitivity in later sections where we consider statistical models

[16] All self-assessments have been recoded so that high values refer to higher self-assessed sensitivity and higher (less available) cognitive state.

Figure 5.5: Weighted percentage of item nonresponse by self-assessment of cognitive state and sample



including these variables as predictors.

This section examined the descriptive statistics for item nonresponse. About three quarters of item nonresponse in the GGP questionnaire stems from skipped items, while "don't know" and refusals are less common. Item nonresponse is concentrated in a number of extreme individuals rather than being uniformly distributed across respondents. The final part of the section examined the bivariate relations between item nonresponse and explanatory variables. The presented plots suggested a dichotomization of the items' intrusiveness and the threat of disclosure before entering them into statistical models as predictors.

Figure 5.6: Weighted percentage of item nonresponse by self-assessment of topic sensitivity and sample (1/2)

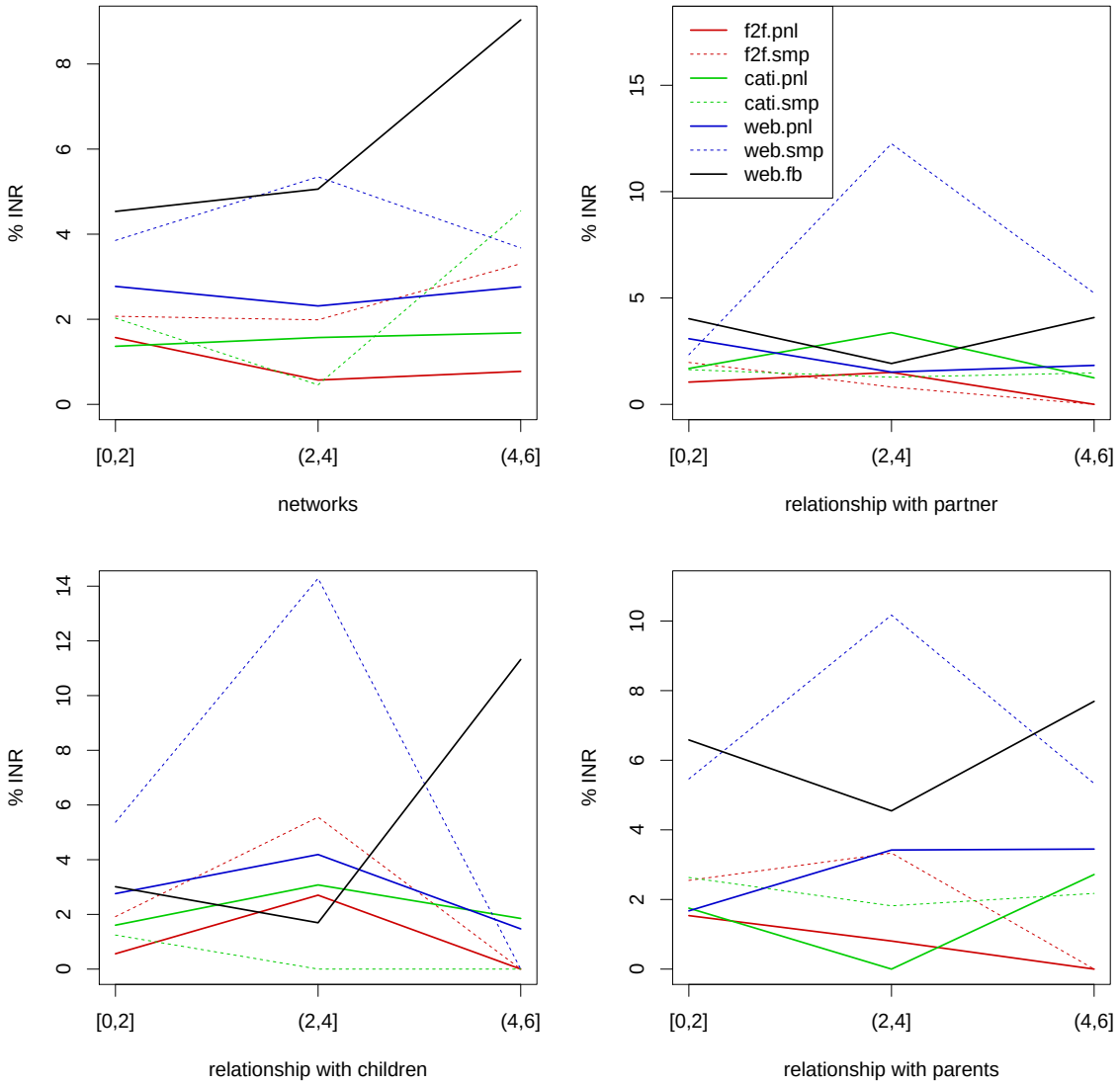
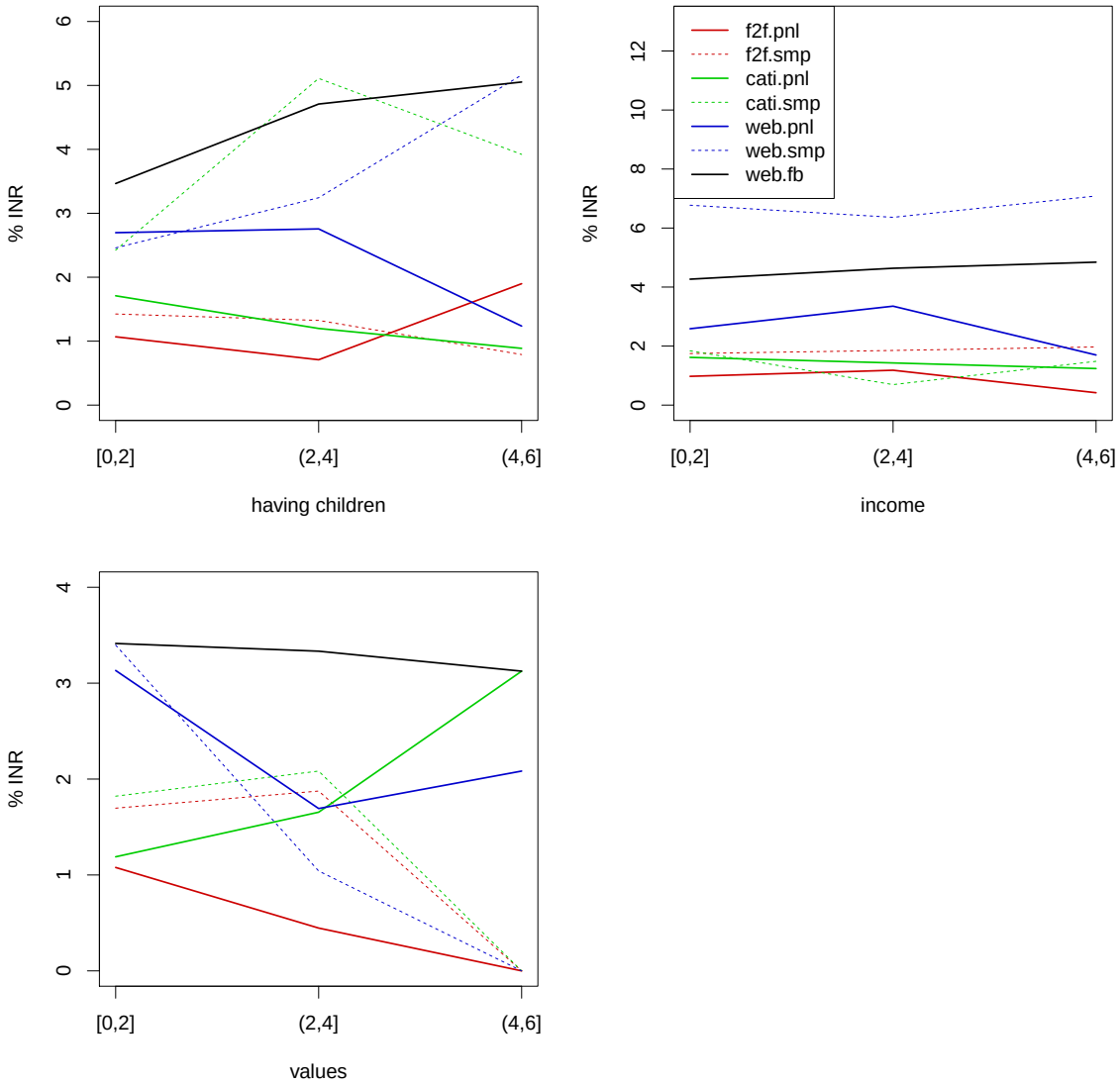


Figure 5.7: Weighted percentage of item nonresponse by self-assessment of topic sensitivity and sample (2/2)



5.2 A separate model for each sample

Before fitting a model to the combined data stemming from rounds 1 and 2, we fit a model to each of the seven samples *separately*. The lme4 package (Bates et al. 2012) for R (R Core Team 2013) was used to estimate the parameters of all generalized linear mixed models for item nonresponse in this chapter. This software package uses the penalized iteratively reweighted least squares (PIRLS) algorithm to estimate model parameters. This computational method is efficient for models with crossed grouping factors and is as such appropriate for fitting item response models (see Doran et al. 2007 for details).

The results of fitting the models for each sample separately are given in Table 5.8, but are only intended as a preliminary analysis of the data. The sample sizes reported in Table 5.7 do not support fitting complex models with many parameters. This warning applies particularly to samples from round 2 of data collection. The model estimates in Table 5.8 are not reported to draw conclusions, but merely to illustrate general trends before proceeding to model the entirety of the data.

Table 5.7: Sample size at each level by sample

	measurement occasions	items	respondents	interviewers
f2f.pnl	49550	478	206	16
f2f.smp	23547	477	107	6
cati.pnl	48606	493	209	18
cati.smp	12202	467	59	9
web.pnl	55974	495	228	
web.smp	9524	450	45	
web.fb	39211	491	262	

The sample size for face-to-face and CATI data from round 2 does not seem to be sufficient to fit a model that includes interviewer random effects. In both aforementioned cases, the interviewer residual variation is estimated at zero. The R^2 is therefore estimated at 1.00 for the f2f.smp model, while the R^2 cannot be estimated for the cati.smp model, as the interviewer-level residual variation was also estimated at zero in the baseline model for cati.smp[17].

In the cati.smp sample, all interviewers (for whom we were able to obtain information)

[17]The R^2 statistics reported in Table 5.8 were calculated according to Equation (3.2) by comparing the residual variation at a particular level to the corresponding residual variation in the baseline model (baseline models not shown). See Section 5.3 for details on how the residual variation and R^2 were computed in the presence of multiply imputed missing values on predictors.

Table 5.8: Generalized linear mixed model for item nonresponse on each sample separately; models for the web have crossed random effects for items and respondents; models for face-to-face and CATI administration have an additional interviewer random effect.

	f2f.pnl		f2f.smp		cati.pnl		cati.smp		web.pnl		web.smp		web.fb	
	est	se	est	se	est	se	est	se	est	se	est	se	est	se
(Intercept)	-7.49	0.38	-9.72	0.88	-7.17	0.30	-7.12	0.48	-5.25	0.45	-5.00	0.39	-5.40	0.19
Interviewer														
male [MI]	-0.40	0.45	-0.46	0.73	-0.20	0.47								
age10 [cn, MI]	-0.05	0.17	-0.74	0.73	-0.55	0.69	-0.41	0.62						
education [cn, MI]	-0.15	0.12	0.15	0.32	-0.09	0.13	0.21	0.60						
log(experience) [cn, MI]	0.05	0.19	-2.81	2.94	0.08	0.12	0.21	0.48						
Respondent														
male	-0.10	0.18	0.00	0.21	-0.07	0.13	0.07	0.33	0.07	0.17	-1.07	0.48	0.22	0.22
age10 [cn]	0.23	0.07	-0.03	0.07	0.07	0.05	0.25	0.11	0.25	0.07	0.56	0.14	-0.16	0.07
education [cn]	-0.01	0.05	-0.09	0.06	0.00	0.04	0.00	0.09	-0.03	0.05	-0.39	0.10	-0.03	0.05
attitude toward surveys [cn, MI]	-0.09	0.15	-0.22	0.13	-0.01	0.11	-0.21	0.20	-0.10	0.16	-0.56	0.31	-0.09	0.14
Item														
intrusiveness [dch]	0.87	0.39	2.13	0.90	1.17	0.35	1.38	0.56	1.12	0.21	1.52	0.39	1.19	0.22
disclosure [dch]	0.95	0.45	0.28	1.15	0.66	0.43	-0.15	0.77	0.69	0.29	1.15	0.40	0.80	0.27
overclaiming [cn]	-0.34	0.24	-1.43	0.73	-0.71	0.25	-0.26	0.38	-0.28	0.13	-0.13	0.20	-0.48	0.14
log(n. words) [cn]	0.28	0.21	0.13	0.52	-0.18	0.20	0.25	0.31	0.09	0.11	0.54	0.18	0.09	0.12
log(n. alternatives) [cn]	0.67	0.35	1.33	0.89	0.45	0.34	0.47	0.53	0.09	0.18	0.21	0.28	0.26	0.19
input numeric	2.56	0.64	4.11	1.67	2.69	0.60	2.93	0.99	0.97	0.32	0.87	0.60	2.36	0.35
input string	1.62	0.77	2.91	1.93	1.81	0.70	1.40	1.11	1.05	0.39	0.96	0.63	2.71	0.40
radio yes	-0.90	0.65	-0.21	1.58	-0.27	0.57	-1.07	1.20	0.62	0.30	0.85	0.51	1.34	0.31
Item-by-respondent interaction														
partner activity [MI]	0.12	0.18	0.41	0.18	0.50	0.14	0.58	0.28	0.72	0.11	0.60	0.25	0.23	0.16
personal information [MI]	0.26	0.08	-0.08	0.08	0.10	0.05	0.41	0.11	0.12	0.07	0.17	0.31	0.03	0.04
rel. quality [MI]	-0.17	0.19	0.21	0.32	0.26	0.22	0.15	0.26	0.27	0.09	-0.24	0.39	-0.19	0.12
HH finances [MI]	0.41	0.13	0.17	0.14	0.07	0.14	0.66	0.21	0.35	0.12	0.07	0.20	0.50	0.11
networks [MI]	0.17	0.11	0.04	0.17	0.23	0.06	0.21	0.14	0.03	0.06	-0.04	0.12	0.05	0.05
rel. partner [MI]	0.10	0.11	-0.81	0.44	0.04	0.11	-0.14	0.12	0.12	0.04	-0.03	0.08	0.12	0.04
rel. children [MI]	-0.13	0.30	-11.04	924.85	0.16	0.15	0.12	0.32	0.00	0.10	-0.05	0.27	0.01	0.08
rel. parents [MI]	-0.02	0.23	0.12	0.19	0.09	0.09	-0.06	0.15	-0.04	0.08	-0.13	0.15	0.06	0.05
having children [MI]	0.02	0.05	-0.08	0.27	0.05	0.06	0.43	0.11	-0.04	0.04	0.40	0.08	0.11	0.04
income [MI]	0.30	0.05	0.25	0.08	0.18	0.05	0.02	0.10	0.15	0.05	-0.03	0.08	0.07	0.04
values [MI]	-0.02	0.31	-0.29	0.95	-0.87	0.66	-1.00	1.12	-0.60	0.30	0.36	0.10	0.30	0.07
$\epsilon_{ivr.}$	0.52		0.00		0.30		0.00							
$\epsilon_{resp.}$	0.86		0.75		0.64		0.79		1.12		1.29		1.37	
ϵ_{item}	1.84		3.59		1.70		2.03		1.05		1.34		1.16	
$R^2_{ivr.}$	0.06		1.00		-0.03		–							
$R^2_{resp.}$	0.17		0.22		0.08		0.39		0.27		0.51		0.04	
R^2_{item}	0.37		0.78		0.60		0.84		0.58		0.44		0.61	

were female. The interviewer sex predictor therefore needed to be excluded from the model for this sample as it was constant across all measurement occasions. Another artifact of small sample size is the conspicuously high magnitude and standard error of the relationship with children predictor in the f2f.smp model. In this sample, *all item nonresponses* occurred at the same[18] value of the relationship with children predictor.

Despite the estimation problems, the estimates in Table 5.8 show certain consistent features. We consider a fixed effect statistically significant if the magnitude of the effect is at least twice the size of its standard error (this approximately corresponds to an alpha level of 0.05).

- Interviewer characteristics are not significant predictors in any of the four models in which they were included.
- The effect of respondent age[19] is either positive (older respondents produce more item nonresponse) or non-significant. A notable exception is the Facebook sample, where the effect of age is significant and has the opposite direction.
- The effect of respondent education is negative or non-significant across the seven models (higher age is associated with less item nonresponse). The same applies for the respondent's attitude toward surveys: respondents with a more positive attitude toward surveys produce less item nonresponse.
- The expert ratings of item sensitivity perform well as predictors and have the expected direction. Item intrusiveness has a positive and significant effect on item nonresponse across all seven models (highly intrusive items induce more nonresponse). The effect of the threat of disclosure is positive in all models but one (threatening items induce more nonresponse), while the effect of the item's potential for overclaiming is negative across all models (items that allow overclaiming are associated with less item nonresponse).
- The logarithm of the number of words in an item and the logarithm of the number of answer alternatives have insignificant effects in most models. The direction of the effect of measures of item complexity, however, is consistently positive (with a single exception).
- Open-ended items (with manual input of either numbers or text) are associated with more item nonresponse than closed-ended items.
- Items that asked the respondent to respond "yes" or "no" by means of radio but-

[18] All item nonresponses occurred at the *zero* value. The relationship with children predictor was coded zero if 1) the item did not concern the relationship with children, or 2) if the item did concern this topic, but the respondent relied that answering to this topic is not sensitive at all for them.

[19] Note that the respondents' and interviewers' age was divided by 10 before it was entered into the models. This transformation allows us to make better use of the two decimal places that are used to report the results in the tables, e.g., the estimate for the age10 predictor is 0.23 with a standard error of 0.07 in the f2f.pnl sample. Had we used the untransformed age, the estimate would have been 0.02 with a standard error of 0.01.

tons have a significant positive effect (are associated to more item nonresponse) in web administration. In face-to-face and CATI models the effect is negative and non-significant.

- The effect of the item-by-respondent interactions is either non-significant or positive. The positive effects indicate that the item topics the respondents assessed as more sensitive and as being in a higher (less accessible) cognitive state are associated to more item nonresponse, as expected.

We will now proceed to fit statistical models to several samples at once. This will not only increase the total sample size but will also enable us to compare the occurrence of item nonresponse across modes. We will exclude the Facebook sample from all subsequent models, as it has been demonstrated to be incomparable to the other samples.

5.3 Models without interviewer level

The fact that the questionnaire was self-administered in web mode has a bearing on how the data should be modeled. In CATI and face-to-face modes, the respondents were nested in interviewers. The interviewer therefore constitutes an additional level that is lacking in web mode. In this section we will omit the interviewer random effects and interviewer-level predictors. We will model the full data from rounds 1 and 2, but will for now disregard the information we have on the interviewers. In Section 5.4, we will proceed by dropping the web respondents from the analysis. Modeling only the part of the data that stems from CATI and face-to-face interviews will allow us to include interviewer random and fixed effects in the model. Finally, in Section 5.5 we will consider an approach that circumvents the problem and allows the inclusion of both the interviewer level and the data gathered on the web.

The sample size at each level of the model is given in Table 5.9. The number of measurement occasions and respondents in Table 5.9 equals the sum of the corresponding cells in the top six rows of Table 5.7. The number of distinct items is not equal for all samples because of the routing in the GGP questionnaire. Certain items were only administered to respondents in a very specific life situation, i.e., they were only administered if *several* conditions (on preceding filter items) were met. If no respondent in the dataset was in a position to be administered such an item, then the item itself is correspondingly excluded from the dataset.

We begin the analyses in this section by fitting three simple models.

Model A.0 is the *baseline* model containing only the random intercepts for items

Table 5.9: Sample size at each level for combined data from rounds 1 and 2

measurement occasions	items	respondents
199403	512	854

and respondents but no fixed effects.

Model A.1 contains the random effects and fixed effects for 1) the mode of administration (represented by two dummy variables with face-to-face as the reference category) and 2) the round of data collection (round 2 is the reference category).

Model A.2 contains the random effects, as well as the mode and round of data collection fixed effects, and adds the *interaction* of the fixed effects.

Model A.0 will be used as the baseline model when computing the proportion of explained variation for all models in this section. The purpose for fitting both models A.1 and A.2 is to determine whether the differences in item nonresponse between the six samples from rounds 1 and 2 can be explained just by taking into account information on the mode and the round of data collection, but not on their interaction (model A.1). The more complex model (A.2) includes the interaction and corresponds to representing the six samples with five dummy variables. Table 5.10 gives the results of the comparison of model fit for the three models.

Table 5.10: Comparison of model fit for models A.0, A.1, and A.2

	Df	AIC	BIC	logLik	Chisq	Chi Df	Pr(>Chisq)
model A.0	3	27960	27991	-13977			
model A.1	6	27863	27924	-13926	102.89	3	0.00
model A.2	8	27864	27946	-13924	3.04	2	0.22

The three models are shown in the rows of Table 5.10 in increasing order of complexity as reflected by the number of model parameters (Df). The fit of the models can be evaluated with regard to different criteria. Akaike's information criterion (AIC) and Schwartz's Bayesian information criterion (BIC) introduce different penalties for the additional model parameters. Both are computed on a scale where a smaller value indicates better fit. BIC, in contrast to AIC, also involves the sample size, implying that differences in likelihood of two competing models need to be viewed not only relative to the difference in the number of parameters but also relative to the number of observations. With a larger sample size, a more drastic decrease in likelihood is required before a complex model will be preferred over a simpler model (Molenberghs and Verbeke 2004).

The p-values in the last column of Table 5.10 refer to the likelihood ratio test. The

highly significant p-value in the second row indicates that model A.1 is clearly preferred over the baseline model. The insignificant p-value in the bottom row, on the other hand, suggests that the decrease in loglikelihood for model A.2 does not outweigh the additional complexity represented by the interaction term. This is in line with the values of AIC and BIC, which also prefer model A.1 over model A.2. The figures in Table 5.10 therefore suggest that the interaction term between mode and panel can be left out. That is, the GGP data can be modeled by considering the separate effects of the mode of administration and round of data collection but disregarding their *joint* effect on item nonresponse.

We now gradually increase the complexity of the model for item nonresponse by adding predictors at different levels. All subsequent models in this section build upon model A.1 in that they exclude the interaction of the mode of administration and the round of data collection.

Model A.3 adds respondent-level characteristics. In order to test Hypotheses 2a.3 and 2a.4, the interaction of the indicator for web mode with respondent age and education is also included.

Model A.4 builds upon model A.3 and adds item facets and their interactions with web mode.

Model A.5 adds respondents' self-assessments of cognitive state and sensitivity to certain item topics along with their interactions with web mode.

Table 5.11 gives a summary of estimates for the aforementioned statistical models. Some predictors have additional information given in square brackets next to their names in the first column. Continuous predictors like respondent age were *centered* (they have [cn] added to their names) before being entered into the model as predictors: the mean across all measurement occasions on a particular continuous predictor was subtracted from the value on this predictor for each measurement occasion. Item intrusiveness and the threat of disclosure were *dichotomized* (denoted by [dch]) for the reasons outlined in Section 5.1. Finally, the respondent's attitudes toward surveys and self-assessments were multiply imputed (denoted by [MI]) as described in Section 4.4.

Models A.3, A.4, and A.5 involve multiply imputed predictors. Each of the models therefore needed to be fitted to each of five imputed datasets, resulting in five sets of parameter estimates for each model[20]. The fixed effects displayed in Table 5.11 were obtained by pooling the estimates according to Rubin's rules (Rubin 1987) as implemented in the mice package for R (van Buuren and Groothuis-Oudshoorn 2011). The

[20]Fitting each model five times substantially increased the estimation time. Estimating a model with all predictors on the full dataset (e.g. models A.6 and C.6) five times took in excess of 20 hours on a computer with Intel i7 2.20 Ghz processor and 6 GB RAM.

Table 5.11: Generalized linear mixed models for item nonresponse with crossed random effects for respondents and items; data from rounds 1 and 2

	model A.0		model A.1		model A.3		model A.4		model A.5	
	est	se	est	se	est	se	est	se	est	se
(Intercept)	-5.21	0.10	-5.18	0.13	-5.35	0.15	-6.47	0.19	-6.58	0.20
(mode) cati			0.55	0.12	0.48	0.12	0.48	0.12	0.43	0.12
(mode) web			1.08	0.12	1.15	0.12	1.70	0.15	1.67	0.15
panel			-0.73	0.11	-0.51	0.12	-0.53	0.12	-0.54	0.12
Respondent										
male					-0.08	0.09	-0.07	0.09	-0.08	0.09
age10 [cn]					0.18	0.04	0.15	0.04	0.15	0.04
age10 [cn] × (mode) web					0.10	0.07	0.14	0.07	0.14	0.07
education [cn]					-0.01	0.03	0.00	0.03	-0.01	0.03
education [cn] × (mode) web					-0.07	0.05	-0.08	0.05	-0.07	0.05
attitude toward surveys [cn, MI]					-0.21	0.08	-0.22	0.08	-0.16	0.08
Item										
intrusiveness [dch]							1.37	0.23	1.26	0.23
intrusiveness [dch] × (mode) web							0.05	0.12	0.08	0.13
disclosure [dch]							0.33	0.26	0.17	0.27
disclosure [dch] × (mode) web							0.87	0.15	0.69	0.18
overclaiming [cn]							-0.46	0.13	-0.46	0.14
overclaiming [cn] × (mode) web							0.29	0.09	0.28	0.09
log(n. words) [cn]							0.03	0.12	-0.02	0.12
log(n. words) [cn] × (mode) web							0.15	0.07	0.15	0.07
log(n. alternatives) [cn]							0.67	0.20	0.69	0.20
log(n. alternatives) [cn] × (mode) web							-0.62	0.12	-0.56	0.12
input numeric							2.70	0.36	2.56	0.36
input numeric × (mode) web							-1.96	0.22	-1.74	0.23
input string							1.91	0.43	1.84	0.43
input string × (mode) web							-0.90	0.25	-0.73	0.26
radio yes							-0.62	0.36	-0.59	0.36
radio yes × (mode) web							1.36	0.21	1.32	0.22
Item-by-respondent interaction										
partner activity [MI]									0.30	0.08
partner activity [MI] × (mode) web									0.15	0.10
personal information [MI]									0.14	0.04
personal information [MI] × (mode) web									-0.15	0.06
rel. quality [MI]									0.12	0.09
rel. quality [MI] × (mode) web									0.02	0.09
HH finances [MI]									0.25	0.07
HH finances [MI] × (mode) web									0.02	0.10
networks [MI]									0.21	0.05
networks [MI] × (mode) web									-0.13	0.05
rel. partner [MI]									0.03	0.06
rel. partner [MI] × (mode) web									0.05	0.07
rel. children [MI]									0.08	0.11
rel. children [MI] × (mode) web									-0.12	0.13
rel. parents [MI]									0.05	0.06
rel. parents [MI] × (mode) web									-0.09	0.08
having children [MI]									0.10	0.03
having children [MI] × (mode) web									-0.11	0.04
income [MI]									0.23	0.03
income [MI] × (mode) web									-0.06	0.04
values [MI]									-0.34	0.21
values [MI] × (mode) web									0.56	0.21
$\epsilon_{resp.}$	1.29		1.20		1.14		1.15		1.13	
ϵ_{item}	1.64		1.64		1.64		1.30		1.30	
$R^2_{resp.}$			0.14		0.23		0.21		0.23	
R^2_{item}			0.00		0.00		0.38		0.37	

pooling of variance components is not as straightforward, however, because in order for Rubin's rules to apply, the quantity in question must have at least an approximately normal distribution. Variance components have a right-skewed distribution which is why the square root transformation was first applied to obtain standard deviations that have a distribution closer to normal. The average across the five standard deviations (one for each MI dataset) is reported as the pooled value ($\epsilon_{\text{resp.}}$ and ϵ_{item}) in the bottom part of Table 5.11. The proportion of explained variation (R^2) at each level was then calculated according to Equation (3.2) by comparing the residual variation at a particular level to the corresponding variation in the baseline model (model A.0).

Fourteen percent of the variation on the respondent level is explained by the mode of administration and round of data collection (model A.1). Adding respondent characteristics as predictors increases $R^2_{\text{resp.}}$ to 0.23. The respondent-level predictors do not explain any variation on the item level, as one might expect. Adding item facets as predictors increases the proportion of explained variation to 0.38 on the item level, but *decreases* the R^2 on the respondent level. This is unusual from the viewpoint of classical regression where (the unadjusted) R^2 can only *increase* by adding predictors to the model.

We will outline a possible explanation for why this happened in the model for item nonresponse. Before describing the hypothetical scenario, we need to reiterate how the response variable was coded and the bearing this has on the interpretation of respondent random effects. Because item nonresponse was coded as a 1-response (as opposed to a substantive response, which was coded zero), high values of respondent random effects correspond to a higher probability of item nonresponse. High values of respondent random effects therefore signify *low* motivation.

One hypothetical scenario in which controlling for item facets lowers the respondent-level R^2 is the following. Let us assume that low-motivation respondents were administered a higher proportion of easy-to-answer items (low-difficulty and low-sensitivity items). In model A.4, where the effect of item facets has been taken into account, these low-motivation respondents get assigned high values of random effects, which moves them further from the other respondents and accordingly increases the respondent-level residual variation. In model A.3, on the other hand, these low-motivation respondents' random effects are lower because they produced less item nonresponse, due to the fact that they received a higher proportion of easy-to-answer items. The increase in respondent-level residual variation (and associated decrease in R^2) can therefore indicate true variation between respondents that was masked in model A.3, which did not control for item facets (see Gelman and Hill 2006, 480-481 for another example and discussion).

Two more models will be considered in this section:

Model A.6 is a more parsimonious version of model A.5. In order to reduce the number of model parameters that need to be estimated, we *remove insignificant item-by-respondent predictors* along with their corresponding interactions with web mode. We will retain in the model those respondent self-assessments where either the baseline effect or the interaction with web is significant at the 0.05 level (the magnitude of the effect is at least twice as large as the corresponding standard error). We thus exclude the self-assessment of cognitive costs for the item related to *relationship quality* and self-assessments of sensitivity to topics that concern *relationships with the respondent's partner, children*, and *parents* (along with the corresponding web mode interactions).

Model A.7 adds two additional predictors to model A.6, which are described in detail below. The purpose of including these predictors is to attempt to explain the respondent's probability of item nonresponse on the currently administered item by means of the respondent's previous record of item nonresponse.

Items in a questionnaire usually follow a particular *order*, which thus constitutes an additional organizing principle in the data[21]. When such an *a priori* ordering of items exists, it is possible for the probability of a 1-response to an item to depend on the responses to preceding items. We wish to examine the effect that previous item nonresponses have on the respondent's tendency to produce item nonresponse to the current item they have been administered. We introduce two predictors into the model with the aim of modeling two different aspects of the respondent's previous record of item nonresponse.

- The (logged) cumulative number of item nonresponses in the interview up to (but excluding) the current item. This is a measure of the respondent's *overall* past tendency to produce item nonresponse[22].
- The indicator for *serial* item nonresponse conveys the information on whether the answer to the item *directly preceding* the current one was an item nonresponse. The serial indicator thus captures the respondent's tendency to produce item nonresponses one after the other.

[21] In order to account for context effects (see e.g. Schwarz and Sudman 1992) the order of items is sometimes randomized. The extent of randomization is usually limited to changing the order of statements in a particular battery, as the items' *content* limits the order in which the items can be administered. A filter item, e.g., must always precede dependent items that will only be administered if a particular response is given to the filter item. Because of the elaborate routing scheme, *the order of the items was fixed* in the GGP questionnaire, i.e., there was no randomization of item order.

[22] The expression $log(0)$ is undefined. A small constant (0.5) is added to the respondent's cumulative number of item nonresponses so that the logarithm can be evaluated even if the respondent's current cumulative is zero.

It is clear from the way these predictors are defined, that their values differ both from respondent to respondent and change as the particular respondent proceeds through the questionnaire. These predictors are therefore defined at the lowest level—the measurement occasion level (see Figure 3.12, page 78).

Table 5.12 gives the results of fitting models A.6 and A.7. Both indicators of previous item nonresponse have positive and highly significant effects. Because of the logit link function, the regression coefficients in the models for item nonresponse are somewhat challenging to interpret. If the respondent produced an item nonresponse on the preceding item (versus the scenario where they gave a substantive answer), then the logit probability of item nonresponse on the current item increases by 3.79. Interpretation is usually aided by referring to *odds ratios*, which can be computed by exponentiating the effect (as shown in Table 5.12 in the column entitled exp(est)). Item nonresponse on the previous item thus increases the odds of item nonresponse on the current item *by a factor of 44*. The strongest effect of the included predictors by far is that of the serial indicator.

In a model with measures of previous item nonresponse as predictors, however, a respondent's random effect can no longer be interpreted as involving the person's motivation. To illustrate this point, we will consider a respondent who produces a great number of item nonresponses *in a series* during the course of the interview. In the baseline model A.0, this respondent's random effect would be estimated to a high value, reflecting the respondent's lack of motivation to optimize. Upon adding respondent-level predictors to the model, this respondent's (lack of) motivation can be calculated by multiplying the vector of regression coefficients by the values on the predictors for this particular individual and adding this respondent's residual. The random effect, in other words, reflects the respondent's motivation, but is adjusted for predictors in the model.

When the serial indicator is included as a predictor in the model, however, the high number of item nonresponses for this particular respondent is explained by the fact that they occurred in a series. Despite the fact that the respondent produced a great number of item nonresponses, the residual for this individual is shrunk toward zero. The value of the random effect is therefore no longer related to the respondent's motivation upon including measures of previous item nonresponses as predictors in the model.

The estimates of model A.7 indicate that measures of previous item nonresponse do explain a great amount of variability in item nonresponse on the current item. The estimates, as mentioned, are positive and high for both added predictors, especially the serial indicator. The inclusion of these predictors is accompanied by a dramatic

Table 5.12: Generalized linear mixed models for item nonresponse with random effects for respondents, interviewers, and items; data from rounds 1 and 2

	model A.6					model A.7				
	est	se	exp(est)	sig		est	se	exp(est)	sig	
(Intercept)	-6.55	0.19	0.00	0.00	**	-6.75	0.19	0.00	0.00	**
(mode) cati	0.44	0.12	1.55	0.00	**	-0.08	0.06	0.92	0.19	
(mode) web	1.70	0.15	5.45	0.00	**	0.55	0.11	1.74	0.00	**
panel	-0.54	0.12	0.58	0.00	**	-0.26	0.06	0.77	0.00	**
Respondent										
male	-0.08	0.09	0.92	0.40		0.01	0.05	1.01	0.79	
age10 [cn]	0.15	0.04	1.17	0.00	**	-0.04	0.02	0.96	0.06	
age10 [cn] × (mode) web	0.14	0.07	1.15	0.04	*	0.05	0.03	1.05	0.12	
education [cn]	0.00	0.03	1.00	0.87		0.04	0.02	1.05	0.01	*
education [cn] × (mode) web	-0.07	0.05	0.93	0.15		-0.06	0.02	0.94	0.01	*
attitude toward surveys [cn, MI]	-0.16	0.08	0.85	0.04	*	-0.04	0.04	0.96	0.26	
Item										
intrusiveness [dch]	1.26	0.23	3.51	0.00	**	0.68	0.28	1.98	0.01	*
intrusiveness [dch] × (mode) web	0.11	0.13	1.12	0.39		-0.07	0.14	0.93	0.61	
disclosure [dch]	0.19	0.27	1.21	0.49		-0.17	0.33	0.84	0.60	
disclosure [dch] × (mode) web	0.67	0.17	1.96	0.00	**	0.70	0.19	2.01	0.00	**
overclaiming [cn]	-0.40	0.14	0.67	0.00	*	0.23	0.16	1.26	0.15	
overclaiming [cn] × (mode) web	0.30	0.09	1.35	0.00	**	0.29	0.10	1.34	0.00	*
log(n. words) [cn]	-0.02	0.12	0.98	0.84		-0.17	0.15	0.84	0.24	
log(n. words) [cn] × (mode) web	0.14	0.07	1.15	0.05	*	-0.05	0.08	0.95	0.49	
log(n. alternatives) [cn]	0.70	0.20	2.02	0.00	**	0.17	0.24	1.18	0.48	
log(n. alternatives) [cn] × (mode) web	-0.57	0.12	0.56	0.00	**	-0.38	0.13	0.68	0.00	*
input numeric	2.58	0.36	13.18	0.00	**	2.17	0.44	8.74	0.00	**
input numeric × (mode) web	-1.78	0.23	0.17	0.00	**	-1.26	0.25	0.28	0.00	**
input string	1.84	0.43	6.31	0.00	**	1.92	0.52	6.82	0.00	**
input string × (mode) web	-0.76	0.26	0.47	0.00	*	-0.17	0.28	0.84	0.53	
radio yes	-0.60	0.36	0.55	0.10		-0.64	0.43	0.53	0.14	
radio yes × (mode) web	1.36	0.22	3.88	0.00	**	1.07	0.24	2.91	0.00	**
Item-by-respondent interaction										
partner activity [MI]	0.30	0.08	1.35	0.00	**	0.20	0.08	1.23	0.01	*
partner activity [MI] × (mode) web	0.14	0.10	1.15	0.18		-0.02	0.11	0.98	0.87	
personal information [MI]	0.14	0.04	1.15	0.00	**	0.07	0.03	1.07	0.02	*
personal information [MI] × (mode) web	-0.15	0.06	0.86	0.02	*	-0.06	0.05	0.94	0.25	
HH finances [MI]	0.25	0.07	1.28	0.00	**	0.23	0.07	1.25	0.00	**
HH finances [MI] × (mode) web	0.02	0.10	1.02	0.80		0.00	0.10	1.00	0.97	
networks [MI]	0.20	0.05	1.23	0.00	**	0.11	0.05	1.12	0.02	*
networks [MI] × (mode) web	-0.13	0.05	0.87	0.01	*	-0.04	0.06	0.96	0.43	
having children [MI]	0.10	0.03	1.11	0.00	*	0.05	0.04	1.05	0.16	
having children [MI] × (mode) web	-0.10	0.04	0.91	0.02	*	-0.04	0.04	0.96	0.41	
income [MI]	0.23	0.03	1.26	0.00	**	0.21	0.03	1.23	0.00	**
income [MI] × (mode) web	-0.06	0.04	0.94	0.11		-0.04	0.04	0.96	0.30	
values [MI]	-0.35	0.21	0.71	0.10		-0.31	0.21	0.74	0.14	
values [MI] × (mode) web	0.55	0.21	1.74	0.01	*	0.42	0.21	1.51	0.05	
Previous item nonresponse										
log(cum. INR) [cn]						1.23	0.03	3.42	0.00	**
serial INR						3.79	0.07	44.08	0.00	**
$\epsilon_{resp.}$	1.14					0.19				
ϵ_{item}	1.31					1.65				
$R^2_{resp.}$	0.23					0.98				
R^2_{item}	0.37					0.00				

decrease in respondent-level residual variation. The proportion of explained variation at the respondent level increases to 98%.

Most fixed effects in the model decrease in magnitude when measures of previous item nonresponse are added. We have already alluded to why this occurs for respondent-level predictors; if respondents, e.g., with a more negative attitude toward surveys produce more item nonresponse, this is reflected in a negative effect of this predictor in model A.6. Upon including measures of previous item nonresponse in the model, however, part of this variation is explained by the fact that some item nonresponses occur in series, lowering the magnitude of the attitude toward surveys predictor.

The decrease in magnitude for item-level predictors occurs because items of a similar kind (a similar level of intrusiveness, etc.) sometimes appear in clusters in the questionnaire. A series of item nonresponses to highly intrusive items is explained by the items' intrusiveness in model A.6. Upon including measures of previous item nonresponse in model A.7, the magnitude of the intrusiveness effect decreases because the item nonresponses are now partially explained by the fact that they occurred one after the other.

The results for Model A.7 indicate that respondents tend to produce item nonresponses in series, and that a high number of previous item nonresponses increases the probability that the respondent will not respond to the current administered item. We will refrain from drawing conclusions beyond these from model A.7, due to issues we have described above.

Models described in this section exclude all interviewer information and therefore fail to control for a possible interviewer effect on item nonresponse. For this reason, we will interpret effects and evaluate hypotheses on the basis of models that do take into account at least the information on which respondents were interviewed by the same interviewer. These models are considered in the following two sections.

5.4 Models with interviewer level, excluding web mode

In order to include interviewer random and fixed effects in the models for item nonresponse, the models in this section will be fitted to the data stemming from face-to-face and telephone interviews, *excluding all web respondents*. Table 5.13 gives the sample size at each level.

Because the narrowed dataset allows us to include interviewer-level information, we will focus in this section on interpreting 1) random effects and residual variations at

Table 5.13: Sample size at each level for combined data from rounds 1 and 2, excluding web mode

measurement occasions	items	respondents	interviewers
133905	503	581	36

each level, and 2) the fixed effects of interviewer-level predictors. We will defer the interpretation of the remaining fixed effects to Section 5.5, where we fit models to the full data, without excluding web respondents. Similarly to the previous section, the model in this section is built up by gradually including sets of predictors. Tables 5.14 and 5.15 give the results.

Model B.0 is the baseline model with random effects for respondents, interviewers, and items, but without fixed effects.

Model B.1 adds dummy predictors for CATI mode (face-to-face is the reference category) and the round of data collection to the baseline model.

Model B.2 adds interviewer characteristics to model B.1.

Model B.3 adds respondent characteristics.

Model B.4 adds item facets as predictors.

Model B.5 adds item-by-respondent interactions.

Model B.6 is a more parsimonious version of model B.5. The same item-by-respondent predictors are found to be insignificant[23] as in Section 5.3, and are removed to reduce the number of model parameters that need to be estimated. The self-assessment of sensitivity to items that concern *values* is also insignificant, but is retained in the model to keep it comparable to the corresponding model from the previous section.

Model B.7 adds measures of previous item nonresponse to model B.6.

In terms of item nonresponse, there appears to be more variation among items than among respondents. Table 5.11 in the previous section shows that, in the baseline model, the estimated residual standard deviation is 1.64 at the item level and 1.29 at the respondent level. These figures are different in Table 5.14 because the model was fit to a different dataset (excluding web respondents). Still, the estimate of the residual standard deviation at the item level is the largest (2.28), indicating that there is more variation among items than among respondents (0.90). The variation among interviewers is lower still, with an estimated residual standard deviation of 0.56.

The distribution of random effects is often depicted by plotting the ordered estimated

[23]These are the self-assessment of cognitive costs for items related to relationship quality and self-assessments of sensitivity to items that concern relationships with the respondent's partner, children, and parents

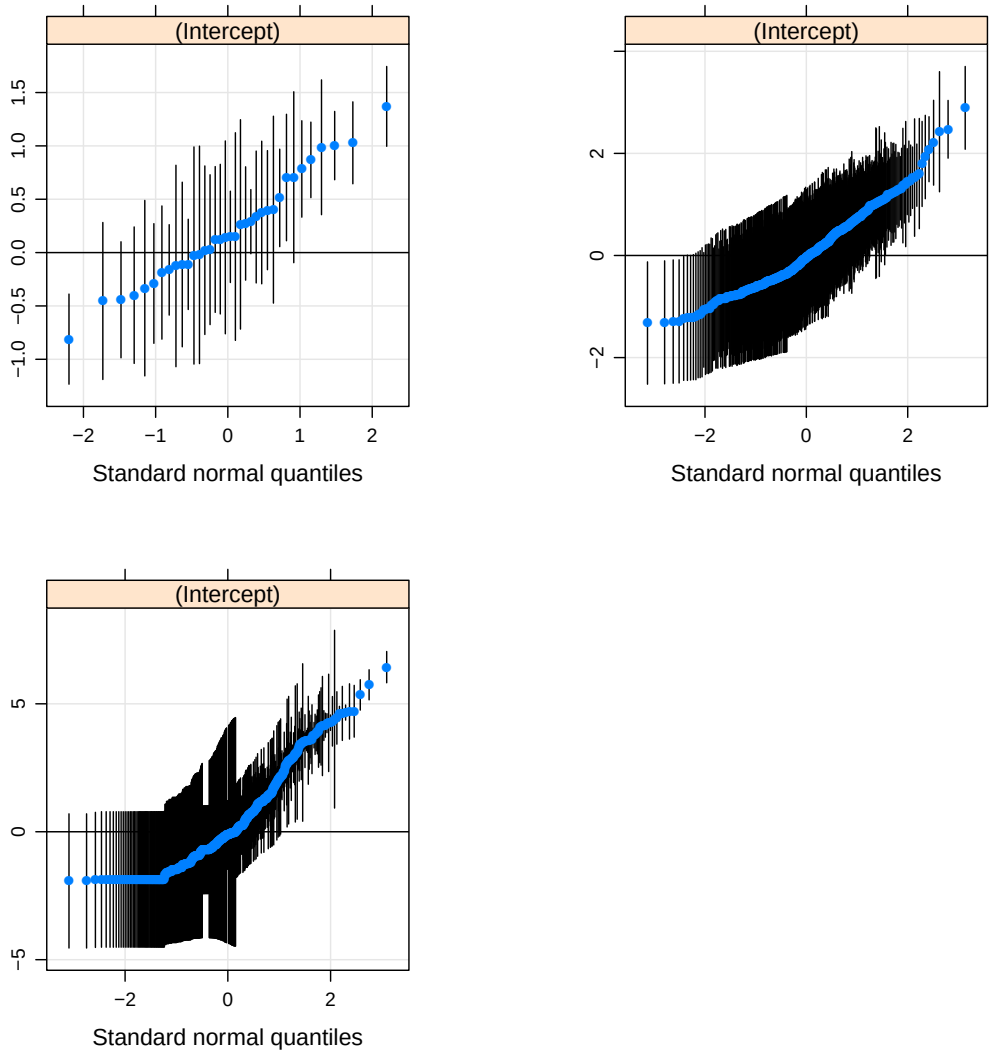
Table 5.14: Generalized linear mixed models for item nonresponse with random effects for respondents, interviewers, and items; data from rounds 1 and 2, excluding web mode

	model B.0		model B.1		model B.2		model B.3		model B.4		model B.5	
	est	se	est	se	est	se	est	se	est	se	est	se
(Intercept)	-6.25	0.17	-5.69	0.23	-5.35	0.24	-5.51	0.24	-6.59	0.30	-6.64	0.30
(mode) cati			0.24	0.24	-0.29	0.29	-0.31	0.28	-0.31	0.28	-0.30	0.28
panel			-0.86	0.12	-0.83	0.12	-0.64	0.12	-0.64	0.12	-0.66	0.12
Interviewer												
male [MI]					-0.49	0.28	-0.46	0.27	-0.45	0.27	-0.44	0.27
age10 [cn, MI]					-0.24	0.15	-0.25	0.15	-0.25	0.15	-0.23	0.15
education [cn, MI]					-0.16	0.08	-0.16	0.08	-0.16	0.08	-0.15	0.08
log(experience) [cn, MI]					0.11	0.10	0.10	0.10	0.10	0.10	0.10	0.10
Respondent												
male							-0.04	0.09	-0.04	0.09	-0.05	0.09
age10 [cn]							0.13	0.03	0.13	0.03	0.13	0.03
education [cn]							-0.02	0.03	-0.02	0.03	-0.02	0.03
attitude toward surveys [cn, MI]							-0.17	0.07	-0.17	0.07	-0.12	0.07
Item												
intrusiveness [dch]									1.28	0.30	1.16	0.29
disclosure [dch]									0.94	0.34	0.69	0.34
overclaiming [cn]									-0.60	0.18	-0.57	0.18
log(n. words) [cn]									0.13	0.16	0.08	0.16
log(n. alternatives) [cn]									0.85	0.27	0.91	0.26
input numeric									3.27	0.50	3.19	0.49
input string									2.20	0.59	2.10	0.58
radio yes									-0.24	0.48	-0.29	0.47
Item-by-respondent interaction												
partner activity [MI]											0.28	0.08
personal information [MI]											0.07	0.03
rel. quality [MI]											0.00	0.12
HH finances [MI]											0.29	0.07
networks [MI]											0.22	0.05
rel. partner [MI]											0.01	0.07
rel. children [MI]											0.04	0.12
rel. parents [MI]											0.05	0.06
having children [MI]											0.06	0.03
income [MI]											0.21	0.03
values [MI]											-0.32	0.22
$\epsilon_{ivr.}$	0.56		0.58		0.48		0.46		0.47		0.47	
$\epsilon_{resp.}$	0.90		0.84		0.84		0.81		0.81		0.81	
ϵ_{item}	2.28		2.29		2.29		2.29		1.69		1.65	
$R^2_{ivr.}$			-0.05		0.29		0.32		0.32		0.31	
$R^2_{resp.}$			0.14		0.14		0.21		0.20		0.21	
R^2_{item}			0.00		0.00		0.00		0.45		0.48	

Table 5.15: Generalized linear mixed models for item nonresponse with random effects for respondents, interviewers, and items; data from rounds 1 and 2, excluding web mode

	model B.6					model B.7				
	est	se	exp(est)	sig		est	se	exp(est)	sig	
(Intercept)	-6.64	0.30	0.00	0.00	**	-7.50	0.27	0.00	0.00	**
(mode) cati	-0.30	0.28	0.74	0.28		-0.35	0.16	0.71	0.03	*
panel	-0.66	0.12	0.52	0.00	**	-0.24	0.08	0.79	0.00	*
Interviewer										
male [MI]	-0.44	0.27	0.65	0.11		-0.19	0.16	0.83	0.23	
age10 [cn, MI]	-0.23	0.15	0.80	0.12		-0.05	0.09	0.95	0.54	
education [cn, MI]	-0.15	0.08	0.86	0.05	*	-0.04	0.04	0.96	0.41	
log(experience) [cn, MI]	0.10	0.10	1.10	0.32		0.01	0.06	1.01	0.88	
Respondent										
male	-0.05	0.09	0.95	0.59		0.02	0.06	1.02	0.68	
age10 [cn]	0.13	0.03	1.14	0.00	**	-0.05	0.02	0.95	0.02	*
education [cn]	-0.02	0.03	0.98	0.37		0.05	0.02	1.05	0.00	*
attitude toward surveys [cn, MI]	-0.12	0.07	0.89	0.09		0.00	0.04	1.00	0.96	
Item										
intrusiveness [dch]	1.15	0.29	3.15	0.00	**	0.70	0.35	2.02	0.04	*
disclosure [dch]	0.69	0.34	1.99	0.04	*	0.24	0.40	1.27	0.55	
overclaiming [cn]	-0.57	0.18	0.57	0.00	*	-0.01	0.21	0.99	0.96	
log(n. words) [cn]	0.08	0.16	1.08	0.62		-0.15	0.19	0.86	0.43	
log(n. alternatives) [cn]	0.91	0.26	2.49	0.00	**	0.68	0.31	1.98	0.03	*
input numeric	3.20	0.49	24.61	0.00	**	3.29	0.58	26.96	0.00	**
input string	2.10	0.58	8.14	0.00	**	2.55	0.68	12.82	0.00	**
radio yes	-0.28	0.47	0.75	0.54		-0.38	0.55	0.68	0.49	
Item-by-respondent interaction										
partner activity [MI]	0.28	0.08	1.33	0.00	**	0.21	0.08	1.23	0.01	*
personal information [MI]	0.07	0.03	1.08	0.04	*	0.02	0.03	1.02	0.48	
HH finances [MI]	0.28	0.07	1.33	0.00	**	0.27	0.07	1.31	0.00	**
networks [MI]	0.22	0.05	1.25	0.00	**	0.14	0.05	1.15	0.00	*
having children [MI]	0.06	0.03	1.06	0.10		0.02	0.04	1.02	0.64	
income [MI]	0.21	0.03	1.24	0.00	**	0.19	0.03	1.21	0.00	**
values [MI]	-0.32	0.22	0.73	0.15		-0.32	0.22	0.73	0.15	
Previous item nonresponse										
log(cum. INR) [cn]						1.44	0.04	4.24	0.00	**
serial INR						2.78	0.11	16.17	0.00	**
$\epsilon_{ivr.}$	0.47					0.23				
$\epsilon_{resp.}$	0.81					0.00				
ϵ_{item}	1.65					1.96				
$R^2_{ivr.}$	0.31					0.83				
$R^2_{resp.}$	0.21					1.00				
R^2_{item}	0.48					0.26				

Figure 5.8: Caterpillar plots of random effects from model B.0 for the interviewer (top left panel), respondent (top right), and item (bottom left)



random effects against quantiles of the standardized normal distribution. The precision of the point estimates is indicated by a vertical line that extends 1.96 standard deviations in each direction from the point estimate. Such plots are often referred to as "caterpillar plots" because of their appearance. Figure 5.8 gives the plots for the random effects from the baseline model B.0. The lower left panel shows that many items have the same minimal value of the random effect. This occurs because many items do not have any item nonresponse whatsoever and thus share the same value of the random effect. The noticeable difference in the precision of the random effects (as indicated by the vertical lines) is due to differing sample size per item: some items were administered to all respondents (and have narrow intervals), whereas other items were only administered to respondents in a specific life situation (and have wider intervals).

The mode of administration and round of data collection explain 14% of variation at the respondent level (Model B.1). Adding these predictors to the model *increases* the residual variation at the interviewer level, indicating that a certain amount of variation among interviewers was masked in model B.0, where these predictors were not included (see the discussion in Section 5.3 for an explanation of how this can happen). The resulting *negative* value of R^2 at the interviewer level illustrates the difficulties of defining a meaningful measure of explained variation for multilevel models. The mode of administration and round of data collection do not explain any variation at the item level.

Adding interviewer characteristics in model B.2 increases the interviewer-level R^2 to 29% and leaves the proportion of explained variation at the respondent and item levels unchanged. Adding respondent characteristics (model B.3) increases the respondent-level proportion of explained variation to 21% and also increases the interviewer-level R^2 to 32%. The item facets added in model B.4 explain 45% of item variation, while including self-assessments (model B.5) increases the item-level R^2 further to 48%. Model B.6 (see Table 5.15) differs from model B.5 in that it removes insignificant item-by-respondent interactions, which does not affect the estimates of the residual variation.

Both indicators of previous item nonresponse have positive and highly significant effects leading to the same conclusions as the results of the corresponding model in Section 5.3: 1) item nonresponses tend to occur in series, and 2) a high number of previous item nonresponses increases the probability that the respondent will not respond to the currently administered item. Including predictors of previous item nonresponse in the model, as mentioned, introduces interpretational difficulties and decreases the magnitude of most other fixed effects in the model. We will thus interpret the effects of interviewer characteristics by referring to model B.6.

The fixed effect of most interviewer characteristics is not significant at the 0.05 level. This is hardly surprising as the sample size at this level is very low: 36 interviewers conducted all the interviews. The interviewer characteristics, as mentioned, explain about a third of the variation at the interviewer level, however, indicating that interviewer characteristics do have substantial predictive power for item nonresponse. Education is the only significant interviewer-level predictor. The negative sign indicates that more highly educated interviewers interviewed respondents that produced less item nonresponse. The effect of interviewer sex and age are negative, indicating that the respondents interviewed by females and younger interviewers produced more item nonresponse. The p-values for interviewer sex and age effects are 0.11 and 0.12 respectively, indicating a notable risk that the direction of these effects might merely

be an artifact of small sample size.

In this section we have applied the generalized linear mixed model for item nonresponse to a narrowed dataset excluding web respondents. This allowed us to include interviewer random and fixed effects. The estimates of the residual standard deviation in the baseline model indicate that the largest variation with regard to item nonresponse in the data is among items, followed by variation among respondents and interviewers. Including interviewer characteristics as predictors explains a substantial proportion (about a third) of the interviewer-level variation. This is noteworthy because authors who studied the effect of interviewers on item nonresponse found differences between interviewers, but were unable to explain the differences in terms of interviewer characteristics (Pickery and Loosveldt 1998, 2001). Even though interviewer characteristics explained a substantial part of the interviewer-level variation in the GGP data, only one out of four interviewer-level predictors has a statistically significant effect. We attribute this to small sample size at the interviewer level.

5.5 Models with interviewer level including web mode

The GGP data have a peculiar nesting structure, owing to the fact that the data were collected in three modes of administration, including web mode. The respondents in face-to-face and CATI modes are nested in interviewers, while web respondents are not, since they filled out the questionnaire without the aid of an interviewer. Such data are said to be *partially nested*. Other instances of such data arise, e.g., in trials where the effect of group therapy is compared to individual therapy (Korendijk et al. 2008).

Korendijk et al. (2008) showed that one way to fit a multilevel model to partially nested data is to treat the non-nested individuals (in our case web respondents) as *being in clusters of size one*. Such "clusters," of course, do not imply that there is dependency between units within the cluster, since each cluster consists of one unit only. Regarding non-nested individuals this way merely assigns a cluster identification to each individual, thus allowing us to fit an ordinary multilevel model. Korendijk et al. (2008) in a simulation study showed that this approach leads to unbiased estimates for all fixed effects and their standard errors.

This approach is applied in the present section to include in the statistical models both 1) random effects for interviewers and 2) the web data. Each web respondent will therefore be treated as if they had been interviewed by their own interviewer. We consider models in this section to be more adequate for the problem at hand than

corresponding models from Section 5.3, as the latter models completely disregard the clustering of respondents in interviewers for face-to-face and CATI modes.

Table 5.16 shows the sample size at each level. The figures for measurement occasions, items, and respondents are the same as in Table 5.9. The number of interviewers, on the other hand, is 309. This is the sum of the 36 actual interviewers in face-to-face and CATI mode along with the "web interviewers" that have been added for each web respondent (208 for web.pnl and 45 for web.smp).

Table 5.16: Sample size at each level for combined data from rounds 1 and 2; each web respondent has been assigned a unique interviewer.

measurement occasions	items	respondents	interviewers
199403	512	854	309

We will gradually build our model by adding sets of predictors in a similar vein as we did in the previous sections:

Model C.0 is the baseline model with random effects for respondents, interviewers, and items, but no fixed effects. As outlined above, each web respondent will be ascribed a unique interviewer.

Model C.1 adds dummy predictors for mode of administration and round of data collection to the baseline model.

Model C.2 adds respondent characteristics and interactions with web mode for respondent age and education.

Model C.3 adds item facets and their interactions with web mode.

Model C.4 adds item-by-respondent predictors and their interactions with web mode.

Model C.5 removes those item-by-respondent interactions from model C.4 that have an insignificant main effect and interaction with web mode. The same item-by-respondent predictors are found to be insignificant, as in Section 5.3, and have been removed to reduce the number of model parameters that need to be estimated.

Model C.6 adds measures of previous item nonresponse to model C.5.

Tables 5.17 and 5.18 give the model estimates. A glance at the variance components in the bottom of the tables reveals that treating web respondents as if each were interviewed by their own interviewer did not have an effect on the estimates of the item-level residual standard deviation and R^2: their values are approximately the same as in the corresponding models in Tables 5.11 and 5.12. The approach taken in this section did, however, affect respondent- and interviewer-level variations. The estimate of the interviewer-level residual standard deviation is higher than its respondent-level

counterpart, which is the opposite of what was found in Section 5.4. We will therefore refrain from interpreting any quantities related to variance components in this section.

Table 5.18 shows the results of models C.5 and C.6. Since including measures of previous item nonresponse as predictors affects others fixed effects in the model (as discussed in Section 5.3), we consider model C.5 rather than model C.6 as the final model for item nonresponse. We will evaluate operational hypotheses based on estimates of fixed effects from model C.5.

The ordering of the modes according to item nonresponse does not change when the model takes into account possible confounding effects. The estimate for web mode is 1.50 and highly significant, while the effect of CATI mode is positive (0.11) and insignificant. The exponentiated value of the web effect gives the odds ratio: the odds of item nonresponse are 4.47 times higher in web mode than in face-to-face interviewing. The data therefore support Hypotheses 1a.1 and 1a.2 (see Section 4.5).

The effect of the panel indicator is negative (-0.72) and significant. The odds of item nonresponse in the first round of data collection (the panel) are lower by a factor of 0.49 when compared to the second round of data collection (the sample). We believe that this is because panel members were already accustomed to being interviewed and well versed in proceeding through the phases of the question-answer process. They were already familiar with the way in which survey questions are usually worded, the various forms of scales to which their answers would need to comply, etc. Another possible cause as to why panel members produced less item nonresponse than respondents sampled from the population is that panel members were given monetary rewards for partaking in surveys.

The effect of respondent sex is insignificant in all fitted models. The effect of respondent age is positive (0.13) and highly significant. The model also contains a significant interaction between respondent age and web mode, which is why the effect of age should be interpreted along with the interaction. The exponentiated age coefficient (1.14) informs us that in face-to-face and telephone interviews the odds of item nonresponse increase by about 14% for each additional ten years of the respondent's age[24]. The positive web interaction adds 0.02 to the baseline effect. The exponentiated effect ($\exp(0.01 + 0.02) = 1.03$) therefore conveys the message that the age effect is even stronger in web mode, with the odds of item nonresponse increasing by about 18% for each additional ten years of the respondent's age. We consider this as evidence in support of Hypotheses 2a.1 and 2a.1.

[24]Let us re-iterate at this point that the age predictor in all statistical models is the respondent (or interviewer) age *divided by 10*. This simple transformation allows us to more precisely convey the magnitude of the effect and its standard error using two-decimal precision.

Table 5.17: Generalized linear mixed models for item nonresponse with random effects for respondents, interviewers, and items; data from rounds 1 and 2

	model C.0		model C.1		model C.2		model C.3		model C.4	
	est	se	est	se	est	se	est	se	est	se
(Intercept)	-4.89	0.12	-4.84	0.28	-5.02	0.27	-6.14	0.30	-6.26	0.29
(mode) cati			0.17	0.35	0.15	0.33	0.14	0.32	0.11	0.31
(mode) web			0.85	0.27	0.94	0.26	1.49	0.27	1.48	0.26
panel			-0.89	0.11	-0.69	0.11	-0.70	0.11	-0.72	0.11
Respondent										
male					-0.04	0.08	-0.03	0.08	-0.04	0.08
age10 [cn]					0.16	0.03	0.13	0.03	0.13	0.03
age10 [cn] × (mode) web					0.12	0.07	0.16	0.07	0.16	0.07
education [cn]					-0.03	0.03	-0.02	0.03	-0.03	0.03
education [cn] × (mode) web					-0.05	0.05	-0.06	0.05	-0.06	0.05
attitude toward surveys [cn, MI]					-0.18	0.07	-0.18	0.07	-0.13	0.07
Item										
intrusiveness [dch]							1.37	0.23	1.27	0.23
intrusiveness [dch] × (mode) web							0.06	0.12	0.09	0.13
disclosure [dch]							0.33	0.27	0.17	0.27
disclosure [dch] × (mode) web							0.87	0.15	0.69	0.18
overclaiming [cn]							-0.47	0.13	-0.47	0.14
overclaiming [cn] × (mode) web							0.30	0.09	0.29	0.09
log(n. words) [cn]							0.02	0.12	-0.03	0.12
log(n. words) [cn] × (mode) web							0.16	0.07	0.16	0.07
log(n. alternatives) [cn]							0.67	0.20	0.69	0.20
log(n. alternatives) [cn] × (mode) web							-0.62	0.12	-0.57	0.12
input numeric							2.70	0.36	2.57	0.36
input numeric × (mode) web							-1.95	0.22	-1.74	0.23
input string							1.93	0.43	1.86	0.43
input string × (mode) web							-0.91	0.25	-0.74	0.26
radio yes							-0.62	0.36	-0.58	0.36
radio yes × (mode) web							1.36	0.21	1.31	0.22
Item-by-respondent interaction										
partner activity [MI]									0.29	0.07
partner activity [MI] × (mode) web									0.16	0.10
personal information [MI]									0.13	0.03
personal information [MI] × (mode) web									-0.15	0.06
rel. quality [MI]									0.13	0.09
rel. quality [MI] × (mode) web									0.01	0.09
HH finances [MI]									0.24	0.07
HH finances [MI] × (mode) web									0.04	0.10
networks [MI]									0.20	0.05
networks [MI] × (mode) web									-0.13	0.05
rel. partner [MI]									0.03	0.06
rel. partner [MI] × (mode) web									0.05	0.06
rel. children [MI]									0.07	0.11
rel. children [MI] × (mode) web									-0.11	0.13
rel. parents [MI]									0.04	0.06
rel. parents [MI] × (mode) web									-0.08	0.08
having children [MI]									0.09	0.03
having children [MI] × (mode) web									-0.10	0.04
income [MI]									0.23	0.03
income [MI] × (mode) web									-0.06	0.04
values [MI]									-0.34	0.21
values [MI] × (mode) web									0.56	0.21
$\epsilon_{ivr.}$	1.02		0.94		0.88		0.85		0.82	
$\epsilon_{resp.}$	0.92		0.87		0.82		0.84		0.84	
ϵ_{item}	1.65		1.65		1.65		1.31		1.31	
$R^2_{ivr.}$			0.16		0.26		0.31		0.35	
$R^2_{resp.}$			0.11		0.20		0.16		0.16	
R^2_{item}			0.00		0.00		0.37		0.37	

Table 5.18: Generalized linear mixed models for item nonresponse with random effects for respondents, interviewers, and items; data from rounds 1 and 2

	model C.5					model C.6				
	est	se	exp(est)	sig		est	se	exp(est)	sig	
(Intercept)	-6.23	0.29	0.00	0.00	**	-6.71	0.21	0.00	0.00	**
(mode) cati	0.11	0.32	1.12	0.72		-0.17	0.12	0.84	0.17	
(mode) web	1.50	0.26	4.47	0.00	**	0.55	0.14	1.74	0.00	**
panel	-0.72	0.11	0.49	0.00	**	-0.31	0.06	0.74	0.00	**
Respondent										
male	-0.04	0.08	0.96	0.63		0.02	0.05	1.02	0.63	
age10 [cn]	0.13	0.03	1.14	0.00	**	-0.04	0.02	0.96	0.09	
age10 [cn] × (mode) web	0.16	0.07	1.18	0.02	*	0.05	0.03	1.05	0.17	
education [cn]	-0.03	0.03	0.97	0.34		0.04	0.02	1.04	0.01	*
education [cn] × (mode) web	-0.05	0.05	0.95	0.27		-0.06	0.03	0.94	0.02	*
attitude toward surveys [cn. MI]	-0.13	0.07	0.87	0.06		-0.02	0.04	0.98	0.49	
Item										
intrusiveness [dch]	1.26	0.23	3.53	0.00	**	0.69	0.28	1.99	0.01	*
intrusiveness [dch] × (mode) web	0.11	0.13	1.12	0.38		-0.07	0.14	0.94	0.63	
disclosure [dch]	0.19	0.27	1.21	0.49		-0.18	0.33	0.84	0.59	
disclosure [dch] × (mode) web	0.68	0.17	1.96	0.00	**	0.69	0.19	2.00	0.00	**
overclaiming [cn]	-0.41	0.14	0.67	0.00	*	0.24	0.17	1.27	0.15	
overclaiming [cn] × (mode) web	0.31	0.09	1.36	0.00	**	0.30	0.10	1.35	0.00	*
log(n. words) [cn]	-0.03	0.12	0.97	0.81		-0.18	0.15	0.83	0.21	
log(n. words) [cn] × (mode) web	0.15	0.07	1.16	0.04	*	-0.05	0.08	0.95	0.55	
log(n. alternatives) [cn]	0.71	0.20	2.03	0.00	**	0.19	0.24	1.20	0.44	
log(n. alternatives) [cn] × (mode) web	-0.58	0.12	0.56	0.00	**	-0.39	0.13	0.68	0.00	*
input numeric	2.58	0.36	13.22	0.00	**	2.20	0.44	9.07	0.00	**
input numeric × (mode) web	-1.78	0.23	0.17	0.00	**	-1.28	0.25	0.28	0.00	**
input string	1.86	0.43	6.42	0.00	**	1.96	0.52	7.10	0.00	**
input string × (mode) web	-0.77	0.26	0.46	0.00	*	-0.19	0.28	0.82	0.49	
radio yes	-0.59	0.36	0.55	0.10		-0.63	0.44	0.53	0.15	
radio yes × (mode) web	1.35	0.21	3.86	0.00	**	1.07	0.24	2.91	0.00	**
Item-by-respondent interaction										
partner activity [MI]	0.29	0.07	1.33	0.00	**	0.22	0.08	1.24	0.01	*
partner activity [MI] × (mode) web	0.15	0.10	1.16	0.15		-0.02	0.11	0.98	0.85	
personal information [MI]	0.13	0.03	1.14	0.00	**	0.06	0.03	1.06	0.05	
personal information [MI] × (mode) web	-0.15	0.06	0.86	0.02	*	-0.06	0.05	0.94	0.29	
HH finances [MI]	0.24	0.07	1.27	0.00	**	0.23	0.07	1.26	0.00	**
HH finances [MI] × (mode) web	0.04	0.10	1.04	0.71		-0.01	0.10	0.99	0.95	
networks [MI]	0.20	0.05	1.22	0.00	**	0.11	0.05	1.12	0.01	*
networks [MI] × (mode) web	-0.13	0.05	0.88	0.01	*	-0.04	0.06	0.96	0.47	
having children [MI]	0.09	0.03	1.10	0.00	*	0.04	0.03	1.04	0.29	
having children [MI] × (mode) web	-0.09	0.04	0.91	0.03	*	-0.02	0.04	0.98	0.62	
income [MI]	0.23	0.03	1.25	0.00	**	0.21	0.03	1.23	0.00	**
income [MI] × (mode) web	-0.06	0.04	0.94	0.12		-0.04	0.04	0.96	0.28	
values [MI]	-0.35	0.21	0.71	0.10		-0.30	0.20	0.74	0.14	
values [MI] × (mode) web	0.56	0.21	1.75	0.01	*	0.41	0.21	1.50	0.05	
Previous item nonresponse										
log(cum. INR) [cn]						1.23	0.03	3.42	0.00	**
serial INR						3.76	0.07	42.84	0.00	**
$\epsilon_{ivr.}$	0.83					0.28				
$\epsilon_{resp.}$	0.84					0.00				
ϵ_{item}	1.32					1.66				
$R^2_{ivr.}$	0.34					0.93				
$R^2_{resp.}$	0.16					1.00				
R^2_{item}	0.36					-0.01				

The data do not, however, provide sufficient support for Hypotheses 2a.3 and 2a.4, which postulate that more educated respondents should produce less item nonresponse, especially in web mode. The direction of the respondent education effect and its interaction with web mode is in line with the hypotheses, but the standard errors of the effects are too large to reject the null hypotheses.

The p-value of the effect of the respondent's attitude toward surveys on item nonresponse falls just short of the 0.05 threshold for significance. The effect is significant, however, in model C.3, which does not include item-by-respondent interactions. Respondents with a more positive attitude toward surveys therefore produce less item nonresponse, but this may partly be because respondents with a more positive attitude toward surveys also found the questionnaire items less sensitive to answer. This effect is captured in models C.4 and C.5 that add respondent-by-item interactions to the fixed effects. The effect of the respondent's attitude toward surveys consequently decreases in magnitude. Even though the p-value slightly exceeds the threshold of 0.05 we will consider the data to provide limited support of Hypothesis 3a.

We hypothesized that sensitive and threatening items would induce more item nonresponse and made the additional hypothesis that this effect would be less pronounced in web mode, because the absence of the interviewer would make it easier for the respondent to answer items concerning taboo topics or requiring the respondent to admit to having acted in counternormative ways or to holding such opinions. The main effects of the expert ratings of item sensitivity are accord with the first part of the hypothesis. Most interactions with web mode, however, have the opposite sign than predicted.

The effect of item intrusiveness is positive and significant. The odds of item nonresponse are higher by a factor of 3.53 if the item was rated as *highly*[25] intrusive. The effect of the interaction of item intrusiveness with web mode is not significant. The main effect of the threat of disclosure is also insignificant, but the interaction with web mode is. The exponentiated sum of the main effect and the interaction ($\exp(0.19 + 0.68) = 2.39$) informs us that in web mode the odds of item nonresponse are 2.39 times higher if the item was rated as posing a high threat of disclosure.

For the item's potential for overclaiming, both the main effect and the interaction with web mode are significant and have the expected direction. For every unit increase in an item's potential for overclaiming (this item facet was rated on a scale from 1 to 3), the odds of item nonresponse decrease by a factor of 0.67 in face-to-face and CATI modes. In web mode, this effect is partly neutralized by the interaction. The odds of

[25]The average intrusiveness rating was *dichotomized* before being entered into the model. This applies also to the average rating of the threat of disclosure (see Section 5.1 for details).

item nonresponse in web mode decrease only by a factor of 0.90 ($\exp(-0.41 + 0.31)$) for every unit increase in an item's potential for overclaiming.

The items' complexity as measured by the length of the wording and number of answer alternatives was hypothesized to be positively related to item nonresponse. The additional hypothesis states that this effect would be even more pronounced in web mode, as the respondents would have no help from the interviewer when faced with a difficult item. The model estimates do not support the hypotheses. The main effect of the length of wording is not statistically significant, while the interaction with mode is both significant and positive. The results therefore indicate that longer wordings induce more item nonresponse on the web. The main effect of (the logarithm of) the number of an item's answer alternatives is positive and significant, while the interaction with web mode is negative and significant. The main effect is therefore congruent with the hypothesis that complex items induce more item nonresponse, while this effect is *weaker* (rather than stronger) in web mode.

Open-ended items induce more item nonresponse, especially in interviewer administered interviews. The main effect for both numeric and string input items is positive while the corresponding interactions with web mode are negative. In face-to-face and CATI mode, the odds for item nonresponse are 13.22 times higher for numeric input items than for closed-form items. On the web, the odds for nonresponse are 2.22 ($\log(2.58 - 1.78)$) times higher for numeric input items. String input items have odds of item nonresponse 6.42 times higher than closed-form items in face-to-face and CATI modes, and 2.97 ($\log(1.86 - 0.77)$) times higher in web mode.

As mentioned, when web respondents were asked to use radio buttons to choose between "yes" and "no" on a long list of items, many respondents marked with "yes" those items that applied, but left the remaining radio buttons unmarked. The predictor named "radio yes" controls for this effect and has the expected direction. In web mode, the odds of item nonresponse on such batteries of items are 2.14 ($\exp 1.53 - 0.59$) times higher than on other items.

According to Hypotheses 5a.1 and 5a.2, the self-assessments of cognitive state should have positive main effects and positive interactions with web mode. The main effects for partner activity, personal information, and household finances are positive and thus in line with the hypothesis. The effects of the interactions with web mode, however, are insignificant, except for *personal information*, which has a negative significant effect in discord with the hypothesis. We note here that the effect of the self-assessment of relationship quality was removed from the model along with its web interaction because it was not significant in model C.4.

Hypotheses 4a.1 and 4a.2 predicted that the self-assessments of the sensitivity of certain item topics would have positive effects, while the corresponding interactions with mode should be negative. The estimates for items concerning networks, having children, and income conform to the hypotheses (though the web interaction for income is not significant). Items concerning values, however, have an insignificant main effect and a positive and significant interaction with web mode, which is at odds with the hypothesis. We note that the predictors for three additional item topics (relationship with partner, children, and parents) were removed from the model along with their interactions with web mode, because their effects were not significant in model C.4. We provide a further discussion of the results in the next section.

5.6 Evaluation of hypotheses and discussion

The fact that web respondents are not nested within interviewers has a bearing on how the data can be modeled and introduces difficulties with our analysis of the GGP data. The simplest approach, that involves summarily disregarding all interviewer-related information, is explored in Section 5.3. We do not, however, draw any conclusions on the basis of such models, precisely because they do not control for a possible interviewer effect on item nonresponse.

In order to explore the effect of interviewer-level predictors, Section 5.4 describes how the statistical models were fit to a narrower dataset excluding web respondents. This also allowed us to examine the size of the variance components in relation to each other. The results indicate that the largest variability with regard to item nonresponse is among items, followed by respondent and interviewer levels. The predictor variables explain a substantial part (a third) of the variation among interviewers, despite the fact that most interviewer characteristics do not reach significance at the alpha=0.05 level. We attribute the lack of significance to the small sample size at the interviewer level.

The final goal of the analysis is to model the full data from rounds 1 and 2 without excluding web respondents or interviewer information. Including web respondents in the analysis, of course, requires that we exclude interviewer-level predictors, but if possible we would like to retain the interviewer random effects as a control in the model. In order to do so, the data in Section 5.5 are modeled as if each web respondent had been interviewed by their own interviewer. This approach leads to incorrect variance component estimates, but produces correct estimates of fixed effects. We consider this approach superior to the exclusion of all interviewer information, which is why we test the hypotheses on the basis of estimates from model C.6 rather than A.6.

The effect of the mode of administration lends credence to the hypotheses: the more direct the contact between the interviewer and respondent, the less item nonresponse. Even after possible confounding variables have been taken into account, the model predicts the least item nonresponse in face-to-face mode and the most in web mode.

We hypothesized that age and education can serve as proxies for respondent cognitive ability and that item nonresponse will therefore be more common among older and less educated respondents, especially so in web mode. The model estimates support the hypothesis for respondent age. The direction of the effects for education, on the other hand, agrees with the hypothesis, but the effects do not reach the prescribed level of statistical significance.

Respondents with a more positive attitude toward surveys in general were found to produce less item nonresponse, supporting Hypothesis 3a.

We also find support for Hypothesis 4a.1, which states that more sensitive items should be related to more item nonresponse. The main effects of corresponding predictors are either in accordance with the hypothesis and significant, or fail to reach the prescribed level of statistical significance.

The interactions with web mode, however, are not as simple as Hypothesis 4a.2 postulated:

- The interaction effect of item intrusiveness and web mode is insignificant: there does not seem to be any difference across modes with regard to the magnitude of the effect of item intrusiveness on item nonresponse.
- The interaction of web mode and the item's threat of disclosure is positive and significant, rather than negative as Hypothesis 4a.2 predicted: the effect of an item's threat of disclosure is *stronger* with web administration.
- Finally, the interaction with the potential for overclaiming is positive and significant, conforming with the hypothesis: the effect of an item's potential for overclaiming is *weaker* with web administration.

Items with threatening response options and items that allow the respondent to overclaim therefore both induce *more item nonresponse in web mode*. A common explanation for both effects could be that the respondent feels more concern for self-presentation in the presence of an interviewer. With the interviewer present, the respondent is less likely to produce item nonresponse to items that allow overclaiming, as skipping over or refusing to answer such an item could be seen as admitting to a "sin of omission" (not having acted in socially desirable ways, Bradburn et al. 1978). With web administration the respondent feels less concern for self-presentation, pro-

ducing more item nonresponse to such items, and so the effect of the item's potential for overclaiming is weaker in web mode.

Similarly, if one (or several) of an item's answer alternatives asks the respondent to admit to having acted in counternormative ways or holding such opinions (our definition of the threat of disclosure), the web respondent can simply skip over such an item or refuse to respond. With the interviewer present, however, this might be seen as *implicitly admitting* that the threatening response alternative is true. When administered such an item by an interviewer, the respondent might therefore *choose a white lie over item nonresponse*. This lowers the effect of the threat of disclosure for face-to-face and CATI modes.

The respondents were asked to rate how sensitive they would find answering items concerning certain topics for the purposes of the survey (and not, e.g., in a casual conversation with friends). The effects of such respondent-specific measures of sensitivity are generally in line with Hypotheses 4a.1 and 4a.2. The exception are items concerning values: such items induce more item nonresponse in web-mode, rather than less. Three out of seven sensitivity self-assessments had insignificant main effects and interactions with web mode and were removed from the final model.

We also hypothesized that items that are complex or concern topics with which the respondent is less familiar would induce more item nonresponse, especially in web mode, where the respondent cannot receive any help from the interviewer. The results, however, do not support this hypothesis. The length of the item wording was only found to be positively related to item nonresponse in web mode, while a higher number of answer alternatives is related to more item nonresponse in face-to-face and CATI modes (but not web). Self-assessments of cognitive state have main effects that accord with Hypothesis 5a.1, while their interactions with web mode either do not reach statistical significance or have the opposite direction (items concerning personal information).

Open-ended items were found to be related to more item nonresponse, while this effect is weaker in web-mode. This direction of the main effects is expected as open-ended items require more effort on the part of the respondent. We have no explanation for the negative sign of the interaction with web, however, as we would expect even more item nonresponse to open-ended items in web mode, where the respondent rather than the interviewer must manually enter the answer.

6 Breakoff analysis

In the present chapter we will analyze the occurrence of breakoff, beginning with the effect of the mode of administration on breakoff. We then focus our analysis on the Facebook sample and apply the survival analysis methods introduced in Section 3.3. Cox models of increasing complexity are fitted to the data and described in Sections 6.3 and 6.4. The final section explores the interplay of item nonresponse and breakoff.

6.1 Mode of administration and breakoff

As mentioned, we encountered almost no breakoff in the first two rounds of data collection. The frequency and proportion of breakoff for each sample is displayed in Table 4.8, which is repeated here (Table 6.1) for convenience. We suspect that this conspicuous lack of breakoff, even in web mode, is a consequence of the advance letters that were sent, informing the sample persons that the GGP interview would take about an hour. Consequently, those sample persons who lacked the motivation to expend this amount of time may have refused to cooperate altogether, resulting in unit nonresponse rather than breakoff.

Table 6.1: Frequency and proportion of breakoff by sample

	n	n breakoff	% breakoff
f2f.pnl	206	0	0.0
f2f.smp	107	1	0.9
cati.pnl	209	14	6.7
cati.smp	59	6	10.2
web.pnl	228	12	5.3
web.smp	45	5	11.1
web.fb	262	158	60.3
Total	1116	196	17.6

The effect of mode on breakoff was analyzed by fitting a logistic regression model to the GGP data from rounds one and two. The dataset is comprised of 854 respondents who were coded 0 for completed interview and 1 for breakoff. The Facebook sample was excluded because it is not comparable to other samples, as Facebook respondents were not informed as to the duration of the interview.

The distinction between panel members of round 1 and respondents sampled from the population in round 2 is a nuisance that needs to be taken into account. The percentages in Table 6.1 show that the breakoff rate was consistently lower in round 1 than in round 2. This is not surprising, since panel members, as previously mentioned, were already well-versed in filling out questionnaires and were also financially motivated to partake in surveys.

Two logistic regression models were fit:

1. A model with mode of administration and an indicator identifying the panel.
2. The same model with the addition of an interaction between the two predictors.

Table 6.2 shows the results of the comparison of fit for the two models. The insignificant p-value (0.45) indicates that the slight improvement in deviance is not worth the increase in the model's complexity. Both Akaike's and Schwartz's information criteria also prefer the simpler model. These results suggest that the GGP breakoff data can be modeled by considering only the separate effects of the mode of administration and the round of data collection while disregarding their *joint* effect on breakoff.

Table 6.2: Comparison of model fit for logistic regression of breakoff with and without the panel-by-mode interaction; the data exclude the Facebook sample.

	AIC	BIC	Residual Dev.	Df	Deviance	Pr(>Chi)
no interaction	287.90	314.90	279.90			
with interaction	290.28	318.78	278.28	2	1.62	0.45

Table 6.3 shows the estimates for the logistic regression model without the interaction. Looking at the breakoff data in Table 6.1, it is obvious that the effect of CATI and web mode (as compared to the face-to-face baseline) is positive, while the effect of the panel indicator is negative. Fitting the statistical model provides the significance test for these effects. The effects of web and CATI are highly significant, while the effect of the panel is marginally significant, with a p-value of 0.06.

Table 6.3: Logistic regression of breakoff on mode of administration and panel indicator; the data exclude the Facebook sample.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-5.35	1.02	-5.25	0.00
(mode) cati	3.33	1.03	3.23	0.00
(mode) web	3.18	1.04	3.06	0.00
panel	-0.68	0.37	-1.85	0.06

Contrary to Hypothesis 1b.2, the effect of CATI in the model for breakoff is higher than

the effect of web administration. In order to test whether this difference is statistically significant, we specified the contrasts for the mode of administration predictor as shown by Table 6.4. Rather than using so-called *treatment contrasts* as before, we thus compared the average effect of CATI and web modes against face-to-face (this is the first contrast shown in the first column of Table 6.4), as well as web against CATI (the second contrast).

Table 6.4: Contrasts for the mode of administration predictor

	cati&web vs f2f	web vs cati
f2f	-2	0
cati	1	-1
web	1	1

The results of re-fitting the model with these contrasts for the mode of administration variable are shown in Table 6.5. The first contrast is positive and significant as expected, while the second is negative and not statistically significant. This indicates a lack of evidence in support of the hypothesis captured by the second contrast: that the effects of web and CATI on breakoff are unequal.

Table 6.5: Logistic regression of breakoff on mode of administration and panel indicator and contrasts for mode of administration as defined in Table 6.4; the data exclude the Facebook sample.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.18	0.42	-7.52	0.00
(mode) cati&web vs f2f	1.08	0.34	3.19	0.00
(mode) web vs cati	-0.08	0.17	-0.44	0.66
panel	-0.68	0.37	-1.85	0.06

In this section we analyzed the data from rounds 1 and 2 and found the least breakoff in face-to-face administration. The breakoff rates in web and telephone modes were found to be to be very similar, with slightly more breakoff in CATI mode (this effect is not statistically significant). The data therefore provide evidence in support of Hypothesis 1b.1 but are not in line with Hypothesis 1b.2.

6.2 Survival analysis

We now proceed to analyze the data from the additional Facebook sample. Slovenian Facebook users aged eighteen or older were invited by advertisement to partake in a demographic survey. Unlike in the first two rounds of data collection, respondents were not informed in advance of how long the questionnaire would take to fill out. We

suspect that this is the reason why about 60% of respondents (see Table 6.1) broke off. A breakoff rate this high is, of course, a reason for concern in any substantive analysis one might want to perform on the dataset. From our survey methodology standpoint, however, it is beneficial, as it allows us to investigate which item and respondent characteristics influence breakoff.

In the remainder of this chapter we will be interested in the respondent's "longevity" when proceeding through the questionnaire. More specifically, we will be interested in the *number of items* that the respondent was willing to answer before deciding to terminate the interview. We will attempt to identify respondent and item characteristics that are related to breakoff by resorting to *survival analysis* methods. This approach to analyzing breakoff has already been utilized by a number of authors before us (Galesic 2006; Matzat et al. 2009; Peytchev 2009). Before continuing, we discuss the reasons for preferring survival analysis to classical statistical methods such as regression.

All respondent interviews eventually end in one of two ways: the respondent either 1) reaches the final item, or 2) prematurely terminates the interview. Survival analysis offers a natural way of differentiating between these two outcomes by defining the former respondents as censored and the latter as having experienced the event (breakoff). This is not a typical example of censoring, though, as censoring cannot occur at any given time. Even if a particular respondent avoided a number of items, e.g. on their partner (if they do not have a partner) or children (if they have no children), they still had to fill in at least 200 items to complete the questionnaire (see Table 4.9). For the GGP data, no censoring is possible under this threshold.

Because all censored respondents have a high value for the "time variable" (the total number of items), we could also use a simpler approach to analyzing breakoff. We could disregard the difference between breakoffs and completing respondents and regress[26] the respondents' total number of items on respondent characteristics.

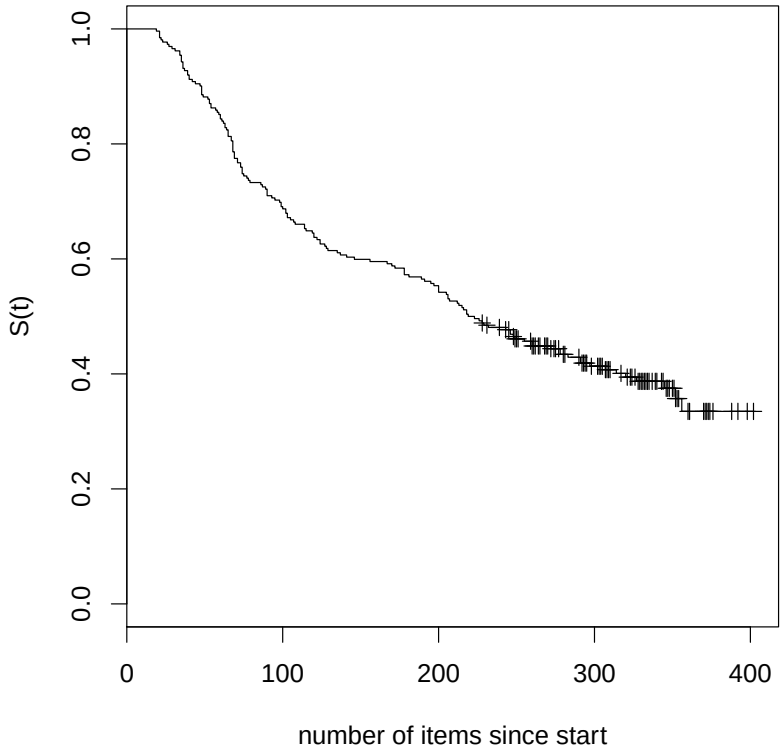
Such a regression model could not accommodate item-level predictors, though, at least not without making unprincipled simplifications. In order to include item-level information we would need to aggregate item facets to the respondent level by computing averages and proportions (e.g. the respondent-level proportion of items that require string input) and including them as predictors. The Cox model in its extended form, on the other hand, seamlessly accommodates item facets as time-varying covariates. Despite the fact that censoring is rather atypical, we thus argue that survival analysis methods fit the problem of analyzing breakoff better than classical regression methods.

[26]Strictly speaking, Poisson regression should be used because the outcome is a *count* variable.

All analyses in this chapter were performed with the survival package (Therneau 2013, see also Therneau and Grambsch 2000) for R. Unlike in the previous chapter on item nonresponse, *required items were not excluded* from the analyses. To reiterate, we will consider the *number* of items from the start of the interview as the time variable in the survival analysis, rather than the actual time in minutes. Breakoff is the event of interest and completing respondents are considered censored.

Figure 6.1 shows the survival curve. Censored units are represented by plus symbols and, as mentioned, all appear after item 200, because no respondent could complete the questionnaire without responding to this minimum of items.

Figure 6.1: Kaplan-Meier survival estimates for breakoff in the Facebook sample; censored units are represented by plus symbols.



The survival curve seems to break approximately at item number 100. The survival curve is steeper before this threshold and has a more moderate slope thereafter. This indicates that breakoff is more frequent shortly into the questionnaire and less frequent later. A similar finding is reported by (Galesic 2006).

We will now fit the Cox proportional hazards model to the breakoff data and test the PH assumption as we add new predictors to the model. In those cases where we find violations of the assumption, we will use a graphical method to determine how to modify the model to make the proportional hazards assumption more tenable. In

order to keep this procedure as simple as possible, we will first add predictors without missing values and add multiply imputed predictors in Section 6.4.

6.3 Cox models with complete predictors

The first model we fit involves only time-independent predictors: the respondent's sex, age, and education. The respondent's attitude toward surveys is another time-independent predictor to be added later when we include multiply imputed predictors. Table 6.6 reports the estimates for the first model. The rightmost column (labeled "zph") contains the p-values from the test for the proportional hazards hypothesis (see Section 3.3.5). A significant p-value signifies that the scaled Schoenfeld residuals for the predictor in question are correlated to the transformed survival time, which is a sign that the PH assumption is violated (Kleinbaum and Klein 2005). This is the case for the respondent age variable in Table 6.6.

Table 6.6: Cox regression of breakoff on core respondent demographics; data from the Facebook sample

	est	se	z	Pr(>|z|)		exp(est)	[95% conf.	interval]	zph
male	0.22	0.18	1.23	0.22		1.25	0.88	1.77	0.77
age10	-0.16	0.07	-2.44	0.01	*	0.85	0.75	0.97	0.00
education	-0.13	0.05	-2.69	0.01	**	0.88	0.80	0.97	0.12
c-index	0.61	0.02							
R^2	0.10								

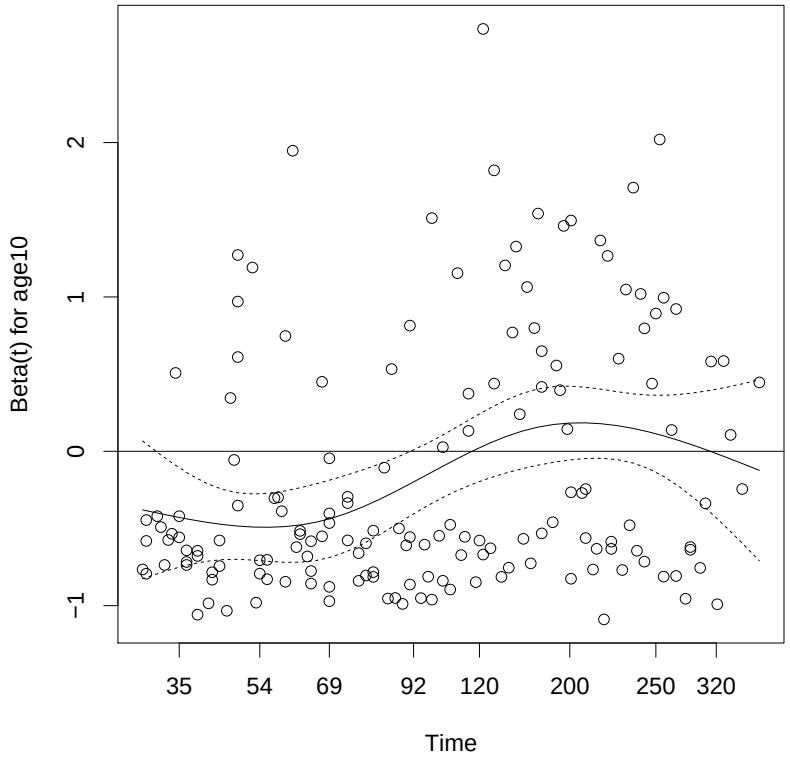
Plotting the scaled Schoenfeld residuals against transformed survival time allows for visual inspection of the PH assumption for a particular predictor. When the assumption is met, there is no correlation between the scaled Schoenfeld residuals and transformed time. Graphically, this means that a plot of the scaled Schoenfeld residuals has no trend. The interpretation of such a plot is greatly facilitated by superimposing a smoothing spline, which should be a horizontal line if the PH assumption is met. Systematic departures from this indicate non-proportional hazards (Grambsch and Therneau 1994).

Figure 6.2 is a diagnostic plot for the respondent age predictor[27]. The solid curve is the superimposed smoothing spline and the dashed lines represent 2 standard error envelopes around the fit. The smoothing spline clearly increases from left to right. The increase is not, however, continuous as the curve is approximately horizontal on the left

[27]In this chapter, too, we use respondents' age *divided by 10* in statistical models in order to make better use of the two decimal places used to report the effect sizes and their standard errors in the tables.

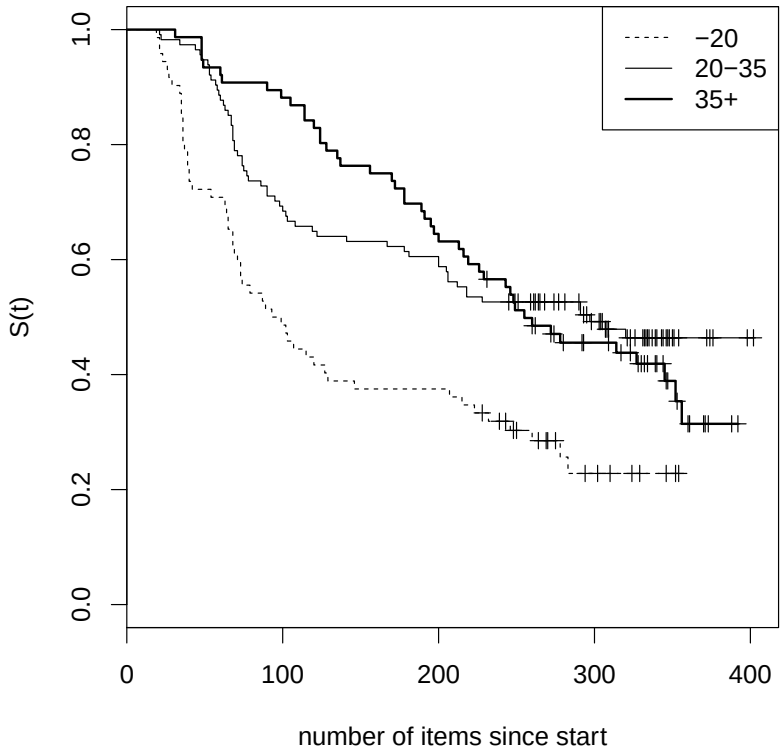
side of the chart and again approximately horizontal on the right, with an apparent increase around item 100. The shape of the curve therefore suggests that the PH assumption might hold for each of two distinct time intervals: before approximately the 100th item and after this threshold.

Figure 6.2: Scaled Schoenfeld residuals for respondent age; smoothing spline for regression coefficient from Cox model superimposed



The offending predictor is a respondent-level characteristic, which is why in this case we can examine another plot to aid interpretation. Figure 6.3 is a plot of survival curves for three age groups of respondents. We categorized the age into three categories of approximately the same size. Because respondents in the Facebook sample were predominantly young persons (see Table 4.11), the highest age group contains persons older than 35. The survival curve for this group decreases in a linear fashion from left to right. The survival curves for the two groups of younger respondents, however, first decrease sharply and then level off somewhat around the 100th item. The implication is that younger respondents break off at a higher rate during the initial phase of the interview and at a lower rate after the 100th item. One possible explanation is that a certain proportion of respondents in the lower age groups started the questionnaire with very low motivation. As these low-motivation respondents broke off shortly into the questionnaire, the more highly-motivated respondents were left, who then broke off at a lower rate.

Figure 6.3: Kaplan-Meier estimates for breakoff by age group in the Facebook sample



Figures 6.2 and 6.3 suggest that the effect of age on the risk of breakoff is negative at the start (older respondents have a lower risk of breakoff), but that this effect changes approximately after the 100th item, after which point the age effect may be zero or even positive. In order to accommodate this time-dependent age effect, we need to extend the Cox model by including in it the interaction of age and a function of the time variable.

Figures 6.2 and 6.3 suggest that a reasonable way of modeling the age effect is to assume that the hazard ratio for age assumes two different values, each value being constant over a fixed time interval. In such a scenario, Kleinbaum and Klein (2005) suggest the use of the so-called *heaviside function* of time in the interaction with the offending predictor. We will therefore include in the model the interaction of respondent age and the heaviside function $g(t)$, defined as:

$$g(t) = \begin{cases} 1 & \text{if } t \geq t_0 \\ 0 & \text{if } t < t_0. \end{cases} \tag{6.1}$$

Informed by the diagnostic charts, we will tentatively set $t_0 = 100$ as the cutoff point. There are two possible parametrizations for the extended model. The first one simply

adds the interaction term for age:

$$h(t, \mathbf{X}) = h_0(t) \cdot \exp\left\{\beta_{\text{sex}}\text{sex} + \beta_{\text{educ.}}\text{educ.} + \beta_{\text{age}}\text{age} + \delta_{\text{age}}\text{age} \cdot g(t)\right\}. \tag{6.2}$$

The parameter β_{age} is the regression coefficient for age when $t < t_0$. The regression parameter for age after the 100th item is calculated as the sum of β_{age} and δ_{age}. In order to simplify interpretation, we will use a slightly different parametrization with *two* heaviside functions $g_1(t)$ and $g_2(t)$:

$$h(t, \mathbf{X}) = h_0(t) \cdot \exp\left\{\beta_{\text{sex}}\text{sex} + \beta_{\text{educ.}}\text{educ.} + \delta_{\text{age1}}\text{age} \cdot g_1(t) + \delta_{\text{age2}}\text{age} \cdot g_2(t)\right\}, \tag{6.3}$$

where

$$g_1(t) = \begin{cases} 1 & \text{if } t < 100 \\ 0 & \text{if } t \geq 100, \end{cases} \tag{6.4}$$

and

$$g_2(t) = \begin{cases} 1 & \text{if } t \geq 100 \\ 0 & \text{if } t < 100. \end{cases} \tag{6.5}$$

Here, $g_1(t)$ is defined to equal 1 for times up to (but excluding) the 100th item, and zero otherwise; while $g_2(t)$ is defined as the converse: equal to zero in the beginning of the questionnaire and 1 after item 100. The parameters δ_{age1} and δ_{age2} can be interpreted directly as the regression parameters for age on the intervals below the 100th item and above this threshold, respectively.

In the survival package in R, estimating the extended model involves switching to a data format where each respondent is represented by several rows instead of one. In the case of model (6.3), two rows per respondent suffice, with one row pertaining to the time interval before the 100th item, and the other row to the time interval after this threshold (if the respondent broke off before the 100th item, this second row is omitted).

Table 6.7 is a fictitious example illustrating how the data were transformed. Instead of the time variable, two columns labeled *start* and *stop* appear marking the limits of the time interval to which the row pertains. The upper row for the first respondent pertains to the interval [1,100) in which the event (breakoff) did not take place. The lower row for the same respondent pertains to the interval [100,210], in which event=1, as the respondent broke off at item 210. Instead of the single age predictor, two columns appear in the data for the extended model. The first column is the respondent age multiplied by the function g_1 and the second column is the age multiplied by g_2 (see definitions above).

Table 6.7: Example data layout for the Cox PH and extended Cox models

Cox PH model data:

resp. ID	time	event	age
1	210	1	41
2	350	0	26
3	80	1	19
...	...	...	...

Extended Cox model data:

resp. ID	start	stop	event	age_lo	age_hi
1	1	100	0	41	0
1	100	210	1	0	41
2	1	100	0	26	0
2	100	350	0	0	26
3	1	80	1	19	0
...	...	...	...	...	...

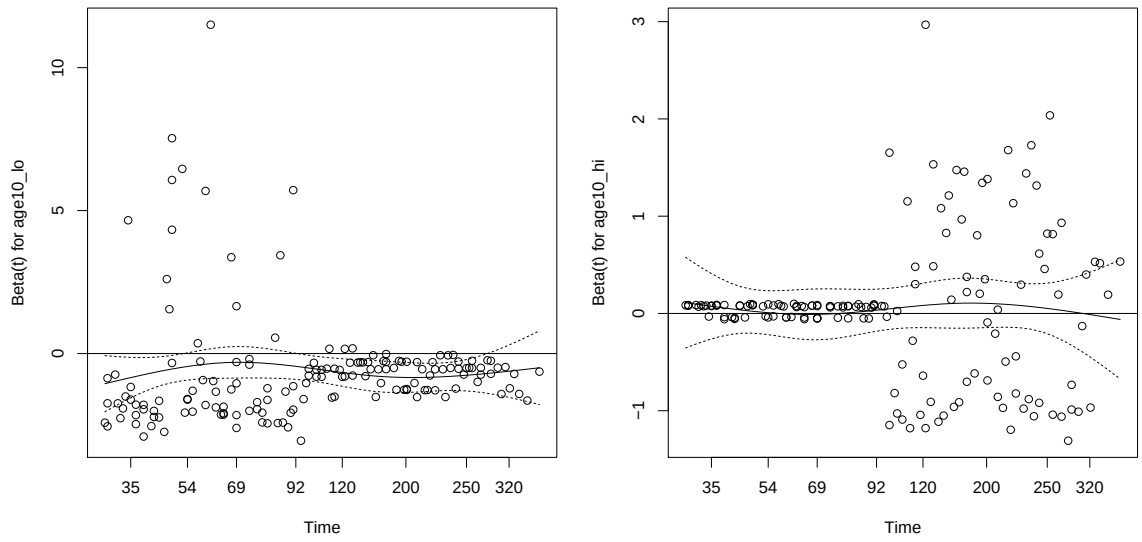
The extended model estimates are shown in Table 6.8. Allowing the regression coefficient for respondent age to assume different values before and after item 100 seems to have resolved the violation of the PH assumption. No p-value in the rightmost column of Table 6.8 is significant. Figure 6.4 shows the diagnostic plots of scaled Schoenfeld residuals for the time-dependent age predictors. The superimposed smoothing curves are now very close to being horizontal, indicating no violations of proportional hazards.

Table 6.8: Cox regression of breakoff on core respondent demographics; the effect of age is modeled separately for $t < 100$ and $t \geq 100$; data from the Facebook sample.

	est	se	z	Pr(>|z|)		exp(est)	[95% conf.	interval]	zph
male	0.23	0.18	1.30	0.20		1.26	0.89	1.79	0.81
age10 ($t < 100$)	-0.62	0.15	-4.28	0.00	**	0.54	0.40	0.71	0.74
age10 ($t \geq 100$)	0.05	0.07	0.69	0.49		1.05	0.92	1.20	0.93
education	-0.10	0.05	-2.13	0.03	*	0.90	0.82	0.99	0.14
c-index	0.64	0.02							
R^2	0.22								

As predicted, the effect of age on breakoff is negative before the 100th item: younger respondents have a higher risk of breaking off shortly into the questionnaire. After this threshold, the effect of age is zero. We will for now refrain from interpreting other model predictors, as all effects are liable to change as we continue to include additional predictors into the model.

Figure 6.4: Scaled Schoenfeld residuals for respondent age before the 100th item (left panel) and after the 100th item (right panel); smoothing spline for regression coefficient from Cox model superimposed



The R^2 estimates that appear in the tables throughout this chapter were calculated according to Equation (3.29) on page 76. The proportion of explained variation increases from 0.10 to 0.22 when the regression coefficient for respondent age is allowed to assume different values before and after the 100th item. The increase in the c-index is not as substantial (from 0.61 to 0.64). We expect to see further increases in the proportion of explained variation and c-index as we add predictors to the model.

Including item facets as predictors in the model

The next batch of predictors that we added were item facets. In order to estimate this model (and all subsequent models in this chapter), the data were converted to respondent-item layout: each respondent is represented by *as many rows as there were items* administered to them (this conversion is illustrated by Table 6.7 for the simpler case of (at most) two rows per respondent). If respondent 1 broke off at the 210th item, there are 210 rows representing him/her. The respondent's sex and education are constant across all rows. The item facets, on the other hand, change from row to row (from item to item)—they will be modeled as time-dependent predictors. The effect of respondent age is also modeled as time dependent: the two resulting predictors change value at the transition from the 99th to the 100th item.

In addition to the item facets that were used to explain item nonresponse in the previous section on item nonresponse, we have also included information on whether the item in question *introduces a new section*. Peytchev (2009) found breakoff to be

more likely at section introductions. Section introductions, according to Peytchev's interpretation, imply a logical break and require the respondent to commit to starting a new part of the survey. We have also included an indicator for whether the item was *required*.

Table 6.9 shows the results. Again, some of the predictors do not satisfy the proportional hazards assumption. The significant p-values in the rightmost column indicate that the Schoenfeld residuals are correlated with time for predictors disclosure[28], input numeric, and section intro.

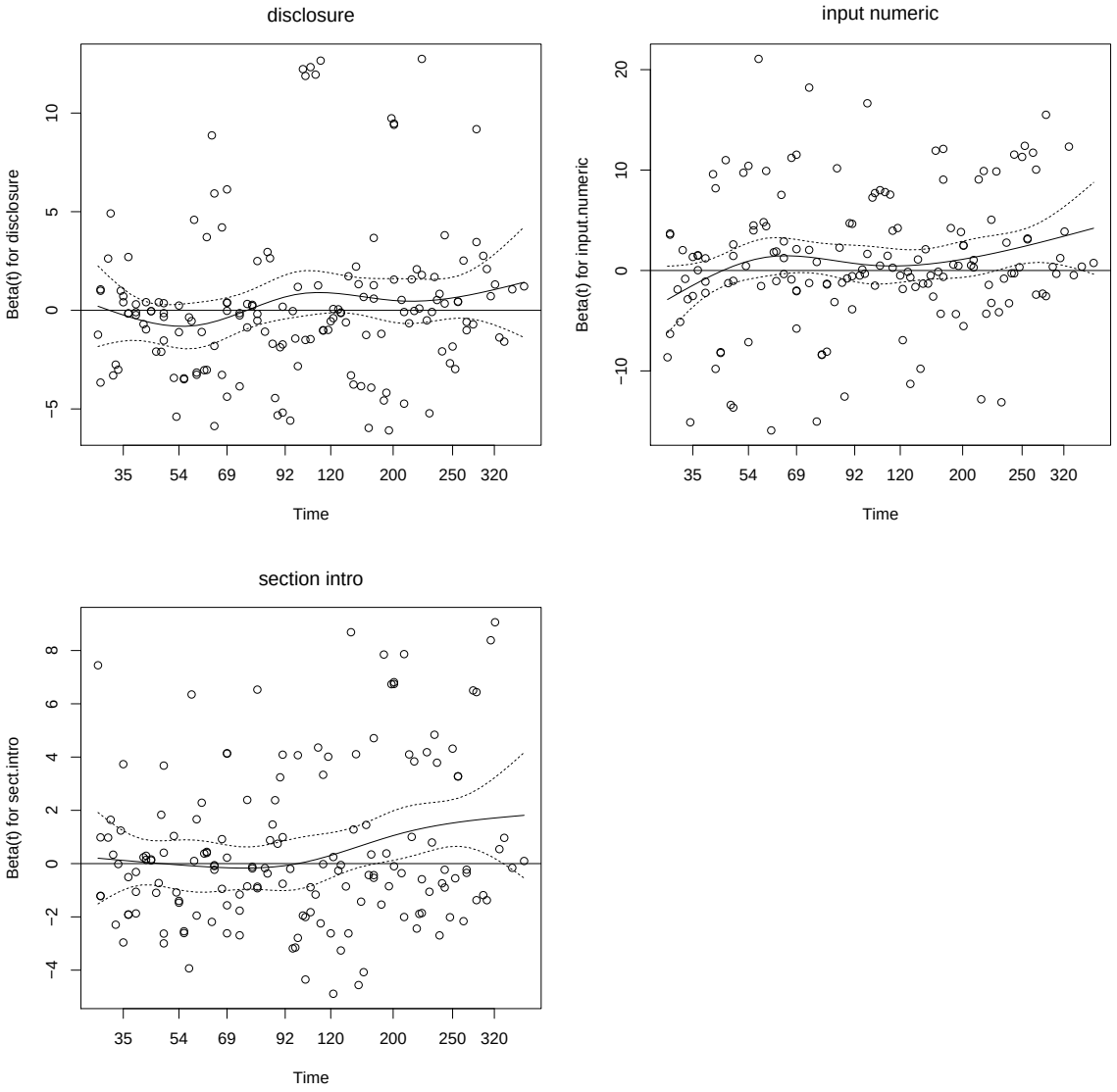
Table 6.9: Cox regression of breakoff on respondent demographics and item facets excluding MI predictors; data from the Facebook sample

	est	se	z	Pr(>|z|)		exp(est)	[95% conf.	interval]	zph
Respondent									
male	0.21	0.18	1.15	0.25		1.23	0.87	1.75	0.98
age10 ($t < 100$)	-0.58	0.14	-4.03	0.00	**	0.56	0.42	0.74	0.61
age10 ($t \geq 100$)	0.03	0.07	0.47	0.64		1.03	0.90	1.19	0.98
education	-0.10	0.05	-2.06	0.04	*	0.91	0.83	1.00	0.24
Item									
intrusiveness	-0.19	0.10	-1.86	0.06		0.83	0.68	1.01	0.68
disclosure	0.24	0.27	0.89	0.37		1.27	0.75	2.13	0.07
overclaiming	-0.61	0.18	-3.30	0.00	**	0.54	0.38	0.78	0.24
log(n. words)	0.34	0.14	2.47	0.01	*	1.40	1.07	1.83	0.19
log(n. alternatives)	0.68	0.18	3.81	0.00	**	1.98	1.39	2.82	0.46
input numeric	0.90	0.43	2.08	0.04	*	2.46	1.05	5.74	0.02
input string	2.71	0.40	6.81	0.00	**	14.98	6.87	32.66	0.52
radio yes	-1.38	1.03	-1.34	0.18		0.25	0.03	1.88	0.90
section intro	0.47	0.22	2.12	0.03	*	1.61	1.04	2.49	0.01
required	0.59	0.26	2.25	0.02	*	1.80	1.08	3.01	0.15
c-index	0.74	0.02							
R^2	0.58								

Figure 6.5 is the diagnostic plot that allows a visual inspection of the proportional hazards assumption. The plot for the section introduction indicator, e.g., suggests that the proportional hazards assumption holds for the first 100 items. After the 100th item, the curve seems to increase in an an approximately linear fashion. We could therefore proceed by including a predictor term like sect.intro $\cdot\, g_2(t)$ to allow a different (but constant) value of the regression coefficient for the interval above item 100, and another term like sect.intro $\cdot\, g_2(t) \cdot t$ to also allow the regression coefficient to increase with time. We are weary, however, of including three parameters to model

[28] Even though the p-value for disclosure is not lower than 0.05, we consider it "marginally significant", and thus a reason for concern.

Figure 6.5: Scaled Schoenfeld residuals for item facets violating the proportional hazards assumption; smoothing spline for regression coefficient from Cox model superimposed



the effect of a single item facet. In order to keep the model as parsimonious as possible and avoid overfitting, we will tentatively model all offending item facets in the same manner as respondent age: by allowing the regression coefficient to assume a different value before item 100 and after this threshold.

The estimates for the fitted model are given in Table 6.10. Partitioning the time scale into two parts and allowing the regression coefficients to assume different values before and after item 100 seems to have resolved the violation of the PH assumption in this case too. We find no significant p-values in the rightmost column of Table 6.10. Figure 6.6 shows the diagnostic plots of the Schoenfeld residuals and tells a somewhat different story. The smoothing splines in some plots still show departures from a horizontal line. The departures, however, are less severe than in Figure 6.8.

Because the statistical test of the correlation of Schoenfeld residuals to transformed time yields no significant p-values, we will consider the non-proportional hazards issue to have been addressed and will not introduce additional parameters into the model.

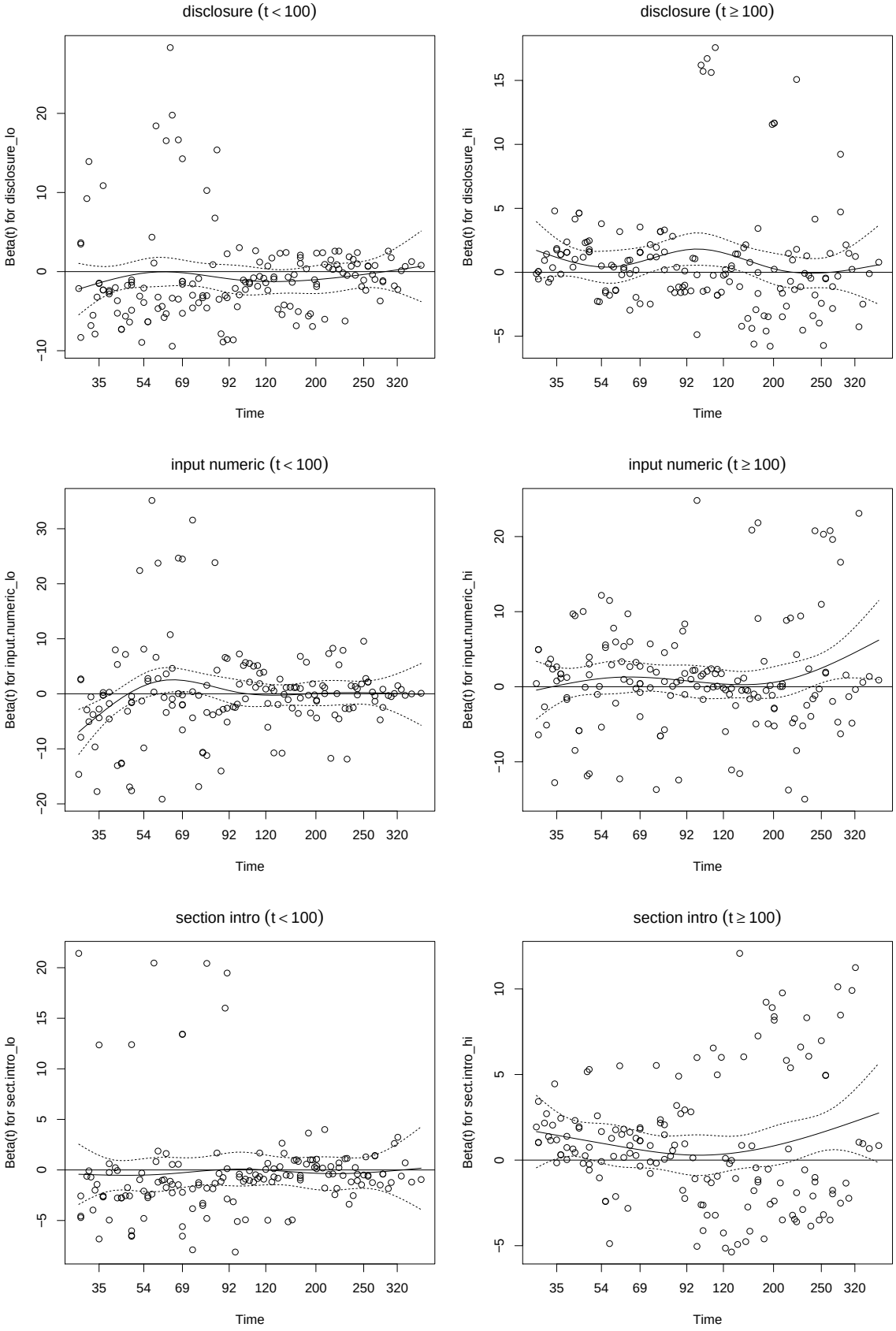
Table 6.10: Cox regression of breakoff on respondent demographics and item facets excluding MI predictors; data from the Facebook sample

	est	se	z	Pr(>|z|)		exp(est)	[95% conf.	interval]	zph
Respondent									
male	0.20	0.18	1.09	0.28		1.22	0.85	1.73	0.98
age10 ($t < 100$)	-0.55	0.14	-3.87	0.00	**	0.58	0.44	0.76	0.78
age10 ($t \geq 100$)	0.03	0.07	0.49	0.62		1.04	0.90	1.19	0.99
education	-0.10	0.05	-2.11	0.03	*	0.90	0.82	0.99	0.22
Item									
intrusiveness	-0.23	0.10	-2.27	0.02	*	0.79	0.65	0.97	0.57
disclosure ($t < 100$)	-0.73	0.42	-1.72	0.09		0.48	0.21	1.11	0.71
disclosure ($t \geq 100$)	0.81	0.29	2.77	0.01	**	2.25	1.27	3.99	0.34
overclaiming	-0.49	0.19	-2.56	0.01	*	0.61	0.42	0.89	0.36
log(n. words)	0.31	0.14	2.22	0.03	*	1.36	1.04	1.79	0.37
log(n. alternatives)	0.55	0.18	3.01	0.00	**	1.73	1.21	2.46	0.38
input numeric ($t < 100$)	0.02	0.53	0.03	0.97		1.02	0.36	2.89	0.32
input numeric ($t \geq 100$)	1.13	0.50	2.26	0.02	*	3.10	1.16	8.26	0.16
input string	2.30	0.41	5.67	0.00	**	9.94	4.49	22.01	0.52
radio yes	-1.33	1.03	-1.29	0.20		0.26	0.04	1.99	0.86
section intro ($t < 100$)	-0.25	0.39	-0.65	0.52		0.78	0.36	1.67	0.80
section intro ($t \geq 100$)	0.96	0.28	3.48	0.00	**	2.61	1.52	4.47	0.66
required	0.52	0.26	1.98	0.05	*	1.68	1.00	2.81	0.13
c-index	0.75	0.02							
R^2	0.63								

The model estimates indicate that the items' threat of disclosure has no effect shortly into the questionnaire, whereas after the 100th item the threat of disclosure increases the risk of breakoff. Similar interpretations apply to items that require numeric input and items that introduce a new section. These item facets have no effect on the risk of breakoff shortly into the questionnaire, whereas after the 100th item they increase the risk of breakoff.

We want to call attention to one particular result reported in Table 6.10. The estimated effect of item intrusiveness is statistically significant and has a *negative* sign, indicating that the risk of breakoff is *lower* at items that are more intrusive. This runs contrary to our expectations, as we hypothesized that intrusive items would induce *more* breakoff. Because we have no explanation for this result, we decided to investigate the matter further by including the *lagged* intrusiveness as a predictor in the model (see Kleinbaum and Klein 2005, 220 for a discussion of lag-time effects). By including the item intrusiveness at different lags in the model we want to investigate

Figure 6.6: Scaled Schoenfeld residuals for item facets violating the proportional hazards assumption; plots for model with two parameters per offending predictor: before the 100th item (left panels) and after the 100th item (right panels); smoothing spline for regression coefficient from Cox model superimposed



whether an intrusive item induces more breakoff *with a delay*.

In order to explore the possibility of positive lagged effects of item intrusiveness, we included this predictor in the model at lags spanning from one to five. Indeed, we found that at lag = 2 the effect of intrusiveness on breakoff is positive and significant. When the unlagged intrusiveness and intrusiveness with lag = 2 are included in the model, the effect of intrusiveness at any other lag (spanning from one to five) is insignificant. The estimates for the model with added intrusiveness at lag = 2 are reported in Table 6.11. The results indicate that the respondents are *less* likely to break off at more intrusive items (the finding already reported by Table 6.10), but that they are *more* likely to break off two items later.

Table 6.11: Cox regression of breakoff on respondent demographics and item facets including the lagged intrusiveness predictor; data from the Facebook sample

	est	se	z	Pr(>|z|)		exp(est)	[95% conf.	interval]	zph
Respondent									
male	0.20	0.18	1.09	0.28		1.22	0.86	1.73	0.98
age10 ($t < 100$)	-0.56	0.14	-3.88	0.00	**	0.57	0.43	0.76	0.78
age10 ($t \geq 100$)	0.04	0.07	0.53	0.59		1.04	0.90	1.19	0.98
education	-0.10	0.05	-2.10	0.04	*	0.90	0.82	0.99	0.23
Item									
intrusiveness	-0.39	0.11	-3.44	0.00	**	0.68	0.55	0.85	0.88
intrusiveness lag2	0.30	0.10	3.18	0.00	**	1.35	1.12	1.63	0.43
disclosure ($t < 100$)	-0.59	0.43	-1.39	0.17		0.55	0.24	1.28	0.80
disclosure ($t \geq 100$)	0.94	0.30	3.14	0.00	**	2.56	1.42	4.61	0.29
overclaiming	-0.50	0.19	-2.60	0.01	**	0.61	0.42	0.88	0.38
log(n. words)	0.33	0.14	2.32	0.02	*	1.39	1.05	1.83	0.40
log(n. alternatives)	0.56	0.18	3.02	0.00	**	1.74	1.22	2.51	0.80
input numeric ($t < 100$)	0.01	0.53	0.01	0.99		1.01	0.35	2.86	0.50
input numeric ($t \geq 100$)	1.08	0.50	2.15	0.03	*	2.95	1.10	7.88	0.30
input string	2.40	0.41	5.90	0.00	**	11.01	4.97	24.42	0.78
radio yes	-1.47	1.03	-1.43	0.15		0.23	0.03	1.73	0.87
section intro ($t < 100$)	-0.41	0.40	-1.03	0.30		0.66	0.31	1.44	0.69
section intro ($t \geq 100$)	0.83	0.28	2.99	0.00	**	2.29	1.33	3.95	0.60
required	0.51	0.26	1.96	0.05	*	1.67	1.00	2.78	0.35
c-index	0.75	0.02							
R^2	0.63								

Including item facets as predictors dramatically increases the model's predictive performance. The c-index increases from 0.64 (see Table 6.8) to 0.75, indicating that model predictions for three quarters of all usable respondent pairs are concordant with actual outcomes. Put more concretely, if the model predicts a certain respondent A to "survive" longer than another respondent B, this prediction will be correct for three quarters of all respondent pairs in which at least one respondent broke off (see

Section 3.3.7 for the definition of the c-index). The proportion of explained variation increases from 0.22 to 0.63, when item facets are added as predictors.

In this section, we have fitted the Cox PH model to time-independent predictors and evaluated the PH assumption. When the assumption was found to be suspect for certain predictors, we demonstrated a graphical approach of how to diagnose the issue and adjust for it. We then expanded the set of predictors with item facets, which required us to use the extended Cox model. The graphical approach to testing the PH assumption was simplified by the fact that in this section we did not include any multiply imputed predictors. The resulting model predicted that more intrusive items lower the risk of breakoff. As we had no explanation for this finding, we explored the possibility of a delayed positive effect of intrusiveness on the risk of breakoff. Indeed we found such a positive and significant effect at lag = 2 indicating that intrusive items increase the risk of breakoff two items later. In the following section we will include multiply imputed predictors in the model.

6.4 Cox models including multiply imputed predictors

Multiply imputed predictors introduce an additional level of complexity into the model. As mentioned, in the presence of multiply imputed predictors, the model has to be fit multiple times, resulting in several (in our case, five) sets of estimates. The estimates are then pooled to obtain a single value for each predictor. The pooling procedure is very clear for the model's fixed effects (Rubin 1987), but procedures for other quantities are less clear.

- We report the *minimum* p-value across the five MI datasets as the "pooled outcome" of the test for the proportional hazards assumption. The rationale is that the minimum p-value reflects the worse-case scenario: if the p-value is greater than some prescribed value (e.g. 0.05) in *each* of the five models, then there is no evidence that the Schoenfeld residuals for the predictor in question are correlated to transformed survival time. We feel that this conservative way of pooling is appropriate for testing the PH assumption.
- The values of the proportion of explained variation were pooled according to the procedure suggested by Harel (2009). We take the square root of the R^2, apply Fisher's z-transformation, and calculate the *simple average* of the five transformed values. We then back-transform the pooled value.
- Because the software returns not only the estimate but also the standard error for the c-index, we applied Rubin's rules to pool the five values of the c-index.

As no consensus has been reached on a procedure for pooling the c-index in the multiple imputation literature, other authors (Clark and Altman 2003) have reported the *median* and the range of the c-index across the MI datasets. In the case of the GGP data, the proportion of missing values is sufficiently low that the median does not differ from our reported value in two-digit precision.

Table 6.12 reports the estimates for the Cox model that includes MI predictors: the respondent's attitude toward surveys and item-by-respondent interactions. The rightmost column gives the minimal p-value (across the five MI datasets) for the test of the PH assumption. No p-value is low enough to cause concern.

Table 6.12: Cox regression of breakoff on respondent demographics and item facets; data from the Facebook sample

	est	se	sig		exp(est)	[95% conf.	interval]	min(zph)
Respondent								
male	0.16	0.18	0.36		1.18	-0.19	0.52	0.91
age10 ($t < 100$)	-0.55	0.14	0.00	**	0.57	-0.83	-0.28	0.78
age10 ($t \geq 100$)	-0.02	0.07	0.79		0.98	-0.15	0.12	0.92
education	-0.09	0.05	0.06		0.92	-0.18	0.00	0.21
attitude toward surveys [MI]	-0.34	0.11	0.00	**	0.71	-0.55	-0.13	0.70
Item								
intrusiveness	-0.45	0.12	0.00	**	0.64	-0.68	-0.22	0.47
intrusiveness lag2	0.29	0.10	0.00	**	1.34	0.10	0.48	0.70
disclosure ($t < 100$)	-0.34	0.43	0.43		0.71	-1.19	0.51	0.96
disclosure ($t \geq 100$)	1.17	0.31	0.00	**	3.23	0.56	1.78	0.12
overclaiming	-0.56	0.20	0.00	**	0.57	-0.94	-0.17	0.23
log(n. words)	0.23	0.14	0.11		1.25	-0.05	0.51	0.61
log(n. alternatives)	0.62	0.19	0.00	**	1.86	0.25	0.98	0.83
input numeric ($t < 100$)	-0.02	0.55	0.97		0.98	-1.10	1.05	0.54
input numeric ($t \geq 100$)	1.23	0.52	0.02	*	3.42	0.22	2.24	0.29
input string	2.38	0.42	0.00	**	10.82	1.56	3.20	0.96
radio yes	-1.25	1.04	0.23		0.29	-3.29	0.78	0.89
section intro ($t < 100$)	-0.40	0.39	0.31		0.67	-1.16	0.37	0.77
section intro ($t \geq 100$)	0.91	0.28	0.00	**	2.48	0.35	1.47	0.51
required	0.48	0.27	0.07		1.62	-0.04	1.01	0.36
Item-by-respondent interaction								
partner activity [MI]	0.26	0.35	0.46		1.30	-0.42	0.94	0.28
personal information [MI]	-0.02	0.13	0.88		0.98	-0.27	0.23	0.86
rel. quality [MI]	-0.02	0.19	0.92		0.98	-0.39	0.36	0.63
HH finances [MI]	-0.26	0.43	0.55		0.77	-1.09	0.58	0.22
networks [MI]	0.36	0.06	0.00	**	1.43	0.23	0.49	0.95
rel. partner [MI]	0.04	0.08	0.61		1.04	-0.12	0.21	0.41
rel. parents [MI]	-0.15	0.16	0.36		0.86	-0.45	0.16	0.23
having children [MI]	-0.02	0.14	0.91		0.98	-0.30	0.27	0.85
income [MI]	-0.03	0.10	0.76		0.97	-0.24	0.17	0.09
values [MI]	0.14	0.16	0.36		1.15	-0.16	0.45	0.31
c-index	0.80	0.02						
R^2	0.73							

The great majority of the item-by-respondent interactions have effects that are not significant at the alpha=0.05 level. Because we want to keep the model as parsimonious as possible, we remove the insignificant predictors. The resulting model is shown in Table 6.13. We consider this model the final model for breakoff: we will interpret the coefficients and evaluate the hypotheses based on this model's estimates.

Table 6.13: Cox regression of breakoff on respondent demographics and item facets; insignificant item-by-respondent interactions removed; data from the Facebook sample

	est	se	sig		exp(est)	[95% conf.	interval]	min(zph)
Respondent								
male	0.17	0.18	0.34		1.19	-0.18	0.52	0.92
age ($t < 100$)	-0.54	0.14	0.00	**	0.58	-0.82	-0.27	0.81
age ($t \geq 100$)	-0.02	0.07	0.81		0.98	-0.15	0.12	0.88
education	-0.09	0.05	0.05		0.92	-0.18	0.00	0.29
attitude toward surveys [MI]	-0.35	0.11	0.00	**	0.71	-0.56	-0.13	0.72
Item								
intrusiveness	-0.48	0.11	0.00	**	0.62	-0.70	-0.25	0.72
intrusiveness lag2	0.29	0.10	0.00	**	1.33	0.10	0.47	0.53
disclosure ($t < 100$)	-0.37	0.43	0.40		0.69	-1.21	0.48	0.87
disclosure ($t \geq 100$)	1.14	0.30	0.00	**	3.14	0.55	1.74	0.18
overclaiming	-0.54	0.19	0.00	**	0.59	-0.91	-0.16	0.48
log(n. words)	0.21	0.14	0.12		1.24	-0.06	0.48	0.53
log(n. alternatives)	0.61	0.18	0.00	**	1.84	0.25	0.97	0.65
input numeric ($t < 100$)	-0.09	0.53	0.86		0.91	-1.13	0.94	0.38
input numeric ($t \geq 100$)	1.13	0.50	0.02	*	3.10	0.15	2.11	0.27
input string	2.38	0.41	0.00	**	10.77	1.58	3.18	0.71
radio yes	-1.27	1.03	0.22		0.28	-3.29	0.76	0.77
section intro ($t < 100$)	-0.39	0.39	0.32		0.68	-1.15	0.38	0.74
section intro ($t \geq 100$)	0.90	0.28	0.00	**	2.46	0.35	1.45	0.63
required	0.47	0.26	0.07		1.60	-0.04	0.98	0.29
Item-by-respondent interaction								
networks [MI]	0.36	0.06	0.00	**	1.43	0.23	0.49	0.91
c-index	0.80	0.02						
R^2	0.72							

The respondent sex has an insignificant effect in all fitted models. The effect of respondent age is negative (-0.54) for the first one hundred items. The exponentiated value of the estimate (the column entitled exp(est)) is commonly referenced in order to aid the interpretation of Cox model estimates. The risk of breakoff during the first one hundred items is lower by a factor of 0.58 for every additional ten years of the respondent's age. After the 100th item, the respondent's age no longer affects the risk of breakoff. The data therefore do not support Hypothesis 2b.1, which predicted that older respondents would be *more* prone to breakoff. As mentioned, we believe that the risk of breakoff is lower for younger respondents because there was a higher proportion of low-motivation respondents among the young ones, and these low-motivation respondents apparently broke off at a high rate during the first 100 items.

We hypothesized that education can serve as a proxy for respondent cognitive ability and that a higher degree of education would be related to less breakoff. The effect of respondent education is negative, in line with Hypothesis 2b.2. The risk of breakoff is lower by a factor of 0.92 for every unit increase in the respondent's education (the education was measured on a scale from 1 to 9).

The respondent's attitude toward surveys also has the hypothesized effect on breakoff, as the risk of breakoff is lower by a factor of 0.71 for every point increase in the attitude toward surveys. The respondent's attitude toward surveys was measured on a scale from 1 to 5, with higher values signifying a more positive attitude toward surveys in general.

Hypothesis 4b predicted item intrusiveness to be related to more breakoff. The model estimates, on the other hand, suggest that the effect of item intrusiveness is more complex. The risk of breakoff at a particular item is *lower* by a factor of 0.62 for every point increase in this item's intrusiveness (intrusiveness was rated on a scale from 1 to 5). The risk of breakoff at lag = 2, however, is *higher* by a factor of 1.33 for every point increase in the item's intrusiveness. A highly intrusive item will therefore temporarily lower the risk of breakoff while increasing the risk of breakoff two items further into the questionnaire.

The effects of the remaining two ratings of item sensitivity have the direction predicted by Hypothesis 4b. The risk of breakoff is lower by a factor of 0.59 for every point increase in the item's potential for overclaiming (overclaiming was rated on a scale from 1 to 3). In order to keep the proportional hazards assumption tenable, we allowed the regression coefficient of the item's threat of disclosure to assume a different value during the first one hundred items and further into the questionnaire. The figures in Table 6.13 indicate that the effect of this predictor is insignificant shortly into the questionnaire and positive after the 100th item. The risk of breakoff after the 100th item is higher by a factor of 3.14 for every point increase in an item's threat of disclosure (the threat of disclosure was measured on a scale from 1 to 3).

Hypothesis 4b also predicted the respondents' self-assessments of the sensitivity of certain item topics to be related to more breakoff. The model estimates for the item-by respondent interactions provide very limited support for this hypothesis. Five out of six[29] predictors have insignificant effects (see Table 6.12). Only the *networks* predictor is significant and has the hypothesized direction; the risk of breakoff is higher at items

[29]The *relationship with children* predictors had to be removed from all models because *all* breakoffs had occurred at items where the value of this item-by-respondent interaction was zero (see Section 4.3, page 99 for details on the coding of this variable). This causes a problem with the estimation procedure: the log-likelihood is maximal at infinity.

that concern social networks for those respondents who replied that items concerning social networks would be more sensitive for them to answer.

Hypothesis 5b predicted difficult items to be associated with more breakoff. We measured an item's difficulty in terms of 1) its complexity, and 2) respondent's familiarity with the item topic. The effect of all item-by-respondent interactions is non-significant (see Table 6.12). All self-assessments of cognitive state were thus removed from the final model. The effect of the wording length is positive, but also fails to reach the prescribed level of statistical significance. The effect of the number of alternatives, on the other hand, is highly significant: the risk of breakoff increases with the number of response alternatives. We consider the data to provide very limited support of Hypothesis 5b.

We allowed the regression coefficient to assume two different values for two more predictors: the indicator for numeric input items and the indicator for items that introduce a new section. The model estimates give similar conclusions about the effect of both variables. During the initial one hundred items, such items have no effect on breakoff. After the 100th item, however, both numeric input items and section introductions increase the risk of breakoff; the former by a factor of 3.10 and the latter by a factor of 2.46.

String input items, unlike numeric input items, increase the risk of breakoff throughout the questionnaire. The risk of breakoff when the respondent is administered a string input item is higher by a factor of 10.77 in comparison to closed-ended items. For comparison with models for item nonresponse, we also included the indicator for *radio yes* items in the model for breakoff. Such items induce more item nonresponse not because they are more difficult or sensitive, but because of the peculiar item format. Such items have no significant effect on breakoff, as one might expect.

Unlike in the item nonresponse analyses in the previous chapter, required items were included in the analysis of breakoff. We would expect such items to *increase* the risk of breakoff, because the error messages insisting that a reply be given are a nuisance that might cause the respondent a degree of frustration. The effect of required items on breakoff is, indeed, positive but, with a p-value of 0.07, does not quite reach the threshold for statistical significance.

In this section we added multiply imputed predictors to the model. We interpreted the effect of respondent characteristics, item facets, and item-by-respondent interactions on the risk for breakoff, and evaluated our hypotheses. We will continue our discussion of the results in the final chapter of this dissertation, where we draw conclusions from both our analysis of item nonresponse and the analysis of breakoff that have so far

been treated separately. We explore the connection between item nonresponse and breakoff in the next section.

6.5 Item nonresponse and breakoff

The analysis presented in this section will be somewhat exploratory. We will be interested in the interplay of item nonresponse and breakoff. The rationale behind our analysis is based on the work of (Galesic 2006), who, in a web survey of the unemployed, found item nonresponse to be more frequent immediately prior to breakoff.

Galesic divided the questionnaire into blocks and asked the respondents after each block to assess their level of burden and interest in the survey, finding that respondents who break off often report lower interest and higher burden than respondents who continue. Galesic explains the respondents' behavior by referring to decision field theory. At the beginning of the survey, the factors that influenced the respondent to participate (incentives, general interest in the topic, etc.) are still influential, but as the survey continues, the negative aspects of participation (fatigue and boredom) become stronger. The respondent would at this point prefer to stop participating, but will not opt for this change until their preference for breakoff exceeds the inhibitory threshold (Galesic 2006, see also Section 2.3).

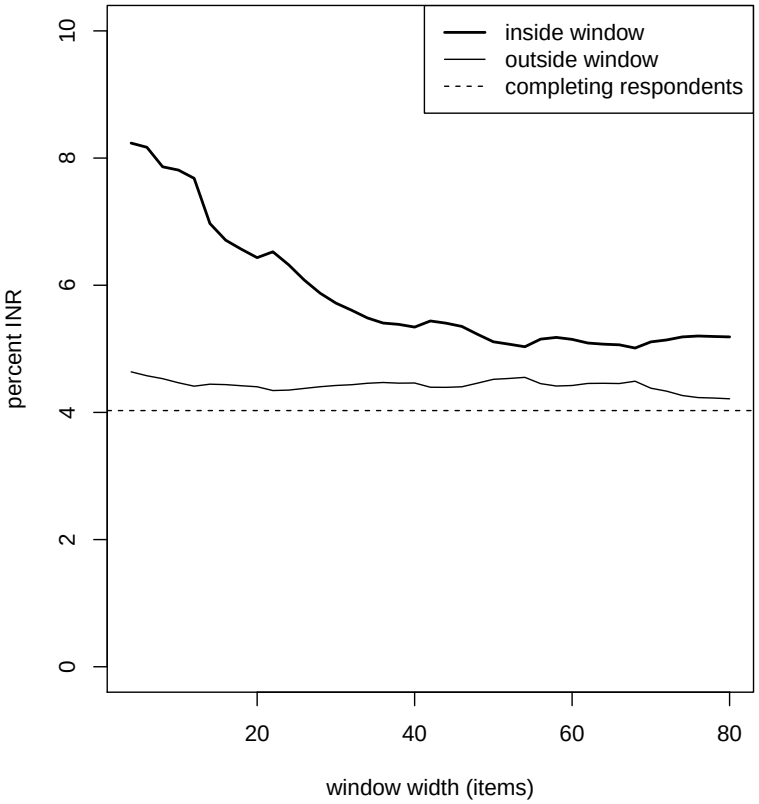
Galesic also found respondents to produce more item nonresponse prior to breakoff. Item blocks immediately preceding breakoff had an item nonresponse rate of 11% while the average item nonresponse rate was 6% for item blocks not immediately preceding breakoff. The item nonresponse rate for completing respondents was lower still at 4%. Galesic's explanation for these findings is that respondents who would already prefer to terminate the interview, but have not yet decided to do so, are characterized by lower motivation. This lack of motivation is then expressed by a lower level of quality in their answers, which is in turn reflected in the item nonresponse rate (Galesic 2006).

We will try to replicate Galesic's results on the Facebook sample. Because the GGP questionnaire was not partitioned into blocks of items, we will define the interval "immediately prior to breakoff" in terms of the *number of items before breakoff*, and will refer to this interval as the *window* before breakoff. A 10-item window therefore refers to the ten items directly preceding the item at which breakoff occurred. Following Galesic's rationale, we would expect the item nonresponse rate to be higher the closer the respondent is to breakoff.

Figure 6.7 allows a graphical evaluation of this hypothesis. The rate of item nonresponse before breakoff was calculated for windows of *differing width*. For respondents

who broke off, we took n items preceding breakoff and calculated the item nonresponse rate by dividing the number of item nonresponses by the number of non-required items in the window. The thick line in Figure 6.7 plots this item nonresponse rate against the window width, n. The curve increases toward the left, indicating that the item nonresponse rate, indeed, is higher closer to breakoff.

Figure 6.7: Weighted item nonresponse percentage, calculated in the window before breakoff; data from the Facebook sample



The unbroken thin line gives the item nonresponse rate for all items outside the window. Because items lying outside the window are the complement of those within the window, the shape of the thin curve somewhat mirrors the shape of the thick one. Finally, the broken line gives the weighted[30] item nonresponse percentage for *completing* respondents. This is independent of the window width and is therefore shown in Figure 6.7 as a constant line. The item nonresponse rate is highest for items preceding breakoff, lower for items not preceding breakoff, and lower still for completing respondents. We have, therefore, found the exact same pattern as Galesic in her dataset.

We additionally wish to explore whether this information can be used in the model for breakoff to improve the prediction of a given respondent's risk of breakoff to the current

[30]See the discussion pertaining to Equation (5.1) on page 122 for an explanation of what we mean by weighting.

item they are being administered. We thus constructed a time-dependent predictor variable reflecting the respondent's tendency to produce item nonresponse in the n items preceding the current one. In order to specify this predictor, we need to set a value for the window width n. Because the item nonresponse rate seems to gradually increase before breakoff, we would prefer a narrower window to exclude items that are not recent and therefore do not reflect the respondent's *current* motivation. On the other hand, the item nonresponse rate in a very narrow window will be more prone to random fluctuations, and as such will not serve well as a predictor of breakoff. We chose a window width of $n = 10$ on the basis of Figure 6.7. The thick curve shows a dramatic increase when window width is decreased to ten items and then levels off somewhat.

The item nonresponse rate in the window preceding the current item heavily depends on its position in the questionnaire: similar items often appear together in the questionnaire, and if the respondent does not respond to one, they are likely not to respond to other similar items either. We mentioned this in Section 5.3 as the respondents' tendency to produce item nonresponses in series. We will apply the *square root transformation* to the number of recent item nonresponses, believing as we do the difference between low counts of nonresponse (e.g., the difference between one item nonresponse and no item nonresponse) in the 10-item window to be more important than the difference between high counts (e.g., between 8 and 9 item nonresponses). We also want to adjust for required items in the window, which is why we define the predictor as:

$$\sqrt{\frac{n}{n-r} \cdot k}, \tag{6.6}$$

where n is window width (in the number of items), k is the number of item nonresponses in the window, and r is the number of required items in the window. Equation (6.6) therefore gives the square root of the *number* of item nonresponses in the n-item window, *adjusted for required items.*

Table 6.14 shows the estimates for the Cox model with the added recent item nonresponse predictor. The effect of the measure of recent item nonresponse is positive, as expected: the risk of breakoff is higher if the respondent produced more item nonresponse in the ten items preceding the current one. While we already suspected this to be the case after examining Figure 6.7, fitting the statistical model with this predictor provides a significance test for the effect of recent item nonresponse. The highly significant p-value indicates that the probability that the effect's direction is an artefact of small sample size is very low. The addition of the measure of recent item nonresponse to the model increases the proportion of explained variation from 0.72 to 0.75

Table 6.14: Cox regression of breakoff on respondent demographics and item facets with the additional predictor for recent item nonresponse; data from the Facebook sample

	est	se	sig		exp(est)	[95% conf.	interval]	min(zph)
Respondent								
male	0.13	0.18	0.46		1.14	-0.22	0.49	0.74
age10 ($t < 100$)	-0.53	0.14	0.00	**	0.59	-0.80	-0.26	0.78
age10 ($t \geq 100$)	-0.01	0.07	0.90		0.99	-0.14	0.13	0.85
education	-0.09	0.05	0.05	*	0.91	-0.18	0.00	0.30
attitude toward surveys [MI]	-0.34	0.11	0.00	**	0.71	-0.55	-0.12	0.77
Item								
intrusiveness	-0.49	0.11	0.00	**	0.61	-0.72	-0.27	0.65
intrusiveness lag2	0.28	0.10	0.00	**	1.32	0.09	0.47	0.42
disclosure ($t < 100$)	-0.39	0.43	0.37		0.68	-1.24	0.46	0.90
disclosure ($t \geq 100$)	1.16	0.31	0.00	**	3.18	0.55	1.76	0.16
overclaiming	-0.52	0.19	0.01	**	0.59	-0.90	-0.15	0.42
log(n. words)	0.21	0.14	0.13		1.24	-0.06	0.48	0.49
log(n. alternatives)	0.64	0.18	0.00	**	1.89	0.28	1.00	0.55
input numeric ($t < 100$)	-0.13	0.53	0.81		0.88	-1.16	0.91	0.32
input numeric ($t \geq 100$)	1.16	0.50	0.02	*	3.17	0.17	2.14	0.23
input string	2.39	0.41	0.00	**	10.96	1.60	3.19	0.73
radio yes	-1.20	1.03	0.25		0.30	-3.22	0.82	0.80
section intro ($t < 100$)	-0.37	0.39	0.34		0.69	-1.13	0.39	0.73
section intro ($t \geq 100$)	0.91	0.28	0.00	**	2.48	0.36	1.46	0.64
required	0.43	0.26	0.10		1.54	-0.08	0.93	0.22
Item-by-respondent interaction								
networks [MI]	0.37	0.07	0.00	**	1.44	0.24	0.49	0.89
Recent item nonresponse								
sqrt(INR last 10)	0.43	0.10	0.00	**	1.54	0.24	0.62	0.18
c-index	0.81	0.02						
R^2	0.75							

In this chapter we have first examined the differences in breakoff rates across modes of administration by analyzing the data from the first two rounds of data collection. We then proceeded to analyze the Facebook sample by applying methods of survival analysis. The final section explored the effect of recent item nonresponse on the risk of breakoff on the currently administered item. We move on to draw conclusions from both the analysis of item nonresponse and breakoff in the final chapter.

7 Conclusion

In this final chapter, we first provide an overview of the material presented in Chapters 2 through 4: our theoretical discussion of item nonresponse and breakoff, the statistical models that were applied, and the methodology of our empirical research. Because heretofore item nonresponse and breakoff have been discussed separately, the second section provides a joint interpretation of the findings for item nonresponse and breakoff. We conclude by highlighting the original contributions of this dissertation and its implications for survey methodology and finally by acknowledging the study's limitations and suggesting lines for further research.

7.1 Overview

The literature on the question-answer process (e.g. Sudman and Bradburn 1974; Sudman et al. 1996; Schwarz and Sudman 1996; Tourangeau et al. 2000) explains the survey interview in terms of concepts from cognitive psychology has been very influential for survey methodology. The authors generally agree that the question-answer process consists of a series of processes that the respondent perform in order to reply to a questionnaire item: comprehension, retrieval, judgment and formatting. Some later authors argue this progression through the phases is merely an ideal that is rarely achieved in practice because most respondents become increasingly fatigued and disinterested in performing the four processes carefully and comprehensively. As this happens, respondents are likely to shift their response strategy to what Krosnick (1991) terms *satisficing*.

The question-answer literature forms the basis of more specific frameworks in survey methodology; authors refer to it to explain respondent behavior like item nonresponse and breakoff. As yet, the only theoretical framework aimed specifically at explaining item nonresponse is Beatty and Herrmann's *response decision model* (2002). The authors argue that three factors drive the respondent's decision whether to respond or not: 1) the availability of the requested information (what they term the respondent's *cognitive state*), 2) the respondent's perception of the required level of accuracy (*adequacy judgments*), and 3) the decision on what to report (*communicative intent*). Simplifying this model somewhat by ignoring adequacy judgments, Beatty and Herrmann claim that the respondent who has been administered a survey question has two decisions to make: whether they *can* respond and whether they *will* respond. Item

nonresponse results when the respondent decides negatively in either case.

Breakoff, in contrast to item nonresponse, came under more intense study with the proliferation of internet surveying when breakoff rates in web surveys rose high enough to become a serious problem. Another reason for the lack of scholarly attention to breakoff is the fact that so far no conceptual framework has been put forward to tackle the specific problem of breakoff, forcing authors to borrow from theoretical frameworks developed for other survey phenomena like item nonresponse and unit nonresponse (Peytchev 2009).

In an attempt to elucidate the respondent's decision to break off, Galesic (2006) resorts to decision field theory and the concept of the *inhibitory threshold*. The respondent is seen as continuously re-evaluating their original decision to partake in the survey (Peytchev 2009). As the survey progresses, the factors that influenced the initial decision to cooperate lose their influence and negative aspects of participation like fatigue and boredom set in and become stronger. As this happens, the respondent's preference to stop participating grows, but according to arguments from the decision field theory, the respondent will not make the decision to break off until this change in motivation exceeds the inhibitory threshold. Galesic (2006) finds that item nonresponse is more common immediately prior to breakoff as compared to earlier in the questionnaire.

Although Galesic does not refer to Krosnick directly, we argue that her notion of the intermediary period of low motivation before the respondent decides to break off is very similar to Krosnick's concept of satisficing. We propose that the difference stems from the fact that Krosnick made his account for interviewer-administered surveys where ending the interview involves terminating the interaction with the interviewer. We argue that the inhibitory threshold is therefore much higher in interviewer-administered surveys: in contrast to web surveys, the end of the questionnaire in interviewer administered surveys is likely to be reached before the inhibitory threshold is exceeded. One of the forms of satisficing, according to Krosnick, is item nonresponse whereby the respondent spares themselves the cognitive effort required to reply to an item. We argue that this is the reason why Galesic (2006) observed an increase in the frequency of item nonresponse immediately prior to breakoff. The above argumentation forms a conceptual interconnection between item nonresponse and breakoff and is one of the contributions of this thesis to the field of survey methodology.

In addition to giving an account of theoretical frameworks, Chapter 2 also reviewed previous research and identified common correlates of item nonresponse and breakoff. Item nonresponse is more common with items dealing with sensitive topics and items including threatening response options, typical examples being items on income and

sexual behavior. Item nonresponse rates have been found to be higher when the questionnaire is self-administered (as compared to interviewer-administered), except when the items deal with sensitive topics: in that case, survey methodologists recommend self-administration to reduce item nonresponse. Respondents' age and education have been used as proxies for their cognitive ability: older and less educated respondents have been found to produce more item nonresponse. Researchers have also considered the effect of interviewers on item nonresponse and have often found that interviewers differ substantially with regard to how much item nonresponse their respondents produce. Most studies have found, however, that interviewer characteristics have little or no explanatory power in models of item nonresponse.

Breakoff has been found to be dramatically more common in web surveys as compared to interviewer-administered surveys, where it is most often negligible. Research on breakoff, has thus concentrated almost exclusively on web surveys. Contrary to researchers' initial expectations, progress bars have been found to increase breakoff. Breakoff has been found to be more common with long-worded items, open-ended items, and items that introduce a new section of the questionnaire. The respondent's age and education, like in studies on item nonresponse, have been used as proxies for respondent cognitive ability in studies of breakoff. More highly educated respondents have been found to break off less frequently. The effect of respondent age, however, was not found to be consistent with several studies reporting *more* breakoff among younger respondents. Galesic (2006) designed a study to allow the investigation of respondents' professed interest and burden on breakoff. Respondents who broke off were often found to express lower interest and higher burden than respondents who continued.

Chapter 3 started with an informal account of multilevel models that are applied whenever the data have a hierarchical, nested, or clustered structure. Survey data have measurement occasions nested within respondents, who are further nested within interviewers. The effect of items can, in principle, be modeled with fixed effects (item indicators) or random effects. If item indicators were to be used in a statistical model, however, item facets could not be included as predictors due to collinearity. We therefore decided to model items with random effects and include item-level predictors to explain part of the variability among items. We thus consider measurement occasions in the survey data to be nested within items and cross-classified with respondents who are further nested within interviewers.

When the response variable is binary, as is the case with item response/nonresponse, a *generalized* linear mixed model is appropriate. Models with the logit link have traditionally been used in item response modeling. Item response theory is a framework for

specifying mathematical functions that describe the interactions of persons and items, and has most often been used in achievement test settings. Typically, the researcher using item response models is interested merely in descriptive measurement of persons and items: the purpose is to measure individuals and items on underlying constructs, and thereupon assign numeric values to them. Actually explaining these values, however, is considered only as the second step, if at all. De Boeck and Wilson (2004b) argue that, even though the philosophical orientations seem to be in conflict, descriptive measurement can be combined with explanatory analysis (including predictors in the models to explain differences) into what they term *explanatory measurement*. A number of explanatory item response models of increasing complexity were presented and discussed in Section 3.2. We apply such models to the problem of item nonresponse in surveys, treating item nonresponse analogously to a correct answer in an achievement test.

Section 3.3 introduced survival analysis methods used to analyze time-to-event data that include *censored* units. In clinical studies, censoring typically occurs because a patient does not experience the event before the study ends, or the patient is lost to follow-up. Interested in analyzing the time to breakoff, however, we define breakoff as the event of interest and consider completing respondents as censored. Section 3.3 explained how survival curves can be estimated and plotted to aid the interpretation of respondents' "survival" in the survey. The Cox proportional hazards model was described and discussed. In addition to including time independent predictors (like respondent characteristics), the *extended* Cox model is able to accommodate time-varying predictors, a feature we employ to include item facets as predictors of breakoff. A discussion of testing the proportional hazards assumption followed, and we expounded upon the approaches for resolving violations.

Chapter 4 discussed a number of different topics. First, the data-collection procedures were described. The data analyzed in the empirical part of the dissertation were collected in a field test of a new Generations and Gender Programme (GGP) questionnaire, which was implemented in three modes: face-to-face, telephone, and web. The data were collected in two rounds in order to maximize the total sample size given the budget. The distinction between the members of a commercial web panel (the first round) and persons sampled from the Slovenian population (the second round) is a nuisance that needs to be taken into account when modeling item nonresponse and breakoff in the GGP pilot data. Even though the GGP interview typically took about an hour, we found almost no occurrence of breakoff in the first two rounds of data collection, surprisingly not even in web mode. Because such data does not allow the investigation of respondent characteristics and item facets connected to breakoff, we initiated a third round of data collection through advertisements on the Facebook

web-page. The Facebook respondents were not told upfront how long it would take to fill out the questionnaire would take, unlike respondents in the first two rounds, who where informed of the expected duration in the advance letter. We suspect that this is partly the reason why breakoff was much more common in the Facebook sample, in which more than half the respondents broke off. Reviewing the demographic structure of the collected samples, the samples from the first two rounds of data collection were found to approximately resemble the overall Slovenian population, while the additional Facebook sample stood out due to an overrepresentation of women and young respondents.

In Section 4.2 we turned to the notion of item sensitivity and decided to measure three distinct aspects of item sensitivity: the intrusiveness of the item topic, the threat of disclosure that some response alternatives can involve, and the item's potential for overclaiming. In the following section, we described how respondents' self-assessments of the sensitivity and difficulty of certain item topics were combined with codes for item topics to form item-by-respondent interactions that were later used as predictors in statistical models. Some predictors had missing values because of nonresponse on corresponding questionnaire items. The missing values were addressed by multiple imputation, as described in Section 4.4. The last section of Chapter 4 put forward operational hypotheses, specifying the results we expected to find upon fitting models to the data on item nonresponse and breakoff.

7.2 Joint interpretation of findings for item nonresponse and breakoff

Chapters 5 and 6 explored item nonresponse and breakoff respectively. Because item nonresponse and breakoff were analyzed by fitting different statistical models to the data, the interpretation of the results for item nonresponse was given without referring to breakoff and vice versa. As explained above, however, we see both item nonresponse and breakoff as expressions of the respondent's lack of motivation to continue providing responses to questionnaire items, with breakoff as the more extreme alternative to item nonresponse. We therefore expect to find a particular predictor (e.g. a respondent characteristic of item facet) to have a similar effect on both item nonresponse and breakoff. In this section we gave a joint interpretation of the results from Chapters 5 and 6, highlighting similarities and differences in the findings. We should reiterate at this point that the Cox models for breakoff were fit to the Facebook data, while the generalized linear mixed models for item nonresponse were fit to data stemming from rounds one and two, containing three modes of administration.

We hypothesized that the inhibitory threshold for item nonresponse and breakoff would be lower when the distance between the respondent and interviewer is higher. The GGP pilot data allow for an examination of this hypothesis, given that the same questionnaire was administered in three different modes. The results, broadly speaking, support our hypothesis. The highest rate of item nonresponse was found for web mode and the lowest rate for face-to-face interviewing. The breakoff rates in the data from the first two rounds, on the other hand, do not quite conform with the hypothesis, with approximately the same proportion of breakoff in web and telephone modes. We would expect the inhibitory threshold for breakoff to be *much* lower in the total absence of an interviewer (web mode), as compared to interviewer-administered questionnaires, and would thus expect the breakoff rate for CATI mode to be closer to the rate for face-to-face interviewing than web administration.

Studies of item nonresponse and breakoff often include respondent demographics as predictors. Because virtually no questionnaire fails to inquire about the respondent's sex, age, and education, such explanatory variables are readily available. The causal mechanism linking item nonresponse and breakoff to respondent demographics is somewhat questionable, however. Authors have proposed that respondent age and education can serve as proxy variables for the respondent's cognitive ability, which is viewed as the underlying causal factor (see Krosnick 1991). The causal mechanism for the effect of gender is more questionable still, which is why we had no prior expectations about the possible effect of respondent gender, which we included in our models merely for exploratory purposes. The results of the analyses show no differences in item nonresponse and breakoff with regard to the respondent's gender.

Model estimates are consistent with the explanation that education can serve as a proxy for cognitive ability. According to the hypothesis, we expect respondents with a higher level of cognitive ability to experience less cognitive burden when answering to questionnaire items. If education can serve as a measure of cognitive ability, we would then expect more highly educated respondents to produce item nonresponse less often and to break off later in the questionnaire, if at all. We also hypothesized that respondents with a lower degree of education would produce more item nonresponse on the web because the web respondent cannot receive any help from the interviewer when faced with difficult items. All corresponding effects have the expected direction. The effect of education and its interaction with web mode, however, do not reach the threshold for statistical significance in the model for item nonresponse.

The aforementioned line of reasoning also posits that cognitive ability diminishes with old age and that therefore the respondent's age should be related to more item nonresponse and breakoff. We added to this the hypothesis that older respondents would

produce more item nonresponse on the web. The results for item nonresponse and breakoff differ widely with regard to age. We find that older respondents, indeed, produced more item nonresponse and that the age effect was even stronger in web mode. Respondent age, however was *negatively* associated to breakoff, contrary to expectations. Upon further investigation, evidence was found that younger respondents broke off at high rates shortly into the questionnaire, whereas after approximately one hundred items, age no longer had an effect on the risk of breakoff.

At this point we should reiterate that the analysis of breakoff was conducted on a different sample (round three) than the analysis of item nonresponse (rounds one and two). We suspect the way respondents for the Facebook sample were recruited to be the reason for the differences in the effect of age between the models. The advance letters and reminders sent to sample persons in rounds one and two gave lent an air of gravity to the survey, mentioning affiliations to the United Nations and the European Union and informing the respondents of the expected interview duration. The additional third round of data collection, on the other hand, had no affiliations other than the University of Ljubljana and did not mention how long filling out the questionnaire would take. Respondents were recruited by means of advertisements on Facebook, whose primary function is entertainment, and the invitation to participate accordingly had a much more informal tone than in the first two rounds. We suspect this attracted many respondents who were only casually interested in seeing what kind of questions would be posed and who quickly lost interest thereafter. The survival curves in Figure 6.3 (page 168) suggest that such low-motivation respondents were especially prevalent among the young. Though the results may be suspect due to low sample size, the item nonresponse model estimates for the Facebook sample (Table 5.8, page 135) also show a *negative* effect of age on item nonresponse, corresponding to the explanation that the younger Facebook users had a more casual approach to filling in the questionnaire.

In order to measure the respondents' attitude toward surveys, we added three items to the questionnaire (see Section 4.4.1). The items were a subset of a longer 16-item instrument with a high level of internal consistency (Stocke and Langfeldt 2004), which is why we also decided to compute a scale variable on the basis of the three items, even though the shorter scale somewhat lacked internal consistency (Cronbach's alpha of 0.61). This variable functioned well as a predictor of both item nonresponse and breakoff. We expected respondents with a more positive attitude toward surveys to be more highly motivated to perform each phase of the question-answer process carefully and comprehensively. The results, indeed, show that respondents with a more positive attitude toward surveys produce less item nonresponse and have a lower risk of breakoff.

In Section 4.2 we discussed the concept of item sensitivity and made the observation that several distinct meanings of the concept have been identified in the survey methodology literature. Rather than narrowing our focus to a single measure, we defined three measures of item sensitivity: the intrusiveness of the item's topic, the answer alternatives' threat of disclosure, and the item's potential for overclaiming. Three independent raters rated each GGP item on each of the three measures of item sensitivity.

The inter-rater agreement was found to be rather low, especially for the ratings of the threat of disclosure, where the value of Krippendorff's alpha was found to be 0.27. We believe the low value of agreement among raters reflects the burden associated with considering *each* answer alternative when rating a great number of items. In a test-run rating of a small subset of items, one rater gave the *highest rating* of the threat of disclosure to the item with the wording "Is your current work contract, if you have any, a permanent contract, a fixed-term contract, or a temporary contract?" because it included the response alternative "no written contract", implying undeclared work. The other raters, however, gave this item the *lowest rating* on the threat of disclosure, because they did not make the connection to undeclared work as they gave their ratings. We believe that this anecdotal example illustrates why inter-rater agreement was low on the threat of disclosure. The experts were asked to rate the whole GGP questionnaire (more than 500 items) and consider the threat involved with each response alternative. This is a daunting task that is difficult to perform comprehensively with such a great number of items.

Research has shown, however, that low values of inter-rater agreement are common when expert raters attempt to identify problematic questionnaire items (Olson 2010; DeMaio and Landreth 2003; Presser and Blair 1994), and Olson (2010) has demonstrated that, despite the lack of reliability, the average expert ratings successfully identify questionnaire items with a higher item nonresponse rate. For this reason, we we included the average rating of the items' threat of disclosure as a predictor in the models, despite its low level of inter-rater agreement. The resulting predictor, as hypothesized, was found to be associated to more item nonresponse, but this effect was only significant in web mode. The item's threat of disclosure was also found to increase the risk of breakoff, but this effect was only significant after the 100th item in the GGP questionnaire. Despite its low inter-rater agreement, the average rating of the threat of disclosure performed well as a predictor and led to reasonable conclusions, which were in accordance with Olson's findings (2010). We believe that higher levels of inter-rater agreement could have nonetheless be reached if raters had been given given a lighter workload, allowing them to consider each item with more care.

We found the effect of the items' potential for overclaiming to be associated with less item nonresponse and breakoff. Following the rationale of (Bradburn et al. 1978), we argue that skipping over, refusing, or breaking off at an item that allows overclaiming could be seen as admitting to a sin of omission—not having acted in socially desirable ways—which is why respondents avoid item nonresponse and breakoff at such items.

The results, furthermore, indicate that the effect of overclaiming is weaker on the web, i.e., there is more item nonresponse on the web for items that allow overclaiming than in interviewer-administered interviews. As mentioned, a similar result was found for the item's threat of disclosure: there was more item nonresponse for threatening items in web administration. This leads us to believe that the respondent's concern for self-presentation could be the common cause for both findings. According to this explanation, the respondent is less likely to produce item nonresponse for items that allow overclaiming in the presence of the interviewer, because skipping over or refusing to answer such an item could be seen as admitting to a sin of omission. Without an interviewer present, however, the concern for self-presentation is weaker and so the item's potential for overclaiming does not lower the respondent's tendency to produce item nonresponse in web mode. Similarly, the web respondent can skip over or refuse to provide an answer to items that include threatening responses. In face-to-face and CATI administration, however, the interviewer might see this as implicitly admitting that the threatening response alternative is true, which is why the respondent might choose a white lie over item nonresponse.

The intrusiveness of the item's topic was found to be related to more item nonresponse, as expected. The effect of intrusiveness on breakoff, however, was found to be more complex than hypothesized. The results convey that a more intrusive item lowers the risk of breakoff while simultaneously increasing the risk of breakoff two items further into the questionnaire. One possible explanation for this finding is that respondents do not want to disclose the information that they regard a particular topic highly intrusive by breaking off at the particular item that concerns such a topic. They actually refrain from breakoff at such items and rather break off shortly after the intrusive item was administered.

We expected items that require more effort to answer to induce more item nonresponse and increase the risk of breakoff. A low-motivation respondent faced with a difficult task might decide to skip the question-answer process altogether and pass over the item, or produce a response like "don't know." If the respondent's preference for discontinuing the interview is already high, such an item can induce breakoff. We attempted to capture the item's complexity through the length of its wording and the number of answer alternatives offered to the respondent. We also included an indicator

of whether the item was open-ended (as opposed to closed-ended) in our models for item nonresponse and breakoff.

The length of the item wording was not found to affect item nonresponse, except in web mode: web respondents were found to produce more item nonresponse to items with long wordings. We attribute the difference between interviewer- and self-administered modes to the visual presentation of the items in web mode. The web respondent must read the item wording themselves, and might therefore prefer to skip an item with a daunting word count spanning across several rows. In face-to-face and CATI modes, on the other hand, the respondent does not know how long the item wording is until the interviewer has finished reading it. The length of the item wording was not found to have a significant effect on the risk of breakoff in interviewer-administered interviews, though the direction of the effect was positive, as hypothesized.

Open-ended items and items with many answer alternatives were found to be associated with more item nonresponse and a higher risk of breakoff, as expected from the explanation that such items require more effort on the part of the respondent. Contrary to expectations, however, the effect of open-ended items and items with many answer alternatives was found to be weaker in web mode: web respondents produced *less* item nonresponse on such items than their face-to-face and CATI counterparts. We would expect the interviewer to motivate the respondent faced with such items to provide an answer despite the additional required effort, which would be reflected in a *positive* interaction term with web mode (more item nonresponse in web mode), where the interviewer is absent, rather than the negative one that was found. We do not have an explanation for the direction of the aforementioned effects in web mode.

Including measures of item difficulty and sensitivity like the ones discussed above as predictors in statistical models assumes that a particular item is perceived in the same way by all respondents, and that it therefore has the same effect on them. Wanting to allow for respondent-specific sensitivity and difficulty effects in the analysis, we added a number of items to the questionnaire, attempting to measure 1) each respondent's sensitivity in answering questionnaire items pertaining to certain topics, and 2) the cognitive state of the respondent's knowledge about certain topics in the GGP questionnaire. In general, we found the expected direction for the effects of such item-by-respondent predictors in the models for item nonresponse, though a number of them did not reach the threshold for significance.

In the Cox model, however, the great majority of item-by-respondent predictors were found to have an insignificant effect on the risk of breakoff. This casts a shadow of a doubt on the results for item nonresponse, too. The additional questions on

respondent-specific sensitivity and cognitive state were asked *before* the GGP items in the Facebook sample (which was used in breakoff analyses) and *after* the GGP items in the first two rounds of data collection (which were used for the item nonresponse analyses). Because the position of the additional questions in the questionnaire is confounded with the round of data collection, we cannot conclude that respondent-specific measures of item sensitivity and cognitive state have a different effect on item nonresponse as compared to breakoff. An equally reasonable explanation given the circumstances is that, in the Facebook sample, the additional items had a hypothetical character (e.g. "How sensitive *would* you find…"), whereas in the first two rounds of data collection, the respondents could *remember* how sensitive they found certain topics and answer accordingly. It is possible, in other words, that we found significant effects in our item nonresponse analysis because respondents in the first two rounds e.g., *remembered* withholding responses to items concerning certain topics due to embarrassment. Respondents in the Facebook sample, on the other hand, needed to employ their imagination to guess what kinds of questions would be posed and assess their sensitivity and cognitive state in this hypothetical scenario.

In Section 5.4 the generalized linear mixed model for item nonresponse was applied to a narrowed dataset excluding web respondents. This allowed us to include interviewer random effects in the model. The results indicate that the largest variation with regard to item nonresponse in the data is among items, followed by variation among respondents and interviewers. We then included interviewer characteristics as predictors in the model, attempting to explain a part of the variation in item nonresponse due to interviewers. Including the interviewers' sex, age, education, and experience explained a substantial proportion—about a third—of the interviewer-level variation. This is noteworthy because authors who studied the effect of interviewers on item nonresponse reported differences between interviewers, but were unable to explain the differences in terms of interviewer characteristics (Pickery and Loosveldt 1998, 2001). The effects of interviewer characteristics (with the exception of interviewer education), however, were not found to have a significant effect on item nonresponse, a finding we attribute to low sample size at the interviewer-level.

We also explored the effect of the respondent's previous record of item nonresponse on the probability of producing nonresponse to the currently administered item. The results show that respondents tend to produce item nonresponse in series, and that having produced (many) item nonresponses to items preceding the current one increases the probability of nonresponse on the current item. These findings should, however, be taken with a grain of salt, because the ordering of the items in the GGP questionnaire was not randomized. This means that the tendency to produce item nonresponse in series could also be due to the fact that many similar items (e.g. items

concerning the same topic) were administered shortly one after the other. If a particular topic was sensitive to the given respondent, for example, this could have caused item nonresponses to all items concerning this topic, and such items were likely to appear one after the other in the questionnaire. In order to study the effect of the respondent's record of previous item nonresponse on the probability of nonresponse to the current item, a study could be devised to address the aforementioned confounding by randomizing the order of the questionnaire items.

Certain filter items in the GGP questionnaire did not allow any item nonresponse. Because they provide no information on the respondents' tendency to produce item nonresponse, such *required* items were excluded from the analysis of item nonresponse. Required items were included in the analysis of breakoff, however, where they were shown to increase the risk of breakoff (though the effect did not quite reach the threshold for statistical significance). This makes sense, because a message requesting a response be given was displayed (or conveyed by the interviewer) each time the respondent tried to skip a required item. Such repetitive reminders could frustrate the respondent—if, e.g., they believed that none of the offered responses fit their situation—to the point of breaking off.

Another predictor that was peculiar to the breakoff analysis was the indicator for section introductions. Items introducing a new section of the questionnaire foreshadow additional material to come and thus provide a natural breaking point in the conversation (Peytchev 2007). Such items were found to have more than twice the relative risk of breakoff by Peytchev (2009). Our findings are similar, with the qualification that section introductions were not found to increase the risk of breakoff shortly into the questionnaire, but rather only after at least a hundred items had been administrated already.

We see both item nonresponse and breakoff as expressions of the respondent's lack of motivation to reply to questionnaire items, with breakoff as the more extreme alternative. According to Galesic's argument (2006), item nonresponse is more common immediately prior to breakoff because the respondent would already prefer to stop responding, but this change in preference has not yet exceeded the inhibitory threshold and found expression in breakoff. As the respondent switches the response strategy to what Krosnick (1991) terms satisficing, the item nonresponse rate correspondingly increases. In Section 6.5 we examined item nonresponse in the n items preceding breakoff and found that the item nonresponse rate increased immediately prior to breakoff in the Facebook sample, thus replicating Galesic's results. We proceeded to demonstrate that a measure of recent item nonresponse could be included as a predictor in the model for breakoff to increase the model's predictive performance.

7.3 Contribution and implications

Our analysis of item nonresponse bears a certain resemblance to the work of Kveder (2005); in his doctoral dissertation, Kveder (2005) applied multilevel models to model item nonresponse in surveys. His work demonstrates that characteristics of the questionnaire item, respondent, and interviewer should *all* be considered to explain item nonresponse: if any of the three levels is omitted from the analysis, conclusions about the effect of the predictors on item nonresponse can be misleading. Because of the sheer amount of data (tens of thousands of respondents from multiple countries and multiple surveys), the response variable in Kveder's models was the item nonresponse *rate* defined at the respondent, interviewer, or questionnaire item level. His analysis, in comparison to ours, thus did not define a *measurement occasion* level and, consequently, Kveder was forced to resort to a two-step estimation procedure in order to consider the effect of predictors at all three levels (item, respondent, and interviewer) on item nonresponse (see Kveder 2005, for details). Our analysis, in contrast, considered a smaller dataset and we were thus able to model item nonresponse at each measurement occasion (rather than the item nonresponse rate), cross-classified between respondent and questionnaire item. This allowed us to apply well known item response models and estimate them in a single step with no computational issues.

The present dissertation makes a number of contributions to the field of survey methodology. The objects of the study are two types of nonresponse that have not been studied as extensively as unit nonresponse. The empirical study thus enriches the extant knowledge regarding factors affecting item nonresponse and breakoff in surveys and has the potential to inform procedures for prevention (e.g. by adapting the questionnaire design) and treatment (e.g. by multiple imputation of missing values) of item nonresponse and breakoff.

The analyses show that characteristics of item, respondent, and interviewer all affect item nonresponse. Our approach demonstrates that considering all three levels simultaneously is feasible and leads to sensible conclusions. We implemented the same questionnaire in three different modes of administration, thus providing an exceptional opportunity to investigate the differential impact across modes of item facets and respondent characteristics on item nonresponse. To the best of our knowledge, no previous study has simultaneously considered the effect of item, respondent, and interviewer on item nonresponse across three different modes of administration.

Our analysis of breakoff applied survival analysis methods to investigate the effect of both respondent characteristics and item facets on the risk of breakoff. Such detailed investigations of breakoff employing survival analysis methods are rare (Galesic 2006;

Peytchev 2009; Matzat et al. 2009). We do not know of any other studies to have checked the proportional hazards assumption in the Cox model for breakoff and adjusted for violations. To the best of our knowledge, our study is the first to include a measure of recent item nonresponse in a model for breakoff and to demonstrate a significant effect[31].

The findings of the study have implications for the field of survey methodology. In the case of web surveys, breakoff could be reduced by continuously monitoring the respondent's item nonresponse rate over the recently administered items. If the item nonresponse rate starts increasing, this could be a sign of the respondent's wavering motivation and thus a warning that breakoff is more likely to occur. If the survey design and software allow, the respondent could then be invited to stop replying, e.g. at the end of the current section, and continue at a later time. If a low-motivation respondent is administered, e.g., an item that introduces a new section, the respondent might take this opportunity to break off and not return.

Items that require more effort on the part of the respondent (open-ended items, items with many response alternatives) and items with threatening response alternatives increase the risk of breakoff. The placement of such items in the questionnaire should be strategic; if the information requested by these items is critically important to the goal of the survey, such items should appear near the start of the questionnaire, when the respondents' motivation is still high. If the requested information is not important, however, such items should come toward the end of the questionnaire. Administering such items increases the risk of breakoff and, if the respondent breaks off, the responses to all subsequent items will be lost.

The results of our analyses have a number of implications for the GGP questionnaire in particular. Missing values due to item nonresponse and breakoff can most simply be avoided by reducing the number of open-ended items[32]. Missing values can, of course, also be avoided by continuing to use face-to-face as the mode of administration despite the high costs. If the decision to employ web administration is made to lower the costs of data collection 1) older respondents should be given the possibility of being interviewed in another mode, 2) early breakoffs should be avoided by shifting difficult and sensitive items (e.g. the modules on fertility and social networks) toward the end of the questionnaire, and 3) the respondent's item nonresponse rate over the recently

[31] Galesic (2006) demonstrated that item response increases immediately prior to breakoff, but did so by comparing the item nonresponse rate for "blocks" immediately preceding breakoff vs. blocks that did not. Galesic's approach thus did not control for possible confounding by item facets and respondent characteristics.

[32] E.g. by skipping the open-ended item on income and proceeding directly to the multiple-choice version of this item, by offering a reasonably short number of answer alternatives for the respondent's parents' occupation rather than asking the respondent to formulate the answer themselves etc.

administered items should be monitored (if the item nonresponse rate suddenly increases, the respondent should be asked to continue filling out the questionnaire at a later time).

7.4 Limitations and suggestions for further research

The present study of item nonresponse and breakoff has a number of limitations that must be taken into account when drawing conclusions from its results. The respondents' self-assessment items were asked *after* the GGP items in the first two rounds of data collection[33]. It is therefore possible that the significant effects of respondent self-assessments on item nonresponse were found because the respondents could *remember* that items concerning particular topics were difficult or sensitive for them. In order to investigate whether respondent self-assessments of item topic difficulty and sensitivity can be useful for identifying problematic items, the study should be repeated with self-assessment items administered *before* the rest of the questionnaire.

The inter-rater agreement was found to be rather low, especially for the ratings of the threat of disclosure. As mentioned, we believe that the low value of the inter-rater agreement reflects the raters' burden associated with considering each answer alternative when rating a great number of items. Higher levels of inter-rater agreement could be reached if raters were given a lighter workload, allowing them to consider each item with more care.

The theoretical framework underlying the statistical models is itself limited in that it reduces the factors influencing item nonresponse and breakoff to respondent motivation and item burden. Such a model excludes factors that might be relevant to a specific mode of administration, e.g. a person filing out a web questionnaire might be disturbed by incoming emails, conversations with other people in the same room etc.

The procedures that were employed to collect the GGP pilot data do not allow the results of the statistical analyses to be generalized to the population. Generalization in the strict statistical sense would require that a population be defined and that a random sample be drawn from it with a non-zero selection probability for each unit in the population. In the first round of data collection the members of a commercial web panel were interviewed; no population was defined and no sampling was used. The second round did define a population (the non-institutionalized Slovenian 18+ population) and employed random sampling, but the sample design was severely complicated

[33]The GGP management was concerned that adding such items to the beginning of the questionnaire could introduce adverse context effects and only agreed to the self-assessment items' being added *after* the GGP items.

by the mixed mode design (see Section 4.1.1 for details). Finally, the additional third round of data collection recruited respondents via advertisements on the Facebook web page, again without defining a population or random sampling.

Most of the study's findings, however, are congruent with expectations derived from theory. Based on the theoretical frameworks discussed, we have no reason to believe that we would have arrived at dramatically different conclusions were the study to be repeated, e.g., on respondents drawn from a population of another European country with an internet penetration rate similar to Slovenia's. We are more reserved about the generalizability of results based on the Facebook sample, however. International academic surveys like the GGP do not typically recruit respondents in such informal manners as advertisements on web pages. We believe that Facebook respondents had a more casual approach to responding to questionnaire items than is common for respondents in academic surveys and that this attitude is reflected, e.g, in the high breakoff rate. The results based on the Facebook sample might therefore be more applicable to web-based polls and commercial surveys.

Another aspect of generalizability refers to the set of questionnaire items that was used in the analyses. The GGP questionnaire consisted of a great number of items widely differing in facets like item topic, response type, scale, etc. The majority of GGP questions inquired into *facts*, however, rather than, e.g., opinions. We thus believe the results would generalize well to other surveys with questionnaires consisting predominantly of factual items. We believe our findings regarding the effect of item facets on item nonresponse and breakoff are also relevant for opinion surveys, but note that they might not generalize as well.

The original idea for this dissertation was to apply a more complex model to the problem of item nonresponse in surveys. According to Beatty and Herrmann, the respondent is faced with two decisions when administered a survey question: whether they *can* respond and whether they *will* respond. Item nonresponse results when the respondent decides negatively in either case (Beatty and Herrmann 2002). Stocke (2006) takes this notion further and speaks of the *cognitive costs* that an item poses, which are affiliated with the former decision, and the *psychological and social costs* that are associated with the latter decision. According to this conceptualization, questionnaire items differ with respect to not one but two latent dimensions: cognitive costs and psychological/social costs. We thus considered a *two-dimensional* IRT model for item nonresponse (see Reckase 2009). In such a model, the probability of a substantive response (the left-hand side) is modeled as the product of two terms:

$$(1 - \pi_{pi}) = f^{-1}_{\text{logit}}(\theta^{(1)}_p - \beta^{(1)}_i) \cdot f^{-1}_{\text{logit}}(\theta^{(2)}_p - \beta^{(2)}_i). \tag{7.1}$$

Here, $\beta_i^{(1)}$ denotes the ith item's cognitive costs and $\beta_i^{(2)}$ denotes the item's psychological/social costs, while the theta parameters are the corresponding dimensions of the respondent's motivation. The model can be further elaborated by including fixed effects (of predictor variables) and additional random effects (for interviewers). Because the two terms on the right-hand side of (7.1) are multiplied, the probability of obtaining a substantive response is low if either term on the right-hand side is close to zero. This occurs if either dimension of the item's costs far exceeds the respondent's motivation. For this reason, the above model is referred to as *non-compensatory* (Reckase 2009). The mathematical form of the model reflects our belief that a high value on one dimension of the respondent's motivation (the motivation to expend cognitive effort) cannot compensate for a lack in the other (the motivation to answer to intrusive or threatening items). We encountered convergence errors when attempting to estimate the parameters of such a model in WinBUGS (Platinovšek 2013). We attribute the estimation problems to the extremely low proportion of item nonresponse in our data. The dataset does include a great number of rows (measurement occasions), but the rows all convey the same message: that respondents in general give substantive responses to nearly all administered items. In order to investigate the applicability of the two-dimensional model (7.1), a questionnaire could be devised that concentrates on sensitive and difficult items, thus most likely resulting in a dataset with a less extreme proportion of item nonresponse.

As we were limited to web respondents from the Facebook sample, our survival analysis of breakoff did not necessitate a consideration of interviewer effects. If we should need to model breakoff in an interviewer-administered survey with survival analysis methods, a class of models known as *shared frailty models* could be considered (Duchateau and Janssen 2008; Hanagal 2011; Wienke 2011). These models introduce a multilevel aspect to the analysis by assigning a common random effect to all individuals in the same group (all respondents interviewed by the same interviewer). An alternative approach, suggested by O'Quigley and Stare, is to avoid complex modeling and additional assumptions of random effects models and simply stratify with regard to group (interviewer). The efficiency gain provided by the added structure of a random effects model was shown to provide only modest efficiency gains for group sizes (interviewer workloads) of five or more (O'Quigley and Stare 2002).

Bibliography

Aalen, Odd O., Ornulf Borgan and Hakon K. Gjessing. 2008. *Survival and Event History Analysis: A Process Point of View*. New York: Springer.

AAPOR. 2009. *Standard Definitions: Final Disposition Codes and Outcome Rates for Surveys*. Deerfield, IL: AAPOR.

Aoki, K. and M. Elasmar. 2000. Opportunities and challenges of a Web survey: A field experiment. In *55th Annual Conference of American Association for Public Opinion Research*. Portland, Oregon.

Bates, Douglas, Martin Maechler and Ben Bolker. 2012. *lme4: Linear mixed-effects models using S4 classes*. Available at: `http://CRAN.R-project.org/package=lme4` (September 3, 2013). R package version 0.999999-0.

Bates, Nancy. 2001. Internet versus mail as a data collection methodology from a high coverage population. In *Proceedings of the Annual Meeting of the American Statistical Association*. Alexandria, VA, USA: American Statistical Association. Available at: `http://www.amstat.org/sections/srms/proceedings/y2001/Proceed/00311.pdf` (September 30, 2013).

Beatty, Paul and Douglas Herrmann. 2002. To answer or not to answer: Decision processes related to survey item nonresponse. In *Survey Nonresponse*, eds. Robert M. Groves, Don A. Dillman, John L. Eltinge and Roderick J. A. Little, 71–86. New York: John Wiley & Sons.

Beatty, Paul, Douglas Herrmann, Cathy Puskar and Jeffrey Kerwin. 1998. Don't Know Responses in Surveys: Is What I Know What You Want to Know and Do I Want you to know it? *Memory* 6 (4): 407–426.

Bell, Ralph. 1984. Item Nonresponse in Telephone Surveys: an Analysis of Who Fails to Report Income. *Social Science Quarterly (University of Texas Press)* 65 (1): 207–215.

Berk, M.L., N.A. Mathiowetz, E.P. Ward and A.A. White. 1987. The effect of prepaid and promised incentives: Results of a controlled experiment. *Journal of Official Statistics* 3 (4): 449–457.

Bishop, George F., Alfred J. Tuchfarber and Robert W. Oldendick. 1986. Opinions on Fictitious Issues: The Pressure to Answer Survey Questions. *Public Opinion Quarterly* 50 (2): 240–250. doi:10.1086/268978.

Bosnjak, Michael and Tracy L. Tuten. 2001. Classifying Response Behaviors in Web-Based Surveys. *Journal of Computer Mediated Communication* 6 (3).

Boyer, Kenneth K., John R. Olson, Roger J. Calantone and Eric C. Jackson.

2002. Print versus electronic surveys: A comparison of two data collection methodologies. *Journal of Operations Management* 20 (4): 357–373. doi: 10.1016/S0272-6963(02)00004-9.

Bradburn, Norman M. 2004. Understanding the Question-Answer Process. *Survey Methodology* 30 (1): 5–15.

Bradburn, Norman M., Lance J. Rips and Steven K. Shevell. 1987. Answering Autobiographical Questions: The Impact of Memory and Inference on Surveys. *Science* 236: 157–161.

Bradburn, Norman M., Seymour Sudman, Ed Blair and Carol Stocking. 1978. Question Threat and Response Bias. *The Public Opinion Quarterly* 42 (2): 221–234.

Börkan, Bengü. 2010. The Mode Effect in Mixed-Mode Surveys: Mail and Web Surveys. *Social Science Computer Review* 28 (3): 371–380. doi:10.1177/0894439309350698.

Browne, William J, Harvey Goldstein and Jon Rasbash. 2001. Multiple membership multiple classification (MMMC) models. *Statistical Modelling* 1 (2): 103–124. doi:10.1177/1471082X0100100202.

Busemeyer, Jerome R. and James T. Townsend. 1993. Decision Field Theory: A Dynamic-Cognitive Approach to Decision Making in an Uncertain Environment. *Psychological Review* 100 (3): 432–459.

Cannell, C.F., P. Miller and L. Oksenberg. 1981. Research on interviewing techniques. In *Sociological Methodology*, ed. S. Leinhardt, 389–437. San Francisco: Jossey-Bass.

Cantor, David, Barbara O'Hare and Kathleen O'Connor. 2008. The use of monetary incentives to reduce non-response in random digit dial telephone surveys. In *Advances in telephone survey methodology*, eds. James M. Lepkowski, Clyde Tucker, J. Michael Brick, Edith de Leeuw, Lilli Japec, Paul J. Lavrakas, Michael W. Link and Roberta L. Sangster, 471–498. New York: Wiley.

Catania, Joseph A., Diane Binson, Jesse Canchola, Lance M. Pollack, Walter Hauck and Thomas J. Coates. 1996. Effects of Interviewer Gender, Interviewer Choice, and Item Wording on Responses to Questions Concerning Sexual Behavior. *Public Opinion Quarterly* 60 (3): 345–375. doi:10.1086/297758.

Clark, Taane G. and Douglas G. Altman. 2003. Developing a prognostic model in the presence of missing data: an ovarian cancer case study. *Journal of Clinical Epidemiology* 56 (1): 28–37. doi:http://dx.doi.org/10.1016/S0895-4356(02)00539-5.

Collett, David. 1994. *Modelling Survival Data in Medical Research*. London: Chapman & Hall.

Conrad, F. G., M. P. Couper, R. Tourangeau and A. Peytchev. 2003. Effectiveness of Progress Indicators in Web Surveys. In *RC33 6th International Conference on Social Science Methodology: Recent Developments and Applications*

in Social Research Methodology, 2004, 333–342. Available at: `www.asc.org.uk/publications/proceedings/ASC2003Proceedings.pdf` (July 12, 2013).

Couper, Mick P. 2000. Web Surveys: A Review of Issues and Approaches. *Public Opinion Quarterly* 64 (4): 464–494. doi:10.1086/318641.

Couper, Mick P., Michael W. Traugott and Mark J. Lamias. 2001. Web Survey Design and Administration. *Public Opinion Quarterly* 65 (2): 230–253.

Cox, David R. 1972. Regression Models and Life-Tables. *Journal of the Royal Statistical Society* 34 (2): 187–220.

Crawford, Scott D., Mick P. Couper and Mark J. Lamias. 2001. Web Surveys: Perceptions of Burden. *Social Science Computer Review* 19 (2): 146–162. doi:10.1177/089443930101900202.

Curtin, Richard, Stanley Presser and Eleanor Singer. 2005. Changes in Telephone Survey Nonresponse over the Past Quarter Century. *Public Opinion Quarterly* 69 (1): 87–98. doi:10.1093/poq/nfi002.

Curtin, Richard, Eleanor Singer and Stanley Presser. 2007. Incentives in random digit dial telephone surveys: A replication and extension. *Journal of Official Statistics* 23 (1): 91–105.

Davern, M., T.H. Rockwood, R. Sherrod and S. Campbell. 2003. Prepaid Monetary Incentives and Data Quality in Face-to-Face Interviews: Data from the 1996 Survey of Income and Program Participation Incentive Experiment. *Public Opinion Quarterly* 67 (1): 139–147. doi:10.1086/346012.

De Boeck, Paul and Mark Wilson. 2004a. Descriptive and explanatory item response models. In *Explanatory Item Response Models*, eds. Paul De Boeck and Mark Wilson, 43–74. New York: Springer.

———. 2004b. *Explanatory Item Response Models*. New York: Springer.

———. 2004c. A framework for item response models. In *Explanatory Item Response Models*, eds. Paul De Boeck and Mark Wilson, 3–41. New York: Springer.

de Heer, W. 1999. International response trends: Results of an international survey. *Journal of Official Statistics* 15 (2): 129–142.

de Leeuw, E. and W. de Heer. 2002. Trends in household survey nonresponse: A longitudinal and international Comparison. In *Survey nonresponse*, eds. R. Groves, D. Dillman, J. Eltinge and R. Little, 41–54. New York: John Wiley.

de Leeuw, Edith D. 1992. Data quality in mail, telephone and face to face surveys. Ph.D. thesis, Vrije Universiteit Amsterdam.

de Leeuw, Edith D., Joop J. Hox and Mark Huisman. 2003. Prevention and treatment of item nonresponse. *Journal of Official Statistics* 19 (2): 153–176.

DeMaio, T. J. and A. Landreth. 2003. Examining expert reviews as a pretest method. In *In ZUMA-Nachrichten Spezial Band 9, questionnaire evaluation standards*, eds. P. Prüfer, M. Rexroth, J. Fowler and F. Jackson, 60–73. Mannheim: ZUMA.

Denniston, Maxine M., Nancy D. Brener, Laura Kann, Danice K. Eaton, Timothy McManus, Tonja M. Kyle, Alice M. Roberts, Katherine H. Flint and James G. Ross. 2010. Comparison of paper-and-pencil versus Web administration of the Youth Risk Behavior Survey (YRBS): Participation, data quality, and perceived privacy and anonymity. *Computers in Human Behavior* 26 (5): 1054–1060. doi: 10.1016/j.chb.2010.03.006.

Denscombe, Martyn. 2009. Item non-response rates: a comparison of online and paper questionnaires. *International Journal of Social Research Methodology* 12 (4): 281 – 291. doi:10.1080/13645570802054706.

Dillman, D.A. 2000. *Mail and Internet Surveys: The tailored design method.* New York: John Wiley.

Dillman, D.A. and J. Tamai. 1988. Administrative issues in mixed-mode surveys. In *Telephone survey methodology*, eds. R. Groves, P. Biemer, L. Lyberg, J. Massey, W. Nicholls II and J. Waksberg, 509–528. New York: Wiley-Interscience.

Dillman, Don A., John L. Eltinge, Robert M. Groves and Roderick J. A. Little. 2002. Survey Nonresponse in Design, Data Collection, and Analysis. In *Survey Nonresponse*, eds. Robert M. Groves, Don A. Dillman, John L. Eltinge and Roderick J. A. Little, 3–26. New York: John Wiley & Sons.

Diment, K. and S. Garret-Jones. 2007. How demograpgic characteristics affect mode preference in a postal/web mixed-mode survey of Australian researchers. *Social Science Computer Review* 25 (2): 420–427.

Dirmaier, Jörg, Timo Harfst, Uwe Koch and Holger Schulz. 2007. Incentives increased return rates but did not influence partial nonresponse or treatment outcome in a randomized trial. *Journal of Clinical Epidemiology* 60 (12): 1263 – 1270. doi: http://dx.doi.org/10.1016/j.jclinepi.2007.04.006.

Doran, Harold, Douglas Bates, Paul Bliese and Maritza Dowling. 2007. Estimating the Multilevel Rasch Model: With the lme4 Package. *Journal of Statistical Software* 20 (2): 1–18.

Duchateau, Luc and Paul Janssen. 2008. *The Frailty Model.* New York, NY: Springer.

Elliott, Marc N., Carol Edwards, January Angeles, Katrin Hambarsoomians and Ron D. Hays. 2005. Patterns of Unit and Item Nonresponse in the CAHPS Hospital Survey. *Health Services Research* 40 (6P2): 2096–2119.

Fowler, Floyd J. and Thomas W. Mangione. 1990. *Standardized Survey Interviewing: Minimizing Interviewer-Related Error.* Newbury Park: SAGE Publications.

Freireich, Emil J, Edmund Gehan, Emil Frei III, Leslie R. Schroeder, Irving J. Wolman, Rachad Anbari, E. Omar Burgert et al.. 1963. The Effect of 6-Mercaptopurine on the Duration of Steroid-induced Remissions in Acute Leukemia: A Model for Evaluation of Other Potentially Useful Therapy. *Blood* 21 (6): 699–716.

Galesic, Mirta. 2006. Dropouts on the Web: Effects of Interest and Burden Experi-

enced During an Online Survey. *Journal of Official Statistics* 22 (2): 313–328.

Gelman, Andrew and Jennifer Hill. 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.

Generations and Gender Programme. 2013. Available at: `http://www.ggp-i.org` (August 6, 2013).

Goldstein, Harvey. 2011. *Multilevel Statistical Models*. 4th ed. West Sussex: John Wiley & Sons.

Goyder, J. 1994. An experiment with cash incentives on a personal interview survey. *Journal of the Market Research Society* 36 (4): 360–366.

Grambsch, Patricia M. and Terry M. Therneau. 1994. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika* 81 (3): 515–526. doi: 10.1093/biomet/81.3.515. Available at: `http://biomet.oxfordjournals.org/content/81/3/515.abstract`.

Groves, R. M. 1989. *Survey Error and Survey Costs*. New York: John Wiley.

Groves, Robert M. and Mick P. Couper. 1998. *Nonresponse in Household Surveys*. New York: Wiley.

Groves, Robert M., Don A. Dillman, Roderick J. A. Little and John L. Eltinge, eds. 2002. *Survey Nonresponse*. New York: John Wiley & Sons.

Groves, Robert M. and Robert L. Kahn. 1979. *Surveys by Telephone*. New York: Academic Press.

Gruskin, Elisabeth P., Ann M. Geiger, Nancy Gordon and Lynn Ackerson. 2001. Characteristics of Nonrespondents to Questions on Sexual Orientation and Income in a HMO Survey. *Journal of the Gay and Lesbian Medical Association* 5 (1): 21–24. doi:10.1023/A:1009586032661.

Guo, Shenyang. 2010. *Survival analysis*. New York: Oxford university press.

Hanagal, David D. 2011. *Modeling Survival Data Using Frailty Models*. London: Chapman and Hall.

Harel, Ofer. 2009. The estimation of R2 and adjusted R2 in incomplete data sets using multiple imputation. *Journal of Applied Statistics* 36 (10): 1109–1118.

Harrell, Frank E. 2001. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. New York: Springer.

Harrell, Frank E., Kerry L. Lee and Daniel B. Mark. 1996. Multivariable prognostic models: issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine* 15 (4): 361–387.

Hayes, Andrew F. and Klaus Krippendorff. 2007. Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures* 1 (1): 77–89.

Heerwegh, Dirk and Geert Loosveldt. 2006. An Experimental Study on the Effects of Personalization, Survey Length Statements, Progress Indicators, and Survey

Sponsor Logos in Web Surveys. *Journal of Official Statistics* 22 (2): 191–210.

———. 2008. Face-to-face versus web surveying in a high-Internet-coverage population: Differences in response quality. *Public Opinion Quarterly* 72 (5): 836–846. doi:10.1093/poq/nfn045.

Hox, J. J., E. D. de Leeuw and I. G. G. Kreft. 1991. The effect of interviewer and respondent characteristics on the quality of survey data: a multilevel model. In *Measurement Errors in Surveys*, eds. P. P. Biemer, R. M. Grovesand L. E. Lyberg, N. A. Mathiowetz and S. Sudman. New York: Wiley.

Hox, Joop J. 2002. *Multilevel Analysis: Techniques and Applications*. New Jersey: Laurence Erlbaum Associates.

Hox, Joop J. and EdithD. Leeuw. 1994. A comparison of nonresponse in mail, telephone, and face-to-face surveys. *Quality and Quantity* 28 (4): 329–344. doi: 10.1007/BF01097014.

James, J.M. and R. Bolstein. 1990. The effect of monetary incentives and follow-up mailings on the response rate and response quality in mail surveys. *Public Opinion Quarterly* 54 (3): 346–361. doi:10.1086/269211. Available at: `http://poq.oxfordjournals.org/content/54/3/346.abstract`.

Janssen, Rianne, Jan Schepers and Deborah Peres. 2004. Models with item and item group predictors. In *Explanatory Item Response Models*, eds. Paul De Boeck and Mark Wilson, 189–212. New York: Springer.

Jäckle, Annette and Peter Lynn. 2008. Respondent incentives in a multi-mode panel survey: Cumulative effects on nonresponse and bias. *Survey Methodology* 34 (1): 105–117.

Klassen, Robert D. and Jennifer Jacobs. 2001. Experimental comparison of Web, electronic and mail survey technologies in operations management. *Journal of Operations Management* 19 (6): 713–728. doi:http://dx.doi.org/10.1016/S0272-6963(01)00071-7.

Klein, David J., Marc N. Elliott, Amelia M. Haviland, Debra Saliba, Q. Burkhart, Carol Edwards and Alan M. Zaslavsky. 2011. Understanding Nonresponse to the 2007 Medicare CAHPS Survey. *The Gerontologist* 51 (6): 843–855. doi: 10.1093/geront/gnr046.

Klein, John P. and Melvin L. Moeschberger. 2003. *Survival analysis: techniques for censored and truncated data*. New York, NY: Springer.

Kleinbaum, David G. and Mitchel Klein. 2005. *Survival Analysis: A Self-Learning Text*. 2nd ed. New York: Springer.

Knäuper, Bärbel, Robert F. Belli, Daniel H. Hill and A. Regula Herzog. 1997. Question Difficulty and Respondents' Cognitive Ability: The Effect on Data Quality. *Journal of Official Statistics* 13 (2): 181–199.

Koch, Achim and Rolf Porst, eds. 1998. *Nonresponse in survey research*. Manheim,

FRG: ZUMA Nachrichten Spezial.

Korendijk, Elly J.H., Cora J.M. Maas, Joop J. Hox and Mirjam Moerbeek. 2008. The Robustness of the Parameter and Standard Error Estimates in Trials with Partially Nested Data. A Simulation Study. In *Proceedings of the 23rd International Workshop on Statistical Modelling*, 299–304. Available at: `http://www.statmod.org/files/proceedings/iwsm2008_proceedings.pdf` (October 2, 2013).

Kreuter, Frauke, Stanley Presser and Roger Tourangeau. 2008. Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity. *Public Opinion Quarterly* 72 (5): 847–865. doi:10.1093/poq/nfn063.

Krippendorff, Klaus. 2004. Reliability in Content Analysis: Some Common Misconceptions and Recommendations. *Human Communication Research* 30 (3): 411–433.

Krosnick, Jon A. 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology* 5 (3): 213–236. doi:10.1002/acp.2350050305.

Krosnick, Jon A. and Duane f. Alwin. 1987. An Evaluation of a Cognitive Theory of Response-order Effects in Survey Measurement. *Public Opinion Quarterly* 51 (2): 201–219. doi:10.1086/269029.

Kupek, Emil. 1998. Determinants of Item Nonresponse in a Large National Sex Survey. *Archives of Sexual Behavior* 27 (6): 581–594. doi:10.1023/A:1018721100903.

Kveder, Andrej. 2005. Multilevel item nonresponse modeling. Ph.D. thesis, Faculty of Social Sciences, University of Ljubljana.

Kwak, Nojin and Barry Radler. 2002. A comparison between mail and web surveys : Response pattern, respondent profile, and data quality. *Journal of Official Statistics* 18 (2): 257–273.

Little, Roderick J. A. 1988. Missing-Data Adjustments in Large Surveys. *Journal of Business & Economic Statistics* 6 (3): 287–296.

Little, Roderick J. A. and Donald B. Rubin. 2002. *Statistical Analysis with Missing Data*. 2nd ed. Wiley-Interscience.

Lord, F.M. and M. Novick. 1968. *Statistical Theories of Mental Test Scores*. Reading, MA: Addison Wesley.

Lorenc, Boris. 2010. Item nonresponse analysis for a mixed-mode survey. In *Official Statistics –Methodology and Applications in Honour of Daniel Thorburn*, eds. Michael Carlson, Hans Nyquist and Mattias Villani, 117–136. Department of Statistics, Stockholm University.

Lozar Manfreda, Katja and Vasja Vehovar. 2002. Do Mail and Web Surveys Provide Same Results? *Development in Social Science Methodology* 18: 149–169. Available at: `http://mrvar.fdv.uni-lj.si/pub/mz/mz18/lozar1.pdf` (September 30, 2013).

Lozar Manfreda, Katja and Vasja Vehovar. 2002. Survey Design Features Influencing

Response Rates in Web Surveys. In *Proceedings of the International Conference on Improving Surveys, 2002*.

Mack, S., V. Huggins, D. Keathley and M. Sundukchi. 1998. Do monetary incentives improve response rates in the survey of income and programme participation? In *Proceedings of the Survey Research Methods Section*. Alexandria, VA: American Statistical Association. Available at: `www.amstat.org/sections/srms/Proceedings/papers/1998_089.pdf` (October 1, 2013).

Matzat, Uwe, Chris Snijders and Wouter van der Horst. 2009. Effects of Different Types of Progress Indicators on Drop-Out Rates in Web Surveys. *Social Psychology* 40 (1): 43–52.

Meulders, Michael and Yiyu Xie. 2004. Person-by-item predictors. In *Explanatory Item Response Models*, eds. Paul De Boeck and Mark Wilson, 43–74. New York: Springer.

Molenberghs, Geert and Geert Verbeke. 2004. An Introduction to (Generalized (Non)Linear Mixed Models. In *Explanatory Item Response Models*, eds. Paul De Boeck and Mark Wilson, 43–74. New York: Springer.

Musch, Jochen and Ulf-Dietrich Reips. 2000. A Brief History of Web Experimenting. In *Psychological Experiments on the Internet*, ed. M.H. Birnbaum. San Diego: Academic Press.

Olson, Kristen. 2010. An Examination of Questionnaire Evaluation by Expert Reviewers. *Field Methods* 22 (4): 295–318. doi:10.1177/1525822X10379795.

Olsson, Ulf. 2002. *Generalized Linear Models: An Applied Approach*. Lund: Studentlitteratur.

O'Quigley, John and Janez Stare. 2002. Proportional hazards models with frailties and random effects. *Statistics in Medicine* 21 (21): 3219–3233. doi:10.1002/sim.1259.

O'Quigley, John, Ronghui Xu and Janez Stare. 2005. Explained randomness in proportional hazards models. *Statistics in Medicine* 24 (3): 479–489. doi:10.1002/sim.1946. Available at: `http://dx.doi.org/10.1002/sim.1946`.

Petrolia, Daniel R. and Sanjoy Bhattacharjee. 2009. Revisiting Incentive Effects: Evidence from a Random-Sample Mail Survey on Consumer Preferences for Fuel Ethanol. *Public Opinion Quarterly* 73 (3): 537–550. doi:10.1093/poq/nfp038. Available at: `http://poq.oxfordjournals.org/content/73/3/537.abstract`.

Peytchev, Andrey A. 2007. Participation Decisions and Measurement Error in Web Surveys. Ph.D. thesis, University of Michigan.

Peytchev, Andy. 2009. Survey Breakoff. *Public Opinion Quarterly* 73 (1): 74–97. doi:10.1093/poq/nfp014.

———. 2011. Breakoff and unit nonresponse across web surveys. *Journal of Official Statistics* 27 (1): 33–47.

———. 2013. Consequences of Survey Nonresponse. *Annals of the American Academy*

of Political and Social Science 645 (1): 88–111. doi:10.1177/0002716212461748.

Pickery, Jan and Geert Loosveldt. 1998. The Impact of Respondent and Interviewer Characteristics on the Number of "No Opinion" Answers. *Quality and Quantity* 32 (1): 31–45. doi:10.1023/A:1004268427793.

———. 2001. An Exploration of Question Characteristics that Mediate Interviewer Effects on Item Nonresponse. *Journal of Official Statistics* 17 (3): 337–350.

———. 2004. A Simultaneous Analysis of Interviewer Effects on Various Data Quality Indicators with Identification of Exceptional Interviewers. *Journal of Official Statistics* 20 (1): 77–89.

Platinovšek, Rok. 2013. Applying multilevel item response theory models to the problem of item nonresponse in surveys. In *9th International Conference Multilevel Analysis, March 27 & 28, 2013*. Utrecht: Universiteit Utrecht, Faculteit Sociale Wetenschappen, Methoden & Technieken.

Presser, Stanley and Johnny Blair. 1994. Survey Pretesting: Do Different Methods Produce Different Results? *Sociological Methodology* 24: 73–104.

R Core Team. 2013. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. Available at: `http://www.R-project.org/` (July 4, 2013).

Rasch, Georg. 1960. *Probabilistic models for some intelligence and attainment tests*. Copenhagen: Danmarks Paedagogiske Institut.

Reckase, Mark D. 2009. *Multidimensional Item Response Theory*. New York, NY: Springer.

Reja, U., K. Lozar Manfreda, V. Hlebec and V. Vehovar. 2003. Open-ended vs. Close-ended Questions in Web Questionnaires. *Advances in methodology and statistics (Metodološki zvezki)* 19: 159–177.

Roster, Catherine A., Robert D. Rogers, Gerald Albaum and Darin Klein. 2004. A comparison of response characteristics from web and telephone surveys. *International Journal of Market Research* 46 (3): 359–374.

Rubin, Donald B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.

Schafer, J.L. 1997. *Analysis of Incomplete Multivariate Data*. Chapman and Hall/CRC.

Schemper, Michael and Janez Stare. 1996. Explained variation in survival analysis. *Statistics in Medicine* 15 (19): 1999–2012.

Schoenfeld, David. 1982. Partial residuals for the proportional hazards regression model. *Biometrika* 69 (1): 239–241. doi:10.1093/biomet/69.1.239. Available at: `http://biomet.oxfordjournals.org/content/69/1/239.abstract`.

Schuman, H. and S. Presser. 1981. *Questions and Answers in Attitude Surveys: Experiments on Question Form, Order, and Context*. New York: Academic Press.

Schwarz, Norbert and Seymour Sudman. 1992. *Context Effects in Social and Psychological Research*. New York: Springer.

———. 1996. *Answering Questions: Methodology For Determining Cognitive And Communicative Processes In Survey Research*. San Francisco: Jossey-Bass Publishers.

Shettle, C. and G. Mooney. 1999. Monetary incentives in U.S. government surveys. *Journal of Official Statistics* 15 (2): 231–250.

Shin, Eunjung, Timothy P. Johnson and Kumar Rao. 2012. Survey Mode Effects on Data Quality: Comparison of Web and Mail Modes in a U.S. National Panel Survey. *Social Science Computer Review* 30 (2): 212–228. doi: 10.1177/0894439311404508.

Singer, Eleanor. 2002. The use of incentives to reduce nonresponse in household surveys. In *Survey nonresponse*, eds. Robert M. Groves, Don A. Dillman, John L. Eltinge and Roderick J. A. Little, 163–178. New York: Wiley.

Singer, Eleanor, Martin R. Frankel and Marc B. Glassman. 1983. The Effect of Interviewer Characteristics and Expectations on Response. *The Public Opinion Quarterly* 47 (1): 68–83.

Singer, Eleanor, John van Hoewyk and Mary P. Maher. 1998. Does the Payment of Incentives Create Expectation Effects? *The Public Opinion Quarterly* 62 (2): 152–164. Available at: `http://www.jstor.org/stable/2749620` (May 3, 2011).

———. 2000. Experiments with Incentives in Telephone Surveys. *The Public Opinion Quarterly* 64 (2): 171–188. Available at: `http://www.jstor.org/stable/3078814` (May 3, 2011).

Singer, Eleanor, John Van Hoewyk, Nancy Geblerand Trivellore Raghunathan and Katherine McGonagle. 1999. The effect of incentives on response rates in interviewer-mediated surveys. *Journal of Official Statistics* 15 (2): 217–230.

Singer, Eleanor and Cong Ye. 2013. The Use and Effects of Incentives in Surveys. *The ANNALS of the American Academy of Political and Social Science* 645 (1): 112–141. doi:10.1177/0002716212458082.

Smyth, Jolene D., Leah Melani Christian and Don A. Dillman. 2008. Does "yes or no" on the telephone mean the same as "chek-all-that-apply" on the web? *Public Opinion Quarterly* 72 (1): 103–113. doi:10.1093/poq/nfn005.

Snijders, Tom A.B. and Roel Bosker. 1999. *Multilevel Analysis: An Introduction to Basic and Advanced Multilevel Modeling*. London: Sage.

Stanton, Jeffrey M. 1998. An empirical assessment of data collection using the internet. *Personnel Psychology* 51 (709–725): 709–725. doi:10.1111/j.1744-6570.1998.tb00259.x.

Statistical office of the republic of Slovenia. 2013. *SI-STAT Data Portal (Education)*. Available at: `http://pxweb.stat.si/pxweb/Database/Demographics/05_`

population/20_Socio-economicPopulation/01_05G20_education/01_05G20_education.asp (May 14, 2013).

Stocke, Volker. 2006. Attitudes Toward Surveys, Attitude Accessibility and the Effect on Respondents' Susceptibility to Nonresponse. *Quality & Quantity* 40 (2): 259–288. doi:10.1007/s11135-005-6105-z.

Stocke, Volker and Bettina Langfeldt. 2004. Effects of Survey Experience on Respondents' Attitudes Towards Surveys. *Bulletin de Méthodologie Sociologique* 81 (1): 5–32. doi:10.1177/075910630408100103.

Strack, F. and L.L. Martin. 1987. Thinking, judging, and communicating: A process account of context effects in attitude surveys. In *Social Information Processing and Survey Methodology*, eds. H.J. Hippler, N. Schwarz and S. Sudman, 123–148. New York: Springer-Verlag.

Suchman, Lucy and Brigitte Jordan. 1990. Interactional Troubles in Face-to-Face Survey Interviews. *Journal of the American Statistical Association* 85 (409): 232–253.

Sudman, Seymour and Norman M. Bradburn. 1974. *Response effects in surveys: A review and synthesis*. Aldine Pub. Co.

Sudman, Seymour, Norman M. Bradburn and Norbert Schwarz. 1996. *Thinking About Answers: The Application of Cognitive Processes to Survey Methodology*. San Francisco: Jossey-Bass Publishers.

Teisl, Mario F., Brian Roe and Michael E. Vayda. 2006. Incentive Effects on Response Rates, Data Quality, and Survey Administration Costs. *International Journal of Public Opinion Research* 18 (3): 364–373. doi:10.1093/ijpor/edh106.

Therneau, Terry M. 1999. *A Package for Survival Analysis in S*. Available at: www.mayo.edu/research/documents/tr53pdf/DOC-10027379 (July 4, 2013).

———. 2013. *A Package for Survival Analysis in S. R package version 2.37-4*. Available at: http://CRAN.R-project.org/package=survival (July 4, 2013).

Therneau, Terry M. and Patricia M. Grambsch. 2000. *Modeling Survival Data: Extending the Cox Model*. New York: Springer.

Thissen, D. and M. Orlando. 2001. Item response theory for items scored in two categories. In *Test scoring*, eds. D. Thissen and H. Wainer. Mahawah, NJ: Lawrence Erlbaum.

Tourangeau, R. and K. Rasinski. 1988. Cognitive processes underlying context effects in attitude measurement. *Psychological Bulletin* 103 (3): 299–314.

Tourangeau, R. and T. Yan. 2007. Sensitive Questions in Surveys. *Psychological Bulletin* 133 (5): 859–888.

Tourangeau, Roger, Lance J. Rips and Kenneth Rasinski. 2000. *The Psychology of Survey Response*. Cambridge University Press.

Truell, Allen D., James E. Bartlett and Melody W. Alexander. 2002. Response rate,

speed, and completeness. *Behaviour Research Methods, Instruments, & Computers* 34 (1): 46–49.

Tu, Su-Hao and Pei-Shan Liao. 2007. Social Distance, Respondent Cooperation and Item Nonresponse in Sex Survey. *Quality & Quantity* 41 (2): 177–199. doi: 10.1007/s11135-007-9088-0.

Tuerlinckx, Francis and Paul De Boeck. 2004. Person-by-item predictors. In *Explanatory Item Response Models*, eds. Paul De Boeck and Mark Wilson, 43–74. New York: Springer.

Tzamourani, P. and P. Lynn. 1999. *The Effect of MOnetary Incentives on Data Quality - Results from the British Social Attitudes Survey 1998 Experiment. CREST Working Paper No. 73.* Tech. Rep.. Oxford: University of Oxford. Available at: `www.crest.ox.ac.uk/papers/p73.pdf` (October 1, 2013).

van Buuren, Stef. 2012. *Flexible Imputation of Missing Data.* Chapman and Hall/CRC.

van Buuren, Stef and Karin Groothuis-Oudshoorn. 2011. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software* 45 (3): 1–67.

Wienke, Andreas. 2011. *Frailty Models in Survival Analysis.* London: Chapman and Hall.

Willimack, D.K., H. Schuman, B.E. Pennell and J.M Lepkowski. 1995. Effects of a prepaid nonmonetary incentive on response rates and response quality in a face-to-face survey. *Public Opinion Quarterly* 59 (1): 78–92. doi:10.1086/269459. Available at: `http://poq.oxfordjournals.org/content/59/1/78.abstract`.

Wolfe, Edward W., Patrick D. Converse and Frederick L. Oswald. 2008. Item-Level Nonresponse Rates in an Attitudinal Survey of Teachers Delivered via Mail and Web. *Journal of Computer-Mediated Communication* 14 (1): 35–66.

Yan, Ting and Richard Curtin. 2010. The Relation Between Unit Nonresponse and Item Nonresponse: A Response Continuum Perspective. *International Journal of Public Opinion Research* 22 (4): 535–551. doi:10.1093/ijpor/edq037.

Subject index

Author index

A Multiple imputation

There is no clear-cut method for determining when the MICE algorithm has converged. The literature suggests to plot values of various parameters against the iteration number, resulting in so-called *traceplots*. The traceplots in this appendix show the mean and standard deviation of the imputed values for each incomplete variable in the MI model. The means for all $m = 5$ streams are plotted in the same chart; the same goes for the standard deviations. On convergence, the different streams should be freely intermingled with each other, showing no definite trends (van Buuren and Groothuis-Oudshoorn 2011).

Figure A.1: Kernel density plot for marginal distributions of observed data (blue) and the five densities per variable calculated from imputed data (red) for the additional Facebook sample

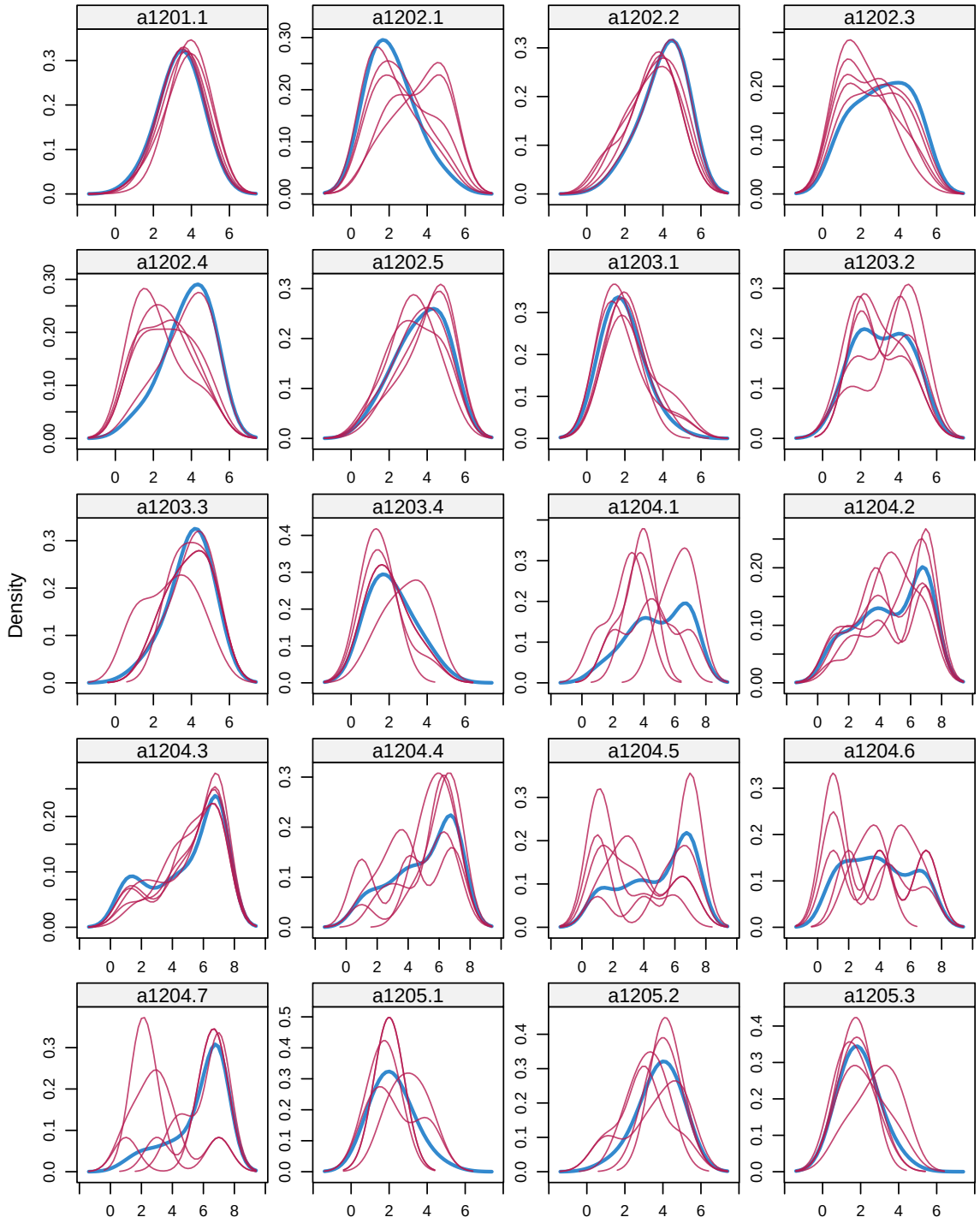


Figure A.2: Traceplots for multiple imputation with predictive mean matching for the MI model for all samples except web.fb (1/2)

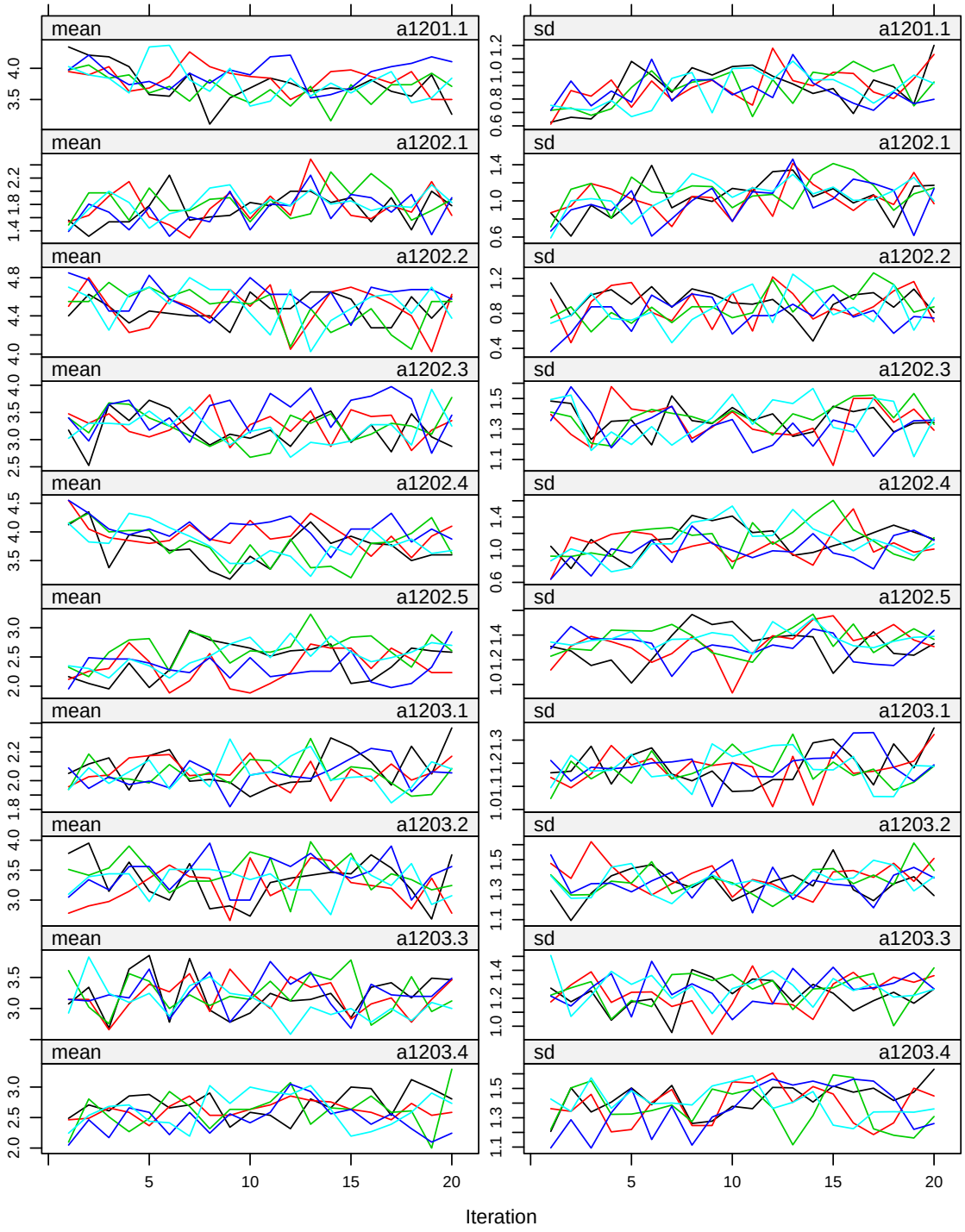


Figure A.3: Traceplots for multiple imputation with predictive mean matching for the MI model for all samples except web.fb (2/2)

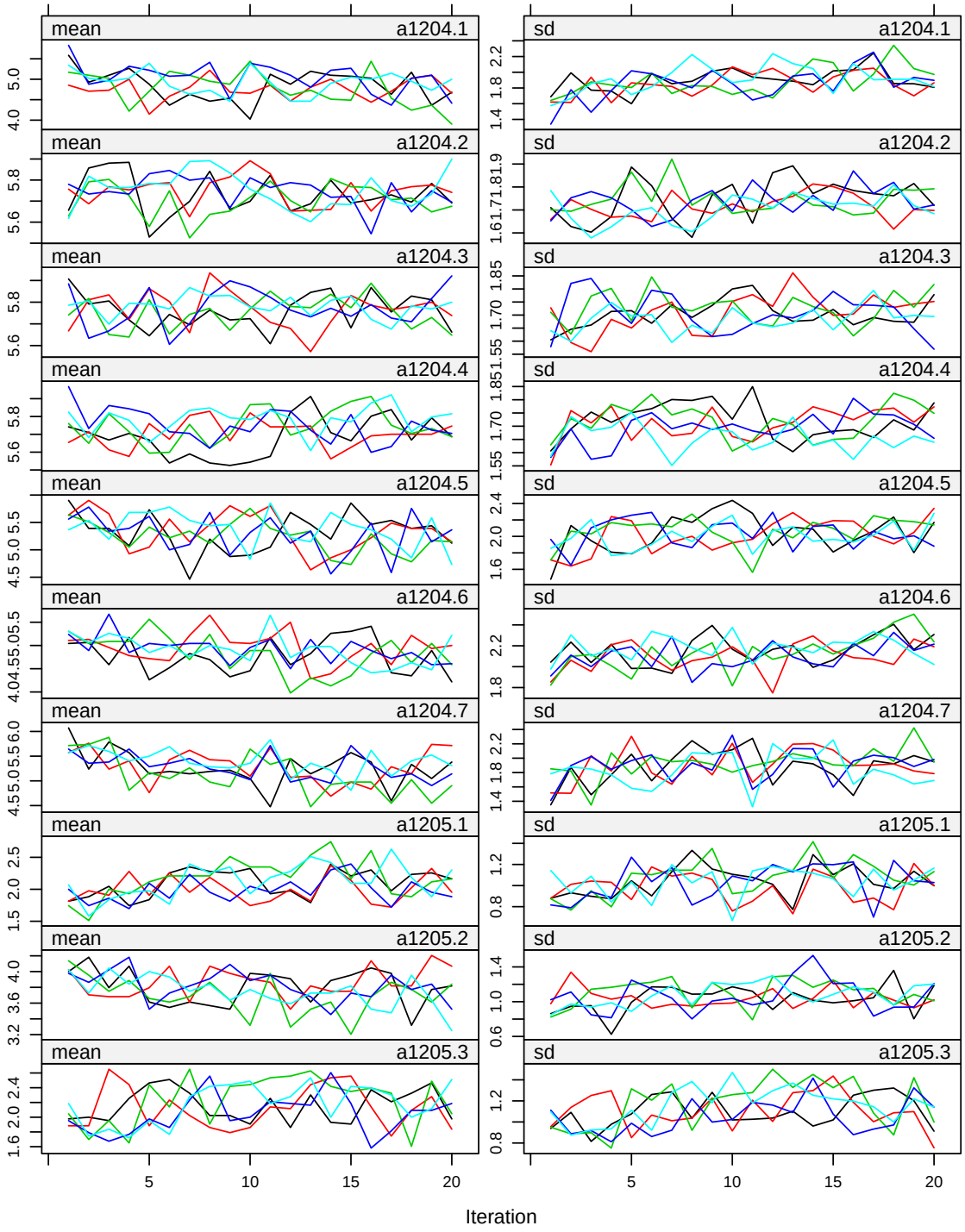


Figure A.4: Traceplots for multiple imputation with predictive mean matching for the MI model for the Facebook sample (1/2)

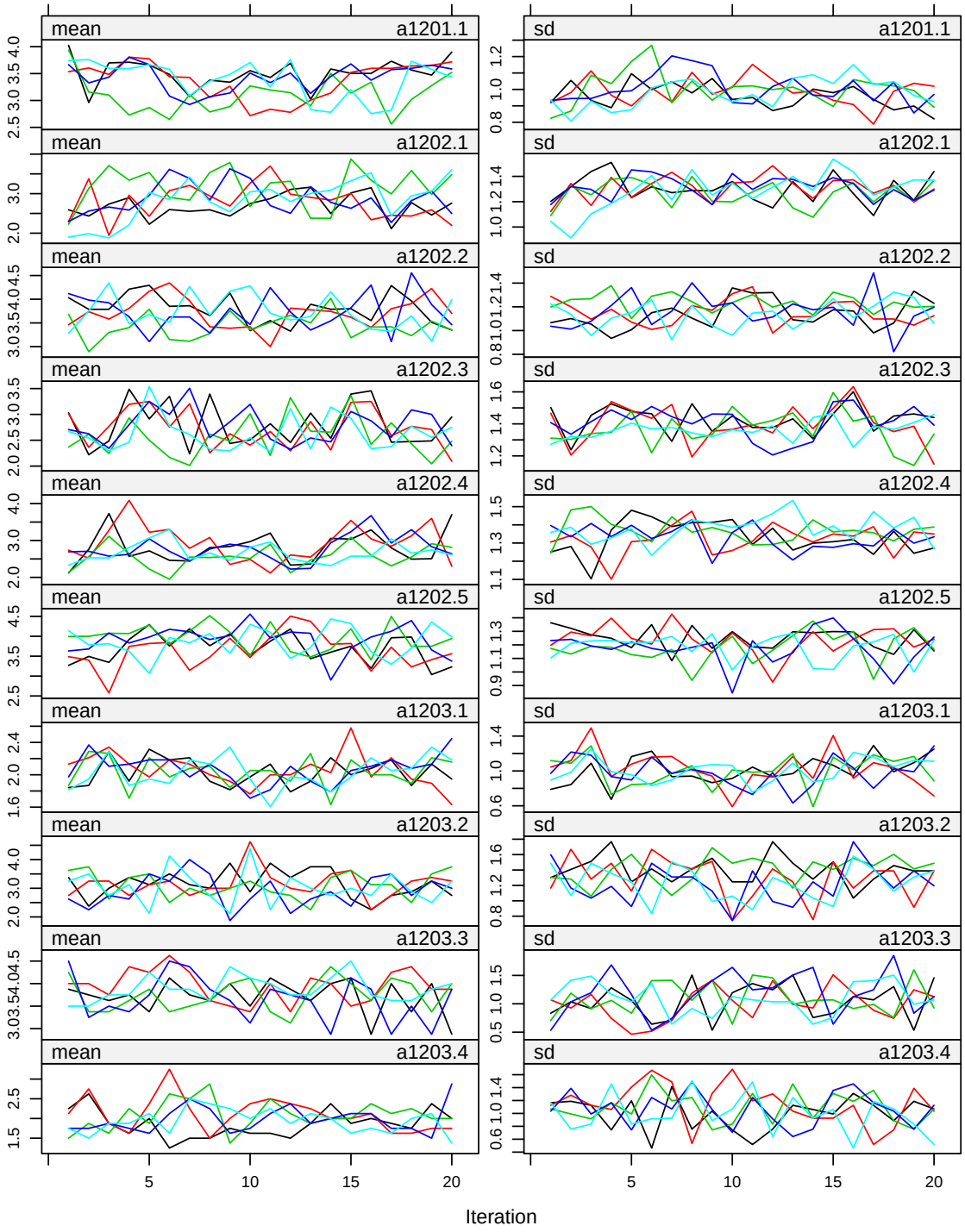


Figure A.5: Traceplots for multiple imputation with predictive mean matching for the MI model for the Facebook sample (2/2)

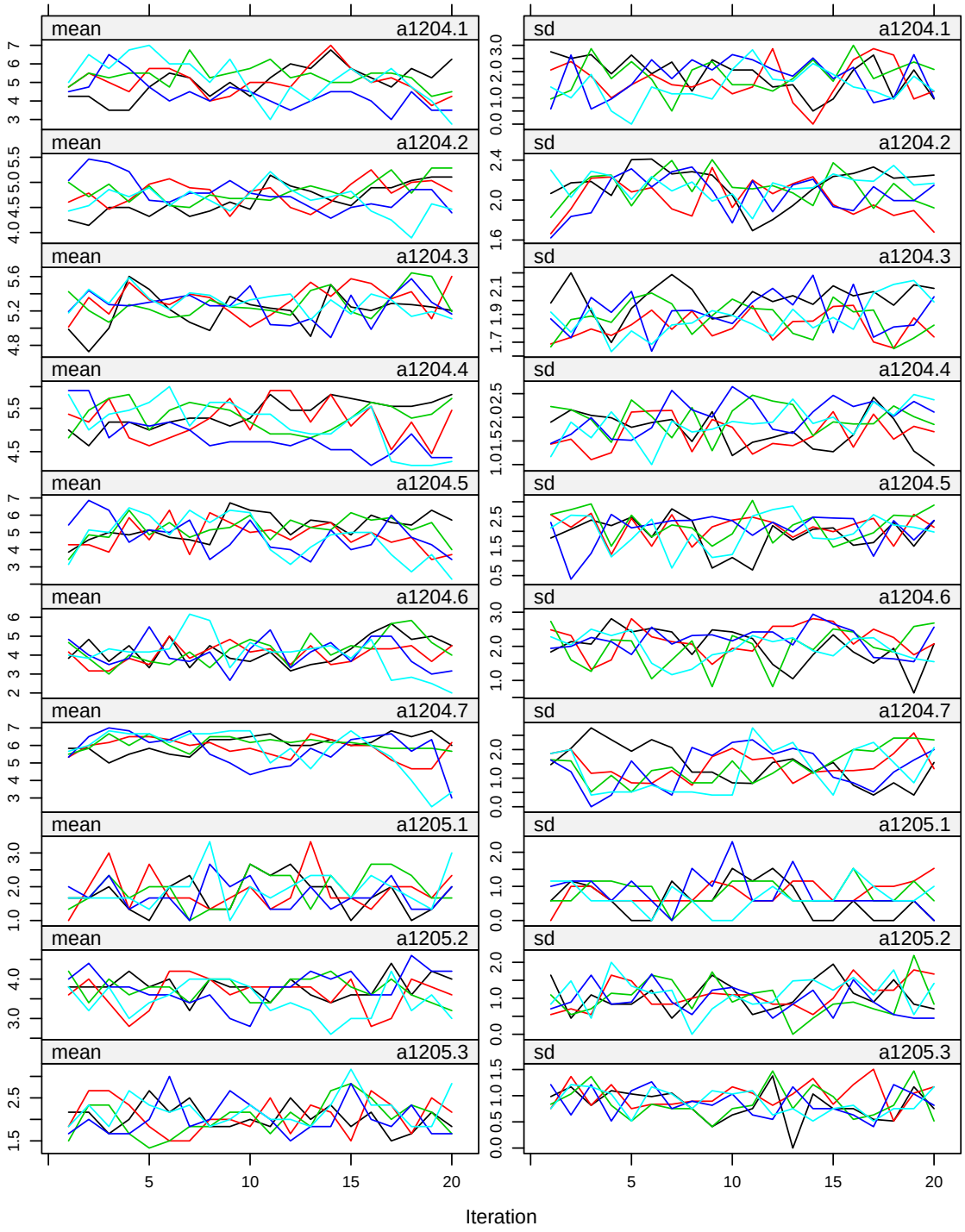
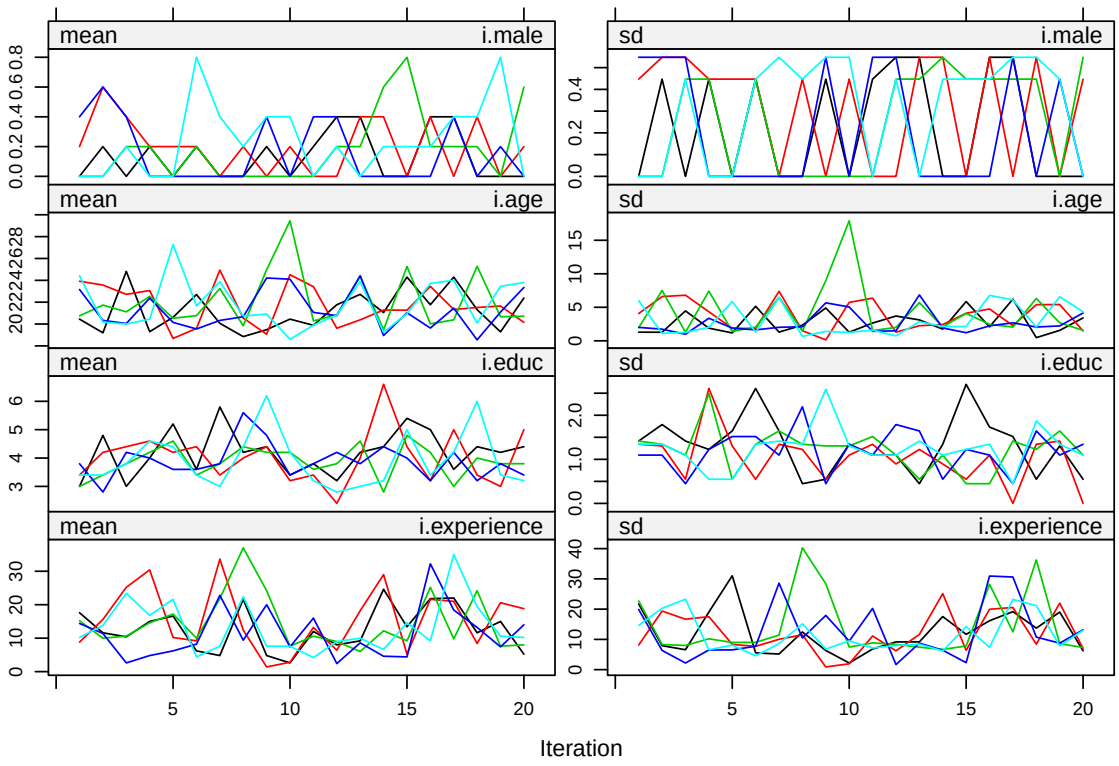


Figure A.6: Traceplots for multiple imputation with predictive mean matching for the MI model for the interviewer-level variables



Povzetek

Neodgovor predstavlja na področju anketne metodologije izrazit problem, saj manjkajoče vrednosti, ki so posledica neodgovora, zmanjšujejo zaupanje v anketne ocene. Manjkajoče vrednosti se lahko pojavijo na različne načine. *Neodgovor enote* se zgodi, ne uspemo pridobiti meritev za celotno vzorčeno enoto, kar se navadno zgodi, ko vzorčena oseba zavrne sodelovanje v anketi ali vzorčene osebe ne uspemo kontaktirati. O *neodgovoru na anketno vprašanje* (angl. *item nonresponse*) govorimo, ko je vzorčena oseba pripravljena sodelovati v anketi, vendar pa podatki za določena anketna vprašanja niso na voljo. O prekinitvi anketiranja (angl. *breakoff, dropout*) govorimo, ko anketiranec[34] začne izpolnjevati anketo, vendar preneha, še preden jo dokonča. V doktorski disertaciji se osredotočamo na obliki neodgovora, ki sta bili doslej v literaturi deležni manj pozornosti: neodgovor na anketno vprašanje ter prekinitev anketiranja.

Teoretično ozadje

Anketni intervju je bil v sedemdesetih letih prejšnjega stoletja konceptualiziran z uporabo pojmov iz socialne in kognitivne psihologije. Tourangeau et al. (2000) delijo proces odgovarjanja na štiri faze: razumevanje (angl. *comprehension*), priklic (angl. *retrieval*), presojanje (angl. *judgment*) in odgovor (angl. *response*). Anketiranec lahko odstopi od idealne poti v katerikoli izmed štirih faz, kar vodi v mersko napako ali neodgovor. Krosnick (1991) imenuje pozorno in izčrpno izvajanje vsake od štirih faz *optimiziranje* (angl. *optimizing*). Po njegovem mnenju se anketiranci kmalu po začetku anketiranja utrudijo ter postanejo nezainteresirani, nepotrpežljivi in raztreseni. Breme procesa odgovarjanja postane previsoko, zato se anketiranci, namesto da bi poskušali ustvariti optimalni odgovor, zadovoljijo z ustvarjanjem zgolj sprejemljivih odgovorov. Krosnick takšno vedenje imenuje *zadovoljevanje* (angl. *satisficing*). Podajanje odgovora "ne vem" vidi kot eno izmed oblik zadovoljevanja.

[34]Izraza *anketiranec* in *anketar* v nadaljevanju uporabljamo v moški obliki za pripadnike obeh spolov.

Literatura o procesu odgovarjanja na anketna vprašanja predstavlja osnovo za teoretične modele za bolj specifične pojave v anketnem procesu. Edini teoretični model, ki ga najdemo v literaturi in je specifično namenjen pojasnjevanju neodgovora na anketno vprašanje, je odločitveni model odgovora na anketno vprašanje (angl. *response decision model*, Beatty and Herrmann 2002). Po tem modelu na odločitev, odgovoriti ali ne, vplivajo trije faktorji: kognitivno stanje (angl. *cognitive state*), sodbe o primernosti (angl. *adequacy judgments*) in sporočanjski namen (angl. *communicative intent*). Beatty and Herrmann (2002) svoj model nekoliko poenostavita (tako da izpustita sporočanjski namen) in trdita, da je anketiranec, ki mu je zastavljeno anketno vprašanje, soočen z dvema odločitvama: ali je *sposoben* odgovoriti na vprašanje ter ali *želi* odgovoriti na vprašanje. Če se anketiranec odloči negativno glede katerekoli izmed omenjenih dveh odločitev, je rezultat neodgovor na anketno vprašanje.

Za razliko od raziskav neodgovora na anketno vprašanje, študije prekinitev anketiranja niso bile zelo pogoste dokler stopnja prekinitev ni postala resen problem, kar se je zgodilo z uveljavitvijo spletnega anketiranja v devetdesetih letih prejšnjega stoletja. V anketno-metodološki literaturi doslej ni bil predlagan teoretični model, ki bi bil namenjen specifično pojasnjevanju prekinitev anketiranja, zato so si avtorji, ki so preučevali prekinitve, morali teoretične modele izposojati z drugih področij. Galesic (2006) pojasnjuje anketirančevo odločitev za prekinitev anketiranja skozi pojem *zadrževalnega praga* (angl. *inhibitory threshold*, Busemeyer and Townsend 1993). Poudarja, da anketiranec pri vsakem postavljenem vprašanju ponovno pretehta svojo odločitev, da sodeluje v anketi. Ko se anketa nadaljuje, faktorji, ki so vplivali na izhodiščno odločitev za sodelovanje, postopno izgubljajo svoj vpliv in v ospredje stopijo negativni vidiki sodelovanja npr. utrujenost in dolgčas. S tem narašča anketirančeva preferenca, da preneha sodelovati v anketi, vendar pa se anketiranec za prekinitev ne odloči vse dokler omenjena sprememba v preferenci ne preseže zadrževalnega praga. Galesic (2006) v svoji raziskavi ugotovi, da je neodgovor na anketno vprašanje pogostejši tik preden pride do prekinitve anketiranja.

Čeprav se Galesic (2006) ne sklicuje neposredno na Krosnicka, trdimo, da je njena ideja vmesnega obdobja nizke motivacije (ko bi anketiranec že rajši prenehal sode-

lovati, ampak se še ne odloči za prekinitev) zelo podobna Krosnickovemu konceptu zadovoljevanja. Razlike lahko izhajajo iz dejstva, da je Krosnick pojem zadovoljevanja opredelil za osebne in telefonske ankete, kjer prekinitev anketiranja zahteva, da anketiranec prekine pogovor s sogovornikom—anketarjem. Trdimo, da je zaradi tega zadrževalni prag ob prisotnosti anketarja občutno višji: za razliko od spletnih anket je ob prisotnosti anketarja bolj verjetno, da bo anketiranec prišel do konca vprašalnika, preden bo dosežen zadrževalni prag. Krosnick kot eno izmed oblik zadovoljevanja omenja neodgovor na anketno vprašanje, s katerim si anketiranec prihrani kognitivni napor odgovarjanja na anketno vprašanje. Trdimo, da je to razlog, da Galesic (2006) opaža porast pogostosti neodgovora na anketno vprašanje tik pred prekinitvijo anketiranja. Tako kot Peytchev (2009) torej vidimo prekinitev anketiranja kot bolj skrajno alternativo neodgovoru na anketno vprašanje. Na odločitev za neodgovor na anketno vprašanje kot tudi na odločitev za prekinitev anketiranja vplivajo isti faktorji: po eni strani anketirančeva motivacija, po drugi pa breme anketnih vprašanj.

Zbiranje podatkov ter anketa *Generations and Gender Programme*

V empiričnem delu disertacije smo neodgovor na anketno vprašanje ter prekinitev neodgovora proučevali na primeru podatkov, pridobljenih s pilotsko anketo *Generations and Gender Programme* (v nadaljevanju GGP). Isti vprašalnik je bil izveden v treh načinih anketiranja: osebno, telefonsko in spletno. Da bi pri danih sredstvih maksimizirali velikost vzorca, smo se odločili podatke zbirati v dveh fazah. V prvi fazi so bili anketiranci člani spletnega panela podjetja Valicon, v drugi fazi pa so bili vzorčeni iz Slovenske populacije. V vsaki od obeh faz zbiranja podatkov smo uporabili vse tri omenjene načine anketiranja.

Čeprav je izpolnjevanje vprašalnika GGP tipičnemu anketirancu vzelo približno eno uro, pa v prvih dveh fazah zbiranja podatkov skoraj ni bilo prekinitev anketiranja; presenetljivo jih ni bilo niti v spletnem načinu. Takšni podatki nam ne omogočajo analizirati, kako so lastnosti anketiranca in anketnih vprašanj povezane z prekinitvijo anketiranja, zato smo se odločili za dodatno zbiranje podatkov. V tej tretji fazi zbiranja podatkov smo anketirance na anketo vabili z oglasi na spletni strani Facebook.

Za razliko od anketirancev v prvih dveh fazah, ki so bili o trajanju ankete obveščeni v vabilu k sodelovanju, anketirancem v tretji fazi zbiranja podatkov nismo vnaprej povedali, kako dolgo bo trajalo izpolnjevanje vprašalnika. Prekinitev anketiranja je bila v tej tretji fazi zbiranja podatkov mnogo pogostejša, saj je anketiranje prekinila več kot polovica anketirancev. Domnevamo, da so se za sodelovanje v prvih dveh fazah zbiranja podatkov odločili zgolj visoko motivirani anketiranci, ki so bili anketo pripravljeni izpolnjevati eno uro. V tretji fazi zbiranja podatkov pa so z odgovarjanjem začeli tudi niže motivirani anketiranci, ki so v nadaljevanju zato pogosto prekinili anketiranje.

Ob pregledu demografske strukture zbranih podatkov ugotovimo, da opazno izstopa vzorec, zbran v dodatni tretji fazi spletnega anketiranja z rekurutacijo preko Facebooka. V omenjenem vzorcu je kar tri četrtine žensk, več kot polovica vseh anketirancev pa je mlajših od 26 let. To demografsko strukturo pripisujemo samo-izbiri, saj v tretji fazi zbiranja podatkov nismo uporabili vzorčenja, pač pa so se uporabniki spletne strani Facebook sami odločili za sodelovanje s klikom na oglas. V primerjavi z omenjenim vzorcem iz tretje faze zbiranja podatkov pa se vzorci iz prvih dveh faz ne razlikujejo občutno od Slovenske populacije polnoletnih oseb.

Ekspertno ocenjevanje anketnih vprašanj

Z anketiranjem smo zbrali podatke o neodgovoru in prekinitvah anketiranja (odvisni spremenljivki, ki ju bomo pojasnjevali v statističnih modelih), kot tudi o značilnostih anketirancev. Po drugi strani pa na podlagi pregledane literature predpostavljamo, da na neodgovor in prekinitev vplivajo tudi značilnosti anketnih vprašanj, zlasti občutljivost anketnih vprašanj. Tourangeau, Rips and Rasinski (2000) trdijo, da se v literaturi pojavljajo trije različni pomeni pojma občutljivosti. Anketno vprašanje lahko anketiranec dojema kot občutljivo, prvič, kadar je tema le-tega *vsiljiva*. Za vsiljivost gre, kadar se anketno vprašanje dotika tem, ki niso primerne v vsakodnevnih pogovorih. Tovrstna vprašanja predstavljajo napad na anketirančevo zasebnost ne glede na pravilni odgovor. Drugi pomen občutljivosti zadeva *nevarnost razkritja* občutljivih informacij, pri čemer anketiranca skrbijo posledice, če se odloči na anketno vprašanje odgovoriti po resnici in priznati npr. kriminalno dejanje, uporabo drog ipd. Tretji

pomen občutljivosti zadeva *družbeno zaželenost*: vprašanje je občutljivo v tem smislu, kadar od anketiranca izvablja družbeno zaželen odgovor. To pojmovanje občutljivosti predpostavlja obstoj jasnih norm glede anketirančevega vedenja ali mnenj.

Različne vidike občutljivosti smo merili z ekspertnim ocenjevanjem vsakega izmed anketnih vprašanj v vprašalniku GGP. Da bi izmerili kar se da različne aspekte občutljivosti anketnih vprašanj, pa smo se odločili obrniti tretji pomen občutljivosti in tako meriti *potencial za pretirano pozitivno predstavitev*. Ta pojem se tako nanaša na družbeno zaželeno vedenje kot npr. udeležbo na volitvah. Pričakujemo, da bo pri vprašanjih, ki zadevajo takšna vedenja ter mnenja, *manj* neodgovora na anketno vprašanje in prekinitev, saj se lahko prekinitev ali neodgovor pri takšnem vprašanju razume kot "greh izostanka" (angl. *sin of omission*, Bradburn et al. 1978).

Trije neodvisni ocenjevalci so za vsako izmed anketnih vprašanj vprašalnika GGP ocenili vsako izmed omenjenih mer občutljivosti. Čeprav so si bili ocenjevalci po poklicnem ozadju podobni (vsi so bili metodologi s Fakultete za družbene vede Univerze v Ljubljani), pa smo na podlagi ocen izračunali dokaj nizko mero strinjanja med ocenjevalci: Krippendorffov alpha je znašal 0.65 za vsiljivost, 0.21 za nevarnost razkritja in 0.51 za potencial za pretirano pozitivno predstavitev. Navkljub nizkim vrednostim Krippendorffovega alphe (opomba velja zlasti za nevarnost razkritja) smo za vsako anketno vprašanje izračunali povprečje vsake od mer občutljivosti po treh ocenjevalcih ter tako dobljene mere uporabili kot pojasnjevalne spremenljivke v statističnih modelih za neodgovor na anketno vprašanje in prekinitev anketiranja. Prejšnje raziskave (Olson 2010; DeMaio and Landreth 2003; Presser and Blair 1994) so namreč pokazale, da so nizke mere zanesljivosti pri ekspertnem ocenjevanju anketnih vprašanj pogost problem ter da je mogoče na osnovi povprečne ekspertne ocene uspešno identificirati anketna vprašanja z visoko stopnjo neodgovora (Olson 2010).

Dodana anketna vprašanja

Da bi lahko bolje pojasnili neodgovor na anketno vprašanje ter prekinitve anketiranja, smo v vprašalnik GGP dodali določena anketna vprašanja. Če bi se zanašali zgolj na

prej omenjene mere občutljivosti anketnih vprašanj, bi predpostavljali, da določeno anketno vprašanje vsi anketiranci razumejo enako in da ima na njih enak učinek. V vprašalnik smo zato dodali anketna vprašanja, v katerih smo anketirance zaprosili, naj ocenijo, kako občutljivo bi bilo za njih odgovarjati na določene teme, ki so se pojavile v vprašalniku GGP. Osredotočili smo se na teme, za katere smo predvidevali, da bi lahko bile občutljive[35].

Z naslednjim sklopom vprašanj, ki smo jih dodali, smo skušali izmeriti anketirančevo kognitivno stanje glede določenih tem, ki so se pojavile v vprašalniku GGP. Anketiranca smo prosili, naj oceni svoje strinjanje s trditvami[36] kot "Znane so mi podrobnosti glede službe/dejavnosti moje/ga partnerja/ke." Če je anketiranec izrazil nestrinjanje s to izjavo, je verjetno, da je informacija, po kateri sprašuje anketno vprašanje, v višjem kognitivnem stanju, zato se bo ta anketiranec moral bolj potruditi pri oblikovanju odgovora. Pri tem anketirancu je zato bolj verjetno, da se bo pri anketnih vprašanjih, ki zadevajo partnerjevo dejavnost, pojavil neodgovor ali prekinitev.

Zadnji sklop vprašanj, ki smo jih dodali, je zadeval anketirančevo splošno naravnanost do anket. Anketna vprašanja smo vzeli iz daljšega merskega inštrumenta, ki je obsegal 16 indikatorjev (Stocke and Langfeldt 2004). Da vprašalnika GGP ne bi močno podaljšali, smo obdržali zgolj tri trditve[37], ter anketirance prosili, naj izrazijo strinjanje z njimi. Za anketirance, ki so bolj pozitivno naravnani do anket, je pričakovati, da bodo bolj skrbno izvajali vsako izmed faz procesa odgovarjanja na anketna vprašanja ter zato zagrešili manj neodgovorov na anketna vprašanj in imeli manjše nagnjenje k prekinitvi anketiranja.

[35]Vključili smo naslednje teme: odnosi anketiranca z drugimi ljudmi ter pomoč, ki jo daje in sprejema od njih; anketirančev odnos do svojega partnerja; anketirančev odnos do svojih otrok; anketirančev odnos do svojih staršev; odločitev imeti otroke; dohodek in imetje anketirančevega gospodinjstva; anketirančeva stališča do vprašanj, kot so poroka, odnosi med spoloma, medgeneracijski odnosi.

[36]Poleg omenjene trditve smo prosili še za oceno strinjanja z naslednjimi: 1) Včasih imam probleme s priklicem informacij kot npr. rojstnih dnevov sorodnikov. 2) Le redko razmišljam o svojih odnosih z drugimi ljudmi. 3) Natančno poznam finančno situacijo in finančne transakcije svojega gospodinjstva.

[37]Obržali smo smo naslednje trditve: 1) Ankete so pomembne za znanost, politiko in gospodarstvo. 2) Ankete me zgolj ovirajo pri tem, da bi počel/a pomembnejše stvari. 3) Pri anketah imam možnost izraziti svoje lastno mnenje.

Večkratno vstavljanje manjkajočih vrednosti

Pri analizi neodgovora na anketno vprašanje se manjkajoče vrednosti nikoli ne pojavijo pri odvisni spremenljivki. Kadar so anketiranci izpustili odgovor na določeno anketno vprašanje, smo to kodirali kot vrednost 1 na odvisni spremenljivki in zato neodgovor za naš namen ne predstavlja manjkajoče vrednosti. Podobna logika velja za analizo prekinitve anketiranja.

S problemom manjkajočih vrednosti pa smo soočeni, kadar pride do neodgovora na anketna vprašanja, ki jih želimo uporabiti kot pojasnjevalne spremenljivke v statističnih modelih. To se je zgodilo v primeru zgoraj omenjenih anketnih vprašanj, ki smo jih dodali v vprašalnik. Do drugega primera manjkajočih vrednosti pa je prišlo, ker nismo uspeli pridobiti informacije o spolu, starosti, izobrazbi in izkušnjah z anketiranjem za vsakega od 36 anketarjev, ki so sodelovali v anketi GGP.

Ad-hoc metode za ravnanje z manjkajočimi vrednostmi kot npr. brisanje enot z manjkajočimi vrednostmi ali vstavljanje povprečja so zelo problematične, saj dajejo nepristranske cenilke z nepristranskimi standardnimi napakami samo v primeru, da držijo zelo stroge predpostavke (Schafer 1997). Metode večkratnega vstavljanja predstavljajo statistično bolj vzdržno alternativo ad-hoc postopkom. Namesto, da bi manjkajočo vrednost vstavili enkrat, jo vstavimo m-krat pri čemer je m ponavadi nizko število kot npr. 5 ali 10. Večkratno vstavljanje je postopek, ki temelji na statističnem modelu: vstavljene vrednosti so oblikovane na osnovi razpoložljivih kovariat. Variabilnost med m vstavljenimi vrednostmi odseva negotovost glede prave vrednosti (Little and Rubin 2002).

Za vstavljanje manjkajočih vrednosti smo uporabili postopek večkratnega vstavljanja z verižnimi enačbami (angl. *Multiple Imputation by Chained Equations*, van Buuren 2012) in vsako manjkajočo vrednost vstavili po petkrat. S tem postopkom smo dobili pet podatkovnih datotek, na katerih smo opravili analize, nato pa smo rezultate vsake od petih analiz ponovno združili v točkovne ocene s pripadajočimi standardnimi napakami.

Statistični modeli za neodgovor na anketno vprašanje ter prekinitev anketiranja

Za analizo neodgovora na anketno vprašanje smo uporabili *posplošene linearne mešane modele* (angl. *generalized linear mixed models*). Raba teh modelov na področju *teorije odgovora na postavko* (angl. *item response theory*) je analogna naši aplikaciji na problem neodgovora na anketno vprašanje. Za teorijo odgovora na postavko je značilna predpostavka, da nemerljive lastnosti oseb ter postavk določajo izid. Bolj konkretno: statistični model predpostavlja, da razlika med *sposobnostjo* osebe in *zahtevnostjo* testnega vprašanja (pri testu znanja) določa verjetnost, da bo dana oseba na zadano vprašanje odgovorila pravilno.

Kot je razvidno iz zgornjega opisa, je pri rabi teorije odgovora na postavko navadno cilj *opisna meritev* (angl. *descriptive measurement*, De Boeck and Wilson 2004b) na eni ali več latentnih dimenzijah. Posameznikom in postavkam torej skušamo pripisati številsko vrednost na teoretičnih konstruktih. Pojasnjevanje teh vrednosti se izvede šele kot drugi korak ali pa sploh ne. Kot nasprotje usmeritve, ki jo imenujeta opisna meritev, De Boeck and Wilson (2004b) postavita pojasnjevalno analizo (angl. *explanatory analysis*), katere cilj je pojasniti odvisno spremenljivko s pojasnjevalnimi spremenljivkami, ki so na voljo. Avtorja trdita, da je obe filozofski usmeritvi navkljub navideznemu navzkrižju moč kombinirati v usmeritvi, ki jo imenujeta pojasnjevalna meritev (angl. *explanatory measurement*). Raba posplošenih linearnih mešanih modelov je primerna tako pod eno kot pod drugo filozofsko usmeritvijo, kot tudi pod njuno kombinacijo (De Boeck and Wilson 2004b).

Kot smo omenili, je naša aplikacija posplošenih linearnih mešanih modelov na problem neodgovora na anketno vprašanje analogna njihovi rabi v teoriji odgovora na postavko: predpostavljamo, da je verjetnost neodgovora na anketno vprašanje določena z razliko med anketirančevo motivacijo ter bremenom zastavljenega anketnega vprašanja. Za razliko od modelov v teoriji odgovora na postavko, kjer je cilj najpogosteje zgolj opisna meritev lastnosti oseb in postavk, pa v naših modelih nastopajo tudi pojasnjevalne spremenljivke, saj nas zanima predvsem, kako je verjetnost neodgovora na anketno vprašanje povezana z lastnostmi anketnih vprašanj, anketirancev ter anketarjev.

Podatki, zbrani z anketo GGP, imajo posebno strukturo gnezdenja, ki jo imenujemo *delno gnezdenje* (angl. *partial nesting*). Gre za to, da so anketiranci pri osebnem in telefonskem anketiranju gnezdeni v anketarjih, spletni anketiranci pa ne, saj so le-ti vprašalnik izpolnili brez pomoči anketarja. Korendijk et al. (2008) so pokazali, da lahko delno gnezdene podatke modeliramo tako, da posameznike, ki niso gnezdeni, obravnavamo kot da so gnezdeni v skupinah z velikostjo ena. V našem primeru to pomeni, da spletne anketirance obravnavamo kot da je vsakega anketiral drug anketar. S tem vsakemu anketirancu pripišemo pripadnost "anketarju," zaradi česar lahko uporabimo navadni večnivojski model. Korendijk et al. (2008) so v svoji simulacijski študiji pokazali, da ta pristop vodi k nepristranskim ocenam fiksnih učinkov (angl. *fixed effects*) ter njihovih standardnih napak. Ko v nadaljevanju interpretiramo učinke lastnosti anketirancev in anketnih vprašanj na neodgovor na anketno vprašanje, se nanašamo na ocene iz modela, kjer smo uporabili pravkar opisani pristop k obravnavanju spletnih anketirancev.

Za analizo prekinitve anketiranja smo uporabili metode analize preživetja, kot že drugi avtorji pred nami (Galesic 2006; Matzat et al. 2009; Peytchev 2009). Metode analize preživetja so namenjene analizi časa, do katerega se zgodi določen *dogodek*, ter omogočajo upoštevanje krnjenih enot (angl. *censored units*) v analizi. V kliničnih raziskavah do krnjenja navadno pride, ker pacient ob koncu raziskave še ni izkusil dogodka (npr. še ni umrl). V naši aplikaciji kot dogodek definiramo prekinitev anketiranja, pri čemer nas zanima *število vprašanj* (in ne npr. čas od začetka intervjuja do dogodka, merjen v minutah), na katera je bil anketiranec pripravljen odgovoriti preden je prekinil anketiranje. Vse anketirance, ki so do konca izpolnili anketo, obravnavamo kot krnjene. Pri tem ne gre za tipičen primer krnjenja, saj se krnjenje ne more zgoditi ob kateremkoli času. Da je lahko dosegel konec vprašalnika GGP, je moral vsak anketiranec namreč odgovoriti na vsaj 200 anketnih vprašanj, torej krnjenje pod to pražno vrednostjo ni mogoče.

Najpogosteje uporabljeni regresijski model v analizi preživetja je Coxov model sorazmernih ogroženosti (angl. *proportional hazards model*). Le-ta nam omogoča, da ocenimo, kako pojasnjevalne spremenljivke vplivajo na tveganje za nastop dogodka,

v našem primeru prekinitve anketiranja. *Razširjeni* Coxov model pa nam omogoča, da poleg časovno neodvisnih učinkov (lastnosti anketiranca) kot pojasnjevalne spremenljivke v model vključimo tudi časovno odvisne spremenljivke (lastnosti anketnih vprašanj). Pri rabi Coxovega modela je ključno, da je zadovoljena predpostavka sorazmernih ogroženosti. Zadovoljenost le-te smo za vsak Coxov model preverili s statističnim testom. Kadar se je izkazalo, da je predpostavka sorazmernih ogroženosti za določeno pojasnjevalno spremenljivko kršena, smo postopali tako, da smo z grafično metodo določili prelomno točko in s tem celotno časovno obdobje razdelili na dva intervala, na katerih je bilo omenjeni predpostavki zadoščeno.

Rezultati

Rezultati analiz potrjujejo, da sta tako neodgovor na anketno vprašanje kot prekinitev anketiranja bolj pogosta v spletni verziji vprašalnika kot pri osebnem in telefonskem anketiranju. Tak rezultat je bil pričakovan in je v skladu s teoretično podlago, ki predpostavlja, da je zadrževalni prag za obe obliki neodgovora višji pri načinih anketiranja, kjer je prisoten anketar.

Raziskave neodgovora na anketno vprašanje ter prekinitve anketiranja navadno kot pojasnjevalne spremenljivke vključujejo tudi anketirančeve demografske lastnosti. V literaturi je pogosta uporaba anketirančeve starosti in izobrazbe kot *proxy* mer anketirančeve kognitivne sposobnosti. Anketiranci z višjimi kognitivnimi sposobnosti naj bi bili tako manj obremenjeni, ko odgovarjajo na anketna vprašanja, in zaradi tega pri njih pričakujemo manj prekinitev ter neodgovora na anketno vprašanje. Rezultati naših analiz so v skladu z opisano logiko, kar se tiče izobrazbe anketiranca: pri bolj izobraženih anketirancih opažamo manj neodgovora na anketno vprašanje ter nižje tveganje za prekinitev anketiranja. Kar se tiče anketirančeve starosti, pa se rezultati za neodgovor in prekinitev razlikujejo. Višja starost je res povezana s pogostejšimi neodgovori na anketno vprašanje. V nasprotju s pričakovanji pa je višja starost povezana z *nižjim* tveganjem za prekinitev. Podrobnejša analiza pokaže, da so mladi anketiranci anketo prekinjali že kmalu po začetku anketiranja. Po približno sto anketnih vprašanjih pa anketirančeva starost nima več učinka na tveganje za prekinitev.

Kot smo omenili, smo v vprašalnik dodali tri vprašanja, s katerimi smo merili anketirančevo splošno naravnanost do anket. Pričakovali smo, da bodo anketiranci, ki so bolj pozitivno naravnani do anket, bolj skrbno izvajali vsako izmed faz procesa odgovarjanja na anketna vprašanja (glej Stocke 2006). Rezultati analiz potrjujejo, da je anketirančeva pozitivna naravnanost do anket povezana z manj neodgovora na anketno vprašanje ter nižjim tveganjem za prekinitev anketiranja.

Kot smo omenili, so trije neodvisni eksperti vsako anketno vprašanje v vprašalniku GGP kodirali na treh merah: vsiljivost teme vprašanja, nevarnost razkritja (občutljivih informacij) ter potencial za pretirano pozitivno predstavitev. Rezultati analiz kažejo, da so anketna vprašanja, ki predstavljajo nevarnost razkritja, povezana z višjo mero neodgovora, vendar pa je omenjeni učinek statistično značilen samo pri spletnem anketiranju. Nevarnost razkritja občutljivih informacij je povezana tudi z višjim tveganjem za prekinitev anketiranja, vendar pa ta učinek ni bil značilen na začetku (približno prvih sto vprašanj) vprašalnika GGP.

Če anketiranec preskoči, zavrne odgovor, ali prekine anketo pri vprašanju, ki omogoča pretirano pozitivno predstavitev (npr. pomoč prijateljem pri skrbi za otroke), je to moč razumeti, kot da anketiranec implicitno priznava, da se ni vedel na družbeno zaželen način (Bradburn et al. 1978). Rezultati analiz so v skladu z opisano logiko: potencial za pretirano pozitivno predstavitev je povezan z manj neodgovora na anketno vprašanje in nižjim tveganjem za prekinitev anketiranja. V osebnem in telefonskem načinu anketiranja je opisani učinek še močnejši (interakcija z načinom anketiranja je statistično značilna): pri anketnih vprašanjih, ki omogočajo pretirano pozitivno predstavitev, je manj neodgovora, če je prisoten anketar (v primerjavi s spletnim samo-anketiranjem).

Rezultati pričakovano kažejo, da je vsiljivost teme anketnega vprašanja povezana z višjo verjetnostjo neodgovora. Vpliv vsiljivosti teme na tveganje za prekinitev pa se izkaže za bolj kompleksnega kot smo pričakovali. Rezultati kažejo, da bolj vsiljivo anketno vprašanje zniža tveganje za prekinitev, vendar obenem poviša tveganje za prekinitev dve anketni vprašanji naprej. Možna razlaga za omenjeni rezultat je, da anketiranci ne želijo razkriti informacije, da jim je določena tema res vsiljiva in zato

pri anketnem vprašanju, ki zadeva takšno temo, ne prekinejo anketiranja. Prekinitve pri vsiljivih vprašanjih se vzdržijo in se za prekinitev odločijo raje kmalu po tem, ko jim je bilo postavljeno vsiljivo vprašanje.

V statističnih modelih smo kot pojasnjevalne spremenljivke vključili tudi objektivne lastnosti anketnih vprašanj kot je število ponujenih odgovorov in dolžina vprašanja (število besed). Rezultati kažejo, da ima dolžina vprašanja statistično značilen vpliv na neodgovor zgolj pri spletnem anketiranju: anketiranci so na spletu zagrešili več neodgovora pri dolgih vprašanjih. Vpliv dolžine vprašanja v Coxovem modelu za prekinitev anketiranja ni bil statistično značilen. Rezultati kažejo, da so odprta vprašanja ter vprašanja z velikim številom ponujenih odgovorov povezana z več neodgovora in višjim tveganjem za prekinitev anketiranja. V nasprotju s pričakovanji pa je ugotovitev, da je ta učinek šibkejši pri spletnem anketiranju: anketiranci na spletu so pri takšnih vprašanjih zagrešili *manj* neodgovora v primerjavi z osebami, ki so bile anketirane osebno ali po telefonu (statistično značilna interakcija z načinom anketiranja).

Da bi v model za neodgovor lahko kot pojasnjevalne spremenljivke vključili tudi lastnosti anketarja, smo analizo ponovili na podatkih brez spletnih anketirancev. Z vključitvijo anketarjevega spola, starosti, izobrazbe ter izkušenj z anketiranjem smo pojasnili približno tretjino variabilnosti na nivoju anketarja. Ta ugotovitev je pomembna, saj so avtorji, ki so raziskovali učinek anketarja na neodgovor na anketno vprašanje, našli razlike med anketarji, vendar pa teh razlik niso uspeli pojasniti z vključitvijo anketarjevih lastnosti v statistični model (Pickery and Loosveldt 1998, 2001). V naših analizah so učinki anketarjevih lastnosti (z izjemo izobrazbe) statistično neznačilni, kar pripisujemo majhni velikosti vzorca na nivoju anketarja.

V Coxov model za prekinitev odgovora smo kot pojasnjevalni spremenljivki vključili še indikatorja za to, 1) ali je bil odgovor na dano vprašanje *obvezen* (program za anketiranje ni dovolil, da se vprašanje preskoči) ter 2) ali je dano anketno vprašanje vpeljalo novo temo. Skladno s pričakovanji se izkaže, da je tveganje za prekinitev v obeh primerih višje. V model za prekinitev smo kot pojasnjevalno spremenljivko vključili tudi mero pogostosti neodgovora na nedavna anketna vprašanja. V skladu s pričakovanji se tudi ta učinek izkaže za pozitiven in statistično značilen: kadar anketiranec pogosto

izpušča odgovor na anketna vprašanja, je tveganje za prekinitev višje. S tem smo replicirali rezultate Galesic (2006).

Omejitve

Pričujoča raziskava ima določene omejitve, ki se jih je potrebno zavedati, ko govorimo o njenih rezultatih. Kot smo omenili, smo izračunali nizko mero strinjanja med ocenjevalci, še zlasti za ocene nevarnosti razkritja občutljivih informacij. Tako nizka mera zanesljivosti najverjetneje odseva visoko obremenitev ocenjevalcev, ki so morali na treh merah občutljivosti oceniti vsakega izmed približno petsto anketnih vprašanj vprašalnika GGP. Menimo, da bi lahko višje strinjanje med ocenjevalci dosegli, če bi jim dali v ocenjevanje manjše število anketnih vprašanj in bi tako lahko občutljivost vsakega pretehtali bolj skrbno.

Teoretični model, ki predstavlja podlago uporabljenim statističnim modelom, ima prav tako svoje omejitve: faktorje vpliva na anketno vprašanje in prekinitev anketiranja smo zreducirali na anketirančevo motivacijo ter breme, ki ga predstavlja anketno vprašanje. Takšen teoretični model zapostavlja faktorje vpliva, ki so morda relevantni za določene načine anketiranja. Osebo, ki izpolnjuje spletni vprašalnik, lahko npr. v vsakem trenutku zmotijo pravkar prispela e-poštna sporočila ali pogovor z drugimi osebami v sobi.

Postopki, s katerimi smo zbirali podatke s pilotsko anketo GGP, ne dovoljujejo, da bi rezultate statističnih analiz posploševali na populacijo. Posploševanje v strogem statističnem smislu zahteva, da opredelimo populacijo in iz nje izberemo slučajni vzorec z ne-ničelno verjetnostjo izbora vsake enote v populaciji. V prvi fazi zbiranja podatkov smo uporabili člane spletnega panela brez opredelitve populacije in vzorčenja. V drugi fazi zbiranja podatkov smo sicer definirali populacijo ter izvedli slučajno vzorčenje, vendar pa je možnost prestopanja med načini anketiranja vzorčni načrt močno zapletla. V tretji fazi zbiranja podatkov smo anketirance rekrutirali z oglasi na Facebooku, ponovno brez slučajnega vzorčenja.

Večina ugotovitev raziskave je kljub temu skladna s pričakovanji, ki sledijo iz teo-

rije, zaradi česar menimo, da bi do podobnih rezultatov prišli tudi, če bi raziskavo ponovili npr. na anketirancih kakšne druge evropske države s stopnjo internetne penetracije podobno kot v Sloveniji. Večje zadržke pa imamo glede možnosti posploševanja rezultatov, ki temeljijo na podatkih, zbranih z anketo na Facebooku. Mednarodne akademske ankete kot anketa GGP navadno ne rekrutirajo anketirancev z neformalnimi načini kot so oglasi na spletnih straneh. Menimo, da so anketiranci, rekrutirani na Facebooku, imeli bolj ravnodušen pristop k odgovarjanju na anketo ter da se to odseva npr. v višji stopnji neodgovora. Rezultati, ki temeljijo na podatkih, zbranih z anketo na Facebooku, so tako morda bolj relevantni za spletne in komercialne ankete.

Izvirni prispevek

Pričujoča doktorska disertacija predstavlja naslednje izvirne prispevke k razvoju področja anketne metodologije. Predmet študije sta dve obliki neodgovora, ki doslej nista bili raziskovani tako obsežno kot neodgovor enote, zato empirična študija razširja obstoječe znanje o faktorjih vpliva na neodgovor na anketno vprašanje ter prekinitev anketiranja. Razumevanje faktorjev vpliva, ki izhaja iz rezultatov študije, se lahko uporabi 1) za preprečevanje neodgovora na anketno vprašanje in prekinitve anketiranja (npr. s prilagoditvijo vprašalnika) ali 2) za izboljšanje postopkov, ki omogočajo analizo podatkov v prisotnosti manjkajočih vrednosti (npr. večkratno vstavljanje manjkajočih vrednosti).

Rezultati analiz kažejo, da na neodgovor na anketno vprašanje hkrati vplivajo lastnosti anketnega vprašanja, anketiranca in anketarja. Pokazali smo, da predstavljeni pristop k analizi omogoča sočasno upoštevanje vplivov lastnosti na vsakem od omenjenih nivojev ter privede do vsebinsko smiselnih rezultatov. Isti vprašalnik smo izvedli v treh načinih anketiranja, kar nam je dalo izjemno priložnost, da proučujemo, kako se vpliv lastnosti anketiranca in anketnega vprašanja na neodgovor na anketno vprašanje razlikuje med načini anketiranja. Kolikor nam je znano, ni pred našo nobena raziskava hkrati obravnavala vpliva lastnosti anketnega vprašanja, anketiranca in anketarja na neodgovor na anketno vprašanje v treh različnih načinih anketiranja.

Za analizo prekinitve anketiranja smo uporabili metode analize preživetja, s katerimi smo lahko sočasno obravnavali tako vpliv lastnosti anketiranca kot tudi vpliv lastnosti anketnih vprašanj na tveganje za prekinitev anketiranja. Tako podrobne študije prekinitve anketiranja so redke (Galesic 2006; Peytchev 2009; Matzat et al. 2009). Ne poznamo drugih raziskav, ki bi za Coxov model za prekinitev anketiranja preverile predpostavko sorazmernih ogroženosti in v primeru kršitev prilagodile model. Kolikor nam je znano, je naša raziskava prva, ki je v modelu za prekinitev anketiranja vključila mero predhodnih neodgovorov na anketna vprašanja ter pokazala statistično značilen vpliv le-teh.