WP1 extension along lines of Neuman-Weiss 1995

Suhas D. Parandekar, Ekaterina Melianova, Artëm Volgin

November 16, 2019

```
> library(dplyr)
> library(sqldf)
> library(XLConnectJars)
> library(questionr)
> library(labelled)
> library(tidyr)
> library(magrittr)
> library(ggplot2)
> library(data.table)
> library(pbapply)
> library(gridExtra)
> library(psych)
> library(stringi)
> library(sjPlot)
> library(sjmisc)

> # wd
> wd <- "C:/Country/Russia/Data/SEASHELL/SEABYTE/edreru/wp1"
> setwd(wd)
> # Data
> df_mincer <- readRDS("df_mincer.rds")
> ########## Aggregating occupations by 2 digits
>
> # First, aggregating military men
> df_mincer$occup <- ifelse(df_mincer$occup == 0, 110, df_mincer$occup)
> # For simplicity those with one digit go to the first category with 2
> # digits (e.g., 1 go to 11)
> df_mincer$occup <- as.numeric(ifelse(df_mincer$occup<10,
+                                     paste0(df_mincer$occup, "1"),
+                                     df_mincer$occup))
> # Leaving only 2 digits
> df_mincer$occup2d <- as.numeric(substr(df_mincer$occup, 1, 2))
> # If N obs is < 30 we technically cannot run a regression
> # Let's aggregate such categories with the respective closest category
> # I detected them manually for the easiness of computations based on this table:
> table(df_mincer$occup2d, df_mincer$YEAR)
> # Aggregating
> df_mincer$occup2d[df_mincer$occup2d == 11|df_mincer$occup2d == 12] <- 112
> df_mincer$occup2d[df_mincer$occup2d == 24|df_mincer$occup2d == 25] <- 245
> df_mincer$occup2d[df_mincer$occup2d == 34|df_mincer$occup2d == 35] <- 345
> df_mincer$occup2d[df_mincer$occup2d == 41|df_mincer$occup2d == 42] <- 412
> df_mincer$occup2d[df_mincer$occup2d == 43|df_mincer$occup2d == 44] <- 434
> df_mincer$occup2d[df_mincer$occup2d == 53|df_mincer$occup2d == 54] <- 534
> df_mincer$occup2d[df_mincer$occup2d == 73|df_mincer$occup2d == 74|
+                   df_mincer$occup2d == 75] <- 7345
> df_mincer$occup2d[df_mincer$occup2d == 81|df_mincer$occup2d == 82] <- 812
> df_mincer$occup2d[df_mincer$occup2d == 92|df_mincer$occup2d == 93|
+                   df_mincer$occup2d == 94|df_mincer$occup2d == 95|
+                   df_mincer$occup2d == 96] <- 923456
> df_mincer$occup2d[df_mincer$occup2d == 61|df_mincer$occup2d == 62] <- 71
> table(df_mincer$occup2d, df_mincer$YEAR)
> # 345 category is too small even within its digit so
> # we need to merge it with another digit
> df_mincer$occup2d[df_mincer$occup2d == 345] <- 412
> # Checking
> tbl <- as.data.frame(table(df_mincer$occup2d, df_mincer$YEAR))
> tiny <- as.numeric(as.character(
+   unique(tbl$Var1[tbl$Freq<30]))); tiny # no categaries with < 30 obs
> # Creating a dummy set for occupations
> dummy_set <- dummy.code(df_mincer$occup2d)
> colnames(dummy_set) <- paste0("occup", colnames(dummy_set), sep = "")
> df <- cbind(df_mincer, dummy_set)
> # Probit regression: developing a female - non-female typology of occupations
>
> # Empty list where the regression output will be written
> probit <- vector("list", length(unique(df$YEAR)))
> for (i in seq(length(probit))){
+   probit[[i]] <- vector("list", length(unique(colnames(dummy_set))))
+ }
> seq_year <- unique(df$YEAR)
> # Looping over each year and occupation
> # takes ~15 sec
> for(i in seq(length(seq_year))){
+   for(j in seq(length(colnames(dummy_set)))){
+     probit[[i]][[j]] <- glm(as.formula(paste0(colnames(dummy_set)[j], "~",
+                                               "female")),
+                            family = binomial(link = "probit"),
+                            data = df[df$YEAR == seq_year[i],])
+   }
+ }
> # Naming
> names(probit) <- seq_year
> # Computng summary
> smry <-  lapply(probit, function(x) {lapply(x, summary)})
> # A table with coefficients for the female variable
> tbl_fem_ <- c()
> for (y in seq_year){
+   for (n in seq(length(colnames(dummy_set)))){
+    tbl_fem_ <- rbind.data.frame(tbl_fem_, cbind.data.frame(
+     "female" = round(smry[[paste0(y)]][[n]]$coefficients[2,1], 2),
+     "p-value" = round(smry[[paste0(y)]][[n]]$coefficients[2,4], 3))
+     )
+   }
+ }
> YEAR <- rep(seq_year, each = ncol(dummy_set))
> occup <- rep(colnames(dummy_set), length(seq_year))
> tbl_fem <- cbind.data.frame(YEAR, occup, tbl_fem_) # final table
> # Female-dominated occupations
> fem_occup_vec <- as.character(unique(
+   tbl_fem[tbl_fem$female > 0 &
+             tbl_fem$'p-value' < 0.05, "occup"]))
> # Non-female occupations
> nonfem_occup_vec <- as.character(unique(
+   tbl_fem[!(tbl_fem$female > 0 &
+             tbl_fem$'p-value' < 0.05), "occup"]))
> # Occupations which are in both categories depending on a wave
> fem_occup_vec[fem_occup_vec %in% nonfem_occup_vec]
> # Let us examine those cases
> both <- tbl_fem[tbl_fem$occup %in%
+                   fem_occup_vec[fem_occup_vec %in% nonfem_occup_vec],]
> # occup14 is insignificant in the majoriy of waves -> let's put it in
> # nonfem_occup
> # occup245 is significant and positive almost each time -> let's put it in
> # fem_occup
> # occup51 is mostly significant and positive -> let's put it in fem_occup
> # occup91 is mostly significant and positive -> let's put it in fem_occup
>
> # Defining a variable with nonfem_occup
> df$occup2d <- as.character(df$occup2d)
> df$fem_occup <- ifelse(df$occup2d %in%
+                             substr(fem_occup_vec[!fem_occup_vec=="occup14"],
+                                    6, nchar(fem_occup_vec[
+                                      !fem_occup_vec=="occup14"])), 1, 0)
> # Looking at the distribution (looks logical)
> table(df$fem_occup)
> table(df$fem_occup, df$female)
> # Filtering the missings left
> df <- df %>%
+   filter(!is.na(wage) & !is.na(edu_4) & wage > 0)
> # Norming wages (to allow comparison)
> # wages_prop <- rio::import("wages_prop.xlsx") # Rosstat statistics
> df <- df %>%
+   group_by(YEAR) %>%
+   mutate(wageBYmed = wage/median(wage))
```

```
> aggregate(wageBYmed ~ YEAR, df, mean)
>

> # Empty list where the regression output will be written
> lm_dep <- vector("list", length(unique(df$YEAR)))
> seq_year <- unique(df$YEAR)
> df$fem_occup <- factor(df$fem_occup,
+                           levels = c(1,0),
+                           labels = c("Female Occupations",
+                                      "Non-female Occupations"))
> # Looping over each year
> for(i in seq(length(seq_year))){
+   lm_dep[[i]] <- lm(log(wageBYmed) ~ edu_4 +
+                                 exper +
+                                 I(exper^2) +
+                                 fem_occup +
+                                 fem_occup*exper +
+                                 fem_occup*I(exper^2) +
+                                 fem_occup*edu_4 +
+                                 exper*edu_4 +
+                                 I(exper^2)*edu_4 +
+                                 fem_occup*edu_4*exper +
+                                 fem_occup*edu_4*I(exper^2),
+                            data = df[df$YEAR == seq_year[i],])
+ }
> names(lm_dep) <- seq_year
> smry_lm_dep <- lapply(lm_dep, summary)

> ################################### Model prediction
> for (i in 1:length(seq_year)){
+   df_year <- as.data.frame(df[df$YEAR == seq_year[i],])
+   pred_y <- exp(predict(lm_dep[[i]], df_year, interval="conf"))
+   df_year <- cbind(df_year, pred_y)
+
+ # Plot
+  p_int <- ggplot(df_year, aes(x = exper, y = fit, group = edu_4, color = edu_4,
+                  linetype = edu_4)) +
+     geom_line(aes(y = fit), size = 1.2) +
+     geom_ribbon(aes(ymin=lwr, ymax=upr, fill = edu_4), alpha = 0.1,
+                 colour = NA) +
+     facet_grid(~ as.factor(fem_occup)) +
+     theme(legend.title = element_blank(),
+          legend.position = "bottom",
+          panel.grid.minor = element_blank(),
+          axis.text.x = element_text(size = 12),
+          axis.text.y = element_text(size = 12),
+          axis.title = element_text(size = 12),
+          legend.text = element_text(size = 12),
+          legend.key = element_rect(size = 12))  +
+     scale_color_manual(values = c("blue", "red", "darkgreen")) +
+     scale_fill_manual(values=c("blue", "red", "darkgreen")) +
+     scale_linetype_manual(values = c("solid", "longdash", "dotted")) +
+     scale_y_continuous(limits = c(0, 3)) +
+     ylab("Monthly wage normed by median") +
+     xlab("Experience")
+
+  ggsave(paste0("p_", seq_year[i], "_int.png"), height = 4, width = 7.5,
+         units = "in")
+  print(i)
+ }
```

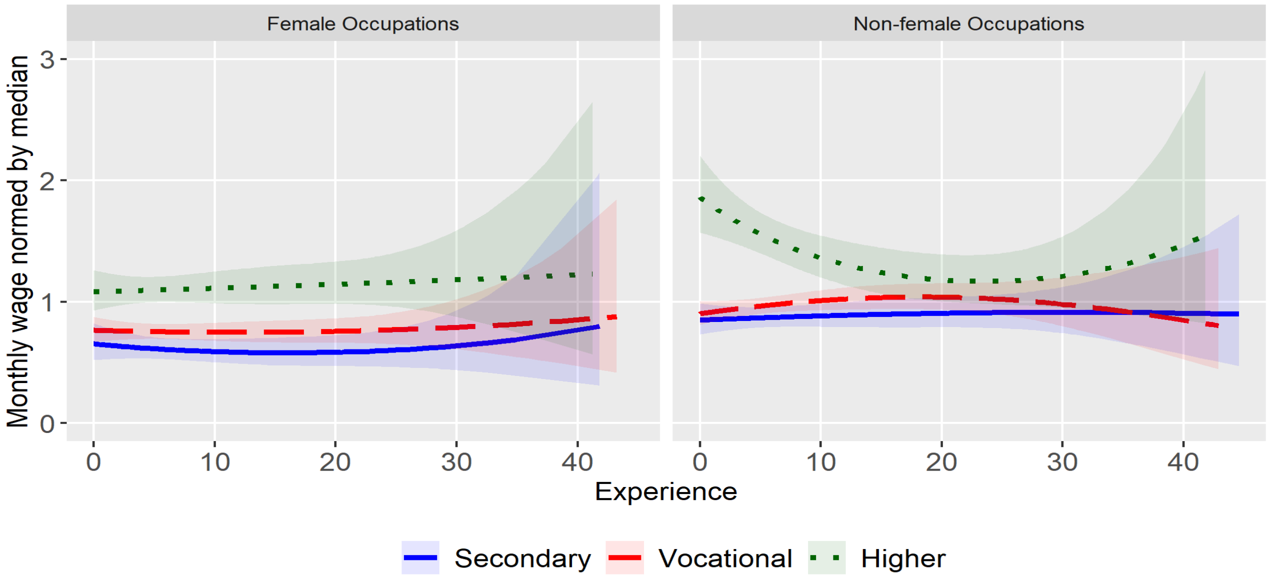

Figure 1: Experience Profiles, RLMS 1994

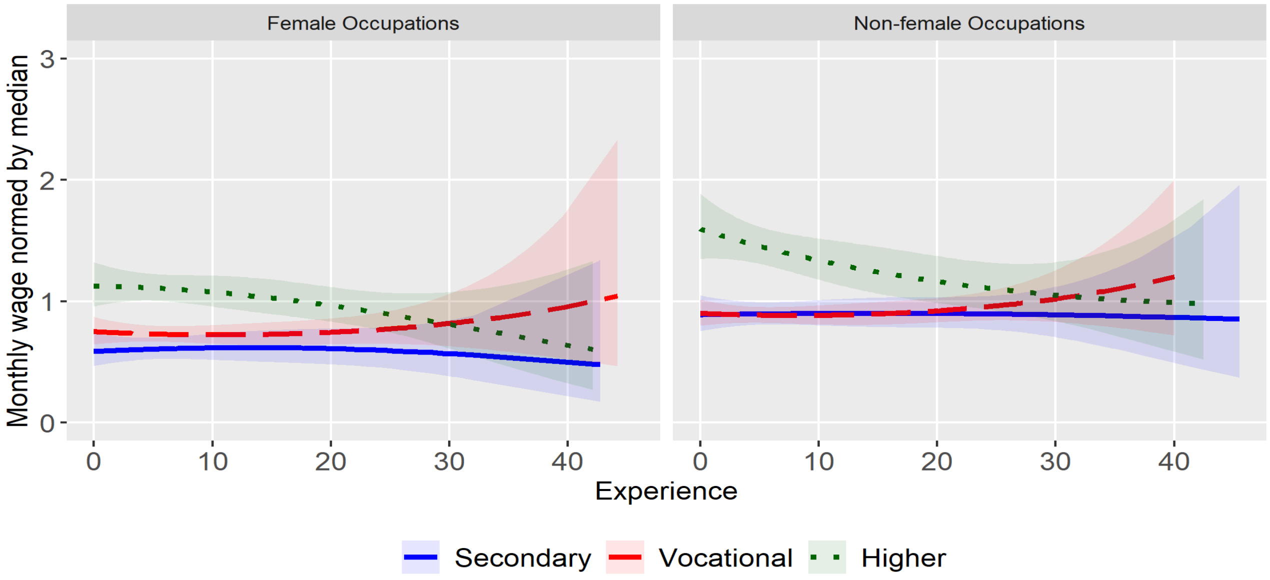

Figure 2: Experience Profiles, RLMS 1995

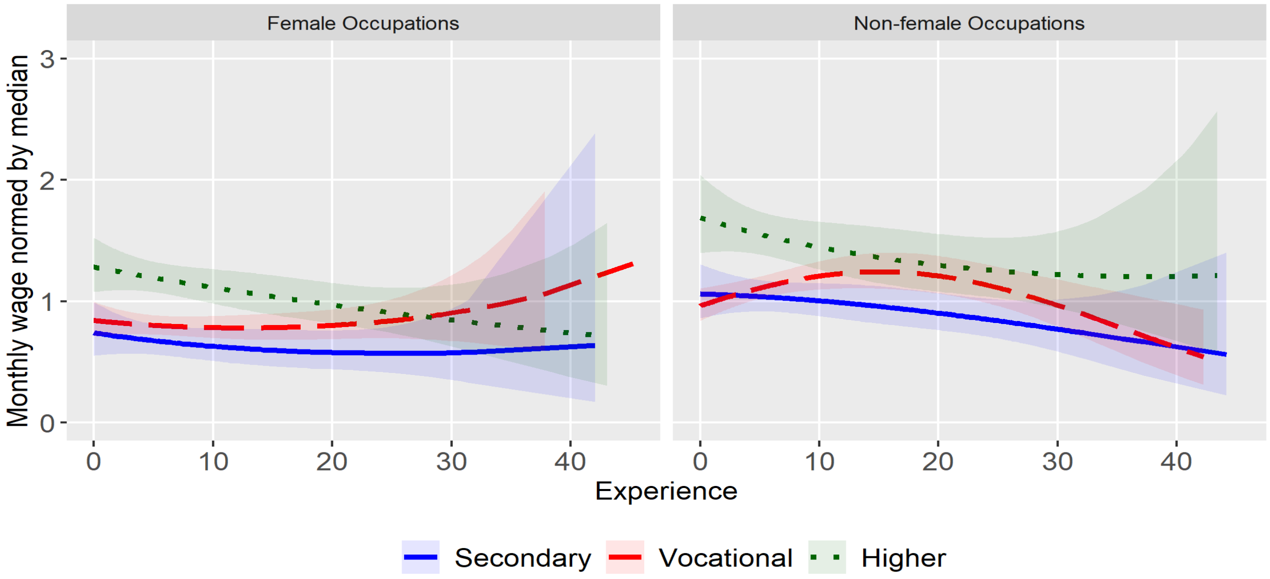

Figure 3: Experience Profiles, RLMS 1996

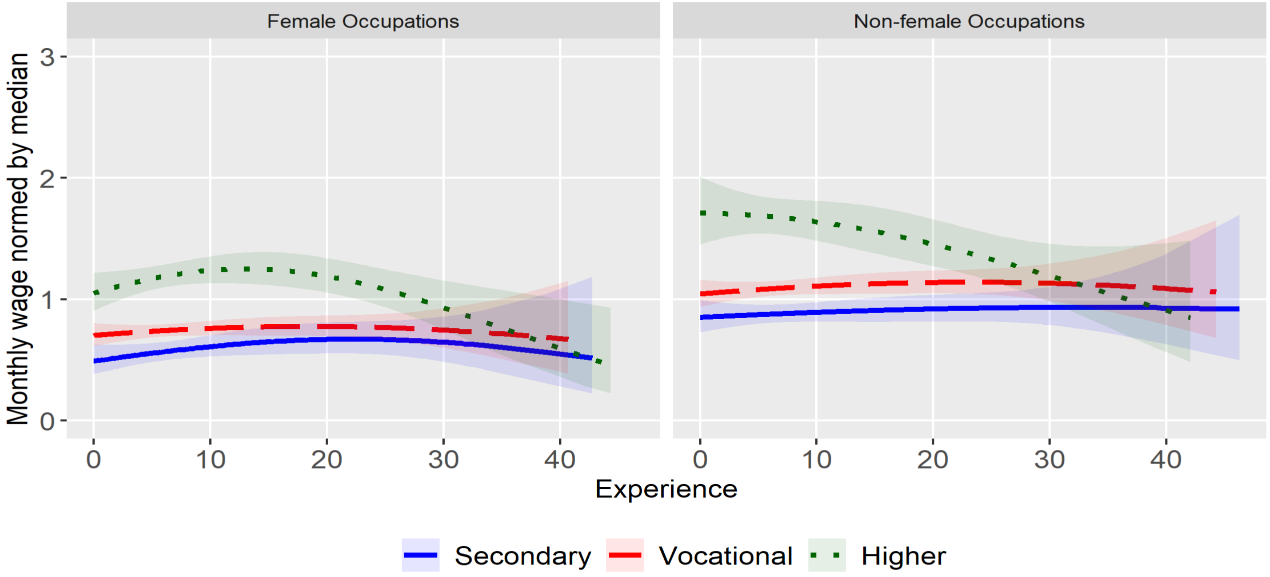

Figure 4: Experience Profiles, RLMS 1998

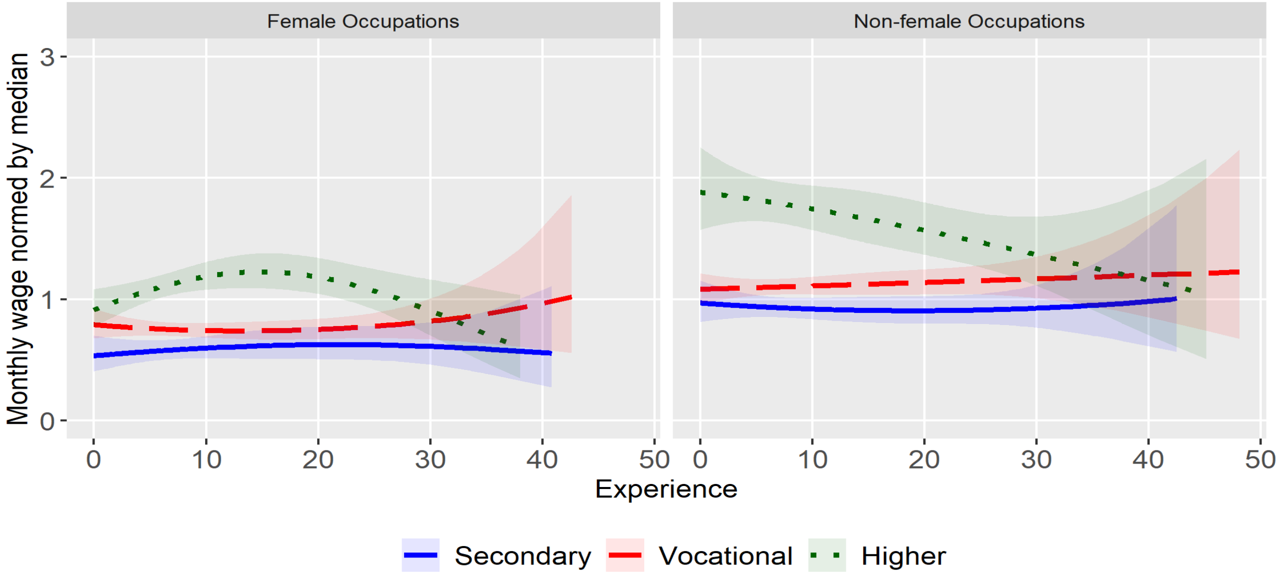

Figure 5: Experience Profiles, RLMS 2000

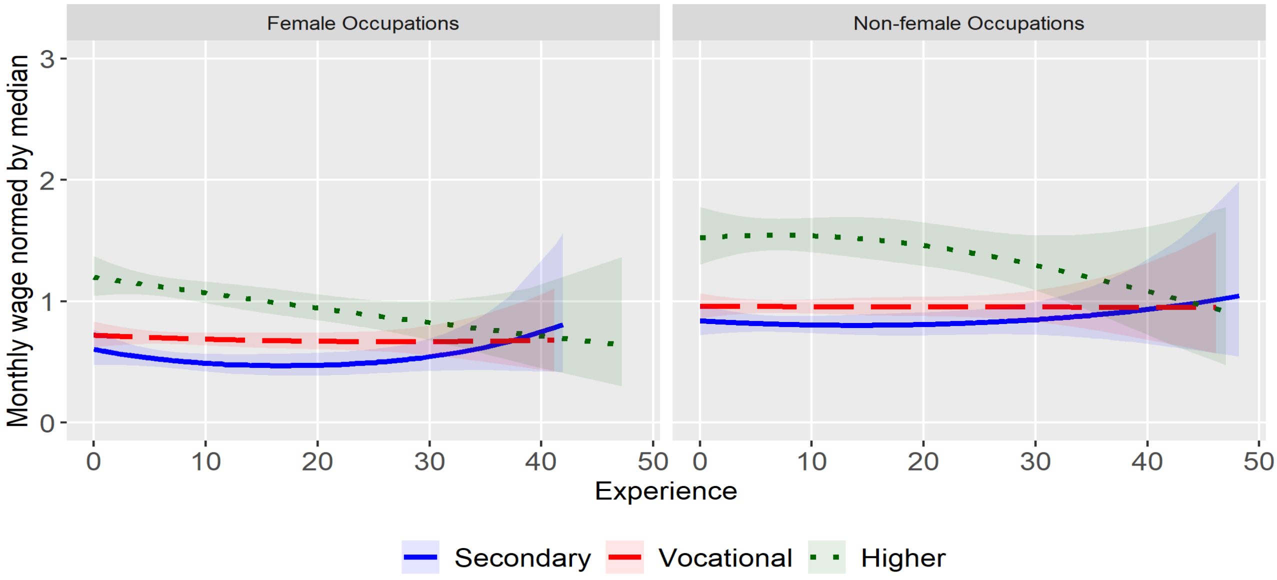

Figure 6: Experience Profiles, RLMS 2001

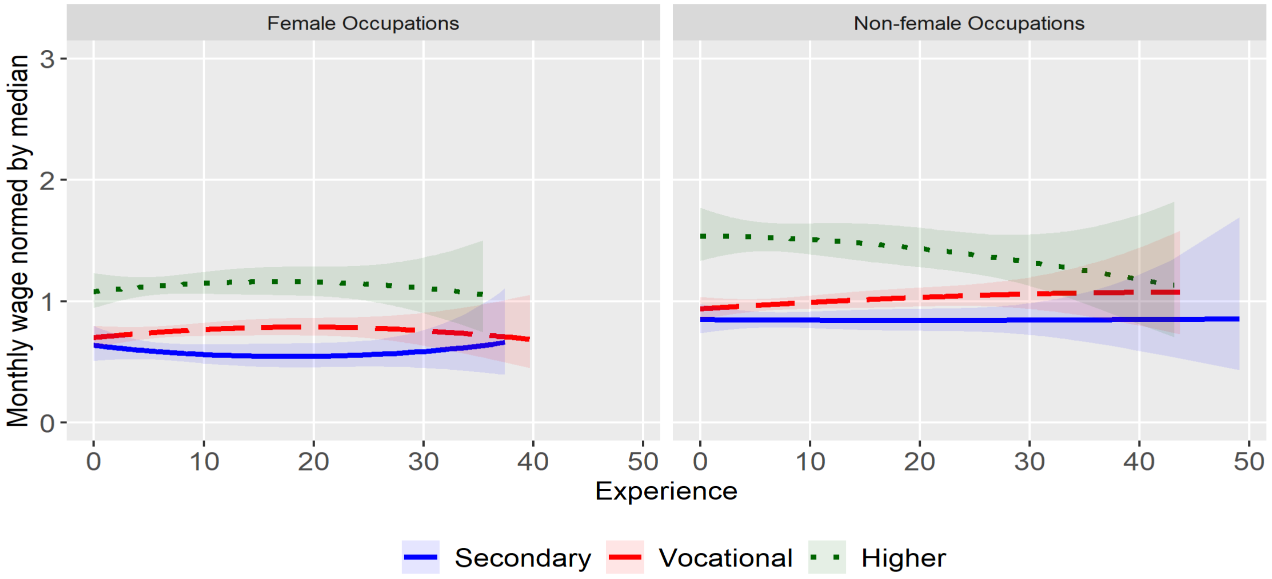

Figure 7: Experience Profiles, RLMS 2002

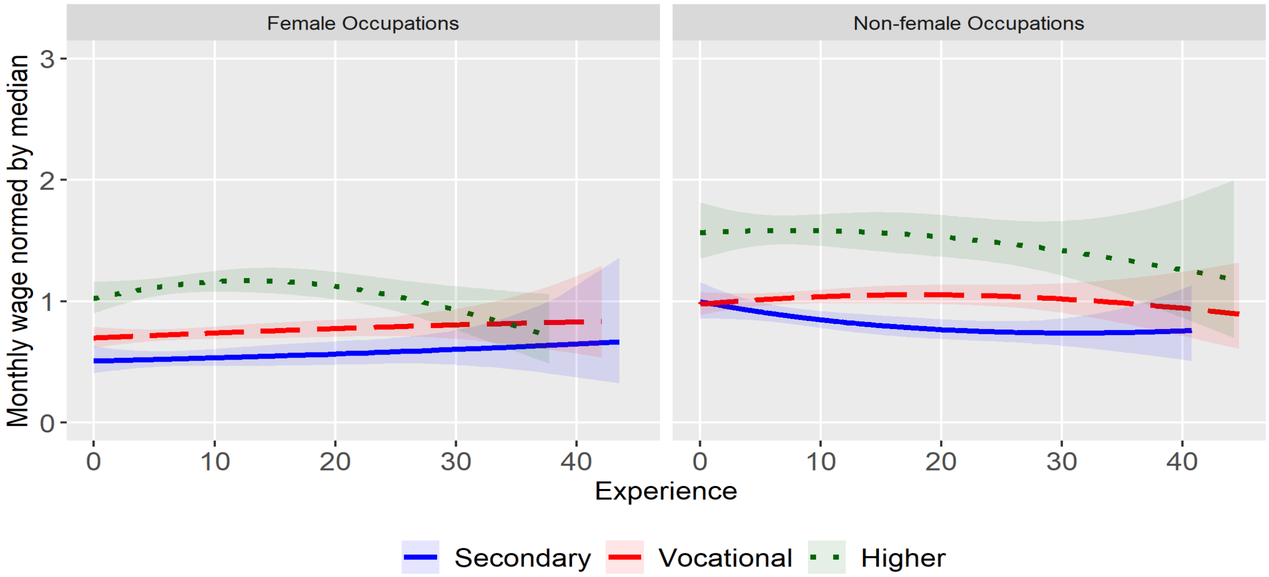

Figure 8: Experience Profiles, RLMS 2003

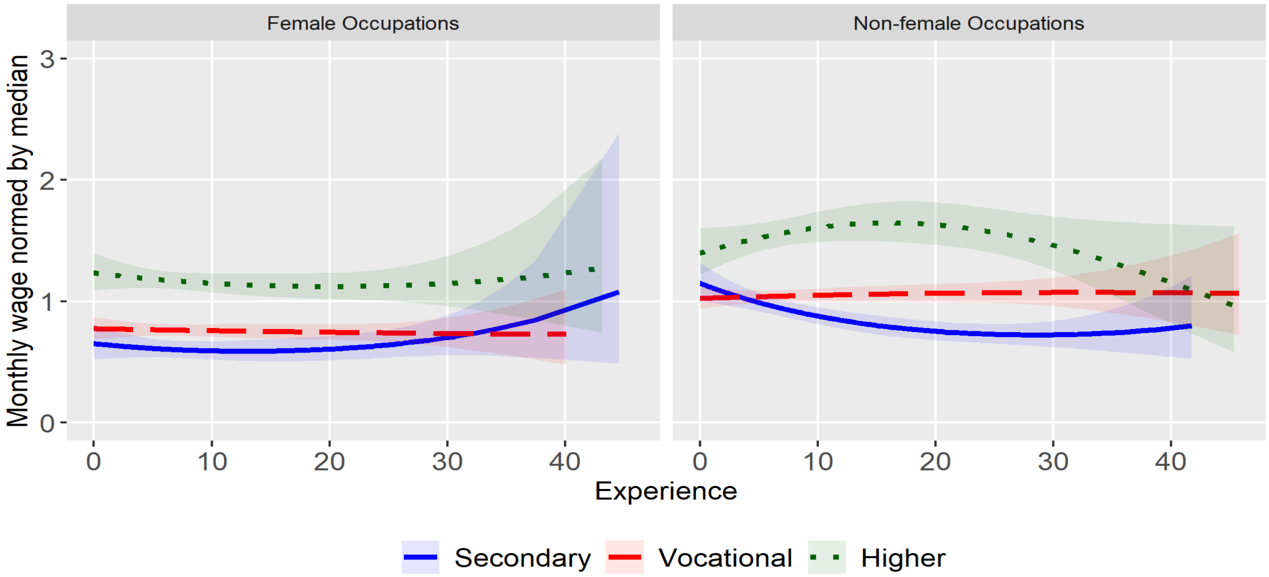

Figure 9: Experience Profiles, RLMS 2004

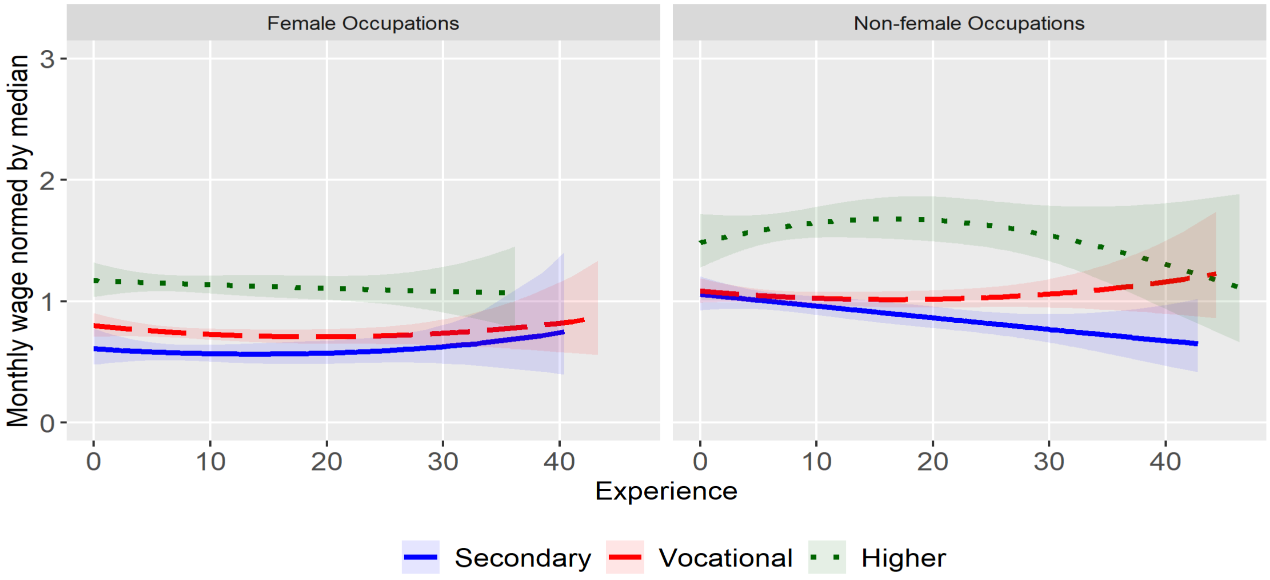

Figure 10: Experience Profiles, RLMS 2005

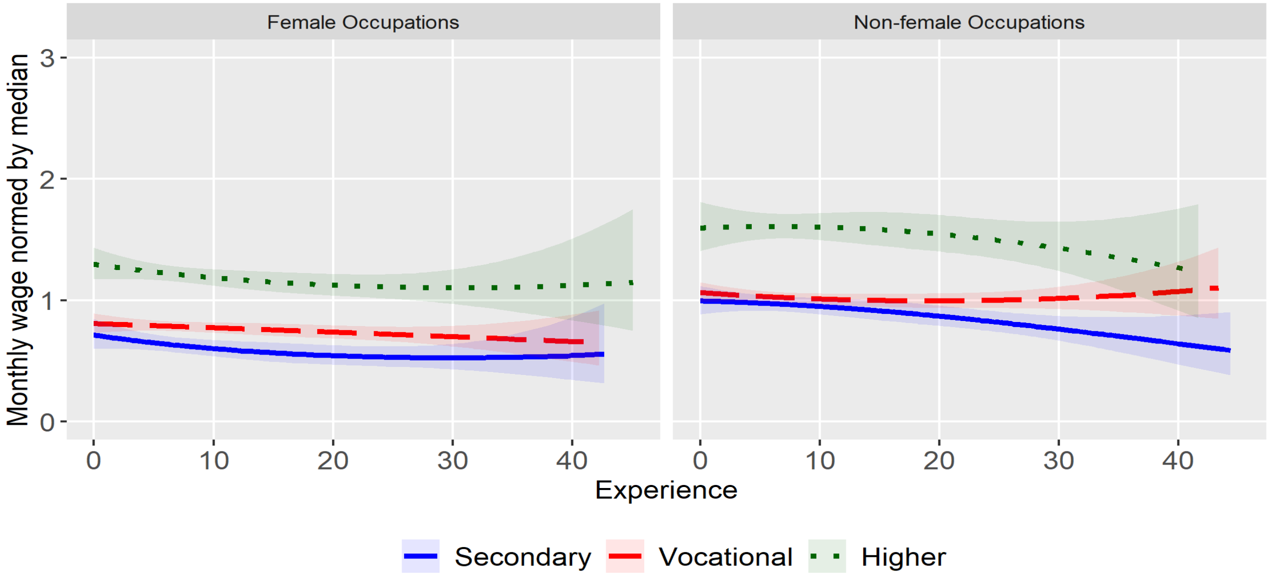

Figure 11: Experience Profiles, RLMS 2006

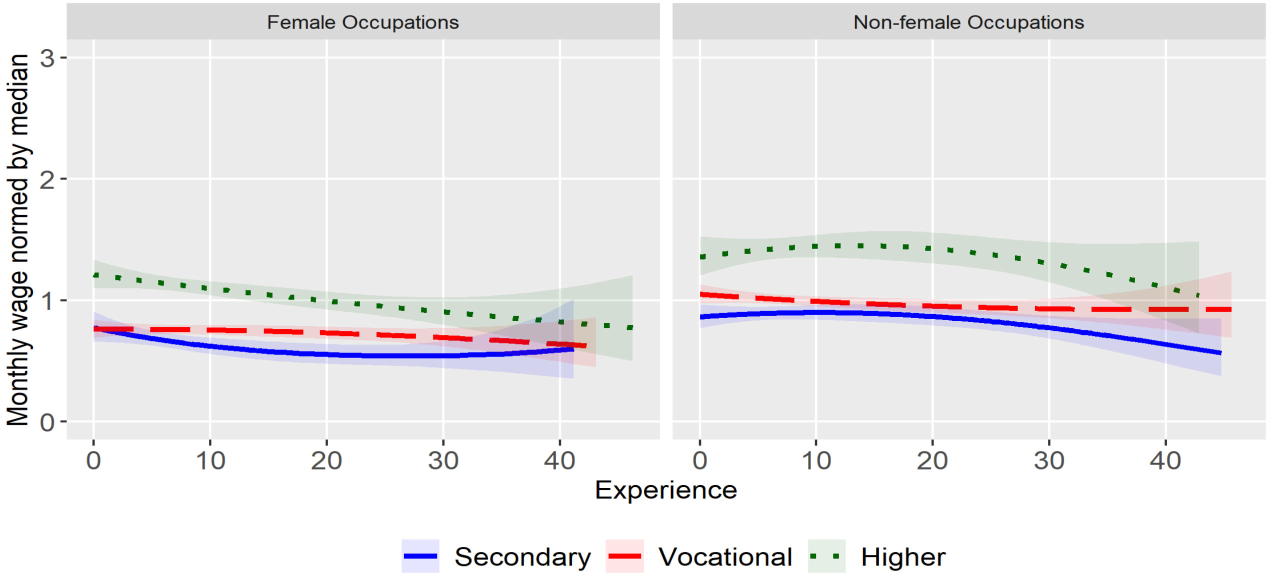

Figure 12: Experience Profiles, RLMS 2007

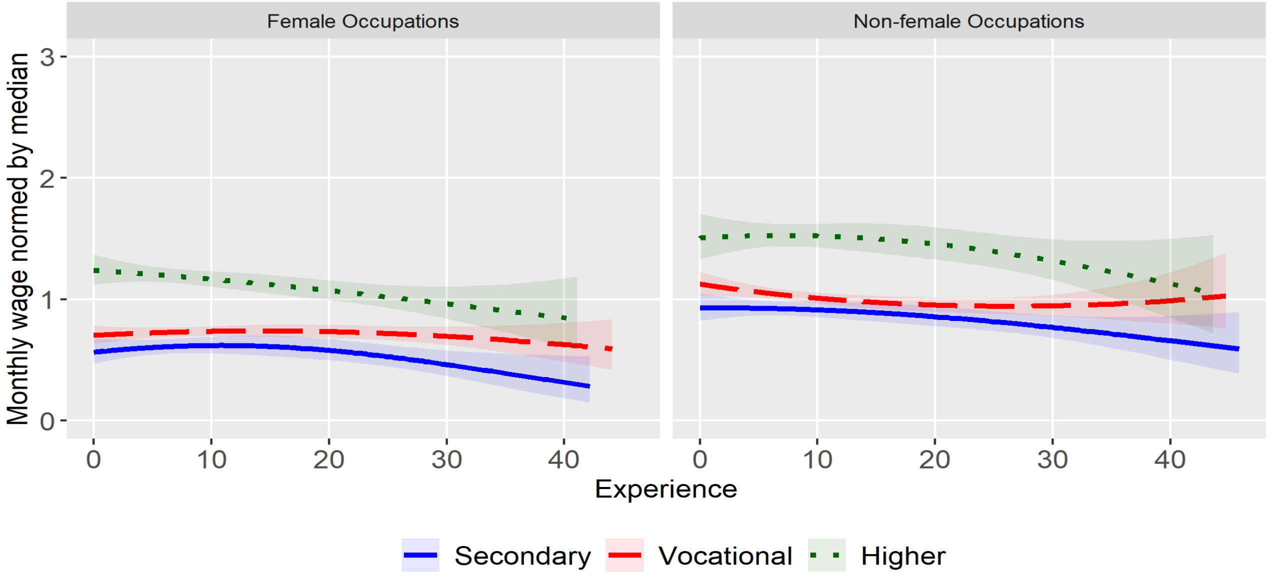

Figure 13: Experience Profiles, RLMS 2008

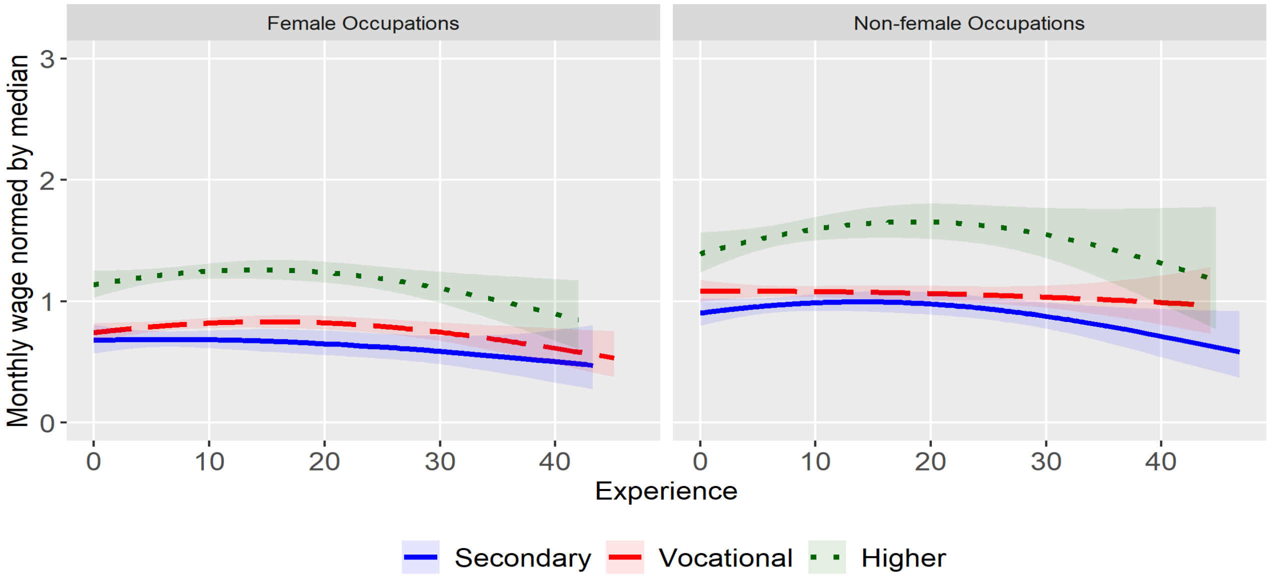

Figure 14: Experience Profiles, RLMS 2009

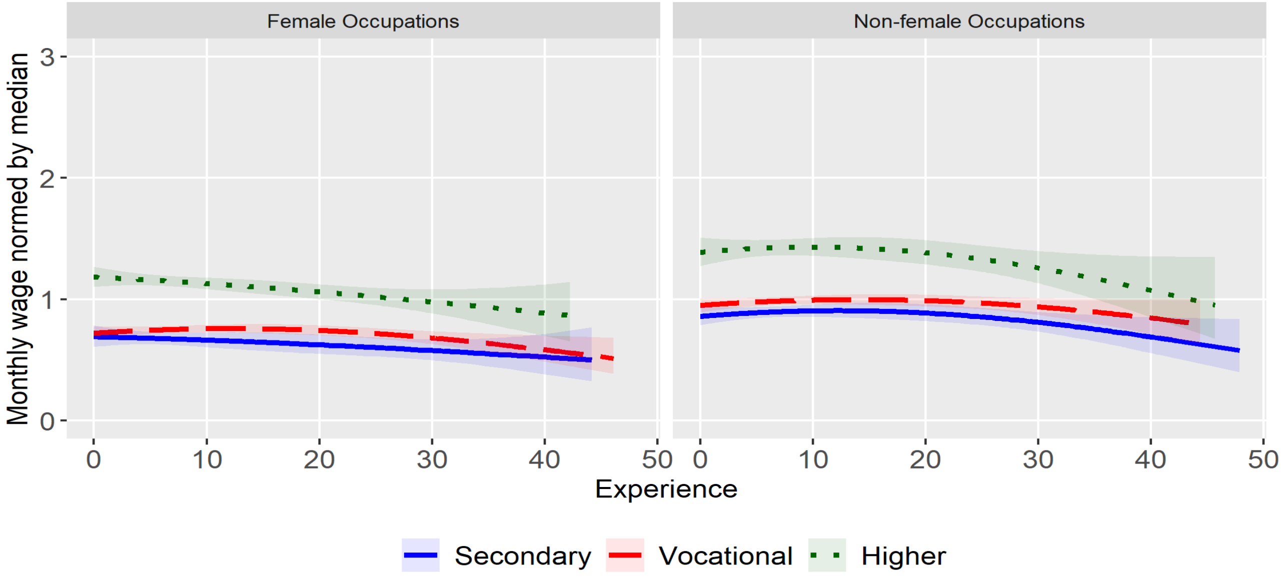

Figure 15: Experience Profiles, RLMS 2010

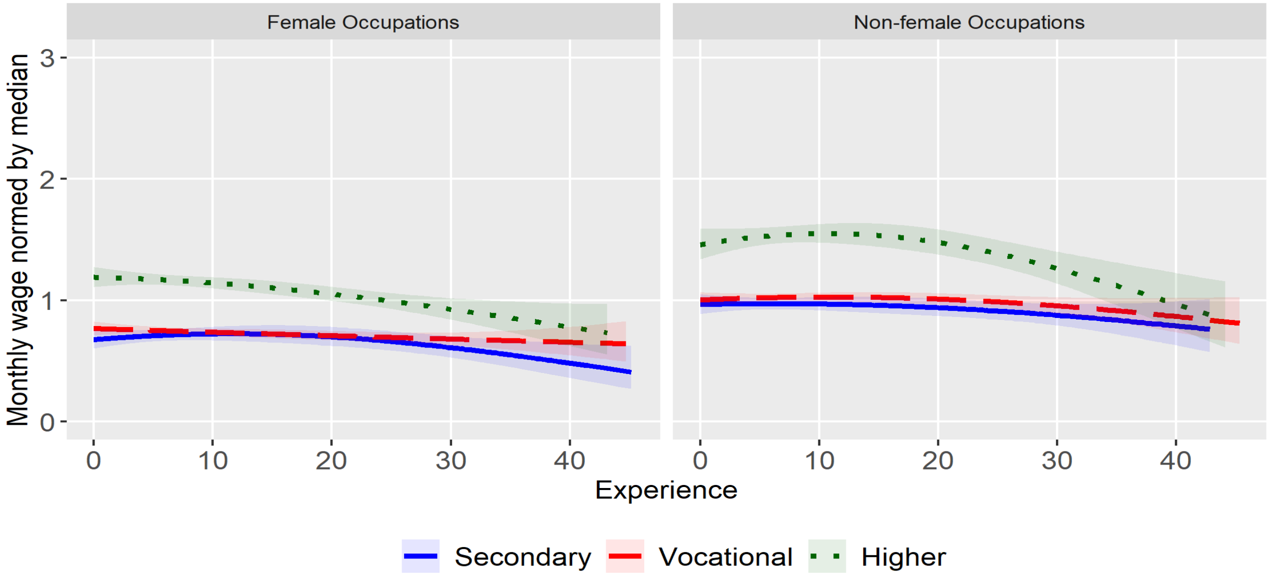

Figure 16: Experience Profiles, RLMS 2011

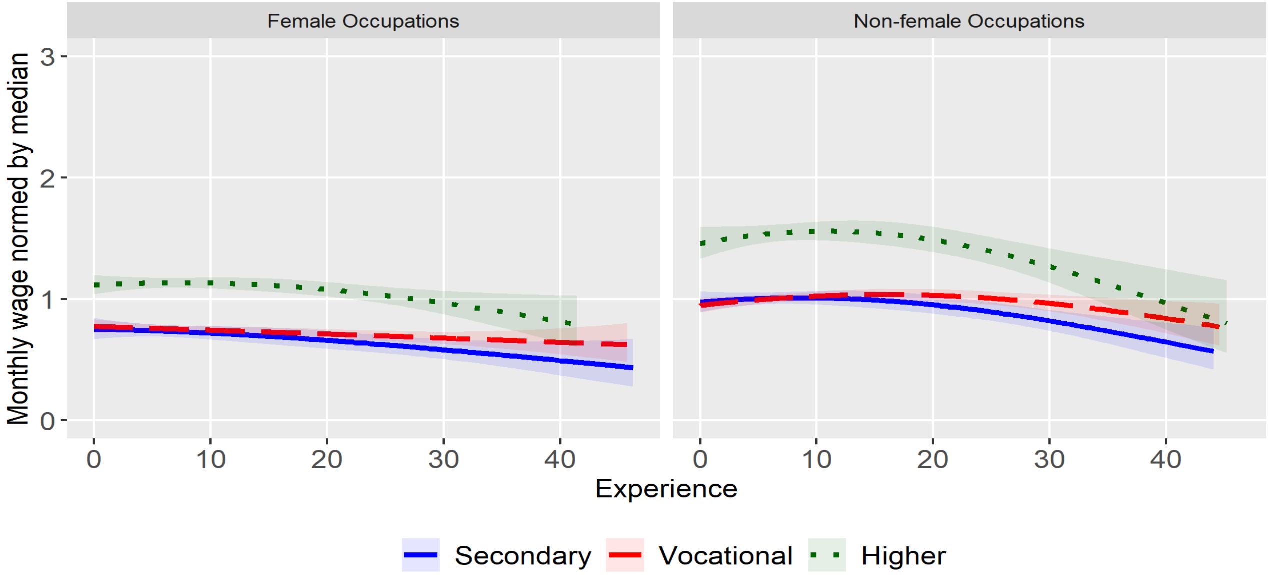

Figure 17: Experience Profiles, RLMS 2012

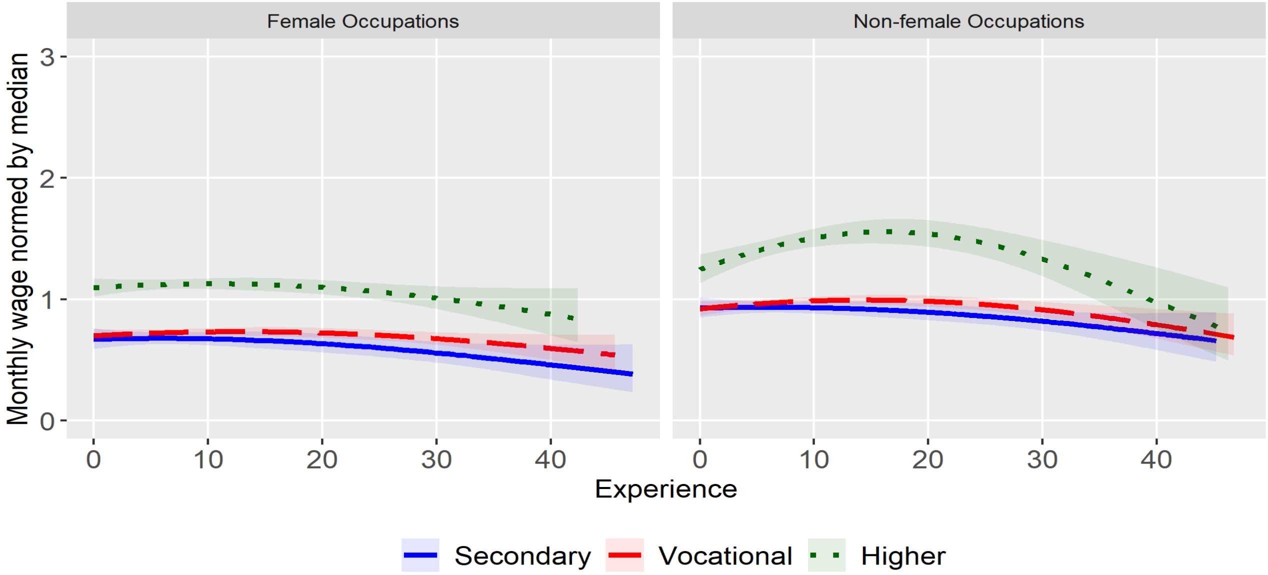

Figure 18: Experience Profiles, RLMS 2013

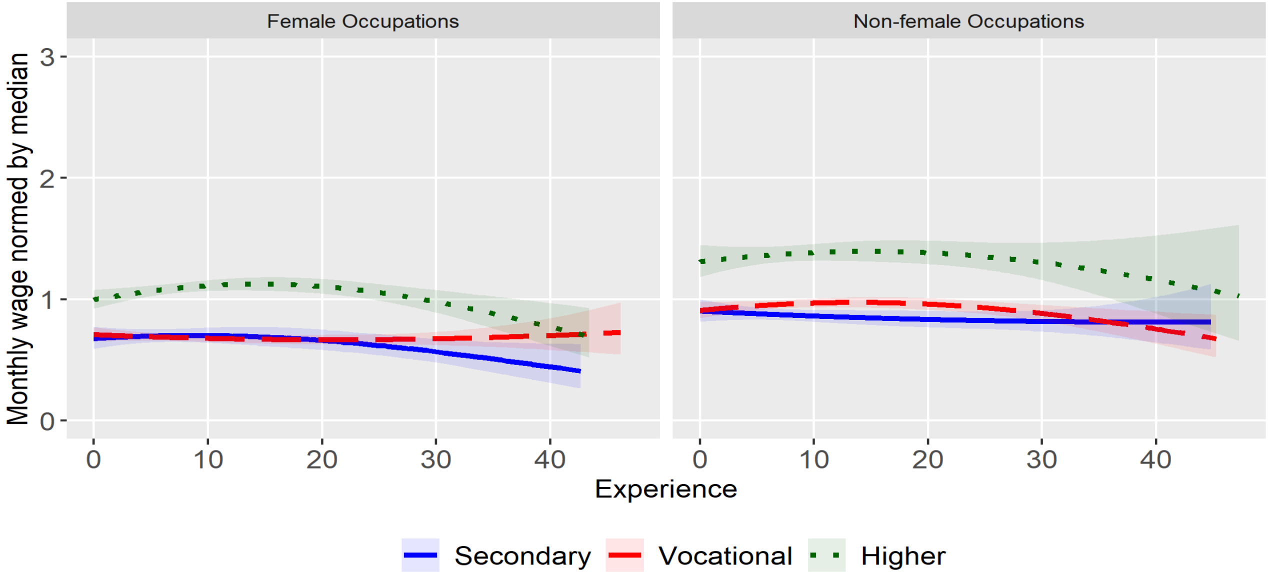

Figure 19: Experience Profiles, RLMS 2014

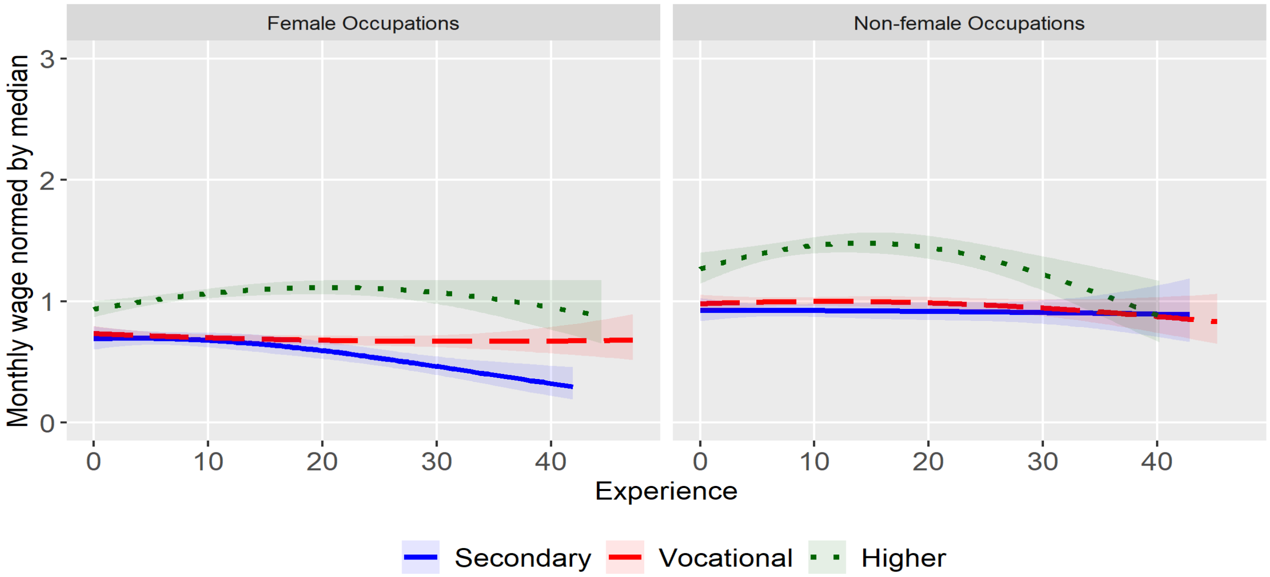

Figure 20: Experience Profiles, RLMS 2015

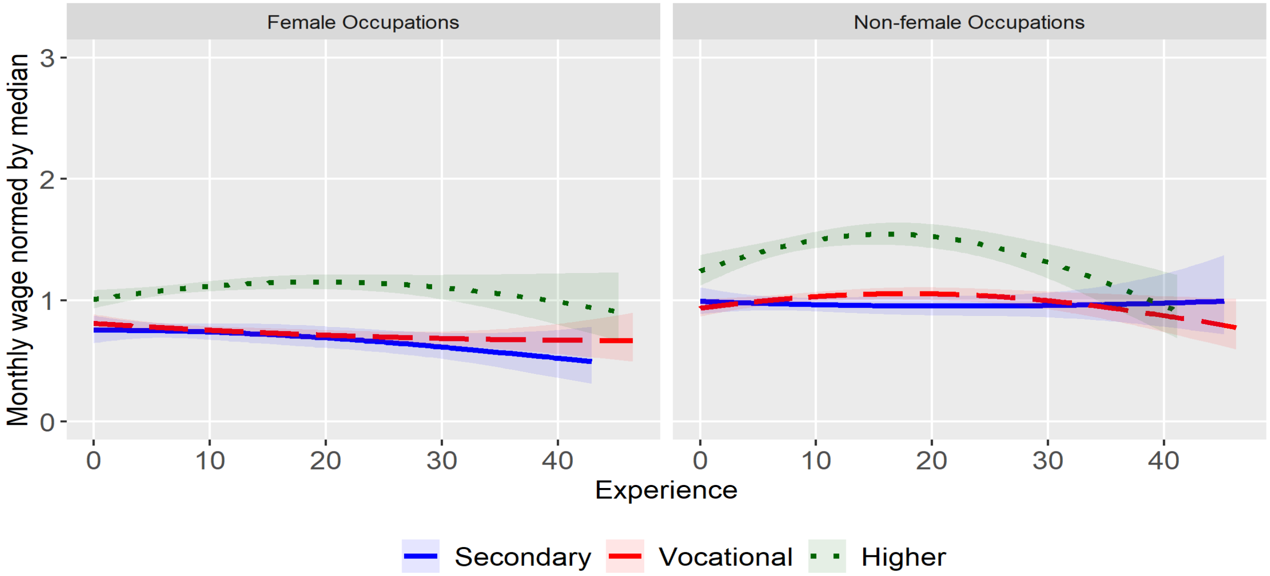

Figure 21: Experience Profiles, RLMS 2016

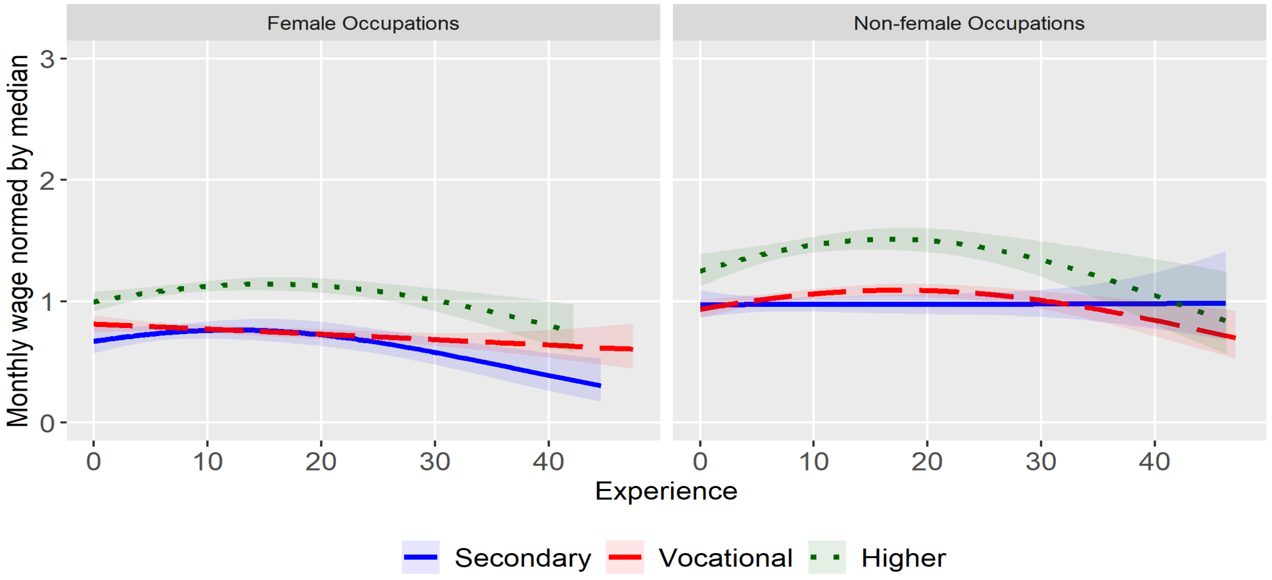

Figure 22: Experience Profiles, RLMS 2017

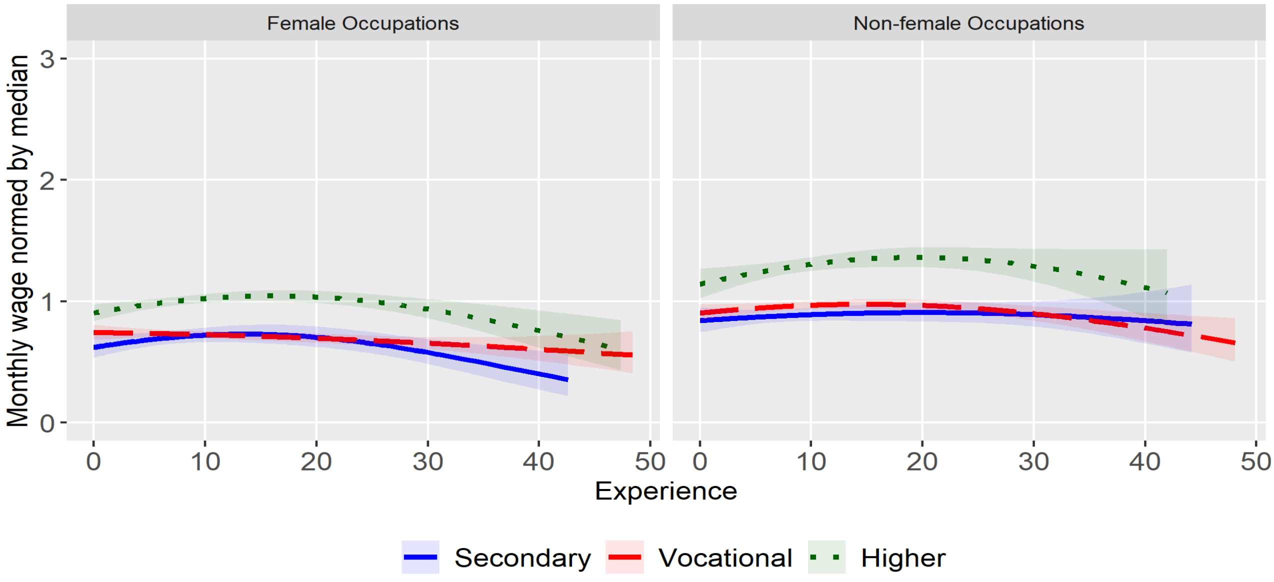

Figure 23: Experience Profiles, RLMS 2018