Appendix C.II: Split-Sample Method - Confirmatory Latent Class Analysis (CLCA)

Adam Garber

Updated: October 2, 2024

```
library(tidyverse)
library(MplusAutomation)
library(rhdf5)
library(here)
library(glue)
library(gt)
library(janitor)
library(reshape2)
library(cowplot)
```

Read data file `n_6000_lca_rep1.dat` (N=6000; Replication 1)

```
lca_data <- read.delim2(here("C1-Simulation", "n_6000_lca_rep1.dat"), sep = "", header = FALSE) %>%
    select(-V18) %>%
    setNames(c("primary", "change", "interupt", "initiat", "engage", "approach",
        "response", "expect", "new", "same", "relative", "objects", "sequence", "trans",
        "avoid", "control", "touch")) %>%
    purrr::modify_if(is.character, as.numeric)
```

```
write_csv(lca_data, here("data", "C2_simulated_N6000R1.csv"))
```

Randomly select rows to split the parent sample equally into exploratory and confirmatory sub-samples

- Exploratory sample (`explr_data`): N = 3,000
- Confirmatory sample (`cnfrm_data`): N = 3,000

```
set.seed(9182024)  # For reproducibility

split_prop <- 0.5

split_data <- lca_data %>%
    mutate(split = ifelse(row_number() %in% sample(1:n(), size = floor(split_prop *
        n())), "exploratory", "confirmatory"))

# Create sub-samples
explr_data <- split_data %>%
    filter(split == "exploratory") %>%
    select(-split)
```

```
cnfrm_data <- split_data %>%
    filter(split == "confirmatory") %>%
    select(-split)
```

Step 1

Estimate exploratory LCA model with data subset `explr_data` (N=3000)

```
exploratoryLCA  <- lapply(4:6, function(k) {
  lca_enum  <- mplusObject(

    TITLE = glue("{k}-Class"),

    VARIABLE = glue(
    "!!! Split Sample Method (STEP-1): Exploratory Model !!!
     categorical = primary-touch;
     usevar = primary-touch;
     classes = c({k}); "),

  ANALYSIS =
   "estimator = mlr;
    type = mixture;
    starts = 500 200;
    processors = 10;",

  OUTPUT = "svalues residual tech11 tech14;",

  PLOT =
    "type = plot3;
    series = primary-touch(*);",

  usevariables = colnames(explr_data),
  rdata = explr_data)

lca_enum_fit <- mplusModeler(lca_enum,
                     dataout=glue(here("C2-Split-Sample", "lca_explr_data.dat")),
                     modelout=glue(here("C2-Split-Sample", "c{k}_explr_data.inp")),
                     check=TRUE, run = TRUE, hashfilename = FALSE)
})
```

Table of Fit

Extract fit data

```
output_lca <- readModels(here("C2-Split-Sample"), filefilter = "explr", quiet = TRUE)

enum_extract <- LatexSummaryTable(output_lca, keepCols = c("Title", "Parameters",
    "LL", "BIC", "aBIC", "BLRT_PValue", "T11_VLMR_PValue", "Observations"), sortBy = "Title")

allFit <- enum_extract %>%
    mutate(CAIC = -2 * LL + Parameters * (log(Observations) + 1)) %>%
    mutate(AWE = -2 * LL + 2 * Parameters * (log(Observations) + 1.5)) %>%
    mutate(SIC = -0.5 * BIC) %>%
    mutate(expSIC = exp(SIC - max(SIC))) %>%
    mutate(BF = exp(SIC - lead(SIC))) %>%
    mutate(cmPk = expSIC/sum(expSIC)) %>%
    dplyr::select(1:5, 9:10, 6:7, 13, 14) %>%
    arrange(Parameters)
```

Create table

```
fit_table <- allFit %>%
    gt() %>%
    tab_header(title = md("**Model Fit Summary Table**")) %>%
    cols_label(Title = "Classes", Parameters = md("Par"), LL = md("*LL*"), T11_VLMR_PValue = "VLMR",
        BLRT_PValue = "BLRT", BF = md("BF"), cmPk = md("*cmPk*")) %>%
    tab_footnote(footnote = md("*Note.* Par = Parameters; *LL* = model log likelihood;
BIC = Bayesian information criterion;
aBIC = sample size adjusted BIC; CAIC = consistent Akaike information criterion;
AWE = approximate weight of evidence criterion;
BLRT = bootstrapped likelihood ratio test p-value;
VLMR = Vuong-Lo-Mendell-Rubin adjusted likelihood ratio test p-value;
*cmPk* = approximate correct model probability."),
        locations = cells_title()) %>%
    tab_options(column_labels.font.weight = "bold") %>%
    fmt_number(c(3:7), decimals = 2) %>%
    sub_missing(1:11, missing_text = "--") %>%
    fmt(c(8:9, 11), fns = function(x) ifelse(x < 0.001, "<.001", scales::number(x,
        accuracy = 0.01))) %>%
    fmt(10, fns = function(x) ifelse(x > 100, ">100", scales::number(x, accuracy = 0.01))) %>%
    tab_style(style = list(cell_text(weight = "bold")), locations = list(cells_body(columns = BIC,
        row = BIC == min(BIC[c(1:3)])  # Change this to the number of classes you are evaluating
),
        cells_body(columns = aBIC, row = aBIC == min(aBIC[1:3])), cells_body(columns = CAIC,
            row = CAIC == min(CAIC[1:3])), cells_body(columns = AWE, row = AWE ==
            min(AWE[1:3])), cells_body(columns = cmPk, row = cmPk == max(cmPk[1:3])),
        cells_body(columns = BF, row = BF > 10), cells_body(columns = T11_VLMR_PValue,
            row = ifelse(T11_VLMR_PValue < 0.05 & lead(T11_VLMR_PValue) > 0.05, T11_VLMR_PValue <
                0.05, NA)), cells_body(columns = BLRT_PValue, row = ifelse(BLRT_PValue <
            0.05 & lead(BLRT_PValue) > 0.05, BLRT_PValue < 0.05, NA))))

fit_table
```

Model Fit Summary Table[1]

Classes	Par	*LL*	BIC	aBIC	CAIC	AWE	BLRT	VLMR	BF	*cmPk*

4-Class	71	−27,013.31	54,595.07	54,369.48	54,666.07	55,376.53	<.001	<.001	0.00	<.001
5-Class	89	−26,202.59	53,117.74	52,834.95	53,206.74	54,097.31	<.001	<.001	>100	1.00
6-Class	107	−26,182.92	53,222.53	52,882.54	53,329.53	54,400.21	0.67	0.17	–	<.001

[1]*Note.* Par = Parameters; *LL* = model log likelihood; BIC = Bayesian information criterion; aBIC = sample size adjusted BIC; CAIC = consistent Akaike information criterion; AWE = approximate weight of evidence criterion; BLRT = bootstrapped likelihood ratio test p-value; VLMR = Vuong-Lo-Mendell-Rubin adjusted likelihood ratio test p-value; *cmPk* = approximate correct model probability.

Save table

```
gtsave(fit_table, here("figures", "C2-fit-table-explr-data.png"))
```

Conditional LCA Plot Function

```
plot_lca_function <- function(model_name, item_num, class_num, item_labels, class_labels = c("Class1",
    "Class2", "Class3", "Class4", "Class5"), class_legend_order = c(1, 2, 3, 4, 5),
    plot_title) {

    mplus_model <- as.data.frame(model_name$gh5$means_and_variances_data$estimated_probs$values)
    plot_data <- mplus_model[seq(2, 2 * item_num, 2), ]

    c_size <- as.data.frame(model_name$class_counts$modelEstimated$proportion)
    colnames(c_size) <- paste0("cs")
    c_size <- c_size %>%
        mutate(cs = round(cs * 100, 2))
    colnames(plot_data) <- paste0(class_labels, glue(" ({c_size[1:class_num,]}%)"))
    plot_data <- plot_data %>%
        relocate(class_legend_order)

    plot_data <- cbind(Var = paste0("U", 1:item_num), plot_data)
    plot_data$Var <- fct_inorder(plot_data$Var)
    plot_data$Var <- factor(plot_data$Var, labels = item_labels)

    pd_long_data <- melt(plot_data, id.vars = "Var")

    # This syntax uses the data.frame created above to produce the plot with
    # `ggplot()`

    p <- pd_long_data %>%
        ggplot(aes(x = as.integer(Var), y = value, shape = variable, colour = variable,
            lty = variable)) + geom_point(size = 4) + geom_line(size = 2) + scale_x_continuous("",
        breaks = 1:item_num, labels = plot_data$Var) + scale_color_viridis_d(end = 0.7,
        alpha = 0.89, option = "H") + labs(title = plot_title, y = "Probability") +
        theme_cowplot() + theme(legend.title = element_blank(), legend.position = "top",
        axis.text.x = element_text(size = 8, vjust = 1))
    p
    return(p)
}
```

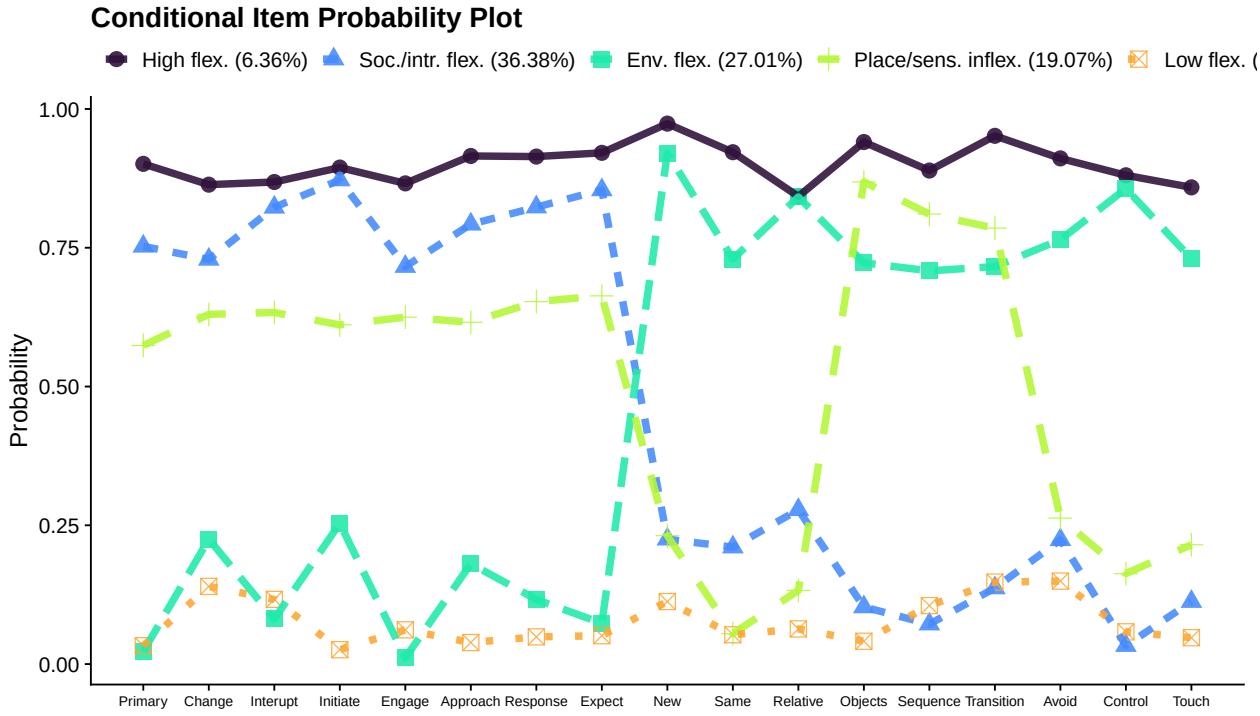
5-Class Probability Plot

Use the `plot_lca` function provided in the folder to plot the item probability plot. Details for this function are written in the document `plot_lca.txt`

Read models for plotting (5-class model)

```
model_c5 <- readModels(here("C2-Split-Sample", "c5_explr_data.out"), quiet = TRUE)
```

```
plot_lca_function(model_name = model_c5, item_num = 17, class_num = 5, item_labels = c("Primary",
    "Change", "Interupt", "Initiate", "Engage", "Approach", "Response", "Expect",
    "New", "Same", "Relative", "Objects", "Sequence", "Transition", "Avoid", "Control",
    "Touch"), class_labels = c("High flex.", "Place/sens. inflex.", "Env. flex.",
    "Low flex.", "Soc./intr. flex."), class_legend_order = c(1, 5, 3, 2, 4), plot_title = "Conditional
```



Save figure

```
ggsave(here("figures", "C5_Plot_Explr-Data.png"), dpi = 300, height = 5, width = 11,
    units = "in")
```

Step 2: Confirmatory analysis

CLCA model estimated from the split sample `cnfrm_data` using start values from exploratory K=5 model

```
m_step1  <- mplusObject(

  TITLE = "Split-Sample (confirmatory sample)",

  VARIABLE =
   "categorical = primary-touch;
    usevar = primary-touch;
    classes = c(5); ",

  ANALYSIS =
   "estimator = mlr;
    type = mixture;
    !starts = 0;
    ! STSEED = 392407; !!! USE SEED TO REPLICATE THESIS RESULTS !!!
   ",

  MODEL =
    "%OVERALL%

     [ c#1*-1.74441 ];
     [ c#2*-0.64595 ];
     [ c#3*-0.29794 ];
     [ c#4*-1.17904 ];

     %C#1%       !!!High_Flex!!!

     [ primary$1*-2.20987 ](t1);
     [ change$1*-1.84601  ](t2);
     [ interupt$1*-1.88724](t3);
     [ initiat$1*-2.13724 ](t4);
     [ engage$1*-1.86574  ](t5);
     [ approach$1*-2.38052](t6);
     [ response$1*-2.36713](t7);
     [ expect$1*-2.45601  ](t8);
     [ new$1*-3.60891     ](t9);
     [ same$1*-2.47183    ](t10);
     [ relative$1*-1.67673](t11);
     [ objects$1*-2.75974 ](t12);
     [ sequence$1*-2.08144](t13);
     [ trans$1*-2.97972   ](t14);
     [ avoid$1*-2.32553   ](t15);
     [ control$1*-1.99748 ](t16);
     [ touch$1*-1.80536   ](t17);

     %C#2%         !!!LocSnsLo!!!

     [ primary$1*-0.29794 ](t52);
     [ change$1*-0.53213  ](t53);
     [ interupt$1*-0.54672](t54);
     [ initiat$1*-0.45183 ](t55);
     [ engage$1*-0.51045  ](t56);
     [ approach$1*-0.47146](t57);
     [ response$1*-0.63201](t58);
```

```
[ expect$1*-0.67809  ](t59);
[ new$1*1.20090      ](t60);
[ same$1*2.84868     ](t61);
[ relative$1*1.87798 ](t62);
[ objects$1*-1.88945 ](t63);
[ sequence$1*-1.45427](t64);
[ trans$1*-1.29703   ](t65);
[ avoid$1*1.03081    ](t66);
[ control$1*1.63541  ](t67);
[ touch$1*1.29558    ](t68);

%C#3%       !!!Env_Flex !!!

[ primary$1*3.76057  ](t35);
[ change$1*1.23660   ](t36);
[ interupt$1*2.42362 ](t37);
[ initiat$1*1.07970  ](t38);
[ engage$1*4.47606   ](t39);
[ approach$1*1.51131 ](t40);
[ response$1*2.02784 ](t41);
[ expect$1*2.53496   ](t42);
[ new$1*-2.44811     ](t43);
[ same$1*-0.99306    ](t44);
[ relative$1*-1.67001](t45);
[ objects$1*-0.95871 ](t46);
[ sequence$1*-0.88720](t47);
[ trans$1*-0.92516   ](t48);
[ avoid$1*-1.17745   ](t49);
[ control$1*-1.78766 ](t50);
[ touch$1*-0.99513   ](t51);

%C#4%         !!!Low_Flex!!!

[ primary$1*3.37636  ](t69);
[ change$1*1.81614   ](t70);
[ interupt$1*2.02582 ](t71);
[ initiat$1*3.62826  ](t72);
[ engage$1*2.72187   ](t73);
[ approach$1*3.20886 ](t74);
[ response$1*2.96584 ](t75);
[ expect$1*2.92061   ](t76);
[ new$1*2.06398      ](t77);
[ same$1*2.88648     ](t78);
[ relative$1*2.69372 ](t79);
[ objects$1*3.14932  ](t80);
[ sequence$1*2.13968 ](t81);
[ trans$1*1.75316    ](t82);
[ avoid$1*1.73877    ](t83);
[ control$1*2.78570  ](t84);
[ touch$1*2.99815    ](t85);

%C#5%       !!!ScIn_Flex!!!
```

```
   [ primary$1*-1.11218 ](t18);
   [ change$1*-0.98964  ](t19);
   [ interupt$1*-1.53552](t20);
   [ initiat$1*-1.91915 ](t21);
   [ engage$1*-0.92521  ](t22);
   [ approach$1*-1.33903](t23);
   [ response$1*-1.53825](t24);
   [ expect$1*-1.76776  ](t25);
   [ new$1*1.23723      ](t26);
   [ same$1*1.32099     ](t27);
   [ relative$1*0.95454 ](t28);
   [ objects$1*2.16596  ](t29);
   [ sequence$1*2.55756 ](t30);
   [ trans$1*1.84039    ](t31);
   [ avoid$1*1.24333    ](t32);
   [ control$1*3.37130  ](t33);
   [ touch$1*2.06513    ](t34);
",

MODELCONSTRAINT = "!!!       THRESHOLD BOUNDARIES            !!!
                   !!! -.85 (THRESHOLD) ~ .70 (PROBABILTY)   !!!
                   !!!  .85 (THRESHOLD) ~ .30 (PROBABILTY)   !!!

   ! LABELS C1-C5 BELOW REFLECT ORDER OF CLASSES IN PLOT (NOT MPLUS C# LABELS ABOVE)
   !    C1   |    C2   |   C3    |    C4   |   C5    !
   !High_Flex|ScIn_Flex|Env_Flex |LocSnsLo |Low_Flex !
   !_________|_________|_________|_________|_________!
    t1 <-.85; t18<-.85; t35> .85;           t69> .85;  !!!U1 !!!
    t2 <-.85; t19<-.85; t36> .85;           t70> .85;  !!!U2 !!!
    t3 <-.85; t20<-.85; t37> .85;           t71> .85;  !!!U3 !!!
    t4 <-.85; t21<-.85; t38> .85;           t72> .85;  !!!U4 !!!
    t5 <-.85; t22<-.85; t39> .85;           t73> .85;  !!!U5 !!!
    t6 <-.85; t23<-.85; t40> .85;           t74> .85;  !!!U6 !!!
    t7 <-.85; t24<-.85; t41> .85;           t75> .85;  !!!U7 !!!
    t8 <-.85; t25<-.85; t42> .85;           t76> .85;  !!!U8 !!!

    t9 <-.85; t26> .85; t43<-.85; t60> .85; t77> .85;  !!!U9 !!!
    t10<-.85; t27> .85; t44<-.85; t61> .85; t78> .85;  !!!U10!!!
    t11<-.85; t28> .85; t45<-.85; t62> .85; t79> .85;  !!!U11!!!
    t12<-.85; t29> .85; t46<-.85; t63<-.85; t80> .85;  !!!U12!!!
    t13<-.85; t30> .85; t47<-.85; t64<-.85; t81> .85;  !!!U13!!!
    t14<-.85; t31> .85; t48<-.85; t65<-.85; t82> .85;  !!!U14!!!
    t15<-.85; t32> .85; t49<-.85; t66> .85; t83> .85;  !!!U15!!!
    t16<-.85; t33> .85; t50<-.85; t67> .85; t84> .85;  !!!U16!!!
    t17<-.85; t34> .85; t51<-.85; t68> .85; t85> .85;  !!!U17!!!

     !!! EQUALITY (& INEQUALITY) CONSTRAINTS !!!
     !  C1vC2  |  C3vC4  |  C4vC5  !
     !_________|_________|_________!
      t1 = t18; t35=-t52;           !!!U1 !!!
      t2 = t19; t36=-t53;           !!!U2 !!!
      t3 = t20; t37=-t54;           !!!U3 !!!
      t4 = t21; t38=-t55;           !!!U4 !!!
```

```
        t5 = t22; t39=-t56;           !!!U5 !!!
        t6 = t23; t40=-t57;           !!!U6 !!!
        t7 = t24; t41=-t58;           !!!U7 !!!
        t8 = t25; t42=-t59;           !!!U8 !!!
        t9 =-t26;                     !!!U9 !!!
        t10=-t27;                     !!!U10!!!
        t11=-t28;                     !!!U11!!!
        t12=-t29;           t63=-t80; !!!U12!!!
        t13=-t30;           t64=-t81; !!!U13!!!
        t14=-t31;           t65=-t82; !!!U14!!!
        t15=-t32;                     !!!U15!!!
        t16=-t33;                     !!!U16!!!
        t17=-t34;                     !!!U17!!!

!-----------------------------------------------------------------!

  ",

  OUTPUT = "tech14;",

  PLOT =
    "type = plot3;
    series = primary-touch(*);",

  usevariables = colnames(cnfrm_data),
  rdata = cnfrm_data)

m_step1_fit <- mplusModeler(m_step1,
                 dataout=here("C2-Split-Sample", "CLCA_SplitSample.dat"),
                 modelout=here("C2-Split-Sample", "CLCA_SplitSample.inp") ,
                 check=TRUE, run = TRUE, hashfilename = FALSE)
```

Step 3: Compare model fit

Conduct the Sattorra-Bentler adjusted Log Likelihood Ratio (LRT) difference test:

- Exploratory model (parent): `c5_explr_data.out` is the 'un-constrained model" with 89 parameters.
- Confirmatory model (nested): `CLCA_SplitSample.out` is the "constrained model" with 61 parameters.

```
split_sample_models <- readModels(here("C2-Split-Sample"), quiet = TRUE)
```

```
# *0 = null or nested model & *1 = comparison or parent model

# Log Likelihood Values
L0 <- split_sample_models[["CLCA_SplitSample.out"]][["summaries"]][["LL"]]
L1 <- split_sample_models[["c5_explr_data.out"]][["summaries"]][["LL"]]
```

```r
# LRT equation
lr <- -2 * (L0 - L1)

# Parameters
p0 <- split_sample_models[["CLCA_SplitSample.out"]][["summaries"]][["Parameters"]]
p1 <- split_sample_models[["c5_explr_data.out"]][["summaries"]][["Parameters"]]

# Scaling Correction Factors
c0 <- split_sample_models[["CLCA_SplitSample.out"]][["summaries"]][["LLCorrectionFactor"]]
c1 <- split_sample_models[["c5_explr_data.out"]][["summaries"]][["LLCorrectionFactor"]]

# Difference Test Scaling correction (Sattorra-Bentler adjustment)
cd <- ((p0 * c0) - (p1 * c1))/(p0 - p1)

# Chi-square difference test(TRd)
TRd <- (lr)/(cd)

# Degrees of freedom
df <- abs(p0 - p1)


# Significance test
(p_diff <- pchisq(TRd, df, lower.tail = FALSE))
```

```
## [1] 1.170139e-317
```

RESULT: The Log Likelihood χ^2 difference test comparing the exploratory and confirmatory LCA models was, $\chi^2(28) = 1588.06, p < .001$. See Reference Here

---

Compare model fit summary statistics: Exploratory & Confirmatory Models

```r
enum_extract1 <- LatexSummaryTable(split_sample_models$c5_explr_data.out, keepCols = c("Title",
    "Parameters", "LL", "BIC", "aBIC", "Observations"))
enum_extract2 <- LatexSummaryTable(split_sample_models$CLCA_SplitSample.out, keepCols = c("Title",
    "Parameters", "LL", "BIC", "aBIC", "Observations"))
```

Calculate indices derived from the Log Likelihood (LL)

```r
allFit <- rbind(enum_extract1, enum_extract2) %>%
    mutate(aBIC = -2 * LL + Parameters * log((Observations + 2)/24)) %>%
    select(1:5) %>%
    mutate(Title = case_when(Title == " 5-Class" ~ "Exploratory Model", Title ==
        " Split-Sample (confirmatory sample)" ~ "Confirmatory Model"))
```

Format fit table

```r
allFit %>%
    gt() %>%
    tab_header(title = md("**Model Fit Comparision Table**"), subtitle = md(" ")) %>%
```

```
    cols_label(Title = "Model", Parameters = md("Par"), LL = md("*LL*"), BIC = md("BIC"),
        aBIC = md("aBIC")) %>%
    tab_footnote(footnote = md("*Note.* Par = Parameters; *LL* = model log likelihood;
     BIC = Bayesian information criterion; aBIC = sample size adjusted BIC."),
        locations = cells_title()) %>%
    tab_options(column_labels.font.weight = "bold")
```

Model Fit Comparision Table[1]

[1]

Model	Par	*LL*	BIC	aBIC
Exploratory Model	89	-26202.59	53117.74	52834.95
Confirmatory Model	61	-26976.57	54441.52	54247.70

[1] *Note.* Par = Parameters; *LL* = model log likelihood; BIC = Bayesian information criterion; aBIC = sample size adjusted BIC.