

## 혼합효과모형(Mixed-Effects Model)을 이용한 실험언어학 데이터 분석 방법 고찰: 자기조절읽기 실험 데이터를 중심으로\*

신정아  
 (동국대학교)

**Shin, Jeong-Ah. 2019. How to analyze experimental linguistic data using a mixed-effects model in R: Focusing on data from a self-paced reading experiment. *Korean Journal of English Language and Linguistics* 19-1, 76-94.** This study examined a practical use of mixed-effects models in R, analyzing accuracy and reading time data from a self-paced reading experiment. It discussed the applications of logistic mixed-effects model for binary data (e.g., accuracy data) and the use of a mixed-effects model for reading time (RT) data, effectively removing outliers within the data set. A sample for mixed-effects model analyses was collected from a previously conducted self-paced reading experiment, involving English reduced relative clauses for 30 advanced and intermediate second language learners. Rationales and guidelines toward selecting the most appropriate mixed-effects model and checking model assumptions were also discussed.

**Keywords:** mixed-effects model, linear mixed model, logistic mixed model, experimental linguistics, psycholinguistics, self-paced reading, reading time, RT data, accuracy

### 1. 서론

2008년 *Journal of Memory and Language* 저널 특별호(Emerging Data Analysis)에는 선형혼합효과모형(linear mixed effect model, LME)을 이용하여 인간의 행동반응 언어 실험을 분석하는 방법에 대한 논문이 소개되었다(e.g., Baayen, Davidson, and Bates 2008). Clark(1973)이 피험자(participant)별 평균을 이용한 분석 결과뿐만 아니라, 실험에 사용된 언어 자료 항목(item)별 평균에 대한 통계 결과를 보고해야 한다고 주장한 ‘언어를 고정효과로 분석하는 오류(language-as-fixed-effect fallacy)’에 대한 논문을 발표한 뒤, 심리언어학 분야 실험 연구들은 항목 분석(item analysis)의 결과, 피험자 분석(participant analysis)의 결과, 그리고 이 두 가지 분석 결과를 함께 고려한 minF값 등을 보고하였다. 왜냐하면, 실험언어학의 실험은 특히 다양한 언어 항목을 사용하기 때문에 언어 항목의 변

\* 이 논문은 2016년 한국연구재단의 국제협력사업(NRF-2016K2A9A2A19939367)과 2018년 동국대학교 우수연구자 지원사업(S-2018-G0001-00022)의 지원을 받아 연구되었음.

이 분포도 보정할 필요성이 대두되었기 때문이다. 즉, 언어 실험은 피험자 당 한 번 이상의 반응(같은 요인의 다른 항목으로)을 얻게 될 뿐만 아니라, 자극 항목별로 두 개 이상의 반응(같은 요인을 다른 피험자로)을 얻을 수 있도록 설계하는 경우가 많기 때문에 피험자 분석뿐만 아니라 항목 분석을 병행하는 것이 꼭 필요하다는 것을 많은 학자들이 인지하게 되었다.

이렇게 기존의 많은 실험언어학 연구들이 피험자를 임의(random) 변인으로 하여 모든 피험자의 평균을 분석하는 피험자 분산 분석(ANOVA)과 언어 항목을 임의변인으로 하여 모든 실험 자료의 평균을 분석하는 항목 ANOVA의 결과를 각각 F1값, F2값으로 따로 보고하였다. 하지만, 최근에는 피험자와 언어 항목을 동시에 임의변인으로 고려하여 한 번에 분석할 수 있는 혼합 모형이 실험 언어학자들의 관심을 받기 시작하였다. 혼합 모형은 각각의 평균값을 이용한 ANOVA에 비해 데이터 손실이 크지 않고, ANOVA 분석으로 적당하지 않은 이분적 데이터도 로지스틱혼합모형(logit mixed-effects model)으로 분석도 가능하다는 장점이 있다(Jaeger, 2008).

혼합효과모형 분석은 보통 오픈소스 통계 프로그램인 R(R Core Team, 2014)을 이용한다. 국내의 실험 언어학 연구에서도 R과 LME 모형을 이용하여 분석한 연구들(e.g., 이준규, 2011, 2016; Lee and Shin, 2016; Seo and Shin, 2016)이 존재하지만, 아직까지는 보편적으로 사용되지 않은 실정이다. 그래서 본 논문은 R에서 LME 모형 분석 방법을 이용하여 언어 실험에서 자주 쓰이는 반응 정답률과 반응 시간을 분석하는 방법을 순서대로 살펴보고자 한다. 특히, 자기조절읽기(self-paced reading) 실험 데이터를 예로 R에서 LME 모형을 이용하여 분석하는 방법을 공유하고자 한다.

## 2. LME 분석 전 데이터 준비 및 점검

### 2.1 데이터프레임 준비

피험자 또는 언어 항목의 평균을 사용하는 기존 방식과는 다르게 LME 모형은 각각의 관찰치(data point)를 모두 사용하는 방식으로, 통계 분석을 하기 전에 결과를 정리할 때 데이터 준비 방법에 대해 먼저 논의할 필요가 있다. 우선, 각각의 관찰치가 각 행을 차지할 수 있도록 아래로 길게 배열하는 것이 필요하다. 보통은 엑셀(Excel) 프로그램을 이용하여 데이터를 정렬하는데, csv 확장명을 사용하면 R 프로그램에서 데이터 불러오기가 용이하다. 데이터 정렬하는데 있어서 중요한 것은 공통적으로 필요한 독립 변수 정보는 주어진 열에 동일하게 나열하고, 각각의 관찰치는 종속변수 열의 각각의 행에 기록하여 아래로 증가하도록 배열한다는 점이다. 독립변수 열의 이름은 보통 “Subject”, “Item”, “Factor1”, “Factor2”으로 하고, 종속변수 열의 이름은 “Response”, “RT(ResponseTime)”, “ACC(Accuracy) 등

으로 표시하여 세로로 긴 형태로 데이터프레임을 정리한다. 이렇게 되면 그림 1과 같이 각각의 관찰치의 종속변수 셀은 다른 관찰치와 다른 값을 가지게 된다.

Subject	Item	Type	Prof	Sentence	Word	RT	Question	ACC
1	1	unambiguous	hi	The tired soldiers beaten in the fierce battle	attacked	1407	Did the soldiers beat anyone?	1
1	2	ambiguous	hi	The brown sparrow noticed on an upper branch	pecked	1572	Did the sparrow notice an insect on the upper branch?	0
1	3	unambiguous	hi	The famous man known among the student	was	522	Was the famous man a stranger to the students?	1
1	4	ambiguous	hi	The small girl lost in the buzzing crowd look	looked	679	Did the small girl lose an item?	1

그림 1. 분석에 필요한 csv데이터프레임 준비

그림 1을 보면, 각각의 열의 이름은 피험자(Subject), 항목(Item), 요인1(Type), 요인2(Proficiency), 실험 문장(Sentence), 목표 단어(Word), 종속변수1(RT), 이해도 측정 질문(Question), 종속변수2(ACCuracy), 순서로 구성되어 있다<sup>1</sup>. 그리고 각각의 행은 각각의 열에 해당되는 정보가 기입되어 있다. 종속변수 1과 2에 해당하는 읽기시간(RT)과 문장 이해도 측정 반응값(ACC)의 관찰치를 제외하고 나머지 정보는 보통 실험자가 실험 설계할 때 미리 준비하는 독립변수 정보이다. 그리고 실험이 끝나면, 각각의 관찰치에 피험자의 번호를 부여한다. 예를 들어, 그림 1의 실험(3장에서 자세히 설명)에는 문장 종류(Type)의 두 레벨, 중의성 불포함(unambiguous) 조건과 중의성 포함(ambiguous) 조건이 있는데, 피험자 1에게 주어진 언어 항목 1(즉, 첫 번째 문장)을 중의성 불포함(unambiguous) 조건으로, 문장 내 목표 단어인 본동사 attacked을 읽는데 1407ms가 걸렸고, 이해도 측정 질문에 정답으로 대답하여 1로 입력되어 있다. 마찬가지로, 피험자 1에게 항목 2를 중의성 포함(ambiguous) 조건으로 실험 문장이 주어졌는데, 문장 내 목표 단어인 본동사 pecked를 읽는데 1572ms가 걸렸다. 그러나 2번 항목의 이해도 측정 질문에 오답으로 대답하여 0으로 입력되어 있다. 이렇게 피험자 1의 전체 결과 밑에 피험자 2의 전체 결과를 나열하는 방식으로, 모든 피험자 결과를 하나의 스프레드시트에 이어 붙인다. 그 다음에는 연구자가 원하는 정보를 보기 위해서 정렬 방식을 바꾸면서 모든 정보가 정확하게 입력되어 있는지 확인한다. 다음 그림 2는 같은 데이터프레임을 언어 항목, 피험자 순으로 정렬한 것이다.

Subject	Item	Type	Prof	Sentence	Word	RT	Question	ACC
1	1	unambiguous	hi	The tired soldiers beaten in the fierce battle	attacked	1407	Did the soldiers beat anyone?	1
2	1	ambiguous	hi	The tired soldiers warned about the dangers	attacked	2347	Were the soldiers not warned of the dangers before attack?	1
3	1	unambiguous	lo	The tired soldiers beaten in the fierce battle	attacked	1518	Did the soldiers beat anyone?	0
4	1	ambiguous	lo	The tired soldiers warned about the dangers	attacked	820	Were the soldiers not warned of the dangers before attack?	1

그림 2. 항목, 피험자 순으로 정렬한 데이터프레임

<sup>1</sup> csv데이터프레임 파일은 <http://bitly.kr/LMEdata>에서 data\_v.csv 다운로드 후 엑셀에서 확인 가능

그림 2에서 알 수 있듯이, 같은 데이터프레임을 항목 1번 문장에 대하여 다른 피험자의 관찰치를 확인할 수 있다. 참고로, 예시로 제시한 실험은 실험 설계부터 균형설계(counterbalanced design)되어, 1번 항목 문장에 대해서 두 가지 조건을 다른 피험자가 보도록 설계되어 있다. 즉, 한 명의 피험자는 두 가지 조건을 다른 항목을 통해서 읽게 되어 있고, 하나의 항목은 두 가지 조건으로 교차해서(crossed) 나누어 있어서 다른 피험자를 통해서 읽게 되어 있다. 그리고 두 가지 조건의 목표단어는 모두 동일하고, 목표 단어 앞부분을 다르게 하여 두 가지로 조건으로 만들어 놓았다.

## 2.2 LME 모형 개요 및 가정

LME 모형은 고정효과(fixed effect)를 보이는 변수뿐만 아니라 임의효과(random effect)를 보이는 변수들까지 포함하므로, 반복측정분산분석의 문제점을 극복할 수 있어서 현재 실험/심리언어학에서 매우 치열하게 논의되고 있다. 고정효과는 보통 실험 설계에서 설정한 변수로, 설정된 레벨이 모집단의 모든 가능한 레벨을 포괄하는 변수에 의한 효과이고, 임의효과는 피험자와 언어 항목 등처럼 설정된 레벨이 모집단의 모든 가능한 레벨을 포괄하지 못하는 변수에 의한 효과를 의미한다(Gelman and Hill 2007, 최재웅, 홍정하 2013 재인용). LME 모형은 고정효과와 임의효과 뿐만 아니라 반복측정 및 범주형/수치형 예측 변수들 다 사용이 가능하며, 계층/내포 데이터, 불균형 관측자료, 결측 데이터(missing data)가 있는 관측 자료도 다룰 수 있다. 그래서 결측 데이터가 있는 경우나 불균형 관측자료일 경우에는 분석의 문제점이 많이 발견되는 반복측정분산분석보다 LME 모형이 실험언어학 데이터 분석에 더욱 적합하다고 할 수 있다.

LME 모형은 서로 다른 변이성을 동시에 모형화하므로, 개별 피험자나 항목별로 각기 다른 회귀선(특정 절편이나 기울기)을 갖는 종속변수를 모형화할 수 있는 장점이 있다. 피험자 또는 항목별로 특정 절편이나 기울기는 정규분포의 임의변수로 모형화되므로 ‘임의절편(random intercept)’, ‘임의기울기(random slope)’라고 부른다. 이렇게 임의절편과 임의기울기까지 모두 최대로 포함한 모형을 ‘최대모형(maximal model)’이라고 부르는데, 최대모형이 더 적합한지 단순모델이 더 적합한지에 대해서는 아직도 의견이 분분하다(Matuschek, Kliegl, Vasishth, Baayen and Bates 2017). Barr, Levy, Scheepers와 Tily(2013) 논문 이후로 2017년 이전까지는 R에서 실행했을 때 수렴(convergence) 문제가 발생하지 않으면 최대모형을 선택하는 것이 일반적인 분석방법이었다. 물론, 실험 설계에서 설정된 요인은 모형을 선택할 때 고려해야한다. 예를 들어, 피험자간 요인(between-subject factor)은 모든 피험자가 동일하게 가지는 요인에 대해 변이성을 볼 수 없게 설계되었으므로 피험자의 임의기울기는 넣지 않는다. 또한, 예측 변수간의 상관관계(correlation)가 높으면 과적합(over-fit)되기 때문에, 경험적으로는 항목의 임의기울기를 모형에 넣지 않거나 복잡한 임의기울기(상호작용까지 포함한 임의기울기)가 없는 모형으로 선택하는 것을 추천한다.

LME 모형은 회귀분석의 기본적인 선형 모형에서 출발한다. 그래서 선형 모형에서 사용하는 가정을 만족해야 한다(Gries, 2013; Winter, 2013). 첫 번째 가정은 선형성이고, 두 번째는 잔차(residual)의 분산의 동질성이다. 세 번째 가정으로, 잔차는 표본에서 표본을 추출하는 모집단에서 정규분포를 따라야 한다. 네 번째 가정은 잔차의 독립성이다. 그리고 마지막 가정으로, 잔차의 다중공선성(multicollinearity)의 불출현이 있다. 이러한 가정의 확인은 모형진단(model-diagnostic) 그래프를 이용하여 검토한다.

R의 설치(<https://www.r-project.org/>에 들어가서 cran을 Korea에서 선택하고 컴퓨터 사양에 맞게 선택하여 다운로드)부터 기초적인 사용법은 Gries(2013)(한국어 번역본, 최재웅, 홍정하 2013)의 제 2장에 쉽게 설명해놓았다. Gries(2013)에서도 제안했듯이, R 프로그램을 설치한 후, RStudio 프로그램을 설치하면 편리하게 이용할 수 있다. 그리고 Winter(2013)의 온라인 LME 튜토리얼을 참고하면 필요한 기능들을 사용할 수 있다.

LME 모형을 이용하여 실험 데이터 분석을 위해서 필요한 R 패키지는 lme4(Bates, Maechler, Bolker, and Walker 2019)와 lmerTest(Kuznetsova, Brockhoff, and Christensen 2019)이다. R Gui 또는 R Studio에서 `install.packages("lme4", "lmerTest")`로 이 두 가지 패키지를 설치한 후, `library(lme4)`와 `library(lmerTest)`로 불러오면 LME 모형을 이용하여 기본적인 분석을 할 수 있다. 그리고 LME 모형의 가정을 확인하기 위한 그래프를 이용하기 위해서 `lattice`(Sarkar 2018) 패키지도 설치하면, Normal QQ-plot 등의 모형진단 그래프를 이용하여 LME 모형의 가정을 위배했는지 여부를 검토할 수 있다(부록의 스크립트 참조).

### 2.3 실험 조건별 코딩

보통 실험언어학에서는 범주형 변인(회귀모형에서는 예측변수)을 많이 사용한다. 예를 들어  $2 \times 2$  실험 설계일 경우, 두 가지 예측변수가 있고, 각각의 예측변수 당 두 가지 레벨이 존재하는 것을 알 수 있다. 즉, 각 조건 당 레벨의 효과를 비교하기 위해서는 대비(contrast) 코딩을 해주어야 하는데, R의 lme4 패키지의 LME 모형 기본 값으로 설정되어 있는 것은 처리대비(treatment contrast, 또는 더미 코딩, dummy coding)이므로 실험의 목적에 맞게 유의해서 코딩을 다르게 해서 넣을 필요가 있다. 보통 두 가지 이상의 예측변수를 가진 실험 설계일 경우는 합계 코딩(sum coding) 또는 효과 코딩(effect coding)을 사용해야 결과를 해석하는데 용이하다. 합계 코딩을 세팅하는 방법으로 `options(contrasts=c("contr.sum", "contr.poly"))`을 사용하는 방법이 있는데, 다른 함수를 사용하지 않아도 되니 편리하지만, 분석 결과의 절편을 사용하여 결과를 해석(최재웅, 홍정하 2013, pp. 386-388 참조)하여야 해서 번거로울 수 있다. 반면, 하나의 예측변수의 두 레벨의 차이를 해석하는데 용이한 코딩 방법은 비교하고자 하는 레벨을 각각 -0.5와 0.5로 코딩하여 사용하는 효과 코딩이다. -1과 1을 사용하는 합계 코딩의 대비의 크기는 2인 반

면, 효과 코딩은 두 레벨의 차이가 1이 되므로 두 레벨의 차이를 바로 알 수 있다는 장점이 있기 때문이다. 만일 결측 데이터가 있을 경우에는 일괄적으로 -0.5와 0.5를 사용하기 보다는 scale() 함수 및 개별 함수(부록의 myCenter 함수 참조)를 사용하여 코딩하여 중심화(centering)하는 것이 더욱 정확한 분석 방법이다.

### 3. 혼합효과모형을 이용한 자기조절읽기 실험 데이터 분석 방법

심리언어학 분야 중 문장 처리 연구는 실시간 언어 처리를 알아보기 위해 자기조절읽기 실험을 많이 이용하였다. 통사적으로 중의적인 구문을 사용하여 문장 처리의 메커니즘을 연구하였는데, 특히 일시적 중의성(temporary ambiguity)을 포함한 영어 축약 관계절(reduced relative clauses)을 실험 자료로 많이 사용하였다(e.g., Juffs, 1998). 본 연구도 통사적으로 중의적인 축약 관계절을 영어학습자들이 어떻게 처리하는지 표본 데이터를 모아 혼합효과 모형을 이용하여 분석하는 방법을 살펴보고자 한다. 특히, Rah와 Adone (2010) 연구에서 사용한 실험 문장을 이용하여 자기조절읽기 실험을 부분적으로 재현(replicate)하여, 중급 또는 고급 수준의 제2 언어 학습자들이 일시적으로 통사적으로 모호한 문장의 이해도 및 읽기 시간이 그렇지 않은 문장과 비교하여 차이가 있는지 알아보는 실험을 예로 들어 LME 모형으로 분석하는 방법을 소개하고자 한다.

표본 데이터는 총 30명(남자 9명)의 IbeX Farm 플랫폼(Drummond 2013)의 자기조절읽기 실험을 통하여 획득하였고, 피험자는 한국(24명), 홍콩(5명), 독일(1명)에서 영어를 제2 언어로 학습하는 학습자이고 평균나이는 25.63(범위: 20~33)이었다. 영어 숙달도는 자가평가(self-rating)를 이용하여 고급(advanced, 18명)과 중급(intermediate, 12명), 두 그룹으로 나누었다. 실험에 사용된 문장은 다음과 같다.

1) The brave fireman rescued from the burning roof saved the helpless people.

Q: Did the fireman rescue someone from the burning roof?

2) The brave fireman seen on the burning roof saved the helpless people.

Q: Did the fireman see someone saving helpless people?

(1)번 문장은 (2)번 문장의 seen과 다르게 rescued 단어 구역에서 일시적으로 통사적 중의성을 포함한다. 본동사로도 해석될 수 있고 축약 관계절로도 해석될 수 있는 이러한 통사적 중의성은 목표구역(p)인 from에서 1차적으로 해결되고, 목표구역(v)인 saved에서 완전히 해결된다. 이렇게 구성한 총 12개의 실험 문항과 실험 문항과 전혀 관계없는 덤(filler) 문장 34개를 포함하여 약 15분 동안 실험을 진행하였다. 실험은 스페이스바(space bar)를

누르면 단어 하나씩 제시되는 자기조절 읽기 실험으로 진행하였다.

피험자 내(within-subject) 변인으로 문장의 종류(Type) 조건을 두 레벨(ambiguous vs. unambiguous)로 하였고, 피험자 간(between-subject) 변인으로 영어 숙달도(Proficiency)를 두 그룹(hi vs. lo)으로 구분하였다. 본 실험의 데이터는 부록 스크립트에 제시한 링크에서 다운로드한 뒤, `data <- read.csv(file=file.choose(), sep = ',', header = T, fill=T)`로 그 파일을 불러들이면 데이터로 사용할 수 있다. LME 모형을 이용한 분석 과정 및 결과는 다음과 같다.

### 3.1 예-아니오 응답의 정답률 분석

실험 연구에서 기본적으로 사용하는 종속 변수 중 하나는 정답률(accuracy rate)이다. 문법판단 과제, 이해도 측정 과제 등에서 예-아니오 응답과 같은 이진(binary) 종속변수를 가진 데이터를 분석할 때는 이진 로지스틱회귀모형(binary logistic regression model)을 사용한다. 로지스틱 연결함수(link function) 변환에 대한 자세한 내용을 이해하는 것은 본 논문의 목적이 아니어서, 이진 종속변수를 분석하는 방법으로 로지스틱회귀모형을 LME 모델 안에서 사용하는 방법만 다루도록 한다.

본 연구의 이해도 측정 결과는, 맞으면 1, 틀리면 0으로 입력되어 있다. 연구 목적에 따라, 즉 실시간 언어 처리를 관찰하기 위해 실험에 집중했는지 여부를 확인하기 위한 용도로 이해도 측정 문항을 사용하는데, 이 때는 정답률이 낮은 피험자의 결과 전체를 분석에서 제외하기도 한다. 보통 매우 쉬운 이해도 측정 문항을 이용해서 70퍼센트 또는 80퍼센트 미만의 정답률을 보인 피험자는 분석에서 제외하기도 하는데, 제2 언어 학습자를 대상으로 하는 실험의 경우에는 이진 문장을 이해했는지 여부를 확인하므로, 실험 목적에 따라 분석에서 제외하지 않기도 한다(Marsden, Thompson, and Plonsky 2018).

본 연구의 목적은 영어 숙달도에 따라 일시적 중의성이 포함된 문장이 그렇지 않은 문장보다 문장 이해도가 감소하는 정도에 차이가 있는지 확인하기 위해 문장 이해 문제의 정확성을 측정하였다. 두 가지 예측 변수인 문장 종류(TYPE)와 영어 숙달도(PROFICIENCY)에 따른 정확도 평균표는 `ACCtable <- tapply(ACC, list(Type, Proficiency), mean);` `ACCtable`로 구할 수 있고 그 결과는 표 1과 같다.

표 1. 실험 조건과 영어 능력 변인에 따른 문장 이해도 평균

	중급학습자	고급학습자
중의성 포함 문장	0.509	0.662
중의성 불포함 문장	0.663	0.780

표 1은 각각의 조건에 해당하는 평균을 보여준다. 이러한 결과가 통계적으로 유의미한 결과인지를 확인하기 위해 LME 모형, 특히 종속 변수가 0 또는 1로 이진 종속 변수를 맞는

로지스틱 혼합 모형으로 분석한다. 모형 선택 과정에서 최대모형에서 변수를 줄여가는 방법과 최소모형에서 변수를 늘여가는 방법 중 Barr 외 3인(2014)을 따라 최대모형에서 변수를 줄여가는 방법을 일반적으로 사용한다. 본 연구의 실험 결과도 최대모형으로 시작하였으나 변수 간 상관성이 높게 나와 과적합된 모형들에서 임의변수의 기울기를 하나씩 줄여나갔다<sup>2</sup>. 고정변수는 실험 설계 목적에 따라 주효과와 상호작용효과까지 포함한 공식으로 사용하였다. 모형 선택은 anova 함수를 이용하여 모형의 변수가 유의미한지를 검정하였고, 가장 설명력이 높은 모형(AIC 또는 BIC 값이 낮은 모형)인 glmer(ACC~Typec\*Profc+(1|Subject)+(1+Typec|Item),data,family="binomial")을 최종 모형으로 선택하였다(모형 선택 과정의 여러 모형의 자세한 공식은 부록 참조). 결과 값을 보고하기 전에, 최종 모형으로 선택된 모형의 가정을 검증해야 한다. 검증 과정은 다음 장에서 자세히 설명하기로 하고, 가정을 위반하지 않은 최종 모형의 결과는 표 2와 같다.

표 2. 문장 이해도 결과

고정효과 (절편)	추정값	표준오차	z값	p값	유의성
문장종류	-0.941	0.393	-2.474	0.013	*
영어숙달도	1.022	0.672	1.401	0.161	
상호작용	-0.471	0.423	-2.416	0.016	*
		0.666	-0.707	0.480	

Note. \* $p < 0.05$

표 2를 보면, 고정효과 결과가 나타나 있는 것을 알 수 있다. 실험 목적에 따라 다르겠지만, 보통 실험언어학 결과는 실험 변인에 따른 결과가 중요하기 때문에, 임의효과 결과는 보고하지 않는다. 같은 이유로 고정효과의 절편값도 해석하지 않고, 실험 변인인 문장종류의 주효과, 영어 숙달도의 주효과 및 문장종류와 영어 숙달도의 상호작용효과를 해석한다<sup>3</sup>. 문장종류의 기울기 추정값은 -0.941로 문장종류의 레벨인 중의성 불포함 문장(unambiguous) 조건을 기준 레벨(-0.5)로, 중의성 포함 문장(ambiguous)를 비교 레벨(0.5)로 효과대비 코딩을 하였기 때문에, 중의성 불포함 문장 조건이 이해도 결과가 더 높지만, 이 값은 통계적으로 유의미하지 않다는 것을 알 수 있다. 그리고 영어 숙달도의 기울기 추정값은 1.022로 중급 학습자는 -0.5, 고급 학습자는 0.5로 효과대비 코딩을 하였으므로, 고급 학습자 그룹이 이해도 결과가 더 높고, 이 값은 0.05 수준에서 유의미하다는 것을 알 수 있다. 그리고 상호작용 효과는 통계적으로 유의미하지 않았다고 해석할 수 있다.

<sup>2</sup> Barr 외 3인(2013)에 따르면, (a)임의효과 사이의 상관관계를 제거하기, (b) 분산의 가장 적은 부분을 야기하는 임의효과를 제거하기, 또는 (c)가장 복잡한 상호작용에 해당하는 임의기울기를 제거하여 수렴할 수 있도록 모형을 단순화할 필요가 있다.

<sup>3</sup> 본 연구에서는 대비코딩인 효과코딩을 사용하였기 때문에, 고정변인의 기울기 추정값을 주효과로 바로 해석할 수 있다. 만일 처리대비인 터미 코딩인 사용할 경우 기울기 추정값으로 주효과를 바로 알 수 없음에 유의해야 한다.



### 3.2 반응/읽기 시간 결과 분석

심리언어학 실험 연구에서 많이 사용하는 종속 변수 중 하나는 반응 또는 읽기 시간이다. 반응 시간을 밀리초(ms) 단위로 해서 측정하여 실시간 언어처리의 양상을 파악한다. 우선, 일반적인 RT분석은 이해도 측정 문항에서 정답으로 응답한 문항만 포함한다. 그래서 오답은 분석에서 제외하고, 얼마나 제거 되었는지 보고 한다. 또한, RT의 LME 모형 분석에 필요한 데이터를 준비하는데 있어서 중요한 사항은 특이값의 처리이다. 기존 연구 방식을 참조하여 미리 정한 특정값보다 큰 수나 작은 수를 특이값으로 처리하는 방식과 피험자 개개인에 따른 특이값을 처리하는 방식, 두 가지 모두가 실험 연구에서 자주 사용되는데, 본 연구에서는 두 가지 방식 모두 사용하여 분석해보도록 한다.

본 연구의 피험자 정답률을 확인해보면, 80%이상 정답률을 보인 피험자는 30명중 6명이 었다. 보통의 반응시간 연구에서 이해도 측정은 피험자가 실험에 집중했는지 여부를 확인하는 용으로 많이 사용된다. 그래서 보통 70% 또는 80%보다 낮은 정답률을 보인 피험자의 결과와 오답으로 기록된 항목은 분석에서 제외하는 것이 일반적이다. 하지만, 연구 주제가 이해도 측정에 관한 것일 경우, 정답률이 낮은 피험자를 분석에서 제외하지 않는다. 본 연구의 분석에서도 오답률이 높은 피험자의 데이터 전체를 제외하지 않고, 대신 오답으로 기록된 항목을 RT분석에서 제외하였다. 분석에서 제외된 오답 항목의 비율은 35.28%으로, 나머지 64.72%의 정답 항목만 RT를 분석하였다. 그리고 많은 RT연구에서 기준값(cut-off value)으로 정한, 2000ms 초과, 200ms 미만의 RT와 각 피험자의 RT평균의  $\pm 2.5$ 표준편차를 기준값으로 정하고 그 큰 기준값보다 더 크거나 작은 기준값보다 더 작은 특이값은 분석에서 제외하였다(특이값을 제외하는 R 코드는 부록 참조). 이렇게 제외된 항목은 정답 항목의 14.59%이다. 특이값을 제외한 데이터만 RT 데이터셋으로 남기고, 그 데이터셋으로 LME 모형 분석을 한다. 이렇게 특이값을 제외한 목표구역( $p$ )의 읽기 시간 평균은 표 3과 같다.

표 3. 목표구역 ( $p$ )의 읽기 시간 평균

	중급학습자	고급학습자
중의성 포함 문장	665.549	640.066
중의성 불포함 문장	567.837	582.824

또한, 문장 읽기 분석에 필요한 데이터를 준비하는데 있어서 중요한 것은 단어별로 길이가 다르기 때문에 길이를 고려한 RT 분석이 중요하다. 종속 변수 RT는 단어 길이를 고려한 RT의 잔차(Residual)나 log변형을 한 logRT의 잔차를 많이 사용한다(Jaeger, 2009)<sup>4</sup>.

<sup>4</sup> Wurm와 Fisiaro(2014)는 잔차 사용을 비판하기도 하였다. 잔차 대신 단어 길이를 변인으로 모델에 넣어서 사용하는 방법도 있다.

RT의 잔차는 예측 변수로 단어 길이를 변인으로 넣은 LME 모형과 residuals 함수를 이용하여 구할 수 있다(`len <- lmer(RT ~ WLength + (1+WLength|Subject), data=raw_data); raw_data$RRT <- residuals(len)`). 잔차를 구하기 위한 모형 공식에서 사용한 데이터셋은 raw\_data<sup>5</sup>로, 목표 문장 외에 덤(filler) 문장까지 포함한 데이터를 사용하는 것을 추천한다. 왜냐하면, 목표 단어의 RT뿐만 아니라 다른 단어의 RT도 잔차를 구하는 모형에 사용하면 데이터의 크기가 커져서 피험자 RT의 잔차를 정확하게 구할 수 있기 때문이다.

그리고 데이터의 분포를 정규 분포에 가깝게 하기 위해 데이터 변형(transformation)을 하는데, RT 데이터를 log변형한 logRT의 잔차를 분석의 종속 변수로 이용하기도 한다. RT의 잔차 모형과 마찬가지로 logRT의 잔차는 단어 길이를 변인으로 넣은 LME 모형을 이용하여 구할 수 있다(`len2 <- lmer(log(RT) ~ WLength + (1+WLength|Subject), data=raw_data); raw_data$logRRT <- residuals(len2)`). 그리고 정규성을 확인하는 방법으로는 목표 단어 구역만 모은 RT.data 데이터셋을 이용하여 `qqmath(~RT, RT.data); qqmath(~RRT, RT.data); qqmath(~logRRT, RT.data)`로 Normal QQ plot을 출력하여 확인할 수 있다. 또한, `densityplot(~RT, RT.data); densityplot(~RRT, RT.data); densityplot(~logRRT, RT.data)`도 사용하여 확인할 수 있다. 두 그래프를 비교하면 logRT의 잔차가 좀 더 정규분포에 가깝게 되는 것을 확인할 수 있다(모형 진단에 대해서는 아래 그림 3과 다음 장의 설명 참조).

목표구역(p)의 logRT의 잔차에 대한 혼합 모형을 구성하는 과정에서 오답 관찰치를 제거하면서 조건 당 데이터의 수가 동일하지 않게 된다. 이러한 경우, 고정변인을 효과대비 코딩을 하기 위해 중심화하는 함수 myCenter(Jaeger, 2009)를 사용하면 더욱 정확하게 분석할 수 있다. 정답률 분석에서 사용하였던 방식으로 중심화된 상호작용 효과(cType\*cProf)까지 포함한 고정효과와 피험자의 임의기울기로 cType을 포함하고, 항목의 임의기울기로 cType\*cProf을 포함한 최대 모형에서부터 변인을 제거하면서 모형을 선택하였다. 임의기울기에 복잡한 변인이 포함된 모형들의 결과는 과적합되어 최종 모형으로 선택된 모형은 `logRRTmodel.p <- lmer(logRRT ~ cType*cProf + (1|Subject) + (1|Item), data=RT.p)`이다. 모형 선택 후 반드시 해야 할 일은 모형에 대한 가정을 검토하는 것인데, 모형 진단 그래프를 이용하여 선형성, 잔차의 등분산, 잔차의 정규성을 검토한다. 우선, `plot(model)`의 결과는 그림 3과 같다.

<sup>5</sup> data\_raw 데이터셋은 <http://bitly.kr/LMEdata>에서 data\_raw.csv 다운로드 후 사용 가능

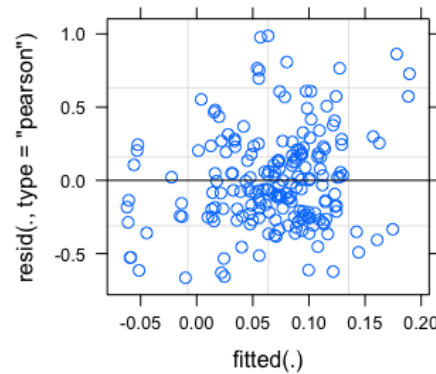


그림 3. 목표구역 (p)의 모형 진단

그림 3에서 보면, 잔차 0을 중심으로 한 직선에는 별다른 특정 구조가 보이지 않는다. 각각의 잔차가 분포되어 있는 것도 점들의 산포가 점차 증가하거나 감소하지 않고 일정하다고 할 수 있으므로 선형성과 잔차의 등분산 가정을 위반하지 않았다고 할 수 있다. 그리고 잔차의 정규성은 `x<-resid(model); qqnorm(x); qqline(x, col = 2)`로 확인할 수 있다. Normal Q-Q 그래프는 그림 4와 같다.

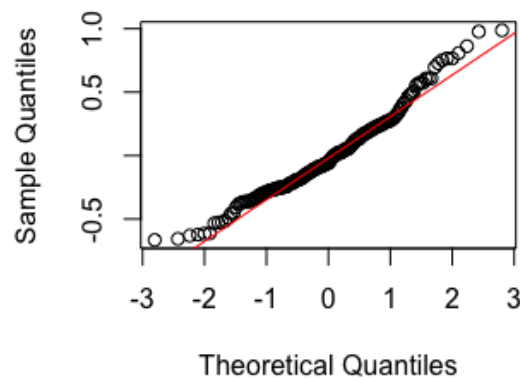


그림 4. 목표구역 (p)의 LME모형의 logRT 잔차의 정규성 진단

그림 4에서 알 수 있듯이, 잔차는 직선에 가깝게 분포되어 있다. 직선에서 벗어난 잔차의 구성은 정규성 가정의 위반을 나타낸다. 그리고 잔차의 독립성 가정은 보통 실험의 피험자를 무작위로 표본을 모집하기 때문에 문제가 되지 않는다. 가장 문제가 될 수 있는 가정인 잔차의 다중공선성(multicollinearity)<sup>6</sup>의 불출현 가정을 검정하기 위해

$\sqrt{\text{vif}(\text{ACCfinal.model})} > 2$ 를 사용하는데, VIF(Variance Inflation Factor, 분산팽창인수)의 루트 값이 2이상일 경우 다중공선성이 발생할 수 있기 때문이다. 선택한 최종 모델은 다중공선성 불출현 가정도 위반하지 않았다. 이렇게 가정 검증을 통해서 최종적으로 선택된 모형의 결과는 표 4와 같다.

표 4. 목표구역 (p)의 읽기 시간 결과

고정효과 (절편)	추정값	표준오차	자유도	t값	p값
문장종류	0.049	0.023	11.207	2.173	0.052
영어숙달도	-0.005	0.039	195.236	-0.121	0.904
상호작용	-0.021	0.044	27.904	-0.484	0.632
	0.040	0.079	90.436	0.502	0.617

표 4를 보면, 고정변인의 효과는 무의미하다는 것을 알 수 있다. 즉, 문장종류에 따라 목표구역 (p)의 읽기 시간은 유의미한 차이가 없었고, 영어 숙달도 및 문장종류와 영어 숙달도의 상호작용도 없었다는 것을 알 수 있다. 문장 종류의 기울기 추정값은 log변형 후 잔차모형을 이용하였기 때문에 기울기를 정확하게 해석하기는 쉽지 않다. 표 3의 값을 이용하여 해석한다면, 중의성 불포함 문장 조건의 상대적으로 높은 RT는 통계적으로 유의미하지 않았다고 해석할 수 있다. 마찬가지로, 중급학습자의 상대적으로 높은 RT도 통계적으로 유의미하지 않았다고 해석할 수 있다<sup>7</sup>.

#### 4. 논의 및 결론

본 연구는 혼합효과 모형을 이용하여 실험언어학에서 자주 사용되는 자기조절읽기 실험 데이터를 분석하는 방법에 대해서 논의하였다. 혼합효과 모형은 실험에서 확인하고자 하는 요인인 고정효과 뿐만 아니라, 피험자와 언어 항목, 이렇게 두 가지 임의변인을 동시에 고려할 수 있기 때문에 각각의 평균값을 이용한 ANOVA에 비해 데이터 손실이 크지 않다는 장점이 있고, ANOVA 분석으로 적당하지 않은 이분적 데이터도 로지스틱혼합모형(logit mixed effect model)로 분석이 가능하다는 점에서 매우 유용하다. 그 외 수치형 변인도 포함할 수 있고, 무응답을 포함한 데이터나 코퍼스와 같이 조건당 불균형적인 데이터에도 사용할 수 있다는 장점이 있으므로 다양한 언어 데이터에 사용가능한 모형이다. 혼합효과모형으로 분석할 때 주의해야 할 점인 효과대비 코딩, 모형 선택, 모형 가정 점검 등을 자기조

<sup>6</sup> 다중공선성은 독립변수들 간에 강한 상관관계가 나타나는 문제를 일컫는다.

<sup>7</sup> 목표구역 (v)의 RT 결과도 목표구역 (p)와 유사하게 같은 모형 선택 과정을 통해서 분석하였는데, 목표구역 (p)의 결과와 비슷하게 목표구역 (v)의 RT에도 문장 종류와 영어 능숙도의 영향이 없었다. 그리고 보통의 자기조절읽기 실험에서는 목표구역 다음에 나오는 영역을 spill-over 구역이라 하여 그 구역도 같은 분석 과정을 통하여 변인의 영향을 확인하는 것이 일반적이다.

절읽기 실험의 문장 이해도 측정을 위한 정오답(Accuracy) 데이터 분석 및 실시간 언어 처리 시간 측정을 위한 읽기 속도(RT) 데이터를 예를 들어 논의하였다.

특히, 모형 선택에 있어서 주의할 점인, 최적의 모형을 선택하는데 주의를 기울일 필요가 있다. 앞서 정답률 분석에서 항목의 임의 기울기를 포함하는 모형으로 최종 모형을 선택하였는데, 그 모형 대신 단순 모형을 선택하였다면 다른 결과를 얻을 수 있기 때문이다. 문장 이해도의 단순 모형(`model <- glmer(ACC~Typec*Profc + (1|Subject)+(1|Item), data,family="binomial")`)을 선택했을 때의 결과는 표 5와 같다.

표 5. 다른 모형(단순 모형)을 잘못 선택시 문장 이해도 결과

고정효과 (절편)	추정값	표준오차	z값	p값
문장종류	0.784	0.320	2.454	0.014
영어숙달도	-0.809	0.264	-3.063	0.002
상호작용	0.751	0.314	2.391	0.017
	-0.486	0.591	-0.822	0.411

표 5를 보면, 문장종류의 효과가 유의미하다는 다른 결과를 볼 수 있다. 단순 모형을 잘못 선택했을 경우, 결과를 잘못 해석할 수 있는 오류를 범할 수 있기 때문에 더욱 주의를 기울일 필요가 있다. 항목의 임의기울기에 문장종류를 포함한 최종 모델에서 문장종류 변수는 항목에 따른 변이성을 포착한 임의기울기에 반영이 되어야 했다. 만일 그렇지 않았다면, 고정효과와 결과로 잘못 해석하여 결과를 완전히 다르게 보고하게 되는 것이다.

읽기 시간(RT) 데이터 분석에 있어서도 주의를 기울여야 할 것은 모형 가정을 만족시키기 위한 점검이 꼭 필요하며, 상황에 따라서는 가정을 충족시키기 위해 데이터의 변형이 필요하다는 것이다. log 변형한 RT의 잔차를 이용하지 않을 경우, 즉 측정한 RT 데이터와 단어 길이를 고려한 RT의 잔차 데이터의 정규성을 진단하면 그림 5와 같다.

그림 5. 목표구역 (p)의 RT 및 단어 길이를 고려한 RT의 잔차의 정규성 진단

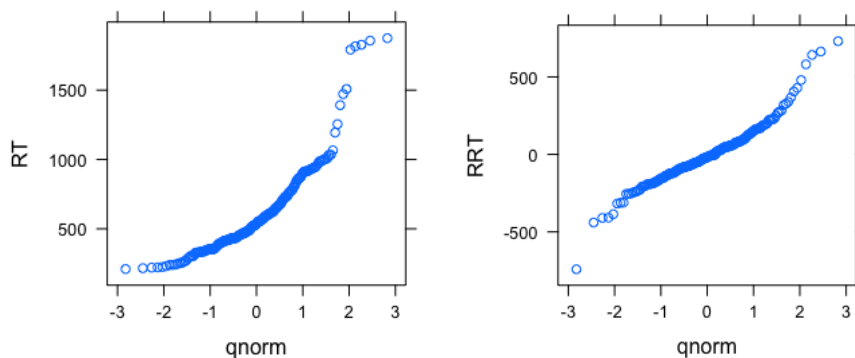


그림 5의 왼쪽 그래프는 RT값을 그대로 사용한 데이터로, 그 값은 곡선을 이루고 있는 것처럼 보인다. 단어 길이를 고려한 RT 잔차를 이용할 경우인 오른쪽 그래프는 RT값 그대로 사용한 경우보다는 직선에 가깝다. 물론, 1500ms 이상의 데이터를 특이값으로 처리하는 방법도 가능하다. 하지만, 정확한 RT 분석을 위해서 단어 길이를 고려하여 RT 잔차를 이용하거나, RT의 log 변형의 잔차를 이용하는 방식을 병행할 필요가 있다.

Journal of Memory and Language와 같은 국제 저명 저널은 언어 실험 결과를 분석할 때 혼합효과모형 선택을 요구하고 있을 정도로, 혼합효과모형이 언어 실험 결과를 해석하는데 매우 설명력이 높다고 할 수 있다. 하지만, 아직 국내의 실험 언어학 분야에서는 여전히 혼합효과모형에 대한 이해가 부족한 실정이다. 모형 선택 과정이나 가정의 점검 등에 대한 명확한 기준이 정립되고, 이 부분에 대한 심도 깊은 연구가 필요하다.

## 참고문헌

- 이준규(Lee, J.). 2011. 반응시간을 이용한 제2 언어 어휘의 연결강도 추정(Estimating the strength of second language word association with reference to reaction times). *《언어학》(Journal of the Linguistic Society of Korea)* 61, 243-261.
- 이준규(Lee, J.). 2016. 제2 언어 머릿속 사전에 관한 심리언어학적 탐색(A psycholinguistic inquiry of the second language mental lexicon). *《언어학》(Journal of the Linguistic Society of Korea)* 74, 51-70.
- 최재웅·홍정하(Choe, J-W. and J. Hong). (역). 2013. 『언어학자를 위한 통계학-R 활용』(*Statistics for Linguistics with R: A Practical Introduction*). 고려대학교출판문화원(Korea University Press).
- Baayen, R. H., Davidson, D. J., and D. M. Bates. 2008. Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language* 59(4), 390-412.
- Barr, D. J., R. Levy, C. Scheepers and H. J. Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language* 68(3), 255-278.
- Bates, D., M. Maechler, B. Bolker and S. Walker. 2019. lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1-21. <http://CRAN.R-project.org/package=lme4>
- Clark, H. H. 1973. The language-as-fixed-effect fallacy: A critique of language statistics in psychological research. *Journal of Verbal Learning and Verbal Behavior* 12(4), 335-359.

- Drummond, A. 2013. *Ibex Farm*. Retrieved from <http://spellout.net/ibexfarm>
- Gelman, A. and J. Hill. 2007. *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.
- Gries, S. 2013. *Statistics for Linguistics with R: A Practical Introduction*. Berlin: De Gruyter.
- Jaeger, F. 2008. Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of Memory and Language* 59, 434–446.
- Jaeger, F. 2009. Centering several variables. Retrieved from <https://hlplab.wordpress.com/2009/04/27/centering-several-variables>
- Jaeger, F. 2008. Modeling self-paced reading data: Effects of word length, word position, spill-over, etc. Retrieved from <https://hlplab.wordpress.com/2008/01/23/modeling-self-paced-reading-data-effects-of-word-length-word-position-spill-over-etc/>
- Juffs, A. 1998. Main verb versus reduced relative clause ambiguity resolution in L2 sentence processing. *Language Learning* 48(1), 107–147.
- Kuznetsova A., P. B. Brockhoff and R. H. B. Christensen. 2019. lmerTest: Tests in Linear Mixed Effects Models. R package version 3.1-0. <http://CRAN.R-project.org/package=lmerTest>
- Lee, J-H. and J-A. Shin. 2016. Syntactic reanalysis and lingering misinterpretations in L2 sentence processing. *Linguistic Research* 33(S). 53–79.
- Marsden, E., Thompson, S., and L. Plonsky. 2018. A methodological synthesis of self-paced reading in second language research. *Applied Psycholinguistics* 1–44.
- Matuschek, H., R. Kliegl, S. Vasishth, H. Baayen and D. Bates. 2017. Balancing Type I error and power in linear mixed models. *Journal of Memory and Language* 94, 305–315.
- R Core Team. 2014. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.R-project.org>
- Rah, A. and D. Adone. 2010. Processing of the reduced relative clause versus main verb ambiguity in L2 learners at different proficiency levels. *Studies in Second Language Acquisition* 32(1), 79–109.
- Sarkar, D. 2018. lattice: Trellis Graphics for R. R package version 0.20–38. <http://CRAN.R-project.org/package=lattice>
- Seo, H-J. and J-A. Shin. 2016. L2 processing of English pronouns and reflexives: An

eye-tracking study. *Korean Journal of English Language and Linguistics* 16(4), 879-901.

Winter, B. 2013. Linear models and linear mixed effects models in R with linguistic applications. arXiv:1308.5499 [<http://arxiv.org/pdf/1308.5499.pdf>]

Wurm, L. H. and S. A. Fisicaro. 2014. What residualizing predictors in regression analyses does (and what it does not do). *Journal of Memory and Language* 72, 37-48.

예시 언어(Examples in): 영어(English)

적용가능 언어(Applicable Language): 영어(English)

적용가능 수준(Applicable Level): 대학 이상 성인(Tertiary)

신정아(Shin, Jeong-Ah)

교수(Professor), 동국대학교(Dongguk Univ.)

영어영문학부(Division of English)

서울특별시 중구 필동로1길 30 (30 Pildong-ro-1-gil, Jung-gu, Seoul)

Tel: 02) 2260-3167

E-mail: jashin@dongguk.edu, jashin@gmail.com

논문 접수(Received): 2019년 2월 14일

논문 수정(Revised): 2019년 3월 10일

게재 확정(Accepted): 2019년 3월 19일



## 부록

## 혼합효과분석 R 스크립트

```

library(lme4); library(lmerTest)
data<-read.csv(file=file.choose(), sep = ',', header = T, fill=T) #data는
http://bitly.kr/LMEdata에서 data_p.csv 다운로드 후 파일 불러오기
data$Subject = factor(data$Subject); data$Item = factor(data$Item)
data$Proficiency <- factor(data$Proficiency, levels=levels(data$Proficiency)[c(2,1)])
attach(data)
mean(age); range(age)
##### Accuracy    분석
ACCtable <- tapply(ACC, list(Type, Proficiency), mean); ACCtable
ACCbar <- barplot(ACCtable, ylab="Accuracy", ylim=c(0, 1), col=rainbow(2),
legend=rownames(ACCtable), args.legend = list(x = 'topright', bty='n', inset=c(-.3,-.4)),
beside=T)

data$Typec <- ifelse((data$Type == 'unambiguous'), -.5, .5) #coding
data$Profc <- ifelse((data$Proficiency == 'lo'), -.5, .5)

#Model selection (모형 안에 함수는 수렴 문제를 일으키지 않기 위해 넣는 기능으로 사용하였
다.)

m.full=glmer(ACC~Typec+(1+Typec| Subject)+(1+Typec*Profc| Item),data,family="binomial"
,control=glmerControl(optimizer="bobyqa",optCtrl=list(maxfun=1e6)))
summary(m.full) #singular fit이므로 over-fitted되서 선택하지 않음

m.subj=glmer(ACC~Typec+(1+Typec| Subject)+(1+Profc| Item),data,family="binomial",contro
l=glmerControl(optimizer="bobyqa",optCtrl=list(maxfun=1e6)))
summary(m.subj) #singular fit이므로 over-fitted되서 선택하지 않음

m.item.full=glmer(ACC~Typec*Profc+(1| Subject)+(1+Typec*Profc| Item),data,family="binom
ial",control=glmerControl(optimizer="bobyqa",optCtrl=list(maxfun=1e6)))
summary(m.item.full) #singular fit이므로 over-fitted되서 선택하지 않음

m.item.1=glmer(ACC~Typec*Profc+(1| Subject)+(1+Typec+Profc| Item),data,family="binomia
l",control=glmerControl(optimizer="bobyqa",optCtrl=list(maxfun=1e6)))
m.item.2=glmer(ACC~Typec*Profc+(1| Subject)+(1+Profc| Item),data,family="binomial",contr
ol=glmerControl(optimizer="bobyqa",optCtrl=list(maxfun=1e6)))
anova(m.item.1, m.item.2) #모델 비교후 1 선택

m.item.3=glmer(ACC~Typec*Profc+(1| Subject)+(1+Typec| Item),data,family="binomial",cont
rol=glmerControl(optimizer="bobyqa",optCtrl=list(maxfun=1e6)))
anova(m.item.1, m.item.3) #1, 3 모델 비교후 3번 선택
anova(m.item.2, m.item.3) #2, 3 모델 비교후 3번 선택

```

```

m.item.0=glmer(ACC~Typec*Profc+(1|Subject)+(1|Item),data,family="binomial",control=glmerControl(optimizer="bobyqa",optCtrl=list(maxfun=1e6))) #단순 모형
anova(m.item.0, m.item.3) #2, 3 모델 비교후 3번 선택
ACCfinal.model <- m.item.3 #최종 모델 3번 선택
summary(ACCfinal.model)

##### 모형 가정 검증
plot(ACCfinal.model) #1.잔차 분산의 동질성
x<-resid(ACCfinal.model); qqnorm(x); qqline(x, col=2) #2.잔차 정규분포
library(car); sqrt(vif(ACCfinal.model))>2 #3.잔차의 공선성(collinearity)의 불출현: 예측변수들의 상호관련성이 없어야한다. 값이 2이상이면 다중공선성 문제

#####RT 분석
RTdata <- subset(data, ACC == 1)
(nrow(data)-nrow(RTdata))/nrow(data) #오답으로 버려진 데이터 비율
RTdata$RT<-as.numeric(as.character(RTdata$RT))
RTda= subset(RTdata, 200<=RT & RT<=2000) # 절대적인 Outlier 제거
#RTda= subset(RTdata, 100<=RT & RT<=2000) # 절대적인 Outlier 제거
(nrow(RTdata)-nrow(RTda))/nrow(RTdata) #제거된 outlier 데이터 비율

## relative outlier 제거하기 (https://github.com/jasongullifer/Simon-Script 참조)
splitlist = split(RTda$RT, list(RTda$Subject),drop = TRUE)
cuthigh = lapply(splitlist, function(x){mean(x)+2.5*sd(x)}) #gives high cutoffs per subject
cutlow = lapply(splitlist, function(x){mean(x)-2.5*sd(x)}) #gives high cutoffs per subject
RTda$cuthigh = unsplit(cuthigh,list(RTda$Subject))
RTda$cutlow = unsplit(cutlow,list(RTda$Subject))
RT<-RTda[(RTda$RT<=RTda$cuthigh & RTda$RT>=RTda$cutlow), ]
(nrow(RTdata)-nrow(RT))/nrow(RTdata) #outlier으로 버려진 데이터 비율

RTmean <- aggregate(RT~Type+Proficiency, data=RT, mean); RTmean
RTsd <- aggregate(RT~Type+Proficiency, data=RT, sd); RTsd

####centering function
myCenter = function(x) x - mean(x, na.rm=T)
emplog = function(p, totalN) {
  y = p*totalN
  return(log( (y+.5)/ (totalN-y + .5) ))
}
emplogweight = function(p, totalN) {
  y = p*totalN
  return(1 / ( 1 / (y+.5) + 1 / (totalN-y + .5) ))
}

RT$cType = myCenter(ifelse(RT$Type == "unambiguous", -.5, .5))
RT$cProf = myCenter(ifelse(RT$Proficiency == "lo", -.5, .5))

#정규성 확인 및 비교

```

```
qqmath(~RT, RT)
qqmath(~RRT, RT)
qqmath(~logRRT, RT)
densityplot(~RT, RT)
densityplot(~RRT, RT)
densityplot(~logRRT, RT)

#모형 선택
logRRTmodel.full<-lmer(logRRT~cType*cProf + (1+cType| Subject)
+ (1+cType*cProf| Item), data=RT,
control=lmerControl(optimizer="bobyqa",optCtrl=list(maxfun=1e6))) #singular fit 이므로 선택
에서 제외

logRRTmodel.1<-lmer(logRRT~cType*cProf + (1+cType| Subject)
+ (1+cType+cProf| Item), data=RT,
control=lmerControl(optimizer="bobyqa",optCtrl=list(maxfun=1e6))) #singular fit 이므로 선택
에서 제외

logRRTmodel.item<-lmer(logRRT~cType*cProf + (1|Subject) + (1+cType| Item), data=RT,
control=lmerControl(optimizer="bobyqa",optCtrl=list(maxfun=1e6))) #singular fit 이므로 선택
에서 제외

logRRTmodel.item.1<-lmer(logRRT~cType*cProf + (1|Subject) + (1+cProf| Item), data=RT,
control=lmerControl(optimizer="bobyqa",optCtrl=list(maxfun=1e6))) #singular fit 이므로 선택
에서 제외

logRRTmodel.simple<-lmer(logRRT~cType*cProf + (1|Subject) + (1|Item), data=RT,
control=lmerControl(optimizer="bobyqa",optCtrl=list(maxfun=1e6)))

summary(logRRTmodel.simple) #모델 선택

####Model assumption 점검
library(car)
attach(RT)
model <- logRRTmodel.simple; plot(model) #1. 선형성 linearity 및 2. 잔차 분산의 동질성
(homogeneity), 등분산성(homoscedasticity)

x<-resid(model); qqnorm(x); qqline(x, col=2) #3. 잔차 정규성

sqrt(vif(model))>2 #4. 잔차의 공선성(collinearity)의 불출현: 예측변수들의 상호관련성이 없어
야한다. 2이상이면 다중공선성 문제이나, 문제 없음.
```