ISSN 2635-4799

# 교육연구와 실천

제10권 제1호 / 2019. 8

성균관대학교 사범대학 교육정책연구원

교육연구와 실천
Journal of Educational Study and Practice
제10권 제1호, 2019, pp. 1-18

# Exploring Bayes Factor Estimates, Type I error rates, and Statistical Power Rates of Bayesian hypothesis testing for mediation

Namwook Koo
(Korea Institute for Curriculum and Evaluation)
Jihye Jeon*
(Sungkyunkwan University)

< Abstract >

Purpose: The purpose of this study was to investigate the validity of the the default Bayesian hypothesis test for mediation which proposed by Nuijten et al. (2015) with Bayes factor estimates, Type I error, and statistical power rates.

Design/methodology/data/approach: This Monte Carlo simulation study investigated estimated Bayes factor with the varied mediation and sample sizes, and compared Bayesian hypothesis test with four different Frequentist tests (Sobel' s test, asymmetric distribution method, bias-corrected bootstrap method) to confirm the validity of the default Bayesian hypothesis test.

Findings/Results: The results showed that when there was no mediation, estimated Bayes factor offered evidence in favor of the null hypothesis that there is no mediation. With medium or large mediation and sample sizes, estimated Bayes factor offered evidence for the alternative hypothesis that there is mediation. However, with small sample size and mediation, estimated Bayes factor did not offered evidence for the alternate hypothesis. Also, the results of Type I error and statistical power rates showed that bias-corrected bootstrap performed best, whereas Bayesian hypothesis test and Sobel' s test produced very conservative Type I error rates and low statistical power rates.

Value: This study found some validity issues of the default Bayesian hypothesis test with small mediation and sample size. Thus, this study suggests using the default Bayesian hypothesis test with large mediation and sample size.

Key words: mediation analysis, default Bayesian hypothesis test, Bayes factor

* Corresponding author: Jihye Jeon(wisdom83@skku.edu)

교육연구와 실천
Journal of Educational Study and Practice
제10권 제1호, 2019, pp. 1-18

# 매개효과 연구에서 베이지안 가설 검정의 베이즈 인자 추정치, 일종오류 및 통계적 검정력 탐색

구남욱(한국교육과정평가원)
전지혜(성균관대학교)*

< 요 약 >

매개효과 연구는 사회과학 분야의 실험 및 비실험 연구에서 유용하게 사용되는 통계 분석방법 중 하나이다. 최근에는 매개효과 연구에서 베이지안 방법이 소개되었으며, Nuijten 외(2015)는 베이지안 가설 검정 방법을 제안하였다. 베이지안 가설 검정 방법의 장점은 베이즈 인자를 통해서 지지하는 영가설 또는 대립가설에 대한 증거를 수량화할 수 있다는 점이다. 이 모의실험 연구에서는 매개효과와 표본의 크기에 따른 베이즈 인자 추정치의 변화를 살펴보았으며, 또한 디폴트 사전분포를 사용한 베이지안 가설 검정 방법의 타당성을 확인하기 위해서 베이지안 가설 검정 방법과 4 종류의 빈도주의적 방법들(Sobel의 검정, 비대칭분포방법, 편향이 수정된 붓스트랩 방법)을 비교하였다. 그 결과 베이즈 인자 추정치는 매개효과가 존재하지 않을 때, 매개효과가 존재하지 않는다는 영가설에 대한 증거를 제공하였다. 또한 매개효과와 표본의 크기가 중간정도이거나 클 때에도 매개효과가 존재한다는 대립가설에 대한 증거를 제공하였다. 그러나 매개효과와 표본의 크기가 작을 때에는, 매개효과가 존재한다는 대립가설에 대한 증거를 제공하지 못하였다. 또한, 일종오류 비율과 통계적 검정력 비율의 결과에서는 편향이 수정된 붓스트랩 방법이 가장 우수한 결과를 산출하였고, 반면 베이지안 가설검정 방법과 Sobel의 검정 방법은 상대적으로 매우 보수적인 일종오류 비율과 낮은 통계적 검정력 비율을 산출하였다. 따라서 이 연구의 결과에 따르면, 매개효과 연구에서 베이지안 가설 검정에 있어서 Nuijten 외(2015)의 디폴트 베이지안 가설 검정 방법은 매개효과와 표본의 크기가 클 때에만 적용하는 것이 바람직하며, 디폴포 베이지안 가설검정 방법에 대한 후속 연구가 필요하다.

**주제어: 매개효과연구 , 디폴트 베이지안 가설검정, 베이즈 인자**

* 교신저자: 전지혜(wisdom83@skku.edu)

# I. 서론

매개효과 연구는 사회과학 분야의 실험 또는 비실험 연구에서 활발히 수행되고 있다. 매개효과는 두 경로계수를 곱하여 추정할 수 있으며, 이러한 매개효과 추정치는 정규분포를 따르지 않는 특성이 있기 때문에, 방법론적인 측면에서 매개효과를 추정하거나 검정하는 새로운 방법을 개발하거나 기존의 방법들을 비교하는 다양한 연구가 수행되고 있다(MacKinnon, Lockwood, Hoffman, West, & Sheets, 2002; MacKinnon, Lockwood, & Williams, 2004; MacKinnon, Fritz, Williams, & Lockwood, 2007; Nuijten, Wetzels, Matzke, Dolan, & Wagenmakers, 2015; Pituch, Stapleton, & Kang, 2006; Pituch & Stapleton, 2008; Preacher, 2011; Yuan & MacKinnon, 2009). 예를 들면, Sobel(1982; 1986)은 다변량 델타(multivariate delta) 방법을 이용하여 매개효과의 분산 추정치를 교정하였고, MacKinnon 외(2002; 2004; 2007)는 비대칭분포(asymmetric distribution)를 활용하여 매개효과를 추정하는 방법을 제안하였으며, Bollen과 Stine(1990), Shrout과 Bolger(2002)는 붓스트랩(bootstrap)으로 매개효과를 추정하는 방안을 제안하였다. 이러한 대표적인 매개효과 추정 방법들은 빈도주의 방법(frequentist methods)에 해당한다. 또한 최근 매개효과 연구에서도 베이지안 방법(Bayesian methods)을 활용하는 연구들(Yuan & MacKinnon, 2009; Nuijten et al., 2015)이 발표되었다. 가설검정에 있어서 빈도주의 방법의 가장 큰 단점은 대립가설이 사실이라는 전제 하에 영가설이 발생할 확률을 추정하여, 이 확률이 일반적으로 유의수준(0.05)보다 작을 때, 영가설을 기각하고 대립가설을 채택한다는 사실이다. 즉, 직접적으로 대립가설이 발생할 확률을 추정하지 못하고, 영가설의 발생 확률을 통해서 대립가설을 기각하는 간접적인 방식을 취함으로써 결과 해석 등이 명확하지 않다. 반면에 베이지안 방법은 직접적으로 대립가설이 발생할 확률을 추정한다. 따라서 가설검정에 있어서도 영가설이 발생할 확률과 대립가설이 발생할 확률을 계산하여, 어떤 가설이 발생할 확률이 상대적으로 더 높은지 비교할 수 있다. 매개효과 연구에서 베이지안 방법을 사용하는 장점으로는 사전정보를 사용하여 추정치의 효율성(efficiency)을 증가시킬 수 있다는 점, 작은 표본으로도 추론이 정확하다는 점 등이 있다(Yuan & MacKinnon, 2009). 베이지안 매개효과 연구(Bayesian mediation analysis)에서 Yuan과 MacKinnon(2009)은 매개효과가 존재한다는 가정 하에서 매개효과(mediated effect)의 사후분포(posterior distribution)와 신뢰구간(credible interval)을 추정하는 방안을 제시하였다. 반면, Nuijten 외(2015)는 매개효과의 존재여부에 대한 베이지안 가설 검정(Bayesian hypothesis test) 방안을 제시하였다. 이 연구의 목적은 매개효과의 가설 검정에 있어서 Nuijten 외(2015)가 제안한 베이지안 방법을 소개하고 이 방법과 빈도주의적 방법들을 일종오류 및 통계적 검정력을 통해서 비교해 보는 것이다. 구체적인 연구문제는 다음과 같다.

첫째, 표본과 매개효과의 크기에 따른 다양한 자료로 을 검정하는데 있어서 베이즈 인자 추정치는 어떠한가?

둘째, 표본과 매개효과의 크기에 따른 다양한 자료로 을 검정하는데 있어서 베이지안 방법과 빈도주의 방법(Sobel의 검정, 비대칭분포 방법, 편향이 수정된 붓스트랩 방법)의 일종오류 비율은 어떠한가?

셋째, 표본과 매개효과의 크기에 따른 다양한 자료로 을 검정하는데 있어서 베이지안 방법과 빈도주의 방법(Sobel의 검정, 비대칭분포 방법, 편향이 수정된 붓스트랩 방법)의 통계적 검정력 비율은 어떠한가?

# Ⅱ. 이론적 배경

## 1. 매개효과 모형

### 가. 모집시기별 대학입학전형에 따른 대학생활적응

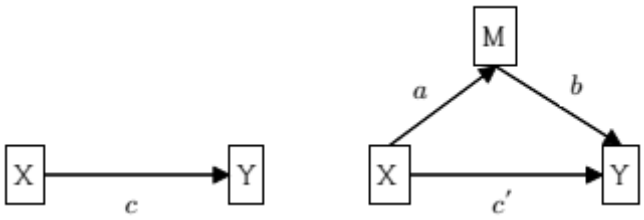
기본 매개효과 모형은 선형 회귀모형으로 나타낼 수 있으며, $X$(독립변인)와 $Y$(종속변인)의 관계는 식 (1)과 [그림 1]에서 $c$로 나타낼 수 있으며, $c$를 매개효과 모형에서는 전체효과(total effect)라고 한다.

$$Y=\beta_1+cX+\varepsilon_1, \quad (1)$$

그리고, $X$와 $Y$의 관계는 $M$에 의하여 매개될 수 있다. 즉, $X$는 $M$에 영향을 미치고, 다시 $M$은 $Y$에 영향을 미치게 된다. 식 (2)에서 $c'$은 $M$을 통제하였을 때, $X$와 $Y$의 관계를 나타내고, $a$는 $X$와 $M$의 관계를, $b$는 $M$과 $Y$의 관계를 나타낸다. 매개효과는 식 (2)와 (3)에서 두 경로 계수인 $a$와 $b$의 곱($=ab$)이며, $Y$에 대한 $X$의 직접효과(direct effect)는 $c'$이다. 그리고 $\varepsilon_1$, $\varepsilon_2$, $\varepsilon_3$은 정규분포를 따르는 잔차를 나타낸다. 또한 전체효과는 직접효과와 매개효과의 합($c=ab+c'$)과 같다.

$$Y=\beta_2+c'X+bM+\varepsilon_2, \quad (2)$$

$$M=\beta_3+aX+\varepsilon_3. \quad (3)$$




[그림 1] 기본적인 매개 모형

## 2. 매개효과연구에서 베이지안 접근 방법

### 가. 베이지안 추론

베이지안 통계에서 불확실성은 확률로 수량화할 수 있다. 또한 모수에 대한 지식은 사전분포(prior distribution: $p(\theta)$)로 나타낼 수 있으며, 이러한 사전분포는 관측된 자료에 의해 업데이트 되어 모수의 사후분포(posterior distribution: $p(\theta|\text{data})$)가 도출된다. 즉, 베이지안 업데이트 과정(Bayesian updating process)에서 연구자는 사전 지식을 토대로 모수의 사전분포를 결정한 후에, 통계적 모형을 적용하여 모수 값($\theta$)들에 따라 자료가 발생할 확률, 즉, 자료의 우도(likelihood of data, $p(\text{data}|\theta)$)를 계산하며, 사전분포는 식 (4)에 제시된 베이즈 규칙(Bayes' rule)에 따라서 사후분포로 업데이트된다(Nuijten et al., 2015; Yuan & MacKinnon, 2009). 베이즈 규칙은 식 (4)와 같다.

$$p(\theta|\text{data}) = \frac{p(\text{data}|\theta)\text{p}(\theta)}{p(\text{data})} \quad (4)$$

위의 식에서 $p(\text{data}) = \int p(\theta)p(\text{data}|\theta)\text{d}\theta$는 정규화 상수(normalizing constant)로 사후분포를 적분한 값이 1이 되게 하는 역할을 한다. $p(\text{data})$는 모수 추정에 영향력이 없기 때문에 식 (4)는 아래의 식 (5)으로 다시 쓸 수 있다.

$$p(\theta|\text{data}) \propto \text{p}(\text{data}|\theta)\text{p}(\theta) \quad (5)$$

위의 식에서 $\propto$는 비례한다는 의미이며, 따라서 식 (6)에서 사후분포는 우도와 사전분포의 곱

에 비례한다는 것을 알 수 있다.

Yuan과 MacKinnon(2009)이 제안한 베이지안 매개효과 추정에서는 식 (2)와 (3)에서 회귀계수들($\beta_2$, $c'$, $b$, $\beta_3$, $a$)과 잔차($\varepsilon_2$, $\varepsilon_3$)의 분산의 사전분포들이 식 (5)을 통하여 사후분포로 업데이트되며, $a$와 $b$의 사후분포를 곱하여 매개효과의 추정치($\hat{a}\hat{b}$)와 95% 신뢰구간(credible interval)을 계산하는 방안을 제안하였다. Yuan과 MacKinnon은 매개효과가 존재한다는 가정 하에, 베이지안 방법을 통하여 매개효과를 추정하는 방법을 제시하였다는 점에서 의미가 있지만, 매개효과를 검정하는 방법에 대해서는 논의하지 않고 후속연구로 제안하였다.

## 나. 매개효과에 대한 베이지안 가설검정

베이지안 가설검정은 두 개의 가설(또는 모형)의 사후확률(posterior probability)을 비교하여, 사후확률이 더 큰 가설을 선택하는 과정이다. 두 개의 가설을 $u_0$와 $u_1$라고 하고, 사전분포는 $p(u_1)$과 $p(u_0)=1-p(u_1)$이고, $p(u_1)/p(u_0)$은 사전 모형의 승산(prior model odds)이다(Nuijten et al., 2015). 사전 모형의 승산은 아래 식 (6)에 의해서 사후 모형 승산으로 업데이트된다.

$$\frac{p(u_1|\text{data})}{p(u_0|\text{data})}=\frac{p(data|u_1)p(u_1)}{p(data|u_0)p(u_0)}. \quad (6)$$

$$\text{Posterior Model Odds} = \text{Bayes Factor} \times \text{Prior Model Odds}. \quad (7)$$

위 식 (7)에서 $p(data|u_1)$과 $p(data|u_0)$는 각각 가설 $u_1$과 $u_0$하에서 data(자료)가 관찰될 확률을 의미한다. 또한, $p(data|u_1)/p(data|u_0)$은 식 (7)에서 베이즈 인자(Bayes Factor)에 해당하며, 자료에 의한 모형의 승산(model odds)의 변화를 나타낸다. 베이즈 인자는 아래의 식 (8)와 같다.

$$BF_{10}=\frac{p(data|u_1)}{p(data|u_0)}. \quad (8)$$

$BF_{10}>1$은 자료가 $u_0$보다 $u_1$에 보다 적합하며, $BF_{10}<1$이면 자료가 $u_1$보다 $u_0$에 보다 적합하다는 의미이다. 즉, $BF_{10}=10$은 자료(data)가 관찰될 확률이 $u_0$보다 $u_1$이 10배 더 높다는 의미이며, $BF_{10}=0.1$은 자료가 관찰된 확률이 $u_1$보다 $u_0$이 10배 더 높다는 의미이다. <표 1>은 베이즈 인자 값에 따른 해석을 제시하고 있으며, 1을 기준으로 1보다 큰 경우와 작은 경우에 각각 대립가설($u_1$)과 영가설($u_0$)을 지지하는 것을 알 수 있다. 따라서 베이즈 인자는 대립가설(alternative hypothesis, $\mu_1$)뿐만 아니라, 영가설(null hypothesis, $\mu_0$)을 채택하는 근거를 양적으로

나타낼 수 있다(Nuijten et al., 2015; Rouder, 2012). 또한 $\mu_0$과 $\mu_1$의 사전확률이 같다고 가정하면, 즉 식 (7)에서 $p(\mu_1)=p(\mu_0)$이면, $p(\mu_1|\text{data})=BF_{10}/(BF_{10}+1)$이 되어 베이즈 인자를 통해서 사후확률을 구할 수 있다(Nuijten et al., 2015).

<표 1> 베이즈 인자 $BF_{10}$에 대한 증거 범주(Jeffrey, 1961: Nuijten et al., 2015에서 재인용)

| 베이즈 인자 $BF_{10}$ | 해석 |
|---|---|
| >100 | $\mu_1$에 대한 극단적인(extreme) 증거 |
| 30-100 | $\mu_1$에 대한 매우 강한(very strong) 증거 |
| 10-30 | $\mu_1$에 대한 강한(strong) 증거 |
| 3-10 | $\mu_1$에 대한 적당한(moderate) 증거 |
| 1-3 | $\mu_1$에 대한 입증되지 않는(anecdotal) 증거 |
| 1 | 증거없음 |
| 1/3-1 | $\mu_0$에 대한 입증되지 않는(anecdotal) 증거 |
| 1/10-1/3 | $\mu_0$에 대한 적당한(moderate) 증거 |
| 1/30-1/10 | $\mu_0$에 대한 강한(strong) 증거 |
| 1/100-1/30 | $\mu_0$에 대한 매우 강한(very strong) 증거 |
| <1/100 | $\mu_0$에 대한 극단적인(extreme) 증거 |

## 다. 선형 모형에서 디폴트 사전분포

베이지안 통계에서 일반선형모형(general linear model)의 사전분포 선택은 중요한 연구주제 중에 하나이며, 연구자가 실질적인 사전정보를 사용할 수 없는 상황에서 선택할 수 있는 사전분포로서 디폴트 사전분포(default prior)에 관한 많은 연구가 수행되어 왔다(Wetzels & Wagenmakers, 2012). 선형 회귀분석에서 가장 널리 적용되는 디폴트 사전분포는 Zellner의 $g$ 사전분포($g$-prior: Zellner, 1986)이다. 이 디폴트 사전분포에서 회귀계수들(regression coefficients, $\beta$)은 정규분포를 따른다. 그리고 오차 분산의 역(inverse of the error variance)인 오차의 정밀도(precision, $\phi$)와 절편(intercept, $\alpha$)은 Jeffreys의 사전분포인 균일분포(uniform distribution)를 따르게 된다.

$$p(\beta|\phi,g,X)=N(0,\frac{g}{\phi}(X^TX)^{-1}) \quad (9)$$

$$p(\phi,\alpha)\propto\frac{1}{\phi}$$

위의 식 (9)에서 $X$는 독립변수들의 행렬을 의미하며, $\phi$는 오차의 정밀도이다. 식 (9)에 따르

면, 자료에서 $\beta$에 관한 정보는 $\frac{1}{\phi}(X^TX)^{-1}$이며, $g$는 $\beta$의 사전분포의 정보를 통제하는 역할을 한다. 즉, $g=1$이면, 사전분포는 관찰된 자료와 같은 가중치를 갖게 되고, $g=3$이면, 1/3의 가중치를 갖게 된다. 따라서 $g$의 선택에 관한 많은 선행연구가 있었다(Wetzels & Wagenmakers, 2012). 예를 들면, $g=n$($n$: 표본의 크기, Kass & Wasserman,1995), $g=k^2$($k$: 모수의 개수, Foster & George, 1995), $g=\max\{n,k^2\}$(Fernandez, Ley, & Steel, 2001) 등이 제안되었다.

Liang 외(2008)는 $g$ 사전분포($g$ priors)의 혼합(mixture)으로 디폴트 JZS(Jeffrey-Zellner-Siow) 사전분포를 제시하였다. 즉, 디폴트 JZS에서 $g$는 역 감마 사전분포(Inverse-Gamma prior)를, $\phi$는 Jeffreys의 사전분포를 따르게 된다.

$$
\begin{aligned}
p(\beta|\phi,g,X) &= \int N(0,\frac{g}{\phi}(X^TX)^{-1})p(g)dg \\
p(\phi) &\propto \frac{1}{\phi} \\
p(g) &= \frac{(n/2)^{1/2}}{\Gamma(1/2)}g^{-3/2}e^{-n/(2g)}.
\end{aligned}
\qquad (10)
$$

식 (10)에서 $g$는 사전분포를 갖게 되어, 하나의 값을 선택할 때, 발생할 수 있는 문제들을 해결할 수 있게 되었다. 디폴트 JZS 사전분포는 Wetzels과 Wagenmakers(2012)의 상관계수 검정, Wetzels, Grasman, 그리고 Wagenmaker(2010)와 Rouder 외(2012)의 분산분석 검정, Nuijten 외(2015)의 매개효과 검정에 사용되었다.

## 라. 매개효과에 대한 베이지안 가설검정

Nuijten 외(2015)는 위의 매개효과를 검정하기 위해서 세 단계를 제시하였다.

(1) 첫 번째 단계: $a$에 대한 증거

매개효과에 대한 가설검증에서 첫 번째 단계는 [그림 1]에서 경로 $a$가 존재한다는 증거로서 $a$에 대한 베이즈 인자 추정치를 계산한다. 이를 위해서 식 (11)처럼 두 가지 가설에 해당하는 선형 모형을 세울 수 있다.

$$
\begin{aligned}
&\mu_0 : M = \beta_0 + \epsilon, \\
&\mu_1 : M = \beta_0 + aX + \epsilon.
\end{aligned}
\qquad (11)
$$

위의 식 (11)에서 $\epsilon$는 정규분포를 따르는 오차항이다. 디폴트 JZS 베이즈 인자(default JZS Bayes factor)는 아래 식 (13)에 의해서 계산되며, 자료가 $u_0$보다 $u_1$을 어느 정도 지지하는지 수량화한 결과이다.

$$BF_{10} = BF_a = \frac{P(D|\mu_1)}{P(D|\mu_0)} = \frac{(n/2)^{1/2}}{\Gamma(1/2)} \times \int_0^{\infty} (1+g)^{(n-2)/2} \times [1+(1-r^2)g]^{-(n-1)/2} g^{(-3/2)} e^{-n/(2g)} dg. \quad (12)$$

위의 식 (12)에서 $n$은 관측치의 수이고, $r^2$은 $u_1$에서 설명력을 나타낸다.

그리고 $u_0$과 $u_1$의 사전확률이 같다고 가정하면($p(u_1) = p(u_0)$), 아래 식 (13)와 같이 베이즈 인자를 통해 사후 확률을 계산할 수 있다.

$$P(\alpha \neq 0|D) = \frac{BF_a}{BF_a + 1} \quad (13)$$

(2) 두 번째 단계: $b$에 대한 증거

매개효과에 대한 가설검증의 두 번째 단계에서는 [그림 1]에서 경로 $b$에 대한 베이즈 인자 추정치를 계산한다. 이를 위해서 식 (14)에서 두 가지의 선형 모형을 세울 수 있다.

$$\begin{aligned} &\mu_0 : Y = \beta_1 + cX + \epsilon, \\ &\mu_1 : M = \beta_2 + c'X + bM + \epsilon. \end{aligned} \quad (14)$$

위의 식 (14)에서 $\epsilon$는 정규분포를 따르는 오차항이다. 디폴트 JZS 베이즈 인자(default JZS Bayes factor)는 자료가 $u_0$보다 $u_1$을 어느 정도 지지하는지 수량화하며, 아래 식 (15)으로 계산한다.

$$BF_{10} = BF_b = \frac{P(D|\mu_1)}{P(D|\mu_0)} = \frac{\int_0^{\infty} (1+g)^{(n-1-p_1)/2} \times [1+(1-r_1^2)g]^{-(n-1)/2} g^{(-3/2)} e^{-n/(2g)} dg}{\int_0^{\infty} (1+g)^{(n-1-p_0)/2} \times [1+(1-r_0^2)g]^{-(n-1)/2} g^{(-3/2)} e^{-n/(2g)} dg}. \quad (15)$$

위의 식에서 $n$은 관측치의 수이고, $r_1$과 $r_0$는 각각 $u_1$와 $u_0$에서 설명력을 의미하고, $p_1=2$과 $p_0=1$는 각각 $u_1$와 $u_0$에서 회귀계수의 수를 나타낸다.

그리고 식 (13)처럼 $p(u_1)=p(u_0)$일 때, 베이즈 인자를 통해 사후 확률을 계산하는 공식은 식 (16)과 같다.

$$P(b\neq 0|D)=\frac{BF_b}{BF_b+1} \quad (16)$$

(3) 세 번째 단계: $ab$(매개효과)에 대한 증거

매개효과 가설검정의 세 번째 단계는 베이즈 인자 추정치를 통해서 산출된 $a$와 $b$의 사후확률을 곱하여 매개효과의 사후확률을 식 (17)과 같이 계산하는 것이다.

$$\text{매개효과에 대한 증거}=P(a\neq 0|D)\times P(b\neq 0|D). \quad (17)$$

위의 식 (17)에서 계산된 매개효과의 사후 확률을 통해 식 (18)과 같이 베이즈 인자의 추정치를 계산할 수 있다.

$$BF_{ab}=\frac{p(a\neq 0|D)\times p(b\neq 0|D)}{1-p(a\neq 0|D)\times p(b\neq 0|D)} \quad (18)$$

위의 식 (18)에서 매개효과가 존재하면 $BF_{ab}$는 1보다 클 것이며, 매개효과가 없다면 $BF_{ab}$는 1보다 작을 것이다.

## 3. 매개효과연구에서 빈도주의적 접근방법

매개효과를 검정하는 가장 대표적인 방법인 Sobel의 검정(Sobel test, 1982)은 추정된 매개효과($\hat{a}\hat{b}$)를 교정된 표준오차로 나누어 $Z$통계치를 구한 후에 표준정규 분포에서 $Z$통계치가 통계적으로 유의미하게 0과 다른지를 검정한다. 이 과정에서 교정된 표준오차는 식 (19)를 통해 계산된다.

$$\hat{\sigma}_{\hat{a}\hat{b}} = \sqrt{\hat{a}^2\hat{\sigma}_{\hat{b}}^2 + \hat{b}^2\hat{\sigma}_{\hat{a}}^2} \quad (19)$$

위의 식(19)에서 $\hat{a}$와 $\hat{b}$는 식 (2)과 (3)에서 회귀계수 추정치이며, $\hat{\sigma}_a$와 $\hat{\sigma}_b$는 $\hat{a}$와 $\hat{b}$의 표준오차이다. 매개효과의 95% 신뢰구간(confidence interval)은 $\hat{a}\hat{b} \pm 1.96 \times \hat{\sigma}_{\hat{a}\hat{b}}$이다. Sobel의 검정은 매개효과의 추정치가 대칭적인 정규 표집분포(normal sampling distribution)를 따른다고 가정하지만, 실제로는 비대칭적인 표집분포를 따르기 때문에, 낮은 통계적인 검정력을 갖는다(MacKinnon et al., 2002).

따라서 MacKinnon 외(2002; 2004; 2007)는 비대칭 표집분포에 의하여 매개효과를 검정하는 비대칭분포방법(asymmetric distribution method)을 제안하였으며, 이 방법에서는 매개효과의 신뢰한계(confidence limits)에서 상한계(upper limit)와 하한계(lower limit)의 기각값(critical value)이 서로 다른 값을 갖게 된다. 또한, Bollen과 Stine(1990), Shrout과 Bolger(2002)은 매개효과연구에서 붓스트랩 방법을 제안하였고, 붓스트랩 방법 중에서도 편향이 수정된 붓스트랩(bias-corrected bootstrap) 방법이 우수한 방법으로 보고되고 있다(MacKinnon et al., 2004; Pituch & Stapleton, 2006, 2008).

# Ⅲ. 연구 방법

## 1. 모의자료 생성

이 연구에서는 통계 프로그램 R 3.0.2(R Development Core Team, 2013)을 통하여 모의자료를 생성하였다. 식 (2)와 (3)에서 독립변수($X$)는 베르누이 분포(Bernoulli distribution)를 따르도록 생성하였으며, 1과 0은 각각 실험집단과 통제집단을 나타낸다. 그리고 종속변수($Y$)와 매개변수($M$)의 오차항들(즉, $e_2$와 $e_3$)은 평균이 0이고 분산이 1인 정규분포를 따르도록 하였다.

## 2. 모의실험의 조건 및 모수 선택

[그림 1]의 매개모형에서 총효과(total effect)는 매개효과(mediated effect or indirect effect)와 직접효과(direct effect)의 합($ab+c'$)이며, 총효과 대비 매개효과의 크기는 $\frac{ab}{ab+c'}$로 나타낼 수 있다(Cheong, 2011). 먼저 매개효과가 없는 경우는 1) $a=0, b=0$, 2) $a=0.4, b=0$, 3)

$a=0, b=0.4$이고, 또한 매개효과가 작은 경우, 중간인 경우, 큰 경우는 각각, 4) $a=0.3, b=0.1$, $(ab=0.03)$ 5) $a=0.3, b=0.4$, $(ab=0.12)$ 6) $a=0.6, b=0.4$ $(ab=0.24)$과 같다. 이 연구에서 직접효과($c'$)는 $c'=0.1$로 고정하였기 때문에 총효과 대비 매개효과의 크기($=\frac{ab}{ab+c'}$)는 매개효과의 크기가 없을 때는 0, 작을 때는 0.23, 중간 정도일 때는 0.55, 그리고 클 때는 0.71의 값을 갖게 된다. 또한 표본의 크기는 20, 40, 80, 160, 320을 고려하였다. 따라서 총 30개의 조건에서 모의자료를 생성하였다.

## 3. 매개효과 검정방법

이 연구에서는 매개효과를 검정하는 데 있어서 베이지안 방법과 빈도주의 방법을 비교한다. 베이지안 방법은 디폴트 JZS 사전분포를 통한 베이즈 인자 추정치를 계산하는 검정방법이고, 빈도주의 방법은 Sobel의 검정방법(Sobel's test; Sobel, 1982), 비대칭분포방법(asymmetric distribution method: Mackinnon et al., 2004; MacKinnon, Fritz, Williams, Lockwood, 2007), 그리고 편향이 수정된 붓스트랩 방법(bias-corrected bootstrap: MacKinnon, Lockwood, & Williams, 2004; Pituch & Stapleton, 2008)이다. 일종오류와 통계적 검정력 비율은 각각의 조건에서 1,000번의 반복을 통해서 베이지안 방법은 베이즈 인자 추정치가 1보다 큰 비율을 계산하였고, 빈도주의 방법은 매개효과의 95% 신뢰구간이 0을 포함하지 않는 비율을 계산하였다. 또한 일종오류 비율은 Bradley(1978)의 준거에 따라 평가하였다. 즉, 추정된 일종오류 비율이 [.025, .075] 구간에 포함되면, 일종오류 비율이 적합하다고 판단하였다.

# Ⅳ. 연구 결과

## 1. 평균 베이즈 인자 추정치

<표 2>은 각 조건에서 1,000번의 반복을 통해 산출된 베이즈 인자 추정치들의 평균을 세 종류의 매개효과의 크기와 다섯 종류의 표본의 크기에 따라서 제시하고 있다. <표 2>에 제시된 평균 베이즈 인자 추정치가 1보다 큰 값을 가질 때, 매개효과가 존재한다는 증거가 되며, 반면에 1보다 작은 값은 매개효과가 존재하지 않는다는 증거이다(Nuijten et al., 2015).

먼저, 매개효과가 존재하지 않는 경우에는 매개효과의 평균 베이즈 인자 추정치가 모두 1보다 작았다. 또한 두 경로계수 $a$와 $b$가 모두 0일 때보다는 $a$또는 $b$가 0일 때 평균 베이즈 인자 추정치가 더 큰 것으로 나타났으며, $b$가 0일 때($a=0.4$, $b=0$)보다 $a$가 0일 때($a=0$, $b=0.4$), 평

균 베이즈 인자 추정치가 좀 더 컸다.

매개효과의 크기가 작을 때에는($ab = 0.03$) 표본의 크기가 160일 때만 평균 베이즈 인자 추정치가 1보다 컸고, 매개효과의 크기가 중간($ab = 0.12$)일 때에는 표본의 크기가 80이상인 조건에서는 평균 베이즈 인자 추정치가 1보다 컸다. 또한 매개효과의 크기가 클 때($ab = 0.24$), 표본의 크기가 40이상인 조건에서 평균 베이즈 인자 추정치는 1보다 컸다. 따라서 매개효과의 크기와 표본의 크기가 증가할수록 평균 베이즈 인자 추정치도 증가함을 알 수 있다. 또한 매개효과가 존재하지 않을 때, 평균 베이즈 인자 추정치는 모든 조건에서 1보다 작았고, 또한 매개효과의 크기와 표본의 크기가 작을 때에도 평균 베이즈 인자 추정치가 1보다 작은 경우가 발견되었다.

〈표 2〉 평균 베이즈 인자 추정치

| | $ab = 0$ | | | $ab \neq 0$ | | |
|---|---|---|---|---|---|---|
| 표본크기 | $a = 0, b = 0$ | $a = 0.4, b = 0$ | $a = 0, b = 0.4$ | $a = 0.3, b = 0.1$ | $a = 0.3, b = 0.4$ | $a = 0.6, b = 0.4$ |
| 20 | 0.083 | 0.115 | 0.188 | 0.112 | 0.209 | 0.434 |
| 40 | 0.053 | 0.103 | 0.188 | 0.101 | 0.281 | 1.249 |
| 80 | 0.032 | 0.095 | 0.186 | 0.105 | 1.073 | 18.406 |
| 160 | 0.020 | 0.126 | 0.158 | 0.395 | 14.587 | 17642.31 |
| 320 | 0.010 | 0.297 | 0.813 | 1.386 | 2762.076 | 4749156820.565 |

## 2. 일종오류 비율

〈표 3〉는 베이즈 방법과 세 종류의 빈도주의적 방법으로 산출된 매개효과의 일종오류 비율을 제시하고 있다. 먼저 두 경로계수가 모두 0인 경우($a = 0$, $b = 0$)에는 모든 방법이 명목상의 일종오류 값보다 작은 부적합한 일종오류 비율을 나타냈다. 또한, $a = 0.4$, $b = 0$일 때, 베이지안 방법과 Sobel의 검정방법은 적합한 일종오류 비율을 산출하지 못했지만, 비대칭 분포방법은 표본의 크기가 160일 때, 그리고 편향이 수정된 붓스트랩 방법은 표본의 크기가 40, 80, 160일 때, 적합한 일종오류 비율을 산출하였다. 그리고 $a = 0$, $b = 0.4$일 때에는 베이지안 방법은 모든 조건에서 부적합한 일종오류 비율을 산출하였고, Sobel의 검정방법은 하나의 조건에서 적합한 일종오류 비율을 산출하였지만, 비대칭분포방법은 모든 조건에서 적합한 일종오류 비율을 산출하였고, 편향이 수정된 붓스트랩 방법도 한 조건을 제외하고는 모두 적합한 일종오류 비율을

산출하였다. 따라서 이 연구에서 비교한 네 가지 매개효과의 검정방법 중에서는 비대칭신뢰한계 방법과 붓스트랩 방법이 상대적으로 베이지안 방법과 Sobel의 방법보다 더 적합한 일종오류 비율을 산출하는 것으로 나타났다.

〈표 3〉 네 가지 방법으로 산출된 매개효과의 일종오류 비율 비교

| | $a=0,\ b=0$ | | | | $a=0.4,\ b=0$ | | | | $a=0,\ b=0.4$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 표본크기 | 베이지안 | Sobel | 비대칭 | 붓스트랩 | 베이지안 | Sobel | 비대칭 | 붓스트랩 | 베이지안 | Sobel | 비대칭 | 붓스트랩 |
| 20 | 0.002 | 0.001 | 0.003 | 0.001 | 0.005 | 0.004 | 0.010 | 0.020 | 0.023 | 0.008 | 0.027 | 0.069 |
| 40 | 0.001 | 0.000 | 0.002 | 0.012 | 0.005 | 0.003 | 0.014 | 0.029 | 0.021 | 0.013 | 0.037 | 0.074 |
| 80 | 0.002 | 0.001 | 0.003 | 0.006 | 0.004 | 0.002 | 0.015 | 0.054 | 0.014 | 0.014 | 0.032 | 0.056 |
| 160 | 0.000 | 0.000 | 0.003 | 0.007 | 0.012 | 0.005 | 0.032 | 0.070 | 0.015 | 0.029 | 0.058 | 0.086 |
| 320 | 0.000 | 0.000 | 0.002 | 0.006 | 0.021 | 0.027 | 0.053 | 0.084 | 0.014 | 0.041 | 0.046 | 0.059 |

*베이지안: 베이즈 인자를 적용한 검정방법, Sobel: Sobel의 검정방법, 비대칭: 비대칭분포방법, 붓스트랩: 편향이 수정된 붓스트랩

## 3. 통계적 검정력 비율

〈표 4〉은 네 가지의 방법으로 산출된 통계적 검정력 비율을 나타낸다. 일종오류 비율의 결과에서 베이지안 방법과 Sobel의 방법은 명목상의 일종오류보다 낮은 부적합한 일종오류 비율을 산출하는 경향이 나타났던 반면에, 비대칭분포와 편향이 수정된 붓스트랩 방법은 베이지안과 Sobel의 방법보다는 더 많은 조건에서 적합한 일종오류 비율을 산출하였다. 통계적 검정력 비율의 결과에서도 비대칭분포방법과 편향이 수정된 붓스트랩 방법이 베이지안과 Sobel의 검정방법보다 더 높은 통계적 검정력 비율을 산출하였다. 특히, 네 가지의 방법 중에서 편향이 수정된 붓스트랩 방법이 모든 조건에서 가장 높은 통계적 검정력 비율을 산출하였다. 매개효과의 크기가 중간 정도일 때($a=0.3,\ b=0.4$), 비대칭분포방법과 편향이 수정된 붓스트랩 방법은 표본의 크기가 360일 때, 0.8에 가까운 통계적 검정력 비율을 나타내었다. 또한 매개효과의 크기가 가장 클 때($a=0.6,\ b=0.4$), 편향이 수정된 붓스트랩 방법은 표본의 크기가 80이상인 경우에 0.8에 가까운 통계적 검정력을 산출하였고, 표본의 크기가 160이상일 때, 모든 방법의 통계적 검정력이 0.9이상인 것으로 나타났다.

〈표 4〉 네 가지 방법으로 산출된 매개효과의 통계적 검정력 비율 비교

| | $a=0.3,\ b=0.1$ | | | | $a=0.3,\ b=0.4$ | | | | $a=0.6,\ b=0.4$ | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 표본크기 | 베이지안 | Sobel | 비대칭 | 붓스트랩 | 베이지안 | Sobel | 비대칭 | 붓스트랩 | 베이지안 | Sobel | 비대칭 | 붓스트랩 |
| 20 | 0.007 | 0.002 | 0.009 | 0.021 | 0.026 | 0.011 | 0.042 | 0.077 | 0.067 | 0.041 | 0.081 | 0.150 |
| 40 | 0.009 | 0.002 | 0.016 | 0.027 | 0.048 | 0.028 | 0.086 | 0.145 | 0.199 | 0.121 | 0.271 | 0.389 |
| 80 | 0.008 | 0.006 | 0.027 | 0.071 | 0.172 | 0.153 | 0.265 | 0.348 | 0.601 | 0.563 | 0.725 | 0.794 |
| 160 | 0.042 | 0.036 | 0.095 | 0.166 | 0.324 | 0.408 | 0.497 | 0.534 | 0.919 | 0.946 | 0.963 | 0.971 |
| 320 | 0.143 | 0.152 | 0.294 | 0.405 | 0.566 | 0.733 | 0.753 | 0.774 | 1.000 | 1.000 | 1.000 | 1.000 |

*베이지안: 베이즈 인자를 적용한 검정방법, Sobel: Sobel의 검정방법, 비대칭: 비대칭분포방법, 붓스트랩: 편이가 수정된 붓스트랩

# Ⅴ. 결론 및 논의

이 연구에서는 매개효과 검정에서 Nuijten 외(2015)가 제안한 베이지안 가설검정 방법을 통하여 표본의 크기와 매개효과의 크기에 따라서 베이즈 인자 추정치가 어떻게 변화하는지 살펴보았고, 또한 베이지안과 빈도주의 방법으로 산출된 일종오류와 통계적 검정력 비율을 비교하였다.

Nuijten 외(2015)는 매개효과 연구에서 매개효과가 존재하지 않는다는 영가설($H_0 : a \times b = 0$)과 매개효과가 존재한다는 대립가설($H_0 : a \times b \neq 0$)이 자료에 적용되었을 때, 베이즈 인자 추정치를 통하여 두 가지 가설 중에서 상대적으로 자료에 더 적합한 가설을 선택하는 방안을 제시하였다. 즉, Nuijten 외(2015)에서 베이즈 인자 추정치는 영가설($H_0 : a \times b = 0$)과 대립가설($H_0 : a \times b \neq 0$) 중에서 상대적으로 자료에 더 적합한 가설을 지지하는 증거이다.

이 연구에서는 각 조건에서 1,000번의 반복을 통해서 생성된 매개효과에 대한 베이즈 인자 추정치의 평균을 제시하고, 어떠한 조건에서 베이즈 인자 추정치 평균이 1보다 큰 값을 갖아서 대립가설($H_0 : a \times b \neq 0$)을 지지하는지 살펴보았다. 그 결과 매개효과가 없는 경우에는 모든 조건에서 베이즈 인차 추정치 평균이 1보다 작았다. 그렇지만, 매개효과의 크기가 작거나 중간인 경우에는 표본의 크기가 각각 320과 80이상일 때, 그리고 매개효과의 크기가 큰 경우에는 표본의 크기가 40이상일 때, 베이즈 인자 추정치가 1보다 컸다. 따라서 매개효과가 존재하지 않을 때에는 베이즈 인자 추정치는 매개효과가 존재하지 않는다는 영가설($H_0 : a \times b = 0$)을 지지하

는 증거가 됨을 확인할 수 있었다. 그렇지만, 매개효과가 존재할 때에는 베이즈 인자 추정치가 그 증거로 올바로 기능하기 위해서는 매개효과의 크기와 표본의 크기가 작지 않아야 함을 확인할 수 있었다. 또한 Nuijten 외(2015)도 매개효과가 존재하지 않을 때에는 베이즈 인자 추정치가 증거를 제공하지만, 매개효과가 존재할 때에는 매개효과와 표본의 크기가 클 때에만 베이즈 인자 추정치가 증거가 되었다고 이 연구와 유사한 결과를 보고하였다. Nuijten 외(2015)와 이 연구의 차이는 이 연구는 각 조건에서 1,000번의 반복을 통해서 생성된 모의자료로부터 산출된 베이즈 인자 추정치의 평균을 사용하였지만, Nuijten 외(2015)는 각 조건에서 하나의 모의자료만을 생성하였다는 점이다.

매개효과 연구에서 베이지안 가설 검정 방법의 가장 큰 장점은 베이즈 인자 추정치를 토대로 영가설과 대립가설 중에서 지지하는 가설에 대한 상대적인 증거를 제공한다는 사실이지만(Nuijten et al., 2015; Rouder & Morey, 2013), 빈도주의 방법과의 비교를 통한 베이지안 가설 검정 방법의 타당성을 탐색하는 연구는 아직까지 찾을 수 없었다. 따라서 이 연구에서는 모의실험을 통해서 Nuijten 외(2015)가 제안한 매개효과에 대한 베이지안 가설 검정 방법과 매개효과에 대한 대표적인 빈도주의 가설 검정 방법들(Sobel의 검정, 비대칭분포방법, 편향이 수정된 붓스트랩 방법)의 일종오류와 통계적 검정력 비율을 비교하였다. 그 결과 일종오류 비율과 통계적 검정력 비율에서 빈도주의 방법인 편향이 수정된 붓스트랩 방법이 가장 우수한 것으로 나타났다. 그 다음 비대칭분포방법이 우수하였다. 반면, 베이지안 방법과 Sobel의 검정방법은 대부분의 조건에서 명목상의 일종오류 비율(0.05)보다도 작은 매우 보수적인 일종오류 비율을 산출하였으며, 또한 이러한 보수적인 일종오류 비율로 인하여 통계적 검정력 비율도 편향이 수정된 붓스트랩 방법과 비대칭분포방법보다 낮은 것으로 나타났다. 이처럼 베이지안 검정방법이 매우 보수적인 일종오류 비율과 낮은 통계적 검정력 비율을 산출한 원인으로는 디폴트 사전분포 때문일 수도 있다. 이 연구의 제한점은 사전분포의 선택에 있다고 할 수 있다. 이 연구에서 사용된 디폴트 사전분포는 Liang 외(2008)가 개발한 다중회귀분석에서의 디폴트 사전분포이며, 사전지식이 없어서 사전분포를 선택할 수 없는 상황에서 적용할 수 있으며, 계산적이 간편하고, 측정단위의 불변성과 일관성의 바람직한 특성을 갖추었다(Rouder & Morey, 2008). 그러나 이 연구에서 사용한 사전분포 외에도 다양한 사전분포를 고려할 수 있으며, 또한 사전지식이 있는 연구자들은 디폴트 사전분포를 사용하기 보다는 그들의 사전지식을 사전분포에 반영하는 것이 보다 바람직할 것으로 판단된다. 따라서 후속연구에서는 베이지안 가설검정에서 다양한 사전분포의 영향력에 대해서 살펴볼 필요가 있을 것이다. 또한 최근에는 통계 프로그램인 R에서 베이지안 통계를 보다 쉽게 활용할 수 있는 다양한 베이지안 패키지들(packages)들이 제공되고 있는 만큼, 일반 사회과학 연구자들도 이전보다 더욱 쉽게 베이지안 통계에 접근할 수 있을 것이다.

# 참고문헌

Bradley, J. V. (1978). Robustness? **The British Journal of Mathematical & Statistical Psychology, 31**, 144-152.

Bollen, K. A., & Stine, R. (1990). Direct and indirect effects: classical and bootstrap estimates of variability. **Sociological Methodology, 20**, 115-140.

Cheong, J. (2011). Accuracy of estimates and statistical power for testing mediation in latent growth curve modeling, **Strcutural Equation Modeling, 18**, 195-211.

Fernandez, C., Ley, E., & Steel, M. (2001). Benchmark priors for Bayesian model averaging. **Journal of Econometrics, 100**, 381-427.

Foster, D., & George, E. (1994). The risk inflation criterion for multiple regression. **The Annals of Statistics, 22**, 1947-1975.

Kass, R. E., & Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. **Journal of the American Statistical Association, 90**, 928-934.

Liang, F., Paulo, R., Molina, G., Clyde, M. A., & Berger, J. O. (2008). Mixtures of g priors for Bayesian variable selection. **Journal of the American Statistical Association, 103**, 410-423.

MacKinnon, D. P., Lockwood, C. M., Hoffman, J. M., West, S. G., & Sheets, V. (2002). A comparison of methods to test mediation and other intervening variable effects. **Psychological Methods, 7**, 83-104.

MacKinnon, D. P., Lockwood, C. M., & Williams, J. (2004). Confidence limits for the indirect effect: Distribution of the product and resampling methods, **Multivariate Behavioral Research, 39**, 99-128.

MacKinnon, D. P., Fritz, M. S., Williams, J., & Lockwood, C. M. (2007). Distribution of the product confidence limits for the indirect effect: Program PRODCLIN. **Behavior Research Methods, 39**, 384-389.

Nuijten, M. B., Wetzels, R., Matzke, D., Doran, C. V., & Wagenmakers, E. J. (2015). A default Bayesian hypothesis test for mediation. **Behavior Research Methods, 47**, 85-97.

Pituch, K. A., Stapleton, L. M., & Kang, J. Y. (2006). A comparison of single sample and bootstrap methods to assess mediation in cluster randomized trials. **Multivariate Behavioral Research, 41**, 367-400.

Pituch, K. A., & Stapleton, L. M. (2008). The performance of methods to test upper-level mediation in the presence of non-normal data. **Multivariate Behavioral Research, 43**, 237-267.

Preacher, K. J., & Hayes, A. F. (2004). SPSS and SAS procedures for estimating indirect effects in simple mediation models. Behavior Research Methods, **Instruments, & Computers, 34**, 717-731.

Rouder, J. N., & Morey, R. D. (2012). Default Bayes factors for model selection in regression. **Multivariate Behavioral Research, 47**, 877-903.

Shrout, P. E., & Bolger, N. (2002). Mediation in experimental and nonexperimental studies: New procedures and recommendations. **Psychological Methods, 7**, 422-445.

Sobel, M. E. (1982). Asymptotic confidence intervals for indirect effects in structural equation models. In S. Leinhardt (Ed.), Sociological Methodology 1982 (pp.290-312). Washington, DC: American Sociological Association.

Sobel, M. E. (1986). Some new results on indirect effects and their standard errors in covariance strcutre models. In N. Tuma (Ed.), Sociological Methodology 1986, (pp.290-312). Washington, DC: American Sociological Association.

Wetzels, R., & Wagenmakers, E. J. (2012). A default Bayesian hypothesis test for correlations and partial correlations. **Psychonomic Bulletin & Review, 19**, 1057-1064.

Wetzels, R., Grasman, R. P. P. P., & Wagenmakers, E. J. (2012). A default Bayesian hypothesis test for ANOVA designs. **The American Statistician, 66**, 104-111.

Yuan, Y., & MacKinnon, D. P. (2009). Bayesian mediation analysis. **Psychological Methods, 14**, 301-322.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. Bayesian Inference and Decision Techniques: **Essays in Honor of Bruno de Finetti**, 233-243.

• 논문 접수 2019년 07월 18일 / 수정본접수 08월 26일 / 게재 승인 08월 30일

• 구남욱 : University of Florida 교육평가 박사학위 취득. 현재 한국교유과정평가원에서 부연구위원으로 재직하고 있으며 국제학업성취도평가에 관심을 가지고 있음.

• 전지혜 : 상명대학교 교육학과 졸업. 성균관대학교에서 교육학으로 석사학위 취득 및 박사과정 수료. 베이지안, 표본조사 설계에 관심을 가지고 있음.

교육연구와 실천
Journal of Educational Study and Practice
제10권 제1호, 2019, pp. 19-39

# Exploring factors effects on university student satisfaction survey

Meounggun Jo*
(Hoseo University)

< Abstract >

This study was conducted to explore the factors that affect university education satisfaction and sub-satisfaction in the recent survey of university education satisfaction conducted by university society. In this study, the average comparison was performed for each group to check if there was a difference in the average of the college education satisfaction score based on gender and college. Studies have shown that there are statistically significant differences in satisfaction between sex and college, and male satisfaction in all areas is higher than that of female students. Also, by college, the number of colleges was high in humanities, social sciences and engineering, while the figure was low in science and technology convergence and arts colleges. What is notable is that the average difference between male and female students in humanities and bio-health colleges was more different than other colleges, while the satisfaction level of art colleges was low on the curriculum, while the satisfaction level of academic achievement was high. It suggested the urgent need to lay the groundwork for providing education services and education courses tailored to students' gender, colleges and universities by compiling the ideals.

Key words: university students satisfaction, university satisfaction

* Meounggun Jo(audqaudq@naver.com)

교육연구와 실천
Journal of Educational Study and Practice
제10권 제1호, 2019, pp. 19-39

# 대학교육만족도에 영향을 미치는 요인 탐색: A대학 사례를 중심으로

조명근(호서대학교)*

< 요 약 >

본 연구는 최근 대학 사회에서 실시하고 있는 대학교육만족도 조사에서 대학교육만족도와 하위 만족도에 영향을 미치는 요인을 탐색하기 위해 수행하였다. 본 연구는 성별과 단과대학에 따른 대학교육만족도 점수의 평균에 차이가 있는지 확인하기 위해 각 집단 별 평균비교를 수행하였다. 연구 결과 성별과 단과대학에 따른 만족도에는 통계적으로 유의한 차이가 있는 것으로 나타났고 모든 영역에서 남학생의 만족도가 여학생의 만족도보다 높은 것으로 나타났다. 또한 단과대학별로는 인문대학, 사회과학대학, 공과대학에서 높은 반면 과학기술융합대학과 예체능대학에서 낮게 나타났다. 특기할만한 점으로는 예술대학의 교육과정에 대한 만족도는 낮은 반면, 학업성취에 대한 만족도는 높게 나타나는 경향을 보였다는 점, 인문대학과 생명보건대학에서는 남학생과 여학생의 평균 차이가 다른 단과대학에 비해 더 차이가 났다는 점이다. 이상을 종합하여 학생의 성별, 단과대학별 맞춤형 교육과정 및 교육서비스 제공의 기틀 마련이 시급하다는 점을 제언하였다.

**주제어: 대학교육만족도, 교육만족도**

* 조명근(audqaudq@naver.com)

# I. 서론

대학은 교육을 제공하고 학생들은 그 서비스를 받는다. 대학 교육 만족도란, 학생들이 대학에서 제공하는 서비스를 받고 그에 대해 느끼는 만족도를 나타내는 것으로, 교육을 서비스의 관점에서 접근한 것이다. 교육부는 인구 절벽과 학생 수 감소라는 시대적 상황 속에서 대학 교육의 수요와 공급의 불일치 현상을 해결하기 위하여 대학구조개혁 평가를 실시하고 있다. 교육을 받는 교육수요자(학생, 산업체, 학부모 등)가 만족할 수 있는 수준의 교육 서비스를 제공하지 못하는 대학에 대한 제재조치를 취하고 있다. 대학도 교육수요자가 만족할만한 교육 서비스를 갖추어야 생존에 유리한 형태로 바뀌고 있다(함은혜, 박상옥, 김은경, 2017).

이와 같은 현실 속에서 대학의 교육서비스에 대한 만족도 조사가 전국적으로 활발히 진행되고 있다. 대학이 제공하는 교육 서비스를 직접 받는 학생들 뿐 아니라 학생들의 학비를 지원하는 학부모, 해당 학생들을 인력으로 활용하게 되는 산업체까지 교육 수요자의 범주에 포함되며 이들의 만족도를 측정하고 불만족 요인을 탐색하여 서비스의 질을 개선하는데 조사의 목적이 있다. 이와 같은 조사가 전국적으로 확대되고 있는 이유는 대학의 서비스 질 개선을 위한 노력의 일환으로 볼 수도 있으나, 대학구조개혁평가의 평가지표 중 하나가 교육수요자 만족도 관리에 포함되어 더욱 부각되기 시작하였다. 대부분의 지방 대학들이 대학구조개혁평가 결과에 따라 존립을 위협받는 상황이 됨에 따라 대학들이 자체적으로 다양한 교육만족도 조사를 수행하기에 이르렀다.

대학 교육 만족도 조사는 대부분 주어진 진술에 동의하는 정도를 리커트 척도를 활용하여 체크하는 형태로 진행된다. 대부분 학교에서 진행하는 교육 수요자 만족도 조사의 경우 실제 전반적 만족도에 영향을 미치는 요인을 사전조사하여 검사에 반영하는 형식 대신 대학 내 각 부서에서 실시하고 있는 프로그램이나 서비스에 대한 설문으로 진행되는데 이는 교육 만족도에 대한 환류체계를 구축하기 위한 것으로 보인다. 이러한 조사를 위해 대학교육만족도를 구성하는 평가 영역들을 먼저 구성하고 이를 구성하는 하위 요인들에 대한 설문 문항을 제작하여 검사를 실시한다. 가령 대학의 교육과정에 대한 만족도를 평가영역으로 삼았다면, 전공 교육과정, 교양 교육과정, 비교과 교육과정으로 하위 구성요인을 설정하고 각각에 대한 교수진이나 교과목에 대한 만족도를 체크하는 형식이다. 또한 조사를 통해 대학교육 만족도를 측정하는 방식으로 크게 두 가지를 선택하는데 대학교육만족도를 구성하는 각각의 요인들의 하위 영역별 문항 점수를 활용하여 전체 만족도를 측정하는 형식을 취하거나 별도로 종속변인으로 활용할 수 있는 전반적 만족도에 대한 문항을 추가하여 설문을 수행하기도 한다. 이런 형태로 전반적 만족도를 산출하게 되면 대학 내 집단 간의 만족도 평균 차이를 추정하는데 활용할 수 있으며, 전반적 만족도를 산출하는 근거로 하위 영역의 점수가 사용되는 것이 충분히 역할을 한다고 판단

할 수 있다(함은혜 외, 2017). 선행연구에 따르면 평가 영역별 점수의 상관은 대부분 .5 이상이며 .7 이상으로 매우 높게 나타나는 경우도 많았다(박제일, 2016; 박혜림, 2015; 정주영, 2013). 또한 대부분의 하위 영역에서 여학생의 만족도가 낮게 나타나는 현상이 공통적으로 발견되기도 하였다(권대봉, 오영재, 박행모, 손준종, 송선희, 2002; 신소영, 권성연, 2014; 최영준, 2013; 한은숙, 김종두, 2003). 이상의 연구결과들을 고찰한 결과 본 연구에서 설정한 구체적인 연구의 문제는 다음과 같다.

첫째, 성별, 학년, 단과대학에 따라 교육만족도 점수의 평균에 차이가 있는가?

둘째, 성별, 학년, 단과대학에 따른 교육만족도 하위 영역의 점수 평균에 차이가 있는가?

셋째, 성별, 학년, 단과대학에 따른 교육만족도 점수의 평균 차이에는 상호작용(interaction effect)이 존재하는가?

# Ⅱ. 선행연구

## 1. 대학교육만족도의 하위 구성요인

대학교육 만족도란 대학에서 교육을 받는 학생들이 느끼는 전반적인 만족도를 총칭하는 말로 사용된다. Astin(1993)은 대학 교육 만족도의 개념을 처음 도입하였는데, 학생들이 경험한 교육에 대해 주관적인 감정을 측정하는 것으로 정의하였다. 대학 내에서의 다양한 교육 경험을 세분화하고 학생들이 느끼는 만족의 정도가 어느 정도인지를 의미 중심으로 정의하는 것이다. Astin은 교육만족도를 두 가지 방식으로 측정하였는데, 대학교육 전반에서 학생들의 만족도를 분석하는 것과 학생들의 요구사항을 세부적으로 설정한 뒤 만족도를 조사하여 우선순위를 밝히는 방식을 활용하였다. 이 때 연구에 포함된 변인은 교수, 교육과정, 수업, 학교생활, 학생지원 서비스, 시설, 교과목 이수 기회 등이다. 윤관호 외(2013)는 학생의 대학 생활 경험에서 교육 공급자인 교수와 직원이 교육의 목표 달성을 위한 지식 서비스를 제공하고, 소비자인 학생이 이를 지원받아 본인의 교육 목표를 달성하는 데 지원되는 물질적, 정신적 지원이라고 명명하였다. 대학교육만족도는 학업의 성패에 중대한 영향을 미치므로(Aldridge & Rowley, 1998) 성공적인 교육 서비스가 무엇인지 예측하는 중요한 지표이며(권대봉 외, 2002), 대학교육의 질 관리를 위한 기본 지표로 활용된다(신소영, 권성연, 2012).

미국에서 처음 시작된 대학교육만족도 조사는 등록률이나 입시 지원자의 규모나 수준, 졸업률 등과 같은 객관적 자료를 바탕으로 하여 학생들의 수준을 가늠하는 간접적인 방법이었다(권대봉 외, 2002). 그러나 이러한 방식은 학생들의 불만족 원인이 무엇인지를 밝히고 그에 대한

피드백을 제공하는 데에 도움을 주기는 어려운 방식이었다. 대학에서 학생들은 교과교육 외에도 최근 대학들이 공을 들이고 있는 비교과 교육 프로그램들을 포함하는 교육과정 이외에도 사회적, 정서적, 환경적 요인을 모두 총합하는 형태로 형성되기 때문에 이를 위한 여러 변인들을 포함시켜 조사를 수행하는 것이 일반적이다.

서울 소재 19개 대학의 대학교육만족도를 조사한 권대봉 외(2002)의 연구에서 대학교육만족도를 측정하기 위한 변인으로 교수와의 관계, 교육의 질, 학습지도와 성적평가의 4개 영역에서의 만족도를 측정하여 합산하였다. 박혜림(2015)은 대학교육만족도조사 문항 개발 연구를 실시하였는데 교육과정, 학생지도, 학생지원, 특성화, 학교이미지, 교육환경, 행정서비스 등을 조사하였다. 또한 박제일(2016)은 전반적 만족도와 서비스에 대한 기대, 경제여건, 전공 만족도, 학업의지, 참여도 등을 조사하였다.

## 2. 대학교육만족도의 중요성 및 영향을 미치는 요인

대학교육만족도는 학생들로 하여금 학생의 동기를 강화시키는 역할을 한다(Elliot & Shin, 2002). 이를 통해 학생들은 학교생활에 적응하고 유지하는 데에 긍정적인 영향을 받게 된다. Levitz, Noel, Richeter(1999)는 학생의 수와 질을 유지하는 데에 학교가 학생들에게 학생 만족이라는 목표를 충족시킨다는 가장 좋은 지표가 될 수 있으며 학교만족도의 중요성을 강조하였다. 또한 Krause(2005)는 중도탈락자 대부분이 낮은 대학교육만족도를 보임을 밝히고 있다. Loreszetti(2009)는 교육 만족 요인이 학생을 유지하는 데에 긍정적 영향을 미치며 교육만족도의 하위 요인들이 학년에 따라 다르게 나타남을 밝혔다. 이 외에도 학생의 배경 요인에 따라 만족도는 조금씩 차이가 있었으나 대부분의 요인들이 학생들의 만족도와 유지에 영향을 미치는 것으로 나타났다(Schreiner, 2009). 뿐만 아니라 Yu와 Kim(2008)은 대학 서비스에 대한 만족도가 대학생활의 질을 결정하는 요인이며 학생들의 학교에 대한 충성도에 영향을 미치는 것임을 밝혔다. 이 외의 여러 연구에서도 대학교육만족도가 학업지속, 충성도, 학교에 대한 태도, 평판 등 다양한 요인에 여향을 미치는 것으로 나타났다(백승학, 2010; 이정원, 임지영, 2008). 이러한 연구들을 통해 대학교육만족도를 적절하게 측정하고 분석하는 것이 중요하며 어떤 요인이 대학교육의 만족도에 영향을 미치는지 파악하는 것도 실제적인 시사점을 줄 수 있을 것으로 판단하였다.

# Ⅲ. 연구 방법

## 1. 연구대상 및 자료 수집 방법

본 연구는 충남지역 A대학에서 실시한 교육수요자만족도조사 중 재학생 만족도 조사 자료를 활용하여 분석을 수행하였다. 6개 단과대학 48개 학과 학생들을 대상으로 비례표집을 수행하였으며, 오프라인 설문지를 활용하여 자료를 수집하였다. 본 조사는 4,556명의 학생들이 설문에 응답하였으며 이는 전체 재학생 11,910명(2017학년도 기준)의 약 38.3%에 해당하는 인원이었다. 최종 분석대상 중 남학생은 2,745명, 여학생은 1,802명이었으며 학년별로는 1학년 1,409명(30.9%), 2학년 1,234명(27.1%), 3학년 1,125명(24.7%), 4학년 이상이 788명(17.3%)으로 나타났다. 단과대학별로는 인문융합대학이 256명으로 5.6%, 사회과학대학이 664명으로 14.6%, 생명보건대학이 536명으로 33.5%, 공과대학이 1,527명으로 33.5%, 과학기술융합대학이 1,179명으로 25.9%, 예체능대학이 394명으로 8.6%를 차지하였다.

〈표 7〉 응답자의 배경 변인

| 구분 | | 빈도 | 비율(%) |
|---|---|---|---|
| 성별 | 남 | 2,475 | 60.4 |
| | 여 | 1,802 | 39.6 |
| 학년 | 1학년 | 1409 | 30.9 |
| | 2학년 | 1234 | 27.1 |
| | 3학년 | 1125 | 24.7 |
| | 4학년 | 787 | 17.3 |
| 단과대학 | 인문융합대학 | 256 | 5.6 |
| | 사회과학대학 | 664 | 14.6 |
| | 생명보건대학 | 536 | 11.8 |
| | 공과대학 | 1527 | 33.5 |
| | 과학기술융합대학 | 1179 | 25.9 |
| | 예체능대학 | 394 | 8.6 |
| 전체 | | 4,556 | 100 |

## 2. 측정도구

본 조사는 대학교육만족도의 평가영역으로 1) 교육과정 및 질 관리 2) 교육환경 및 체계 3)

학습성과 4) 교육목표 및 인재상의 4가지 영역으로 선정하였다. 각 평가영역과 하위영역의 내용은 다음 <표 2>와 같다.

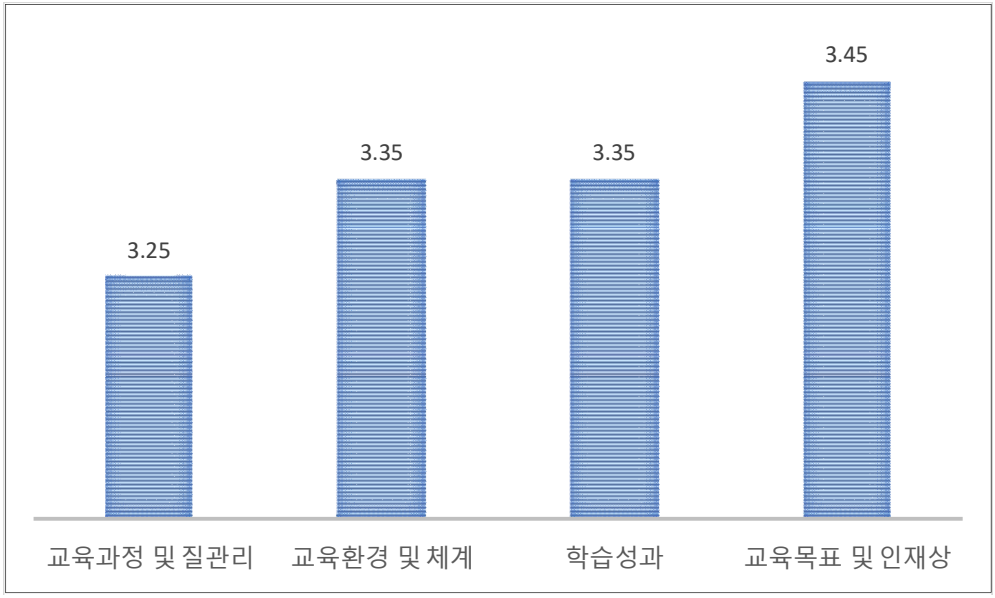
<표 8> 교육만족도 영역별 측정 문항 및 신뢰도 계수

| 평가 영역 | 하위 영역 | 평가내용 | 문항 수 | 신뢰도 계수 |
|---|---|---|---|---|
| 교육과정 및 질관리 | 교육과정 | 교양과목, 전공과목 | 4 | .825 |
| | 교육의 질 관리 | 교양과목, 전공과목, 학습지원 | 11 | .869 |
| | 교수 | 교양과목, 전공과목, 수업준비 등 | 4 | .881 |
| | 비교과 | 운영부서별 만족도 프로그램 | 7 | .929 |
| 교육환경 및 체계 | 교육환경 학생복지 및 편의시설 | 교육환경, 학생복지, 편의시설 만족도 | 18 | .938 |
| | 행정 및 학사서비스 | 행정서비스, 학사서비스 | 10 | .933 |
| 학습성과 | 학습성과 | 역량의 향상 정도 | 13 | .934 |
| 교육목표 및 인재상 | 교육목표 | 적합도, 성취도 | 4 | .918 |
| | 인재상 | 적합도, 성취도 | 2 | .868 |
| | 핵심역량 | 적합도, 성취도 | 16 | .954 |

대학교육만족도 하위영역별 문항점수의 평균, 표준편차, 상관계수는 다음 <표 3>과 같다.

<표 9> 대학교육만족도 평가영역별 평균, 표준편차, 상관계수

| 평가 영역 | 평균 | 표준 편차 | 상관계수 | | |
|---|---|---|---|---|---|
| | | | 교육과정 및 질관리 | 교육환경 및 체계 | 학습성과 |
| 교육과정 및 질관리 | 3.25 | .61 | 1 | | |
| 교육환경 및 체계 | 3.35 | .62 | .769 | 1 | |
| 학습성과 | 3.35 | .65 | .711 | .721 | 1 |
| 교육목표 및 인재상 | 3.45 | .62 | .703 | .647 | .690 |




[그림 2] 대학교육 만족도 영역별 평균

## 3. 분석절차

본 조사는 성별, 학년별, 단과대학별로 학생들의 대학교육 만족도의 평균이 평가 영역별로 서로 다르게 나타나는지를 확인하고 독립변수들 간의 상호작용이 존재하는지를 확인하고자 한다. 이를 위해 먼저 집단 간 평균 비교를 수행하였다. 성별에 따른 대학 교육 만족도 평균 비교를 위해서는 T-test를 활용하였고, 학년과 단과대학별 대학 교육 만족도 평균 비교를 위해서 일원분산분석(one-way ANOVA)을 실시하였고, 집단 간 평균이 통계적으로 유의미한 경우 Scheffe의 사후검정을 활용하여 사후검정을 수행하였다. 또한 상호작용효과를 확인하기 위해 이원분산분석(two-way ANOVA)을 수행하였다.

# Ⅳ. 연구 결과

## 1. 성별

성별에 따른 만족도 평균에 차이가 있는지를 검정하기 위해 T-test를 수행하였으며 그 결과

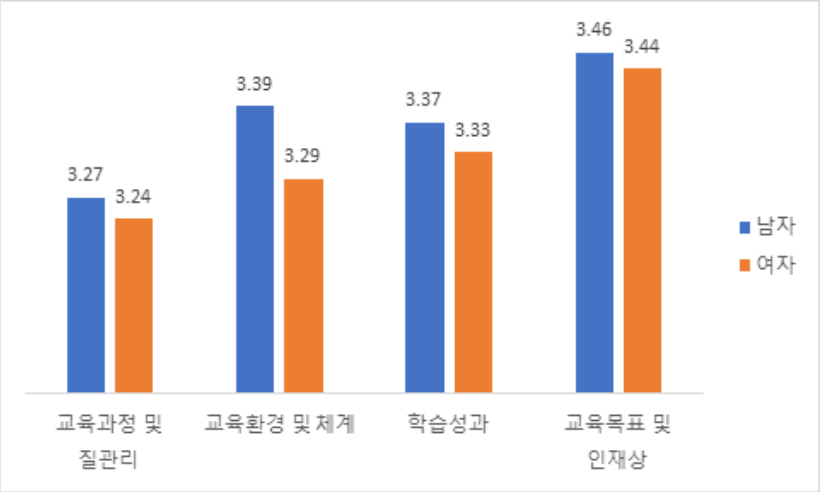
는 다음 <표 4>와 같다. 교육환경 및 체계와 학습성과 영역에서만 통계적으로 유의한 차이가 있었으며 교육과정 및 질 관리나 교육목표 및 인재상에서는 유의한 차이가 없는 것으로 나타났다. 그중에서도 실제적 유의성을 가진 영역은 교육환경 및 체계에 대한 부분으로, 남학생들의 평균이 여학생들의 평균보다 약 0.1점 정도 높은 것으로 나타났다.

〈표 10〉 성별에 따른 대학교육만족도 기술통계

| 성별 | | N | 평균 | 표준편차 | 평균의 표준오차 |
|---|---|---|---|---|---|
| 교육과정 및 질관리 | 남자 | 2745 | 3.2660 | .62471 | .01192 |
| | 여자 | 1802 | 3.2377 | .58782 | .01385 |
| 교육환경 및 체계 | 남자 | 2743 | 3.3910 | .62750 | .01198 |
| | 여자 | 1802 | 3.2922 | .59987 | .01413 |
| 학습성과 | 남자 | 2742 | 3.3692 | .66803 | .01276 |
| | 여자 | 1801 | 3.3277 | .62002 | .01461 |
| 교육목표 및 인재상 | 남자 | 2741 | 3.4638 | .64039 | .01223 |
| | 여자 | 1798 | 3.4415 | .60144 | .01418 |

〈표 11〉 성별에 따른 대학교육만족도 두 집단 평균비교 결과

| 영역 | t | 자유도 | 유의확률 (양측) | 차이의 95% 신뢰구간 | |
|---|---|---|---|---|---|
| | | | | 하한 | 상한 |
| 교육과정 및 질관리 | 1.552 | 4013.625 | .121 | -.00746 | .06419 |
| 교육환경 및 체계 | 5.332 | 3972.646 | .000*** | .06247 | .13512 |
| 학습성과 | 2.141 | 4046.532 | .032* | .00349 | .07955 |
| 교육목표 및 인재상 | 1.193 | 4009.333 | .233 | -.01438 | .05906 |




[그림 3] 성별에 따른 대학교육 만족도 차이

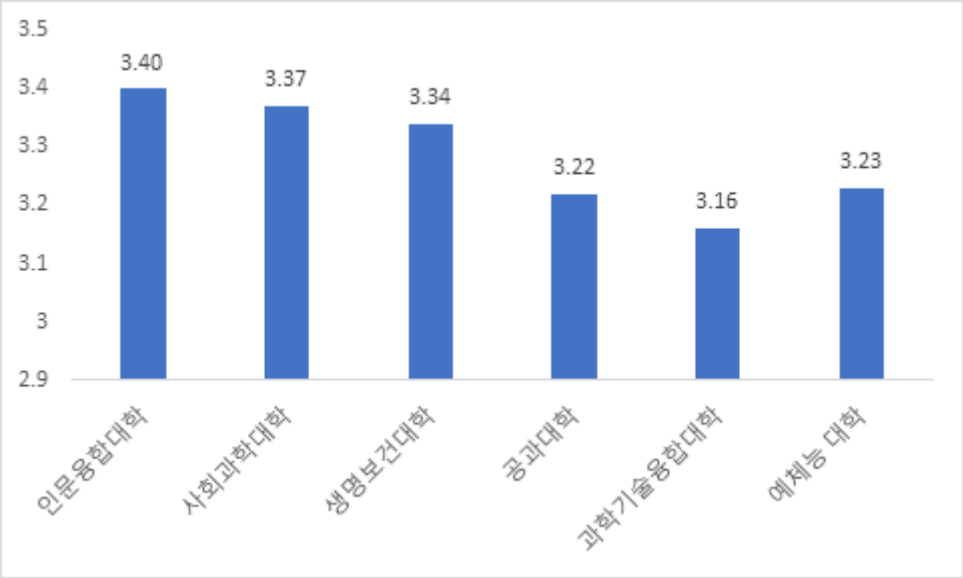
## 2. 단과대학별

<표 12> 단과대학 별 교육과정 질 관리에 대한 만족도 분산분석 표

| | | 제곱합 | 자유도 | 평균제곱 | F | 유의확률 |
|---|---|---|---|---|---|---|
| 교육과정 및 질관리 | 집단-간 (결합) | 30.392 | 5 | 6.078 | 16.598 | .000*** |
| | 집단-내 | 1666.211 | 4550 | .366 | | |
| | 전체 | 1696.603 | 4555 | | | |

<표 13> 단과대학 별 교육과정 질 관리에 대한 만족도

| | 인문융합대학 | 사회과학대학 | 생명보건대학 | 공과대학 | 과학기술융합대학 | 예체능대학 |
|---|---|---|---|---|---|---|
| 교육과정 및 질관리 | 3.40 | 3.37 | 3.34 | 3.22 | 3.16 | 3.23 |




[그림 4] 단과대학 별 교육과정 질 관리에 대한 만족도

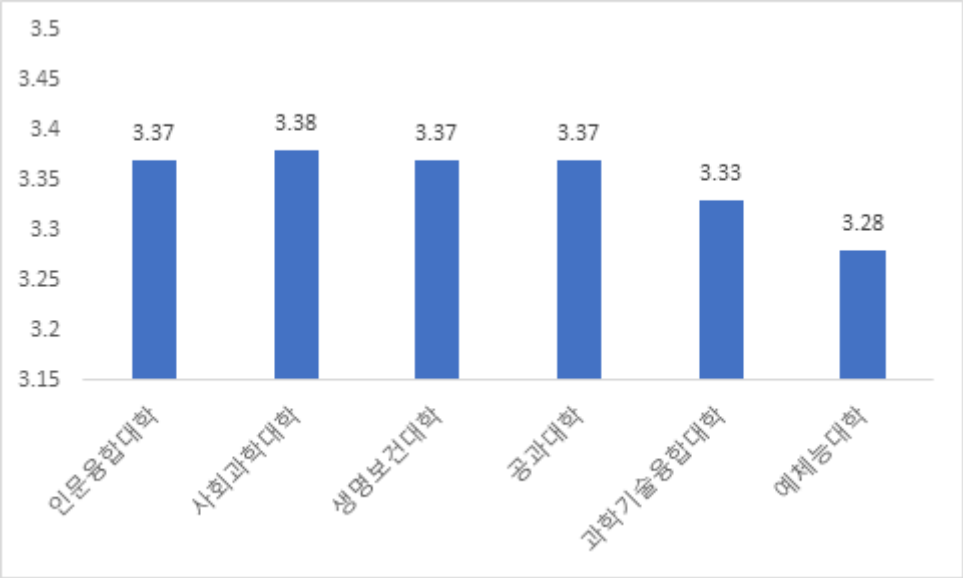
교육과정 및 질 관리 영역의 집단 간 평균차이를 확인하기 위해 일원분산분석을 수행하였으며 그 결과는 <표 6>과 같다. 단과대학별로 통계적으로 유의한 대학 교육 만족도 평균 차이가 있으므로 사후 검정을 수행하엿다. 사후검정결과 인문융합대학, 사회과학대학, 생명보건대학이 공과대학과 과학기술융합대학 대비 높은 만족도 평균을 가지는 것으로 나타났으며, 인문융합대학과 사회과학대학의 경우 예체능 대학 대비 높은 만족도 평균을 가지는 것으로 나타났다.

<표 14> 단과대학 별 교육환경 및 체계에 대한 만족도 분산분석 표

| | | 제곱합 | 자유도 | 평균제곱 | F | 유의확률 |
|---|---|---|---|---|---|---|
| 교육환경 및 체계 | 집단-간 (결합) | 3.882 | 5 | .776 | 2.034 | .071 |
| | 집단-내 | 1735.586 | 4548 | .382 | | |
| | 전체 | 1739.468 | 4553 | | | |

<표 15> 단과대학 별 교육환경 및 체계에 대한 만족도

| | 인문융합대학 | 사회과학대학 | 생명보건대학 | 공과대학 | 과학기술융합대학 | 예체능대학 |
|---|---|---|---|---|---|---|
| 교육환경 및 체계 | 3.37 | 3.38 | 3.37 | 3.37 | 3.33 | 3.28 |




[그림 5] 단과대학 별 교육환경 및 체계에 대한 만족도

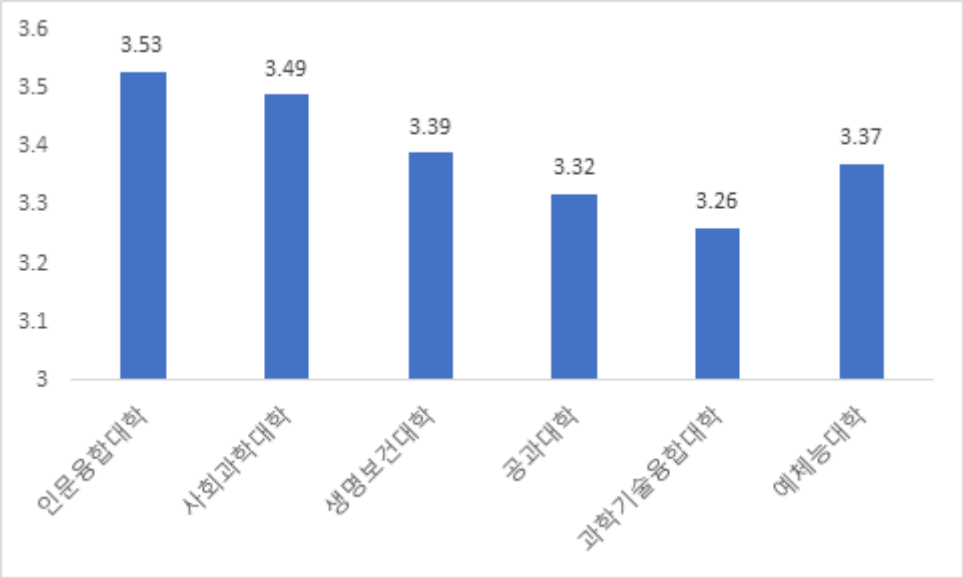
교육환경 및 체계 영역의 집단 간 평균차이를 확인하기 위해 일원분산분석을 수행하였으며 그 결과는 <표 8>과 같다. 단과대학별로 통계적으로 유의한 대학 교육 만족도 평균 차이가 있으므로 사후 검정을 수행한 결과 인문융합대학, 사회과학대학, 생명보건대학이 과학기술융합대학 대비 높은 만족도 평균을 가지는 것으로 나타났다.

<표 16> 단과대학 별 학습성과에 대한 만족도 분산분석 표

| | | 제곱합 | 자유도 | 평균제곱 | F | 유의확률 |
|---|---|---|---|---|---|---|
| 학습성과 | 집단-간 (결합) | 33.708 | 5 | 6.742 | 16.257 | .000*** |
| | 집단-내 | 1885.137 | 4546 | .415 | | |
| | 전체 | 1918.845 | 4551 | | | |

<표 17> 단과대학 별 학습성과에 대한 만족도

| | 인문융합대학 | 사회과학대학 | 생명보건대학 | 공과대학 | 과학기술융합대학 | 예체능대학 |
|---|---|---|---|---|---|---|
| 학습성과 | 3.53 | 3.49 | 3.39 | 3.32 | 3.26 | 3.37 |




[그림 6] 단과대학 별 학습성과에 대한 만족도

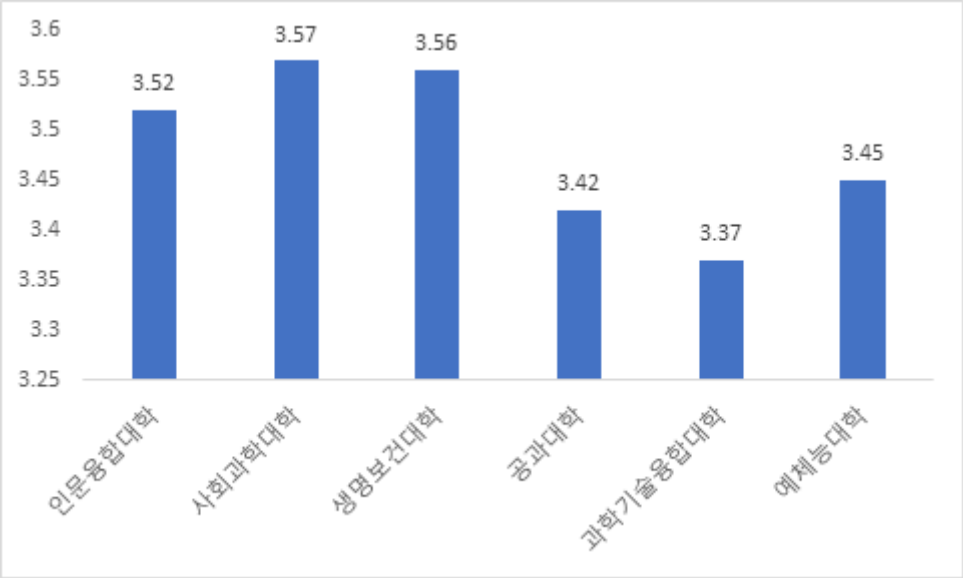
학습 성과 영역의 집단 간 평균차이를 확인하기 위해 일원분산분석을 수행하였으며 그 결과는 <표 10>과 같다. 단과대학별로 통계적으로 유의한 대학 교육 만족도 평균 차이가 있으므로 사후 검정을 수행하였다. 사후검정결과 과학기술융합대학에서의 만족도가 타 단과대학의 만족도에 비해 상대적으로 낮은 것으로 나타났다.

<표 18> 단과대학 별 교육목표 및 인재상에 대한 만족도 분산분석 표

| | | 제곱합 | 자유도 | 평균제곱 | F | 유의확률 |
|---|---|---|---|---|---|---|
| 교육목표 및 인재상 | 집단-간 (결합) | 25.341 | 5 | 5.068 | 13.155 | .000*** |
| | 집단-내 | 1749.817 | 4542 | .385 | | |
| | 전체 | 1775.157 | 4547 | | | |

<표 19> 단과대학 별 교육목표 및 인재상에 대한 만족도

| | 인문융합대학 | 사회과학대학 | 생명보건대학 | 공과대학 | 과학기술융합대학 | 예체능대학 |
|---|---|---|---|---|---|---|
| 교육목표 및 인재상 | 3.52 | 3.57 | 3.56 | 3.42 | 3.37 | 3.45 |




[그림 7] 단과대학 별 교육목표 및 인재상에 대한 만족도

교육목표 및 인재상 영역의 집단 간 평균차이를 확인하기 위해 일원분산분석을 수행하였으며 그 결과는 <표 12>와 같다. 단과대학별로 통계적으로 유의한 대학 교육 만족도 평균 차이가 있으므로 사후 검정을 수행하였다. 사후검정결과 공과대학과 과학기술융합대학에서의 만족도가 타 단과대학의 만족도에 비해 상대적으로 낮은 것으로 나타났다.

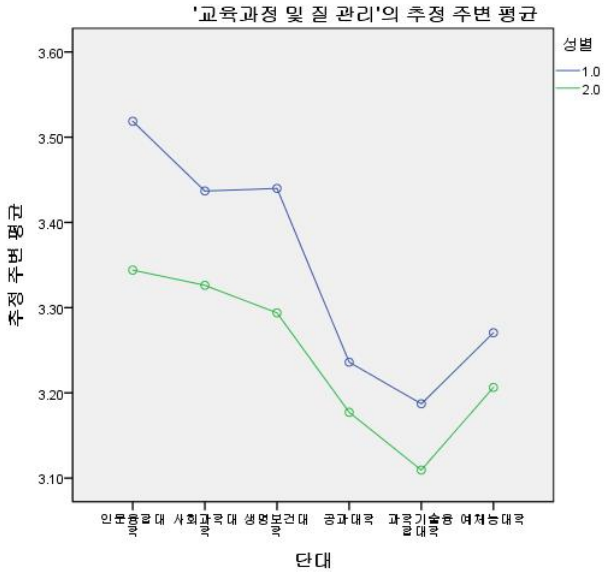
## 3. 상호작용 분석 결과

단과대학과 성별에 따른 대학 만족도의 상호작용이 존재하는지를 확인하기 위하여 이원분산분석을 수행하였다. 종속변인은 대학 만족도의 각 하위 요인들이었으며 독립변인은 성별과 단과대학으로 설정하였다. 일반적으로 여학생들이 공과대학이나 과학기술대학에서 학습하는데 더 어려움을 겪고, 인문대학이나 사회과학대학에서 남학생이 학습에 어려움을 겪는다는 통설을 확인하기 위함이다. 이에 대학만족도 하위 요인에 대한 이원분산분석을 수행하였고 그 결과는 다음 <표 14>에서 <표 17>과 같다. 모든 하위 요인에서 성별과 단과대학의 상호작용효과는 없는 것으로 나타났으며 각각의 주효과만 유효한 것으로 나타났다. 이는 우리가 가지고 있는 통념에서 벗어난 결과이다. 학생들은 성별, 단과대학에 따라 서로 다른 만족도를 보이지만, 두 변인의 상호작용은 없었다. 다양한 하위 영역 중 성별과 단과대학에 따라 가장 큰 차이를 보인 것은 인문대학의 남학생과 여학생의 교육환경 및 체계 영역에서였다. 남학생들은 교육환경 및 체계

에 대해 다른 집단에 비해 가장 높은 만족도 평균을 보인 반면, 여학생들은 하위권에 머무는 것으로 나타났다. 또한, 생명보건대학 학생들에게서도 동일한 현상을 발견할 수 있었는데, 남학생의 평균은 높은 반면 여학생의 평균은 하위권의 만족도를 나타냈다. 반면, 다른 하위 영역들에서는 단과대학별 만족도 하위 영역별 평균 점수에 차이가 있었으나 성별과의 상호작용은 보이지 않는 것으로 나타났다.

〈표 20〉 성별, 단과대학 별 교육목표 및 인재상에 대한 만족도 이원분산분석 표

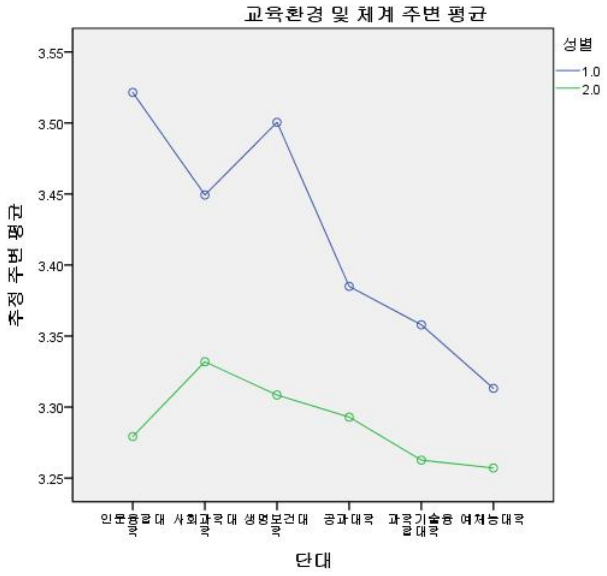
| | | 제곱합 | 자유도 | 평균제곱 | F | 유의확률 |
|---|---|---|---|---|---|---|
| 교육과정 및 질 관리 | 성별*단과대학 | 1.167 | 5 | .233 | .639 | .670 |
| | 단과대학 | 37.118 | 5 | 7.424 | 20.347 | .000 |
| | 성별 | 7.856 | 1 | 7.856 | 21.532 | .000 |




[그림 8] 교육과정 및 질관리 상호작용 도표

〈표 21〉 성별, 단과대학 별 교육목표 및 인재상에 대한 만족도 이원분산분석 표

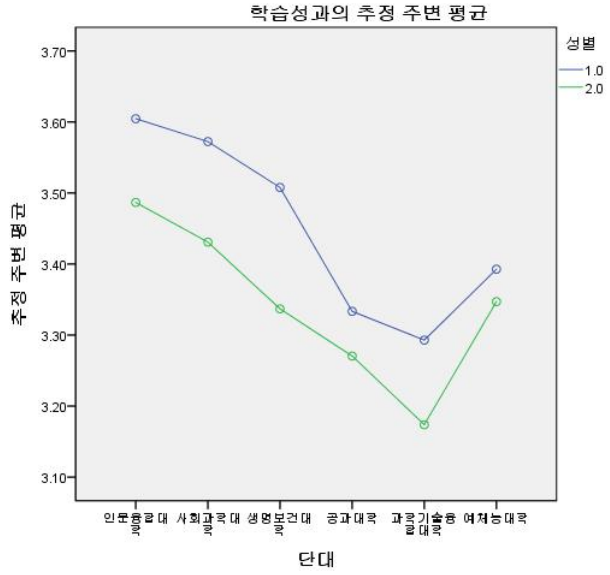
| | | 제곱합 | 자유도 | 평균제곱 | F | 유의확률 |
|---|---|---|---|---|---|---|
| 교육환경 및 체계 | 성별*단과대학 | 2.173 | 5 | .435 | 1.145 | .334 |
| | 단과대학 | 6.267 | 5 | 1.253 | 3.303 | .006 |
| | 성별 | 12.432 | 1 | 12.432 | 32.763 | .000 |




[그림 9] 교육환경 및 체계 상호작용 도표

〈표 22〉 성별, 단과대학 별 교육목표 및 인재상에 대한 만족도 이원분산분석 표

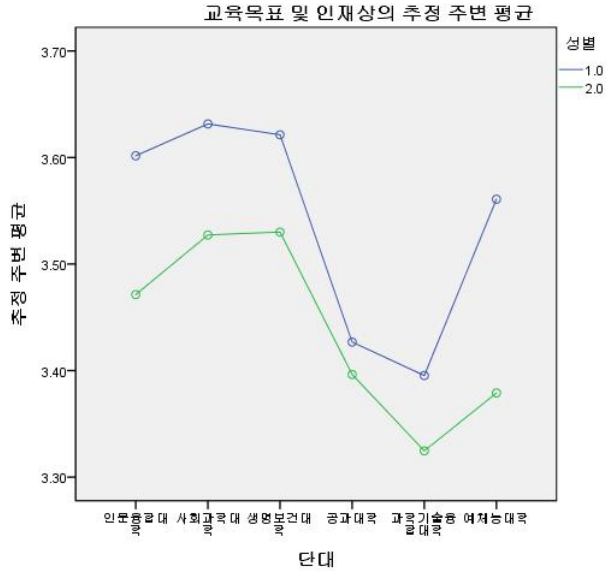
| | | 제곱합 | 자유도 | 평균제곱 | F | 유의확률 |
|---|---|---|---|---|---|---|
| 학습 성과 | 성별*단과대학 | 1.501 | 5 | .300 | .727 | .603 |
| | 단과대학 | 42.238 | 5 | 8.448 | 20.459 | .000 |
| | 성별 | 8.515 | 1 | 8.515 | 20.622 | .000 |




[그림 10] 학습성과 상호작용 도표

〈표 23〉 성별, 단과대학 별 교육목표 및 인재상에 대한 만족도 이원분산분석 표

| | | 제곱합 | 자유도 | 평균제곱 | F | 유의확률 |
|---|---|---|---|---|---|---|
| 학습 성과 | 성별*단과대학 | 1.805 | 5 | .361 | .939 | .454 |
| | 단과대학 | 28.937 | 5 | 5.787 | 15.052 | .000 |
| | 성별 | 7.267 | 1 | 7.267 | 18.900 | .000 |




[그림 11] 교육목표 및 인재상 상호작용 도표

# Ⅴ. 결론

본 연구는 각 대학에서 활발하게 진행되고 있는 대학교육만족도 조사를 수행한 뒤 대학교육만족도를 구성하는 하위 만족도의 평균에 영향을 미치는 요인을 탐색하고자 수행하였다. 이에 성별과 단과대학별로 대학교육만족도의 하위만족도에 집단 간 평균차이가 있는지를 확인하기 위해 평균비교(t-test, 일원분산분석)를 하였고, 성별과 단과대학의 상호작용을 확인하기 위해 이원분산분석을 수행하였다.

분석결과를 요약하면 다음과 같다. 성별에 따른 평균차이는 모든 대학교육만족도의 하위 만족도 요인에서 유의한 것으로 나타났으며 단과대학에 따른 평균차이도 존재하는 것으로 나타났다. 성별과 단과대학에 의한 상호작용은 없는 것으로 나타났다. 이는 사회적 통념 중 하나인 여학생은 이공계 대학에서 낮은 만족도를 보이고, 남학생은 인문계에서 낮은 만족도를 보일 것이라는 통념에 정 반대의 결과를 보여준다. 모든 단과대학에서 남학생의 평균이 여학생의 평균보다 높게 나타났다.

특기할만한 점으로는 첫째, 교육과정 및 질관리, 학습성과, 인재상에 대한 만족도에서는 남녀의 만족도 평균이 비슷한 경향을 보였으나, 교육 환경 및 체계에서는 인문대학과 생명보건대학

에서 남학생과 여학생의 평균 차이가 크게 나타나는 것을 확인할 수 있었으며, 두 단과대학에서 모두 남학생의 평균이 크게 높은 것으로 나타났다.

둘째, 예체능 대학의 학생들의 경우 교육과정 및 질관리 영역에는 낮은 만족도를 나타냈지만 학습성과 영역에서는 높은 만족도를 보이는 것으로 나타났다. 이는 학생들이 학습성과 영역에서 느끼는 만족감에 교육과정이나 질 관리 외의 영역들이 포함되어 있음을 암시하며, 이를 위한 후속 연구가 필요하다. 이상을 종합해 볼 때 대학생들의 성별과 단과대학이 대학교육만족도의 하위 만족도 영역들에 영향을 미치며, 학생들의 소속 단과대학과 성별에 따른 맞춤형 교육서비스 제공을 통해 교육 서비스의 질을 제고할 수 있을 것이다.

# 참고문헌


권대봉(2000). 대학생 교육수요 만족도에 관한 조사연구. 교육인적자원부

권대봉, 오영재, 박행모, 손준종, 송선희(2002). 대학생들의 교육 만족도 결정요인에 관한 탐색적 논의. **교육학연구, 40**(3), 181-202.

김지영, 성창수, 박주연(2017). 대학 창업교육의 고도화를 위한 창의적 문제해결역량교육에 대한 고찰. **벤처창업연구, 12**(2), 65-76.

박제일(2016). 대학만족도 검사의 개발연구. 재활심리연구, 23(4), 843-855.

박혜림(2015). 대학의 특성을 반영한 교육만족도 조사도구 개발. **예술인문사회융합멀티미디어논문지 5**(6), 375-386.

백승학(2010). 대학 만족도가 타인에 대한 입학권유에 미치는 영향: 경로분석적 접근. **한국전문대학교육연구학회논문집, 11**(3), 145-152.

신소영, 권성연(2013). 대학 교육만족도 측정도구 개발 및 타당화 연구. **교육과학연구, 44**(3), 107-132.

신소영, 권성연(2014). 학생 특성별 대학교육 만족요인에 대한 중요도-만족도 분석. **한국교육학연구, 20**(1), 253-280.

이관형, 장용규, 안현미(2000). 교육대학 남학생들의 대학생활 적응과 만족도 연구. **학생생활 연구, 26**. 111-145. 서울교육대학교 학생생활 연구소.

이정원, 임지영(2008). 대학교육 서비스품질, 학교브랜드 및 취업과 학생의 만족도, 충성도의 영향관계. **63차 제주 학술심포지엄.** 570-580.

정주영(2013). 대학교육만족도, 직업가치관, 직업만족도에 관한 구조적 분석. **교육행정학연구, 31**(1), 53-83.

채창균, 최지희, 옥준필(2005). 대졸 청년층의 대학교육만족도, 한국직업능력개발원 연구보고서.

최영준 (2013). 대학교육만족도, 직업가치관, 직업만족도에 대한 구조적 분석, **교육행정학연구, 31**(1), 53-83.

함은혜, 박상옥, 김은경 (2017). 대학교육만족도 영역별 점수의 의미와 활용, **아시아교육연구, 18**(4), 713-738.

한은숙, 김종두(2003). 사범대학생의 교육만족도에 미치는 영향요인 분석. **한국교원교육연구, 20**(3), 313-335.

Aldridge, S., & Rowley, J. (1998). Measuring customer satisfaction in higher education. **Quality Assurance in Education, 6**(4), 197-204.

Astin, A. (1993). What matters in college: Four critical years revisited. San Francisco,

CA:Jossey-Bass Publishers.

Elliot, K. M., & Shin, D. (2002). Student satisfaction: an alternative approach to assessing this important concept. **Journal of Higher Education Policy and Management, 24**(2), 197-209.

Krause, K. L. (2005). Serious thoughts about dropping out in first year: Trends, patterns and implications for higher education. **Studies in Learning, Evaluation, Innovation and Development, 2**(3), 53-68.

Levitz, R. S., Noel, L., & Richter, B. J. (1999). Strategic moves for retention success. **New Directions for Higher Education, 108**, winter, 31-49.

Schreiner, L. A. (2009). Linking student satisfaction and retention. Noel-Levitz Research study.

Yu, G. G., & Kim, J. H. (2008). Testing the mediating effect of the quality of college life in the student satisfaction and student loyalty relationship. **Applied Research Quality Life, 3**, 1-21.

• 논문 접수 2019년 08월 01일 / 수정본 접수 08월 29일 / 게재 승인 08월 30일

• 조명근 : 성균관대학교 교육학과 졸업. 성균관대학교에서 교육학으로 석사 및 박사학위 취득. 현재 호서대학교 더:함 교양대학에서 조교수로 재직하고 있으며 교육수요자만족도, 전공역량, 핵심역량에 관심을 가지고 있음.

교육연구와 실천
Journal of Educational Study and Practice
제10권 제1호, 2019, pp. 41-56

# A Comparison Of Classical Test Theory Equation And Item Response Theory Equation On Anchor Test Of Non-Equivalent Groups

Hwang Sujin*
(Sungkyunkwan University)

< Abstract >

The purpose of this study is to compare of the performance of classical test theory equation and item response theory equation on non-equivalent group anchor test by simulation. Data were generated by package ‘kequate’ in R and compare of 5 types equatings. Finally, to analyze the effect of the conditions in this study, the equated scores were analyzed by package ‘equating’, ‘eqateIRT’, ‘plink’ in R.

The result of this study shows that especially Rasch’s is more useful than any other equatings. And the Tucker methods shows the smallest error in the middle score. On the other hand, tow-parameter IRT mode is large error in most cases.

Key words: Equating, claasical test theory, item response theory, Anchor-test, non-equivalent group

* Sujin Hwang(25-2, Sungkyunkwan-ro, Jongno-gu, Seoul, Korea; s1324@skku.edu)

교육연구와 실천
Journal of Educational Study and Practice
제10권 제1호, 2019, pp. 41-56

# 모의실험을 통한 비 동등집단의 고전적 검사점수 동등화 방법과 문항반응이론 동등화 방법 비교

황수진(성균관대학교)*

< 요 약 >

본 연구는 현행 대학입학수학능력시험 수학과목에서 난이도 차이가 나는 두 유형의 시험을 능력이 다른 집단에 실시했을 때를 고려하여 가장 적합한 동등화방법을 모색하기 위해 모의 실험을 진행하였다. 고전검사이론 동등화 방법 중 선형동등화 방법인 Tucker방법・Levine의 관찰점수 방법, 등백분위 동등화 방법인 빈도추정방법, 그리고 문항반응이론 동등화 방법중 1-모수 방법인 Rasch모형, 2-모수 방법을 비교하고 분석한 후 각 동등화 방법들 중 가장 적합한 방법을 검토하였다. 실험 설계를 살펴보면 검사는 총 30문항으로 20문항은 선택문항, 10문항은 가교문항으로 구성되어 있고, 선택문항의 난이도 2가지 조건과 집단의 능력차이 2가지 조건을 두어 총 4가지의 조건을 설정하였다.

실험에서 고려한 4가지 조건들의 동등화의 결과는 비슷한 맥락으로 나타났다. 전반적으로 IRT 1-모수 방법인 Rasch 모형이 안정적으로 동등화를 하고 있는 것으로 확인되었다. 또한 Tucker방법의 경우 중간점수대에서 다른 동등화 방법들에 비해 가장 동등화 오차가 작았다. 반면 동백분위 동등화방법의 경우 낮은점수와 높은 점수에서 동등화 오차가 컸으며, IRT 2-모수 모형의 경우 동등화 오차가 대부분의 점수대에서 큰 것으로 나타났다. 결과적으로 Rasch모형의 동등화 안정성이 가장 좋았으며, 만약 중간점수대가 모여있는 시험이라면 중간 점수대에서 동등화 오차가 제일 작았던 Tucker방법 또한 좋은 선택지가 될 수 있다. 결과적으로 본 연구는 향후 대학수학능력시험에서 수학과목을 동등화 할 때, 동등화 방법을 고려하는데 작은 시사점을 줄 것으로 기대된다.

**주제어: 동등화, 고전검사이론, 문항반응이론, 가교문항, 비 동등집단**

* 황수진(서울특별시 종로구 성균관로 25-2, s1324@skku.edu)

# I. 서론

매해 실시되는 수능시험은 같은 특성을 측정하기 위하여 동일한 내용으로 구성되어 서로 다른 유형의 문항들로 이루어져 있다. 이러한 경우 매번 치러지는 검사가 공정한지의 여부가 중요해지며, 이를 확보하기 위해 검사에 구성된 문항 내용, 문항 난이도, 신뢰도 등이 통제된 동형검사의 제작이 필수적으로 요구된다.

예를 들어 대학수학능력시험의 성적의 경우 진학 가능한 학교를 결정하는데 지대한 영향을 미치며, 한 해에 한번 치러지기 때문에 시험을 치른 해의 난이도가 높았다고 생각되는 경우에는 진학을 포기하고 다음 해에 치러지는 시험에 응시하려는 학생들도 생긴다. 특히 현재 시행되는 대학수학능력시험 수리영역의 경우 어떠한 난이도의 유형을 선택하느냐에 따라 수험생의 점수가 달라질 수 있다. 실제로 몇몇 대학에서 난이도를 반영할 것이라고 예상하는 자연계열의 경우 난이도 유형을 고려하지 않고 같은 반영비율로 점수를 반영하고 있기 때문에, 난이도 높은 유형을 치러 낮은 점수를 받은 학생은 난이도 낮은 유형을 치른 학생에 비해 높은 능력을 갖고 있다고 하더라도 불리해 진다. 또한 선택과목의 경우에도 매해 치러지는 수능의 난이도가 차이가나면, 같은 능력을 가진 수험생의 점수가 달라질 수 있게 된다.

현재 시험 간 결과해석의 일관성을 유지하기 위하여 교육과정평가원에서는 시험의 난이도를 사전에 분석하고 조정하기 위한 수능모의평가를 사전검사(pilot test)로 실시하여 수학능력시험의 난이도를 일관되게 유지하도록 노력하고 있지만, 실제로 수학능력시험의 통계치를 살펴보면 난이도가 일관되게 유지되고 있지 않다. 이를 구체적으로 살펴보면 2018학년도 수능시험 수리영역의 1등급 비율은 수리가형은 5.13%(8,879명), 수리나형은 7.68%(25,768명)이었고 사회탐구영역에서의 1등급 비율은 사회문화 4.21%(6,289명), 생활과 윤리 6.66%(10,764명), 윤리와 사상 7.97%(2,560명), 경제 11.75%(637명) 등 이었고 과학탐구영역에서의 1등급 비율은 물리1 4.66%(2,691명), 화학1 4.25%(4,240명), 생명과학1 7.52%(11,264명), 물리2 5.04%(143명), 화학2 4.31%(144명), 생명과학2 4.03%(368명), 지구과학2 5.70%(594명) 등이었다. 2019학년도 수능시험 수리영역의 1등급 비율은 수리가형은 6.33%(10,675명), 수리나형은 5.98%(20,368명)이었고, 사회탐구영역에서의 1등급 비율은 사회문화 10.17%(15,240명), 생활과 윤리 8.05%(13,131명), 윤리와 사상 9.88%(3,306명), 경제 6.13%(334명) 등 이었고 과학탐구영역에서의 1등급 비율은 물리1 5.52%(3,210명), 화학1 6.63%(5,774명), 생명과학1 4.98%(7,520명), 물리2 6.32%(185명), 생명과학2 4.98%(423명) 등이었다(교육과정평가원 수능통계, 2019).

즉, 2018학년도에 비해서 2019학년도의 과목별 시험 난이도 차이에 따른 1등급의 비율이 달라졌으며, 선택과목의 난이도 차이에 따라 1등급의 비율도 달라지고 있는 것을 확인할 수 있다. 따라서 난이도의 차이에 따라 특정 시기에 시험을 치르는 사람이 다른 시기에 시험을 치르는

사람보다 유·불리가 달라지고 있으며, 특정과목을 선택하는 사람 또한 선택 과목에 따라 유·불리가 달라지는 상황이 발생하고 있다. 이는 수능시험의 신뢰도와 타당도를 낮추며, 평가의 공정성 또한 의심받게 된다. 하지만 모든 것이 동일한 동형검사를 제작한다는 것은 현실적으로 불가능하기 때문에, 이를 해결하기 위해 완벽히 통제되지 못한 검사의 난이도를 검사가 실시된 후 조정하는 과정인 검사 동등화 작업이 필요하게 된다.

수학능력시험의 경우 시험의 난이도를 조정하기 위하여 6월과 9월에 수능모의평가를 치르고 있으며, 이를 기반으로 시험 간 비교 가능성을 높일 수 있도록 가교역할을 할 수 있는 문항 또는 피험자 설계가 필요한 요소이다. 그러나 검사 동등화는 이론적 근거, 함수의 종류, 자료의 범위, 문항모수의 수 등에 따라 여러 방법으로 구분된다. 이중 한 가지 방법을 선택할 때 동등화 방법의 정확성, 안정성, 사용자 편의성 등을 고려해야 하며, 이 때문에 어떤 것이 가장 좋은 방법인지는 확실하게 말하기 어렵다. 따라서 본 연구는 향후 수학능력시험에서 보다 타당한 동등화 설계를 위한 기초자료를 제공하고자 하는데 그 목적이 있으며, 수학능력시험의 실제 자료 특성을 반영하여 모의자료를 생성하고, 이를 이용해 여러 가지 동등화 방법들을 활용한 결과를 비교·분석하고자 한다.

1985년에 이종성에 의해 처음 소개된 검사동등화 이론은 남현우(1992a; 1992b), 성태제(1991)에 의해 연구가 진행되기 시작하였다. 김현철(1999)은 척도화의 체계적 오차에 대해, 부재율(1999)은 검사동등화할 때의 고려점 들에 대해 설명하고 있다. 2000년대 이후에는 모의실험에 의한 대학입학수학능력시험 선택과목 점수 동등화 방법들의 적합성을 비교하는 연구(김현철, 2003a; 김현철, 2003b)나, 외적 가교검사의 통계적 구성 조건 완화가 검사동등화에 미치는 영향(이현숙·김성훈, 2010), 재검 여부와 재검 횟수에 따른 검사 동등화 차이 연구(김희경, 2007), 가교검사의 길이와 비 동등집단의 능력차이와 분산차이에 대한 연구(조명근, 2013) 등 주로 시뮬레이션이나 다양한 조건에서 어떤 동등화 방법이 더 효율적인지 연구되고 있다. 하지만 대부분의 연구는 고전검사이론들만 비교하거나, 문항반응이론(IRT)의 동등화 방법들만 비교하여 두 이론을 통합하여 비교하는 연구는 부족한 실정이다.

만약 현재 실시하고 있는 대학수학능력시험에 동등화 방법을 적용한다고 가정한다면, 능력차이가 나는 두 집단이 치른 검사가 난이도 차이가 있을 경우 두 검사간의 동등화 방법을 고전검사이론의 동등화 방법들과 문항반응이론의 동등화 방법들을 비교·분석하여 각 상황에서 가장 적합한 동등화 방법을 찾아낼 필요가 있다. 따라서 본 연구에서는 능력이 다른 두 집단에 난이도가 다른 검사를 실시했을 때 적합한 방법을 찾기 위하여, 고전검사이론 동등화 방법 중 선형동등화 방법인 Tucker방법·Levine의 관찰점수 방법, 등백분위 동등화 방법인 빈도추정방법, 그리고 문항반응이론 동등화 방법중 1-모수 방법인 Rasch모형, 2-모수 방법을 비교하고 분석한다.

# Ⅱ. 연구 설계

본 연구는 모의자료를 생성하여 검사조건을 다르게 설정하여 공통문항 비동등 집단의 동등화 작업을 할 경우, 동등화 방법들 간의 동등화 점수 및 동등화 오차에 차이가 있는지 비교하고 상대적으로 어떤 방법이 더욱 적절한지를 파악하는데 그 목적이 있으므로 검사시행 조건들 중 비동등성 정도, 검사의 난이도, 공통문항 수, 표본의 크기 등이 중요하다. 따라서 검사 조건들 중 본 연구의 목적인 집단의 능력차이와 검사의 난이도를 다르게 설정하여 동등화 방법들이 어떻게 수행되는지 살펴보고자 한다. 동등화 방법들은 통계 프로그램 R에 내장된 'equate'패키지와 'mirt' 패키지를 이용하여 동등화 점수를 구한 뒤, 조건에 따른 동등화 방법들의 결과로 나타난 동등화 표준오차(Standard Error of Equating : SEE)를 그래프로 나타내고 비교한다.

## 1. 동등화 설계 및 방법

검사 동등화의 방법을 먼저 선택한 후 자료수집 설계를 하는 것이 아니라, 시행할 자료의 조건에 따라 적절한 동등화 방법이 적용되기 때문에 자료수집 설계는 실제적으로 중요한 문제이다. 동등화 자료수집 설계의 목적은 평가되는 집단들 간의 특성 차이와 검사의 난이도차이 등을 구별할 수 있도록 자료를 수집하고 설계하는 것에 있다. 이를 위해 Kolen과 Brennan(2004)는 두 집단이 동등하도록 검사유형을 관리하는 방법과 동등화될 검사유형에 공통문항을 구성하는 방법이라는 두 가지 관점으로 접근하였다. 이러한 접근들은 단일집단 설계, 교차 설계, 동등집단 설계, 비 동등집단 가교검사 설계로 요약될 수 있다.

이 중 본 연구에서 적용할 비 동등 집단 가교검사 설계(Non-equivalent groups with Anchor Test Design: NEAT)는 검사 비밀 유지가 필요하거나 혹은 다른 실제적인 이유로 다른 유형의 두 검사를 동시에 실시할 수 없을 때 유용한 방법으로, 본 연구에서처럼 두 집단의 능력에 차이가 있는 경우 등 두 집단이 동등하지 않을 때 사용한다. 능력에 차이가 있는 두 집단에 다른 유형의 검사를 실시하게 되면, 피험자가 받는 점수가 두 시험의 유형차이 때문인지 두 집단의 능력 때문인지 판단하기 어려운 문제가 발생하고, 이를 조정하기 위하여 이 설계가 적용된다. 이 설계에서 가교 검사의 역할이 매우 크며, 가교 검사는 능력 차이가 있는 두 집단과 다른 유형의 검사를 조정하는 역할을 하게 된다. 이를 위하여 가교검사의 문항은 다른 유형의 두 검사의 문항에 포함되어 있어야 하고, 다른 두 검사에 포함된 가교검사 문항의 기능은 동일해야하기 때문에 문항 내의 단어, 문항의 순서, 답지의 순서 등은 같게 구성되어야 한다.

검사점수의 동등화 방법은 이론적 근거, 함수의 종류, 자료의 범위, 문항모수의 수 등에 따라

여러 방법으로 구분된다. 이론에 기초하여 나눠보면, 고전검사이론에 바탕을 둔 동등화 방법과 문항반응이론(Item Response Theory : IRT)에 의한 동등화 방법으로 나눌 수 있다(Kolen&Brennan, 2004). 이때 사용되는 함수의 종류에 따라 이때 사용되는 함수에 따라 선형동등화(Linear equating)방법과 동백분위 동등화(Equipercentile equating)방법으로 구분되며, 동등화 함수의 산출에 계산되는 자료의 범위에 따라 이단계 동등화와 합집단 동등화로 구분된다. 그러나 고전검사이론은 검사와 집단의 특성에 영향을 받는다는 한계를 갖는다.

반면 고전검사이론에 비해 문항반응이론은 검사와 집단의 특성에 영향을 받지 않는다. 즉, 문항반응이론은 문항의 모수들이 피험자의 능력수준에 따라 다르게 반응하지 않는 문항 독립성과 피험자의 능력은 어떤 검사를 실시하여도 변하지 않는다는 피험자 독립성을 가진다. 따라서 문항반응이론을 적용한 결과인 능력모수와 문항모수의 추정치는 상호비교 가능하며, 별도의 동등화절차는 필요하지 않다고 본다. 하지만 각기 다른 집단에 두 검사를 실시한 후 문항반응이론을 적용하였을 때 도출된 두 검사의 문항모수와 피험자의 능력모수의 추정치가 다르게 나오는 이유는 문항반응이론의 척도가 지닌 불확정성에 기인한다(Kolen&Brennan, 1995; 박정, 2002). 따라서 이러한 척도를 보정하기 위하여 두 검사에 공통으로 속한 문항의 모수들을 분리하여 추정한 뒤, 이 공통문항의 모수들을 기초로 선형적 관계를 유도하여 변환하게 된다. 이러한 선형적인 관계에서 유도하는 선형변환 동등화 상수를 찾는 방법으로 크게 평균-표준편차 방법, 관찰점수 동등화방법, 검사특성곡선 방법, 그리고 진점수 동등화 방법으로 나눌 수 있다. 본 연구에서는 고전검사이론에서 선형 동등화 방법들 중 Tucker 방법, Levine 방법, 동백분위 동등화 방법 중의 하나인 빈도추정 방법과 문항반응이론에서 난이도 만을 고려한 Rasch 모형(1-모수 모형), 2-모수 모형을 평균-표준편차 방법을 활용하여 동등화한 후 비교하고자 하였다.

## 2. 연구자료

본 연구에서 모의실험을 진행하기 위한 다양한 조건의 모의자료를 생성하기 위하여, 집단 능력모수의 범위는 1992년 2회 실시된 대학수학능력 실험평가의 문항분석을 기초로 산정(황소림, 1993)하였으며, 다양한 능력의 피험자가 충분히 많은 경우에는 피험자의 능력이 정규분포한다(이상하, 박도영, 2012)는 가정하에 집단의 응답반응 자료를 생성하였다. 현재 실시되는 수학능력시험에서 난이도가 차이나면서, 응시 집단의 능력에도 차이가 명확한 과목은 수학과목으로, 난이도가 다른 두 유형의 시험의 표준점수를 대학입학에 활용하고 있다. 수학과목의 두가지 유형의 검사를 동등화했을 때 가장 적합한 방법을 찾고자 수학과목과 비슷하도록 문항수를 30문항, 총점 30점으로 설정하였다. 가교문항 수는 수능의 수학과목에서 두 유형의 공통문항 수가 10문항~14문항 사이로 출제되는 경향(신일용 외, 2010)을 반영하여 최소의 공통문항인 10문항으

로 가정하였다. 집단의 크기는 고전검사이론과 문항반응이론의 동등화 방법들이 최소로 필요로 하는 인원이 1,500명(Kolen&Brennan, 1995; Zeng, 1991)이므로 본 연구에서는 충분히 큰 크기인 3,000명으로 설정하였다.

## 3. 실험조건

### 가. 집단의 능력차이

집단의 능력 차이에 따라 동등화 점수와 오차가 달라지는 지를 비교하기 위하여 검사 Y를 실시한 집단 Q의 능력은 평균 0, 표준편차 1인 정규분포 N(0,1)을 따르는 것으로 가정한다. 반면에 검사 X를 실시한 집단 P의 능력은 다음과 같은 세 가지 정규분포를 따르는 것으로 설정한다. 첫째, 평균 0.5, 표준편차 1인 정규분포 N(0.5,1), 둘째, 평균 1, 표준편차 1인 정규분포 N(1,1)이다.

### 나. 선택문항 특성의 차이

선택문항의 난이도 차이에 따라 동등화 점수와 오차가 달라지는지를 비교하기 위하여 다음과 같이 조건을 설정한다. 본 연구에서 지정한 문항 추측도와 변별도 모수의 범위는 앞서 설정한 피험자 집단 능력 모수 범위를 정할 때 참고한 자료를 바탕으로 설정하였다(황소림, 1993). 문항변별도・문항난이도・문항추측도 모수의 범위는 문항변별도・문항난이도・문항추측도 해석을 위한 언어적 표현상의 기준범위에 의한다(성태제, 1991).

첫째, 선택문항의 추측도의 경우 검사 X와 검사 Y 모두 최소 0.05, 최대 0.25인 균일분포 U(0.05, 0.25)로 가정한다. 선택문항 변별도의 경우 검사 X와 검사 Y는 낮은 수준인 0.5 부터 적절한 수준인 1까지의 균일분포 U(0.5,1)로 가정한다.

둘째, 선택문항 난이도의 경우 검사 Y는 평균 0.0과 표준편차 1인 정규분포 N(0.0,1)로 가정한다. 검사 X는 평균 0.5과 표준편차 1인 정규분포 N(0.5,1)일 때, 매우 어려운 난이도인 1.5과 표준편차 1인 정규분포 N(1.5,1)일 때로 설정한다.

### 다. 가교검사의 특성

공통문항 문항특성의 경우 1992년 실시된 실험평가의 가교문항 문항특성 분석(황소림,1993)을 기초로 설정하였다. 문항추측도는 최소 0.05, 최대 0.25인 균일분포 U(0.05, 0.25), 문항변별도는

최소 0.5, 최대 1인 균일분포 U(0.5, 1), 그리고 문항난이도는 평균 0, 표준편차 1인 정규분포 N(0, 1)로 가정하였다.

〈표 Ⅱ-1〉 실험 조건

| 집단 | 능력 | 검사의 난이도 |
|---|---|---|
| X | $\theta$ =0.3 | A. 보통 난이도(0.5) |
| | | B. 어려운 난이도(1) |
| | $\theta$ =1.0 | C. 보통 난이도(0.5) |
| | | D. 어려운 난이도(1) |
| Y | $\theta$ =0.0 | E. 쉬운 난이도(0) |

## 4. 모의 자료 생성

여러 동등화 방법이 설정된 조건 하에서 어떤 결과를 나타내는지 살펴보기 위하여 통계 프로그램 R에 내장된 'kequate' 패키지를 이용해 다음과 같은 모의자료를 생성한다. 동등화에 사용될 검사의 자료는 Ogasawara(2003)가 제시한 로지스틱함수를 이용한 3모수 IRT모형의 R-code를 이용하여 생성한다. 로지스틱 함수를 적용한 3모수 IRT모형에서 P집단의 $\theta$의 능력을 가진 피험자가 검사 X의 문항 $l$을 맞출 확률은 다음과 같다.

$$P_{Xl}(\theta_P) = c_{Xl} + \frac{1 - c_{Xl}}{1 + \exp[-a_{Xl}(\theta_P - b_{Xl})]}$$

$a_{Xl}$은 문항변별도, $b_{Xl}$은 문항난이도, $c_{Xl}$은 문항추측도

자료생성을 위해 사용된 모수들은 $a_{Xl}$ · $c_{Xl}$은 앞서 설정한 조건의 균일분포에서, $b_{Xl}$는 정규분포에서 추출한다. 문항 또한 앞서 언급한 조건들을 만족하는 자료를 생성하기 위하여, $P_{Xl}(\theta_P)$〉U(0, 1)이면 1, 그렇지 않다면 0으로 설정한다. 여기서 $P_{Xl}(\theta_P)$는 P집단의 $\theta$능력을 가진 피험자가 검사 X의 $l$문항을 맞출 확률을 나타내며, U(0,1)는 최솟값 0과 최댓값 1의 균일분포에서 가져온 난수이다. 검사 Y에 대한 자료생성도 이와 동일하다.

이러한 조건 하에서 생성된 표본의 선택문항과 가교문항에 대한 기초통계는 〈표 Ⅳ-1〉에 제시되어 있다. 집단의 능력에 따라 검사지의 평균이 달라지는 것을 확인할 수 있다. 동형문항인 가교문항의 점수들을 놓고 비교해 보면, 집단의 능력($\theta$)이 1일 때 가교문항의 점수가 약 6.7로

가장 높게 나타났고, 집단의 능력($\theta$)이 0.3일 때 약 5.8로 나타났으며, 집단의 능력($\theta$)이 0인 집단은 약 5.5의 점수를 보였다. 가교검사와 전체 검사의 상관계수는 약 0.8로 나타났으며, 이 수치는 동등화를 진행하기에 적절한 것으로 확인되었다(이용민, 2003).

<표 Ⅱ-2> 선택문항과 가교문항의 기초통계

| 조건 | | 검사지 | 평균 | 표준편차 | 최소값 | 최대값 | 검사와 가교문항간 상관계수 |
|---|---|---|---|---|---|---|---|
| X($\theta$=0.3) | A | 보통 난이도 선택문항 | 11.600 | 3.15 | 2 | 20 | 0.798 |
| | | 가교문항 | 5.834 | 1.96 | 0 | 10 | |
| | B | 어려운 난이도 선택문항 | 8.243 | 2.96 | 1 | 20 | 0.804 |
| | | 가교문항 | 5.765 | 2.00 | 0 | 10 | |
| X($\theta$=1) | C | 보통 난이도 선택문항 | 13.190 | 3.08 | 3 | 20 | 0.803 |
| | | 가교문항 | 6.727 | 1.90 | 0 | 10 | |
| | D | 어려운 난이도 선택문항 | 9.944 | 3.19 | 1 | 20 | 0.777 |
| | | 가교문항 | 6.751 | 1.90 | 1 | 10 | |
| Y($\theta$=0) | E | 쉬운 선택문항 | 11.790 | 3.25 | 2 | 20 | 0.804 |
| | | 가교문항 | 5.476 | 2.00 | 0 | 10 | |

# Ⅲ. 연구결과

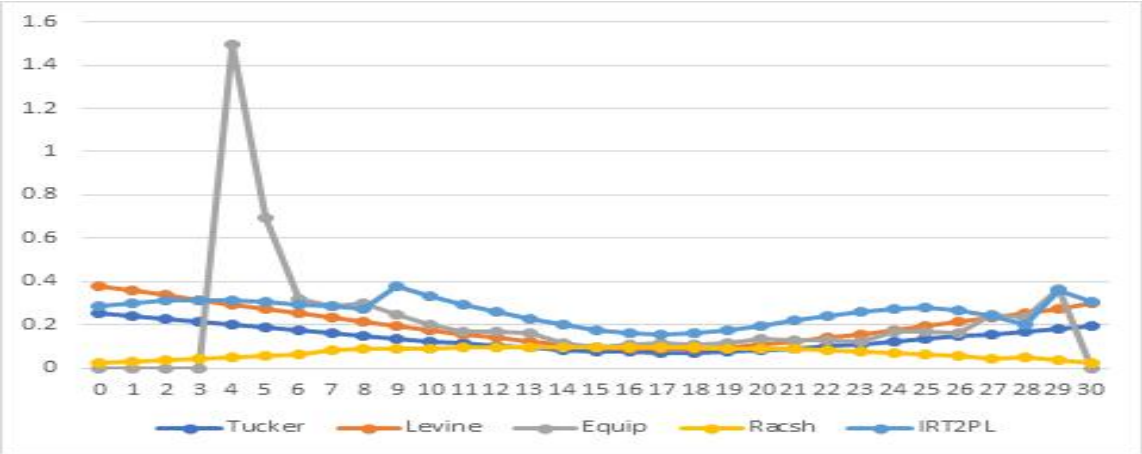
## 1. 집단의 능력($\theta$)의 차이가 작을 때의 동등화 오차

먼저 집단의 능력($\theta$)이 0.3 차이가 있고, 선택문항의 난이도가 0.5정도 차이 일 때, 선택 문항간의 평균을 살펴보았다. X($\theta$=0.3)집단에 실시한 A시험 선택문항의 평균은 11.6이었고, Y($\theta$=0)집단에 실시한 E시험 선택문항의 평균은 11.8이었다. 능력이 조금 더 좋은 X($\theta$=0.3)집단보다 능력이 낮은 Y($\theta$=0)집단 점수가 약 0.2점 더 높은 상황에서 여러 동등화 방법들의 동등화 오차가 <표 Ⅲ-1>에 제시되었다. 전반적으로 IRT 1-모수 방법인 Rasch 모형이 안정적으로 동등화를 하고 있는 것으로 나타났다. 동백분위 동등화방법의 경우 3점~10점 사이의 낮은점수와 23점~29점사이의 높은 점수에서 동등화 오차가 커지는 것을 확인할 수 있었다. 또한 Tucker방법의 경

우 중간점수대인 13점에서 21점 사이에서 다른 동등화 방법들에 비해 가장 동등화 오차가 작은 것으로 나타났다.

<표 Ⅲ-1> 집단의 능력차이가 작을 때 검사점수 동등화 오차(SEE)

| 구분 | Tucker | Levine | 동백분위 | Rasch | 2PL |
|---|---|---|---|---|---|
| 0 | 0.2546675 | 0.3780717 | 0.0000000 | 0.0203525 | 0.2858331 |
| 5 | 0.1872564 | 0.273798 | 0.6961185 | 0.0585382 | 0.3040329 |
| 10 | 0.1243035 | 0.1746839 | 0.2033934 | 0.0915203 | 0.3323108 |
| 15 | 0.0775551 | 0.0979063 | 0.0969722 | 0.0964015 | 0.1778997 |
| 20 | 0.0821905 | 0.1103578 | 0.1336991 | 0.0898634 | 0.1960879 |
| 25 | 0.1329390 | 0.1956887 | 0.1712831 | 0.064565 | 0.2789734 |
| 30 | 0.1968941 | 0.2965321 | 0.0000000 | 0.0263977 | 0.3082282 |



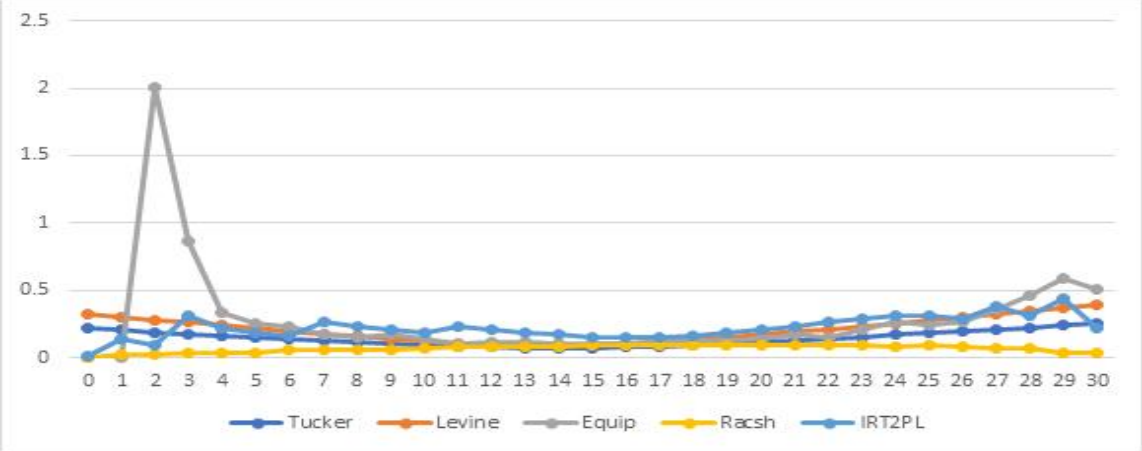

[그림 Ⅲ-12] 집단의 능력차이가 작을 때 검사점수 동등화 오차(SEE)

먼저 집단의 능력($\theta$)이 0.3 차이가 있고, 선택문항의 난이도가 1정도 차이 일 때, 선택 문항간의 평균을 살펴보았다. X($\theta$=0.3)집단에 실시한 B시험 선택문항의 평균은 8.2였고, Y($\theta$=0)집단에 실시한 E시험 선택문항의 평균은 11.8이었다. 능력이 조금 더 좋은 X($\theta$=0.3)집단이 어려운 난이도의 시험을 치게 되어, 능력이 낮은 Y($\theta$=0)집단보다 3.6점 낮았다. 이때 여러 동등화 방법들의 동등화 오차가 <표 Ⅲ-2>에 제시되었다. Rasch 모형이 가장 안정적으로 동등화 하는 것으로 나타났지만, 중간점수 부분에서는 Tucker방법이 Rasch 모형보다 더 작은 동등화 오차를 보였다. 이번에도 동백분위 동등화 방법은 점수가 낮은 경우와 높은 경우에서 가장 큰 동등화

오차를 갖는 것으로 확인되었다.

<표 Ⅲ-2> 검사의 난이도 차이가 클 때 검사점수 동등화 오차(SEE)

| 구분 | Tucker | Levine | Equipercentile | Rasch | 2PL |
|---|---|---|---|---|---|
| 0 | 0.217558 | 0.326534 | 0.0000000 | 0.003837 | 0.014726 |
| 5 | 0.149004 | 0.217811 | 0.2575632 | 0.039159 | 0.181612 |
| 10 | 0.090736 | 0.121923 | 0.1352371 | 0.073580 | 0.189703 |
| 15 | 0.073649 | 0.094318 | 0.1052734 | 0.087986 | 0.156013 |
| 20 | 0.117370 | 0.171853 | 0.1440919 | 0.098069 | 0.206743 |
| 25 | 0.182449 | 0.277088 | 0.2481945 | 0.089844 | 0.312143 |
| 30 | 0.252903 | 0.388109 | 0.5072934 | 0.033094 | 0.219746 |




[그림 Ⅲ-13] 검사의 난이도 차이가 클 때 검사점수 동등화 오차(SEE)

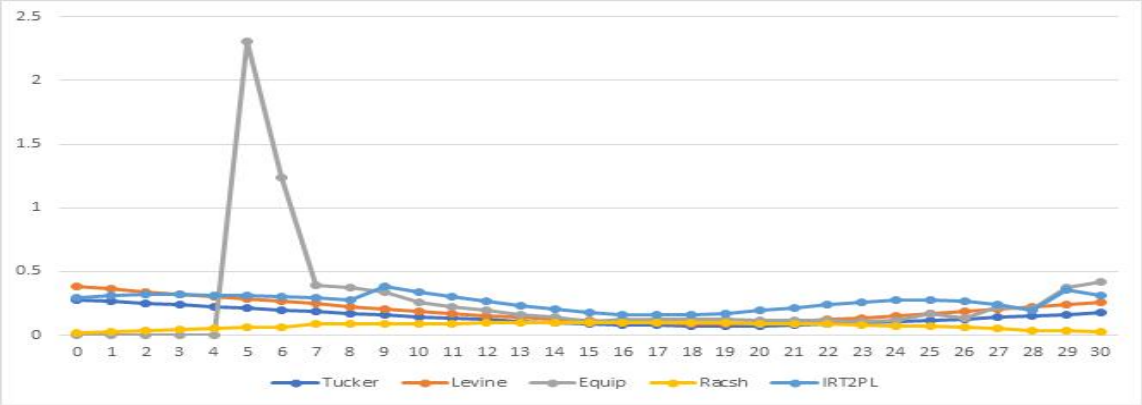
## 2. 집단의 능력($\theta$)의 차이가 클 때의 동등화 오차

다음으로 집단의 능력($\theta$)차이가 크고 선택문항의 난이도가 0.5정도의 차이 일 때 선택문항간 평균을 살펴보면, X($\theta$=1)집단에 실시한 C시험 선택문항의 평균은 13.2로 Y($\theta$=0)집단에 실시한 E시험 선택문항의 평균보다 1.4점 높았다. 능력이 좋은 X($\theta$=1)집단과 능력이 낮은 Y($\theta$=0)집단 점수 차이가 큰 능력 차이에도 불구하고 약 1.4점 차이만 있는 상황에서 여러 동등화 방법들의

동등화 오차가 <표 Ⅲ-2>에 제시되었다. Rasch 모형의 오차가 대부분의 경우 가장 작은 것으로 나타났다. 또한 집단의 능력 차이가 작을 때와 마찬가지로 동백분위 동등화방법의 경우 4점~9점 사이의 낮은 점수와 26점~30점 사이의 높은 점수에서 동등화 오차가 커졌으며, Tucker방법의 경우 중간점수대인 15점에서 21점 사이에서 다른 동등화 방법들에 비해 가장 작은 동등화 오차를 보였다.

<표 Ⅲ-3> 집단의 능력차이가 클 때 검사점수 동등화 오차(SEE)

| 구분 | Tucker | Levine | Equipercentile | Rasch | 2PL |
|---|---|---|---|---|---|
| 0 | 0.280785 | 0.382756 | 0.000000 | 0.022073 | 0.292272 |
| 5 | 0.212072 | 0.284259 | 2.306904 | 0.058951 | 0.310334 |
| 10 | 0.146612 | 0.190110 | 0.262118 | 0.092010 | 0.339205 |
| 15 | 0.091666 | 0.111884 | 0.109409 | 0.097871 | 0.180294 |
| 20 | 0.075475 | 0.099660 | 0.114530 | 0.093043 | 0.193820 |
| 25 | 0.115659 | 0.168475 | 0.166029 | 0.068567 | 0.278560 |
| 30 | 0.177325 | 0.260525 | 0.417394 | 0.024909 | 0.309574 |



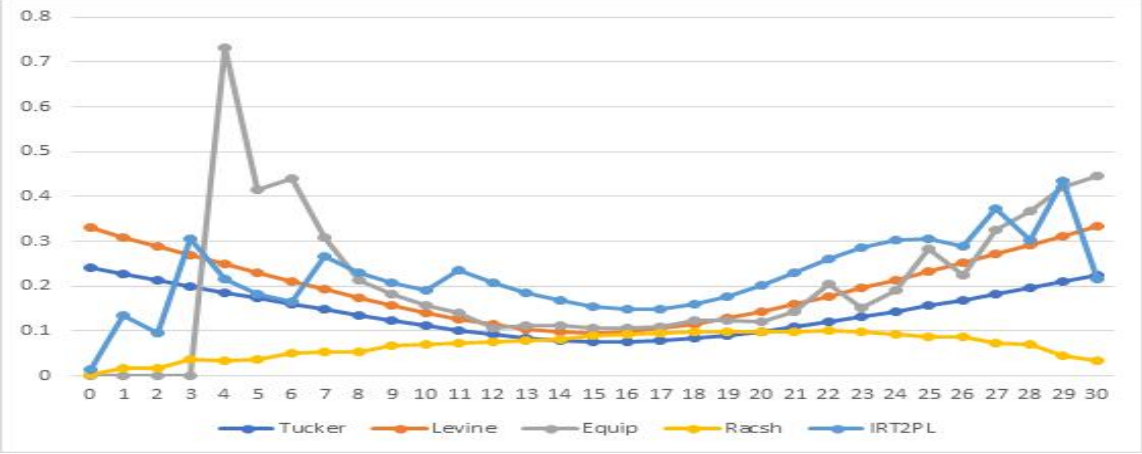

[그림 Ⅲ-14] 집단의 능력차이가 클 때 검사점수 동등화 오차(SEE)

집단의 능력($\theta$)차이와 선택문항의 난이도가 크게 차이가 있을 때를 살펴보면, X($\theta$=1)집단에 실시한 D시험 선택문항의 평균은 9.9로 Y($\theta$=0)집단에 실시한 E시험 선택문항의 평균보다 1.9점 낮았다. 이때 여러 동등화 방법들의 동등화 오차가 <표 Ⅲ-2>에 제시되었다. 앞에 살펴본 조건들에서의 동등화와 마찬가지로 Rasch 모형의 동등화가 가장 안정적인 것으로 나타났다. 또한 Tucker 모형의 경우 14점부터 20점까지 구간에서 가장 낮은 동등화 오차를 보였다. IRT 2-모수

모형의 오차가 대부분 크게 나타났다.

〈표 Ⅲ-4〉 집단의 능력과 검사의 난이도 차이가 모두 클 때 검사점수 동등화 오차(SEE)

| 구분 | Tucker | Levine | Equipercentile | Rasch | 2PL |
|---|---|---|---|---|---|
| 0 | 0.217558 | 0.326534 | 0.0000000 | 0.003837 | 0.014726 |
| 5 | 0.149004 | 0.217811 | 0.2575632 | 0.039159 | 0.181612 |
| 10 | 0.090736 | 0.121923 | 0.1352371 | 0.073580 | 0.189703 |
| 15 | 0.073649 | 0.094318 | 0.1052734 | 0.087986 | 0.156013 |
| 20 | 0.117370 | 0.171853 | 0.1440919 | 0.098069 | 0.206743 |
| 25 | 0.182449 | 0.277088 | 0.2481945 | 0.089844 | 0.312143 |
| 30 | 0.252903 | 0.388109 | 0.5072934 | 0.033094 | 0.219746 |




[그림 Ⅲ-15] 집단의 능력과 검사의 난이도 차이가 모두 클 때 검사점수 동등화 오차(SEE)

# Ⅳ. 결론

본 연구는 현행 대학입학수학능력시험 수학과목에서 난이도 차이가 나는 두 유형의 시험을 능력이 다른 집단에 실시했을 때를 고려하여 가장 적합한 동등화방법을 모색하기 위해 모의 실험을 진행하였다. 실험 설계를 살펴보면 검사는 총 30문항으로 20문항은 선택문항, 10문항은 가교문항으로 구성되어 있고, 선택문항의 난이도 2가지 조건과 집단의 능력차이 2가지 조건을 두어 총 4가지의 조건을 고전검사점수 동등화 방법 및 문항반응이론 동등화 방법을 통해 동등화하여 각 동등화 방법들 중 가장 적합한 방법을 검토하였다.

실험에서 고려한 4가지 조건들의 동등화의 결과는 비슷한 맥락으로 나타났다. 전반적으로 IRT 1-모수 방법인 Rasch 모형이 안정적으로 동등화를 하고 있는 것으로 확인되었다. 또한 Tucker방법의 경우 중간점수대에서 다른 동등화 방법들에 비해 가장 동등화 오차가 작았다. 반면 동백분위 동등화방법의 경우 낮은점수와 높은 점수에서 동등화 오차가 컸으며, IRT 2-모수 모형의 경우 동등화 오차가 대부분의 점수대에서 큰 것으로 나타났다. 결과적으로 Rasch모형의 동등화 안정성이 가장 좋았으며, 만약 중간점수대가 모여있는 시험이라면 중간 점수대에서 동등화 오차가 제일 작았던 Tucker방법 또한 좋은 선택지가 될 수 있다. 결과적으로 본 연구는 향후 대학수학능력시험에서 수학과목을 동등화 할 때, 동등화 방법을 고려하는데 작은 시사점을 줄 것으로 기대된다.

하지만 이 연구는 IRT방법중 3-모수 방법을 비교하지 않았다는 한계점을 갖는다. 또한 대학수학능력시험의 여러 가지 상황을 모두 고려하지 못하여, 좀 더 많은 조건을 고려한 후속연구가 필요하다. 특히 수학시험은 문항의 난이도에 따라 점수가 달라지기 때문에, 이러한 부분을 고려하여 실험을 진행할 필요가 있다. 이와 같은 후속 연구들을 통해, 다양한 상황에서 가장 적합한 방법을 체계적으로 비교하고 검토해야 할 것이다.

# 참고문헌

김현철(1999). 척도화 방법들의 무작위 오차크기의 실증비교. **교육평가연구**, 12(2), 148-168

김현철(2001). 2000학년도 대입수능시험 선택과목 점수에 대한 2단계 선형동등화와 빈도추정 동등화의 결과비교. **교육학연구**, 39(1)

김현철(2003a). 모의실험에 의한 대학입학수학능력시험 선택과목 점수 동등화 방법들의 적합성 비교(1). **교육학연구**, 41(1)

김현철(2003b). 모의실험에 의한 대학입학수학능력시험 선택과목 점수 동등화 방법들의 적합성 비교(2). **교육학연구**, 41(2)

김희경(2007). 재검 여부와 재검 횟수에 따른 검사동등화 차이 연구. **교육평가연구**, 20(4), 213-233

남현우(1992a). 검사동등화에 대한 이론적 고찰. **교육학연구**, 30(2), 205-221

남현우(1992b). 문항모수의 변이에 따른 선형, 동백분위, IRT, 검사동등화 방법의 강인성 비교연구. **교육평가연구**, 5(2), 27-60

남현우(2001). 검사동등화 방법. 교육・심리・사회 연구방법론 총서 시리즈, 교육과학사

박정(2002). 문항반응이론과 선형방법에 의한 대학수학능력시험 선택과목 조정점수 비교. **교육학연구**, 40(5), 161-180

부재율(1999). 검사동등화 시의 고려사항 : 동등성, 동등화 오차, 사전검정 및 사후검정. **교육평가연구**, 12(1), 43-64

성태제(1991). 문항반응이론 입문, 양서원

성태제(1994). 대학별 고사를 위한 문항분석, 표준점수, 검사동등화. 한국통계학회논문집, 1(1), 161-180

신일용, 남진영, 조윤동, 이광상, 김부미(2010). 2012학년도 대학수학능력시험 수리영역 선택과목별 문항 배분 연구. 한국교육과정평가원 연구보고 CAT 2010-7

이기종(2004). 대학수학능력시험 외국어 영역 문항특성 분석. Educational Research, 24(1), 101-116

이상하, 박도영(2012). 문항 예측 정답률과 문항반응이론을 이용한 원점수 분포 추정, **교육평가연구**, 25(3), 459-477

이용민(2003). 모의실험에 의한 문항반응이론 동등화 방법의 적절성 연구. 성균관대학교 석사학위논문

이용민(2010). 모의실험에 의한 커널 동등화 방법과 다른 동등화 방법들의 비교. 성균관대학교 박사학위논문

조명근(2013). 모의실험을 통한 고전적 검사점수 동등화 방법 비교 연구:가교검사 길이와 비동등 집단의 분산차이를 중심으로. 성균관대학교 석사학위논문

황소림(1993). 대학수학능력시험 제 6차, 제7차 실험평가의 문항특성과 피험자 능력점수 동등화. **교육평가연구**, 6(2), 287-314

한국교육과정평가원(2019). 수능통계 연도별 채점현황

Angoff W. H(1984). Scales, norms, and equivalent scores Eduactional testing service. Princeton. N. J

Anthony D. Albano(2014). An R Package for Observed-Score Linking and Equating

BJörn Andersson, Marie Wiberg(2014), IRT Observed-Score Kernel Equating with the R Package Kequate

BJörn Andersson, Kenny Bränberg, Marie Wiberg(2013). Performing the Kernel Method of Test Equating with the Package kequate, **Journal of statistical Software**, 55(6)

Dimitris Rizopoulos(2006). ltm: An R Package for Latent Variable Modeling and Item Response Theory Analyses, **Journal of statistical Software**, 17(5)

Haruhiko Ogasawara(2001), Item Response Theory True Score Equatings and Their Standard Errors, **Journal of Educational and Behavioral Statistics**, 26(1), 31-50

Haruhiko Ogasawara(2003). Asymptotic standard errors of IRT observed-score equating methods, **Psychometrika**, 68, 193-211

Kolen, M. J. , Brennan, R. L.(1995). Test equating, Springer

Kolen, M. J. , Brennan, R. L.(2004). Test Equating, Scaling, and Linking, Springer

Zeng, L. (1991). Standard errors of linear equating for the single-group equating (ACT Research Report Series). American College Testing.

• 논문 접수 2019년 07월 29일 / 수정본 접수 08월 23일/ 게재 승인 08월 30일

• 황수진 : 성균관대학교 교육학과 졸업. 동 대학교에서 교육학으로 석사 및 박사학위 취득. 학업성취, 사교육, 검사점수 동등화 등에 관심을 가지고 있음.