자율평가고사와 전문의 자격고사의 평가도구로써의 타당성 고찰

경상대학교 의과대학, 이비인후과학교실, 연세대학교 이비인후과학교실*·***, 서울대학교 의과대학 이비인후과학교실**

전시영 · 정명현* · 김광현 · 홍원표*****

= Abstract =

Validity of the Intraining Examinations and the Board Examinations - An Experience in the Korean Society of Otolaryngology

Sea Yuong Jeon, M.D., **Myung Hyun Chung,** M.D., **Kwang Hyun Kim,** M.D. and **Won Pyo Hong,** M.D.

The Korean Society of Otolaryngology, Seoul, Korea

The korean society of otolaryngology has had an experience on intraining examination since 1992. We also had the fortieth annual board examination for specialist in 1997. But we have no evidence on the validity of these tests yet. The aim of this study is to examine the validity of the intraining examinations as a tool of formative evaluation, to present a personal progress index demonstrating constructive validity, and to examine the validity of the board examination as a tool of summative evaluation. We did statistic analysis on the consecutive personal scores of 1995 and 1996 intraining examinations, and 1997 written and oral board examinations.

Analysis of the averages, standard deviations, distribution curves, and Wilcoxon singed rank test on the scores of 1995 and 1996 intraining examinations demonstrated the constructive validity. Chi-square test revealed that those who had low scores in intraining examinations of two consecutive years had low scores in 1997 board examinations and personal progress index demonstrated the predictive validity. Correlation and linear regression analysis demonstrated a strong correlation between 1997 written and oral board examination. Analysis of the averages, standard deviations, distribution curves, and Spearman rank correlation coefficient revealed that 1997 written board examination had higher concurrent validity than the that of oral examination.

Key Words: Intraining examination, Board examination, Personal progress index, Constructive validity, Predictive validity, Concurrent validity

본 연구는 대한이비인후과학회 이사회의 자원으로 수련위원회와 고시위원회가 공동 사업으로 완성하였음.

대한이비인후과학회 무임소이사 전 시 영
수련이사 정 명 현
고시이사 김 광 현
이 사 장 홍 원 표

서 론

교육평가는 교육과정에 의하여 교육목표가 어느 정도 성취되었는지를 검정하는 과정이며 평가의 목적에 따라 형성적 평가와 총괄적 평가로 나누어진다. 이비인후과학회에서는 1992년부터 형성적 평가의 도구로서 자율평가고사를 시행하여 왔다. 그리고 총괄적 평가의 도구로서 매년 치르고 있는 전문의 자격고사는 올해 제 40회 고사를 치렀다. 그러나 이러한 시험들이 평가도구로써 얼마나 타당한가에 대한 검토는 아직 보고된 바 없다.

형성적 평가의 도구로서 자율평가고사의 타당성을 검토하고, 형성적 평가의 척도로 이용할 수 있는 개인별 향상지수를 제시하며, 총괄적 평가의 도구로서 1차 및 2차 전문의 자격고사의 타당성을 검토하고자, 1) 자율평가고사는 형성적 평가도구로써 타당한 것인가? 2) 형성적 평가의 한 척도인 개인별 향상지수로부터 총괄적 평가의 성적을 예측할 수 있는가? 3) 1차와 2차 전문의 자격고사 사이의 분별도는 있는가? 4) 1차와 2차의 전문의 자격고사 중 어느 쪽이 총괄적 평가도구로써 더 타당한 것인가? 라는 문제들을 제시하고, 97년 전문의 자격고사를 치른 전공의들의 95년, 96년 자율평가고사 성적과 97년 1차 및 2차 전문의 자격고사 성적에 대한 통계적 분석을 시행하였다.

재료 및 방법과 결과

1997년 제 40회 전문의 자격고사를 치른 전국의 이비인후과 전공의 100명을 대상으로 하였으며, 이들이 전공의 3년차와 4년차 때 치른 95년, 96년 자율평가고사 성적과 수련과정을 마치고 치른 97년 1차 및 2차 전문의 자격고사 성적에 대한 통계적 분석을 시행하였다.

1. 자율평가고사는 형성적 평가도구로써 타당한 것인가?

동일한 집단의 사람들이 같은 수준의 난이도를 가진 시험, 즉 95년, 96년 자율평가고사를 치렀다. 95년에 비해 96년은 1년 더 공부한 만큼 성적이 향상될 것이라는 가설을 설정하였다. 이를 검정하기 위하여 95년, 96년 자율평가고사 성적의 기초 통계량을 분석하고, 95년, 96년 자율평가고사 성적간에 유의한 차이가 있는가를 Wilcoxon 부호순위검정으로 분석하였다.

95년, 96년 자율평가고사 성적의 기초통계량과 Wilcoxon 부호순위검정의 결과는 표 1과 같다. 부호순위검정상 95년에 비해 96년에는 평균성적이 약 10점 증가하는 유의한 향상을 보였다. 성적의 분포는 모두 정규분포를 보이고 표준편차와 왜도는 큰 차이가 없고 첨도만 증가한 것으로 미루어 96년도에는 중위권의 학생수가 많이 증가하였음을 알 수 있다(그림 1). 따라서 95년, 96년 자율평가고사는 동일한 집단의 전공의들이 1년 더 공부한 만큼의 성적 향상을 반영하고 있어, 형성적 평가도구로써 타당하였다고 생각된다.

2. 형성적 평가의 한 척도인 개인별 향상지수로부터 총괄적 평가의 성적을 예측할 수 있는가?

2년간 자율평가고사에서 지속적으로 하위권에 머

표 1. 95년, 96년 자율고사 성적 기초 통계량

기초통계량	평 균	표 준 편 차	왜 도	첨 도
95년도 성적	69.17	8.44	0.20	-0.11
96년도 성적	79.13	9.55	0.23	2.03
부호순위검정		Mean ± SD		Sign Rank
95, 96 성적 차이		9.96 ± 9.95		2167.5(P<0.001)

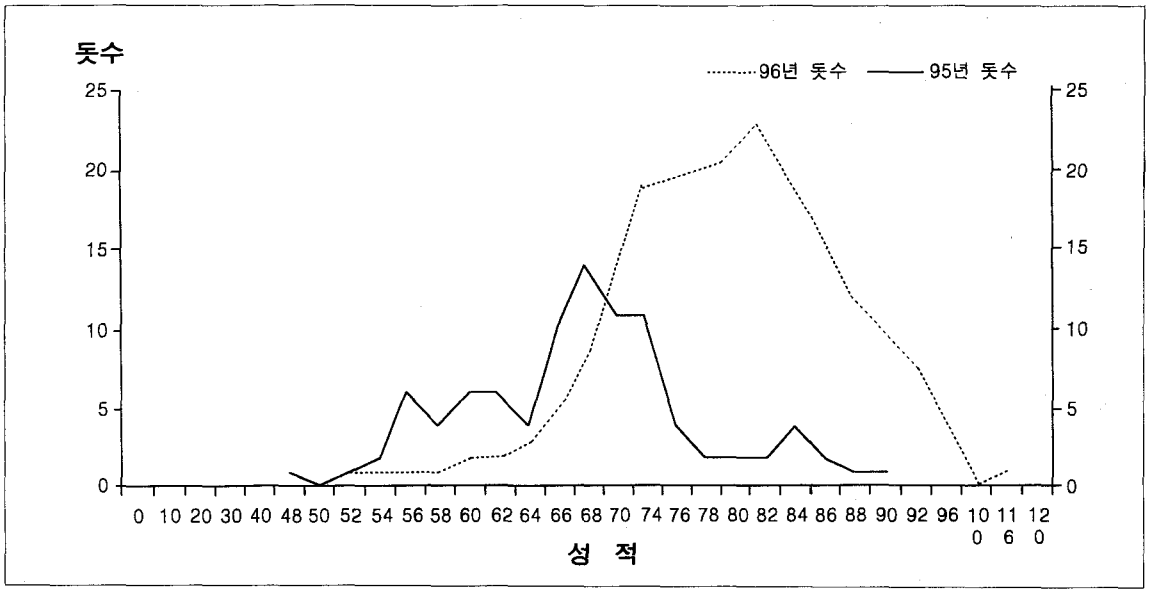

그림 1. 95년과 96년의 성적 분포.

무른 사람은 1차 전문의 자격고사에서도 하위권일 것이라는 가설을 설정하였다. 2년간 지속적으로 하위권에 머묾은 사람을 정의할 수 있는 개인별 향상지수(PPI, personal progress index)[2]는 다음과 같이 구하였다. 2년간 각 연도별 개인 성적을 표준화하여 표준확율변수값(Z)을 구한 다음, 95년과 96년 표준확율변수값을 각각 제곱한 합의 평균(PPI = Z_{95}^2 + Z_{96}^2/2)으로 구하고, 성적의 표준확율변수가 음수, 즉 성적이 평균보다 낮으면서 이 값이 2.25 이상이면 I 영역, 2.25 보다 작고 1 이상이면 II 영역, 1 미만이면 III 영역, 성적이 평균 이상이면 IV 영역으로 나누었다. 1차 전문의 자격고사 성적의 표준확율변수값이 -1.5 SD(standard deviation, 표준편차) 이하이면 I 영역, -1.5 SD보다 크고 -1 SD 이하이면 II 영역, -1 SD보다 크고 0 이하이면 III 영역, 0보다 크면 IV 영역으로 나누었다. 개인별 향상지수에 따른 4개 영역 대비, 1차 전문의 자격고사의 개인별 성적의 표준확율변수값에 따른 4개 영역간의 차이를 chi- square (χ^2) 검정으로 분석하였다.

개인별 향상지수와 1차 전문의 자격고사의 개인별 성적의 표준확율변수값에 따른 4개 영역간 도수와 그 χ^2 값은 표 2와 같았다. χ^2 검정상 개인별 향상지수에 따라 1차 전문의 자격고사의 성적은 통계적으로 유의한 차이를 보였다. 즉 자율평가고사에서 2년간 계속 하위권이 있었던 사람들은 1차 전문의 자격고사에서도 주로 하위권에, 자율평가고사에서 2년간 계속 상위권이 있었던 사람들을 1차 전문의 자격고사에서도 주로 상위권에 머물렀음을 알 수 있었다. 따라서 개인별 향상지수로부터 총괄 평가의 결과를 예측할 수 있다고 생각된다.

3. 1차와 2차 전문의 자격고사 사이의 분별도는 있는가?

동일한 사람이 다른 유형의 시험, 즉 1차 필기고사와 2차 구술고사를 치렀고, 1차 필기고사의 성적에 따라 2차 구술고사의 성적이 나올 것이라는 가설을 설정하였다. 이를 검정하기 위하여 1차 및 2차 전문의 자격고사 성적간의 상관관계를 구하고 선형회귀분석을 시행하였다.

1차 및 2차 전문의 자격고사의 성적간 Spearman 순위상관계수는 0.615($p<0.001$) 이었고, 단순선형회기식은 $97_{2차\ 성적}$ = -13.3063 + 1.1181×$97_{1차\ 성적}$ 이

표 2. 개인별 향상지수와 1차 전문의 자격고사의 개인별 성적의 표준확율변수값에 따른 4개 영역간 도수와 χ^2 값

PPI	1차 전문의 자격고사				
	I	II	III	IV	합계
I	3	0	2	0	5
II	2	3	1	5	11
III	5	1	20	4	30
IV	3	1	15	35	54
χ^2 값	48.068 (P<0.001)				

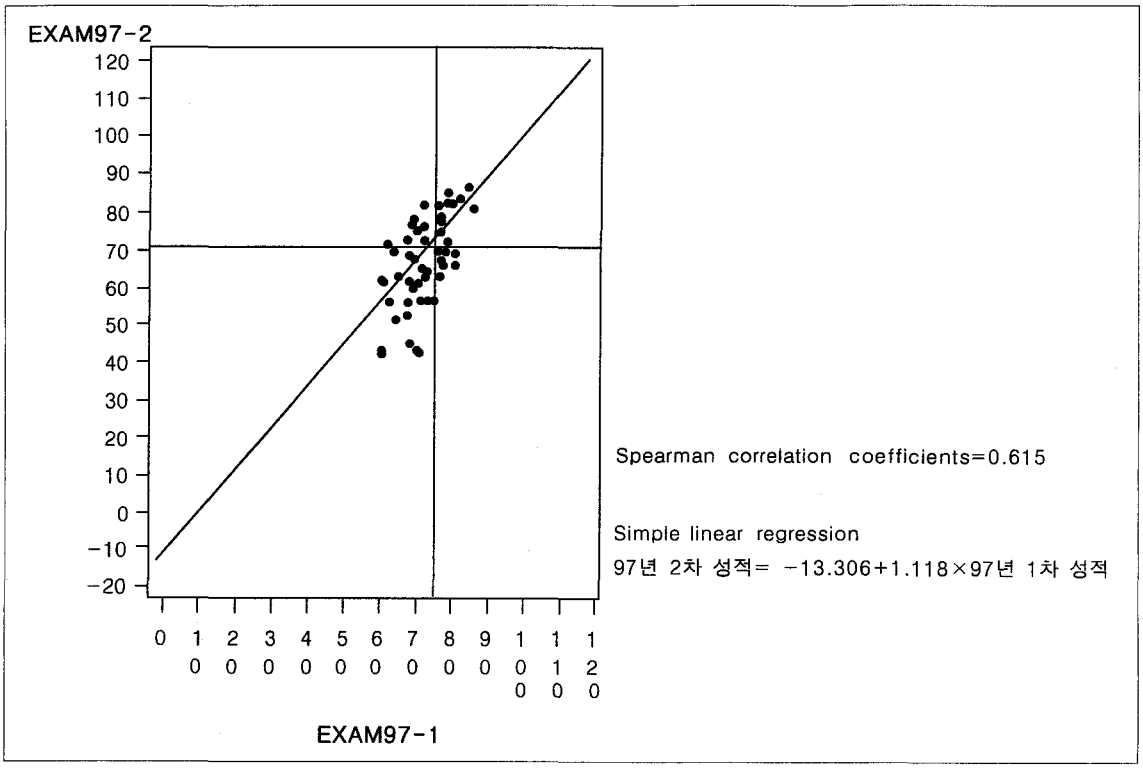

그림 2. 97년 1차와 2차 전공의 자격고사 성적의 상관도.

었다(그림 2). 따라서 1차와 2차의 성적은 비교적 선형인 상관관계를 가지고 scatter plot상의 기울기가 1.12로 두 시험간에 유의한 상관관계를 보이고 있었다. 즉 1차 성적이 좋을수록 2차 성적이 좋았음을 확인할 수 있었다. 따라서 1차와 2차 전문의 자격고사 사이에는 유의한 분별도가 있다고 생각된다.

4. 1차와 2차의 전문의 자격고사 중 어느 쪽이 총괄적 평가도구로써 더 타당한 것인가?

1차 및 2차 전문의 자격고사 성적의 기초 통계량을 분석하여 정규분포를 보이는 쪽이 총괄 평가의 도구로서 보다 타당한 도구라는 가설을 설정하고 1

표 3. 97년 1차 및 2차 전문의 자격고사 성적 기초 통계량

기초통계량	평 균	표준편차	왜 도	첨 도	최대값	최소값
1차 필기고사	74.82	5.54	-0.51	-0.16	86.4	60.85
2차 구술고사	70.35	9.65	-0.88	0.57	86.2	42.25

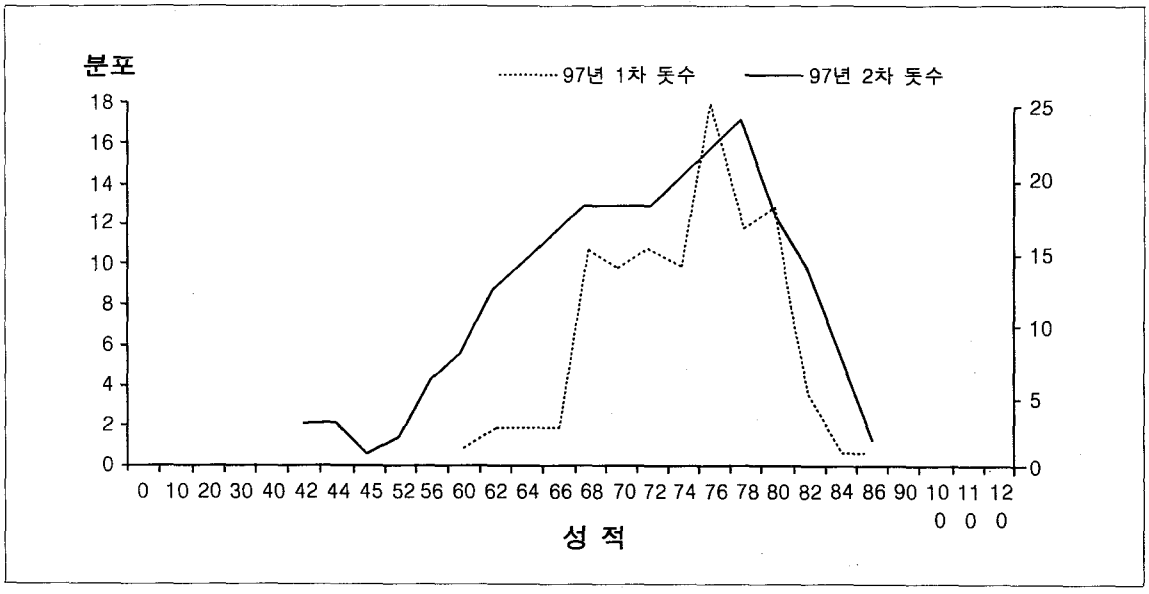

그림 3. 97년 1차와 2차의 성적 분포.

차와 2차 전문의 자격고사의 성적의 기초 통계량을 분석하였다. 또한 5개월 전 치른 96년 자율평가고사의 성적을 보다 잘 반영하고 있는 쪽이 보다 분별도가 큰 시험이라는 가설을 설정하고 96년 자율평가고사 성적 대비, 1차 및 2차 전문의 자격고사 성적간의 상관관계를 구하고 선형회기분석을 시행하였다.

1차 및 2차 전문의 자격고사 성적의 기초통계량은 표 3과 같다. 2차 전문의 자격고사의 성적은 1차 전문의 자격고사 성적에 비하여 왜도의 절대값과 첨도가 증가되었다. 즉 1차에 비해 상위값에 자료가 많이 모여 있는 비대칭형이며 표준편차가 크고 최소값이 낮아 아주 못하는 학생들이 있음을 보여 준다. 따라서 1차 전문의 자격고사가 보다 더 타당한 총괄평가의 도구라고 생각된다(그림 3).

96년 자율고사 성적과 1차 전문의 자격고사의 성적간 Spearman 순위상관계수는 0.530($p<0.001$) 이었고, 단순선형회기식은 $97_{1차\ 성적} = 50.1443 + 0.3118 \times 96_{자율고사}$ 이었다. 96년 자율고사 성적과 2차 전문의 자격고사의 성적간의 Spearman 순위상관계수는 0.324($p<0.001$) 이었고, 단순선형회기식은 $97_{2차\ 성적} = 43.7903 + 0.3357 \times 96_{자율고사}$ 이었다(그림 4). 즉 96년 자율평가 대비 1차와 2차 전문의 자격고사의 scatter plot을 보면 그 기울기는 0.31, 0.36으로 서로 비슷한 수준의 상관도를 보이나, 1차 필기고사가 2차 구술고사보다 더 선형인 상관관계를 보였다. 따라서 1차 전문의 자격고사가 보다 더 분별도가 높은 평가의 도구라고 생각된다.

고 찰

1992년도부터 시행하여온 자율고사평가는 전국적으로 모든 전공의들이 참여하고 많은 전문의들이 출

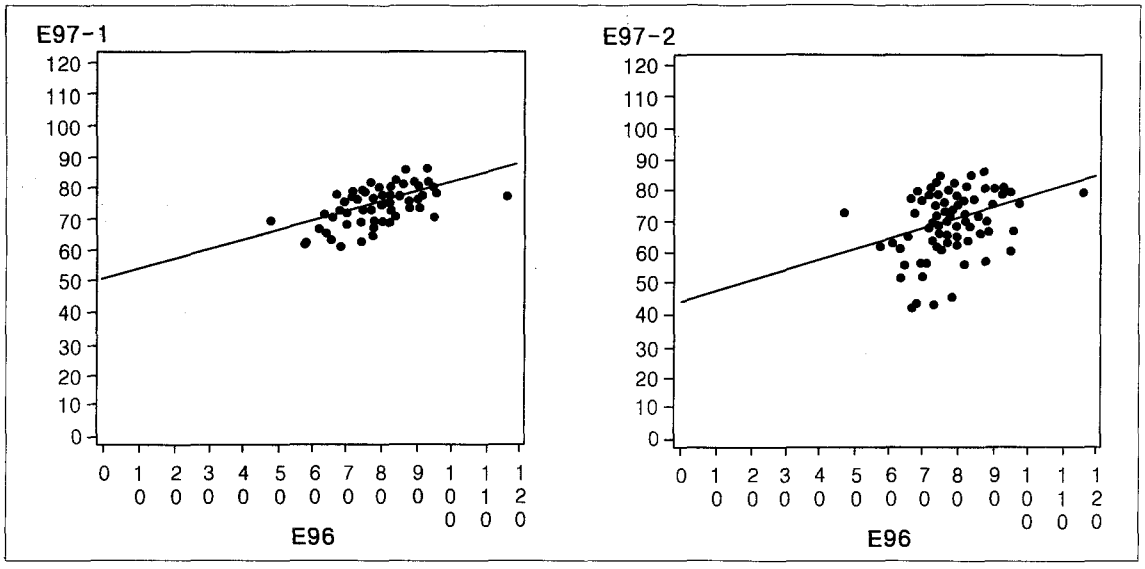

그림 4. 96년 자율평가고사와 97년 1차 및 2차 전공의 자격고사 성적의 상관도.

제자로 참여해온 학회 차원의 대규모 사업이다. 그러나 평가의 대상이 시험을 치르는 전공의들만이 아니라, 평가고사 자체도 평가의 대상이 된다는 것을 간과하여왔다고 생각된다. 매년 시행하여온 평가고사의 문항분석을 통하여 그 난이도와 신뢰도의 검정이 따랐어야 하였으며, 이제 자율평가고사를 시행한 지 5년이 지난 지금은 평가도구로써의 타당성에 대한 검토가 이루어져야 할 시점이라고 생각된다.

자율평가고사를 시작한 92년도에 1년차였던 전공의들은 96년도에 전문의 자격시험을 치렀으므로, 96년도 및 97년도 전문의 자격시험을 치른 전공의 집단의 4년간의 성적추이를 분석하여 형성적 평가도구로써의 자율평가고사의 타당성을 검토하고자 하였으나, 기초 자료의 확보가 여의치 않아 97년도 전문의 자격고사를 치른 전공의 집단을 대상으로 하였으며, 이들이 전공의 3년차와 4년차 때 치른 95년, 96년 자율평가고사 성적과 수련과정을 마치고 치른 97년 1차 및 2차 전문의 자격고사 성적에 대한 통계적 분석을 시행하였다.

95년, 96년 자율평가고사가 형성적 평가도구로써 타당한 것인가를 검토하기 위하여 이들의 난이도와 신뢰도가 같은 수준이라는 가정에서 기초 통계량을 분석하고, 쌍체표본에 관한 검정법인 Wilcoxon 부호순위검정을 시행하였다. 단 95년, 96년 자율평가고사의 난이도와 신뢰도가 같은 수준이라는 가정을 설정하기 위하여서는 95년, 96년 자율평가고사의 문항분석을 통한 난이도와 신뢰도에 대한 검정이 선행되어야 하나 본 연구에서는 그 검정은 수행하지 않았다. 향후 이와 같은 연구에서는 가설 설정의 타당성도 반드시 검정되어야 한다고 생각된다.

형성적 평가의 가장 큰 기능은 성취도에 대한 진단적 기능이며 이를 통하여 학습자와 교육자에 대한 feedback이 이루어져야 한다[3]. 지난 5년간 시행하여온 자율평가고사는 그 기능상 형성적 평가인데 학습자나 교육자에게 feedback되는 구체적인 진단적 척도를 제시하지 못하였다. 최근 교육학회에서는 형성적 평가의 진단적 척도로서 개인별 향상지수가 제시되고 있다. McMaster 의대에서는 1992년부터 의과대학 학생 전체를 대상으로 동일한 문항의 시험(progress test)을 매년 치르고, 개인별 향상 정도를 나타내는 PPI를 통하여 학생들에 대한 재교육의 단서를 제시하고 있다. 즉 3년간 2번 이상 I 영역에 있었거나, 2년간 II 영역에서 I 영역으로 떨어진 경우 재교육을 받도록 feedback 하고 있다[2]. 이비인후과

전공의 수련과정은 4년이며 수련과정 중 재교육의 기회는 없다. 전문의 자격고사에서 실패한 경우 재교육을 받을 수 있는 기회도 없다. 병원마다 수련과정이 다르고 개인별 성취도가 다른 만큼 수련 과정 중 재교육의 기회가 제공되어야 한다고 생각된다. 향후 자율평가고사의 개인별 향상지수가 갖는 진단적 기능이 충분히 검토된 후에는 McMaster 의과대학과 같은 재교육의 기준을 자율평가고사에 도입할 수 있으리라 생각된다.

이비인후과의 전문의 자격고사는 1차 필기고사와 2차 구술고사로 이루어진다. 필기고사는 주로 지식 영역을, 구술고사는 수기와 태도 영역을 평가하는데 더 타당한 도구이다. 그러나 아직까지 1차와 2차 전문의 자격고사 사이의 분별도에 대한 검토는 없었다. 총괄적 평가인 전문의 자격고사는 지식 뿐 아니라 수기 및 태도영역의 성취 수준을 정의하고 있다. 지식 영역의 수준이 높다고 하여 반드시 수기나 태도 영역의 수준이 높아야 하지는 않지만, 지식 영역의 성취 수준은 수기 및 태도 영역의 성취 수준을 잘 반영하여야 한다. 따라서 1차와 2차 전문의 자격고사 사이에는 비교적 유의한 상관관계를 보여야 하는데, 97년도 1차와 2차 전문의 자격고사 성적은 선형회기분석상 서로 아주 높은 상관관계를 보였다. 따라서 97년도 1차와 2차 전문의 자격고사는 상호간 분별도가 높게 잘 치러진 시험이었다고 생각된다.

기초 통계량의 분석이나 96년도 자율평가고사 성적과의 상관관계에 대한 선형회기분석의 결과, 1차 전문의 자격고사가 2차 전문의 자격고사보다 더 분별도가 높은 평가도구로 평가되었다. 그러나 필기고사와 구술고사는 평가 수단이 서로 다르고 또한 평가 영역이 서로 다른 만큼, 어느 한가지 척도만으로 어느 쪽이 총괄적 평가도구로써 더 타당한 것이라고 판단할 수는 없다. 또한 96년도 자율평가고사는 필기시험인 만큼, 96년도 자율평가고사를 기준으로 1차 필기고사와의 비교는 타당할 수 있으나 2차 구술고사와의 비교는 다소 불합리한 점이 있다. 1차와 2차 전문의 자격고사 성적간에는 높은 상관관계가 있는 만큼, 앞으로 몇 가지 문제점들을, 즉 문항수가 적다는 점이나, 평가자에 의한 오차 등의 문제점들을 찾아 개선함으로써 2차 구술고사를 보다 타당한 총괄적 평가도구로 발전시켜야 한다고 생각된다.

요 약

1. 95년, 96년 자율평가고사는 형성적 평가의 도구로서 타당성이 있었다.

2. 2년간 자율평가고사에서 지속적으로 하위권에 머무른 사람은 1차 전문의 자격고사에서도 하위권에 머무는 경향을 보였다. 따라서 지속적으로 하위권에 속한 사람들을 확인할 수 있는 개인별 향상지수는 형성적 평가척도로서 타당성이 있었다.

3. 1차와 2차 전문의 자격고사의 성적은 서로 높은 상관관계를 보여 두 시험간에 유의한 분별도가 있었다. 그러나 총괄적 평가도구로써는 1차 전문의 자격고사가 보다 더 타당성이 있었다.

참 고 문 헌

1. 김용일: 의학교육평가. 서울대학교출판부 1985
2. Keane DR, Blake JM, Norman GR, Barbet C: Introducing Progress Testing in a Traditional Problem Based Curriculum. Internal report McMaster University 1994
3. Norman G: Why Evaulate? Pedagogue. 5:1-7, 1994