교육과정평가연구
The Journal of Curriculum & Evaluation
2007, Vol. 10, No. 1, pp. 47～74

교사의 학생평가 전문성 진단도구 개발

송 미 영(한국교육과정평가원 부연구위원)
김 경 희(한국교육과정평가원 연구위원)

≪ 요 약 ≫

학교 현장에서 교사들에게 수준 높은 전문적 능력을 요구하는 활동인 학생평가에 대한 질적 제고가 더욱 시급해진 현 시점에서 교사가 갖추어야 전문적인 영역을 밝히고, 각 영역별로 각자의 전문성 수준을 가늠해 볼 수 있는 도구를 개발할 필요가 있다. 이에 '교사의 학생평가 전문성 기준'에 제시되어 있는 전문성의 지식과 기술을 교사 스스로 꼼꼼하게 진단하고 점검할 수 있는 문항으로 구성된 '교사의 학생평가 전문성 진단도구'를 개발하였다.

이 진단도구가 교육 현장에서 좀 더 적극적으로 활용되기 위해서 진단도구의 측정학적 양호도, 즉 타당도와 신뢰도를 검증하였다. 교사들은 진단도구를 통하여 학생평가에서 요구되는 전체적인 전문성뿐 아니라 '평가방법의 선정', '평가도구의 개발', '평가실시・채점・성적 부여', '평가결과의 분석・해석・활용・의사소통', '평가의 윤리성'의 5가지 내용 영역별로 전문성을 진단하고 점검할 수 있다. 또한 '교사의 학생평가 전문성 기준'에서 기대하는 전문성의 필수능력의 수준(기준선)을 제시함으로써 교사가 스스로 진단한 결과와 비교하여 자신의 전문성 수준을 파악할 수 있게 하였다.

주제어 : 학생평가, 교사의 전문성, 검사도구, 타당화

Ⅰ. 서론

변화하는 우리의 학교 교육현장은 교사에게 학생평가에 대한 전문성을 요구하고 있다. 학교에서 학생평가 관련 업무는 교사들이 많은 시간을 투자해야 하는 활동 중의 하나이면서, 동시에 수준 높은 전문적 능력을 요구하는 활동이기도 하다. 또한 2005년에 발표된 교육인적자원부의 '학업성적 신뢰성 제고 조치 계획', 중・고등학교에서 서술형・논술형 평가를 50% 이상으로 확대하겠다는 서울시의 '2007학년도 중등 장학계획', 2008학년도 대학입시부터

는 내신성적의 반영비율을 50% 이상까지 확대하겠다는 주요 대학들의 발표 등으로 인해 학교 내에서 실시되는 학생평가의 중요성에 대한 인식이 확대되고, 그 질적 제고에 대한 요구 또한 증가하고 있다.

학생평가(student assessment)는 '학생의 학습과 성취에 대한 교사의 의사결정을 돕기 위하여 정보를 수집하고 해석하여 활용하는 활동'으로 정의할 수 있다(McMillian, 2004). 학생평가는 수업과 학생지도에 관련된 모든 정보를 수집하여 이를 토대로 수업의 질을 개선하려는 활동이다.

수업과 학생지도에 관련된 모든 정보는 수업을 계획하는 단계에서부터 종료까지 수업의 전 과정에서 이루어지는 크고 작은 의사결정을 위해 수집되고 해석되고 활용된다. 따라서 학생평가는 수업이 종료된 후 부가적으로 부여되는 별개의 활동이 아니라 교사의 의사결정과 수업의 모든 과정에 통합된 활동이다. 교사들에게는 항상 교육과정상의 교육목표와 교과 교육목표가 주어지지만 이는 구체적인 수업목표를 개발하기 위한 시작점일 뿐이며, 교사는 학생들이 무엇을 알아야 하며 수업이 끝난 후 무엇을 할 수 있어야 하는지를 분명히 하여 수업을 운영해야 하며, 이에 대해 학생들과 상호 작용할 수 있어야 한다.

이러한 학생평가는 교수·학습 과정의 핵심적인 부분이기 때문에 교수의 효과는 다양한 평가절차를 개발하고 활용하며 평가결과를 해석할 때 요구되는 교사의 전문적인 지식에 의존할 수밖에 없다. 학생평가는 교수활동의 본질적인 부분이며, 훌륭한 교수활동은 훌륭한 학생평가 없이는 얻어질 수 없다(AFT, NCME & NEA, 1990). 따라서 교사의 학생평가 전문성은 교사의 전문성 향상의 가장 핵심적인 부분이라고 할 수 있기 때문에 교사교육의 중요한 내용이 되어야 한다. 이에 한국교육과정평가원은 교사가 학생을 제대로 평가하기 위해 요구되는 평가지식과 기술을 중심으로 교사의 '학생평가 전문성 기준'을 개발하였고, 교사의 학생평가 전문성 향상과 관리를 제안하고 있다.

교직전문가로서의 교사가 학생평가 전문성을 실천하기 위해서는 전문성에 대한 개념적 인식이 전제되어야 한다. '무엇을 알고 할 수 있어야'하고, '어떤 윤리적 자세를 가져야 하는지'에 대한 전문가 수준에서 합의된 기준에 기초한 전문성 훈련과정이 있어야 할 것이다. 따라서 한국교육과정평가원이 교사의 학생평가 활동에서 실질적이고 필수적으로 요구되는 평가 능력 요인을 포괄적으로 반영하는 교사의 학생평가 전문성 기준을 전문가 수준의 합의를 거쳐 개발한 것은 의미 있는 일이다.

교사의 학생평가 전문성은 '학생의 학습과 성취에 관한 평가정보를 수집하고 해석하여 활용할 수 있는 능력'으로 정의되며, 이러한 능력은 평가활동을 하기 위해 요구되는 핵심적인 필수능력이므로 평가 소양이라고 할 수 있다(김경희 외, 2006).

교사의 학생평가에 대한 질적 제고가 더욱 시급해진 현 시점에서 이러한 학생평가 전문성을 구체화하여 교사가 갖추고 신장해야 할 능력 요소를 밝혀서 현직 교사와 예비 교사들에

게 실제적인 교육의 기회를 제공할 필요가 있다. 이러한 필요성에 따라 학생평가 전문성 기준은 교사의 학생평가 전문성의 5가지 능력요소 즉, '평가방법의 선정', '평가도구의 개발', '평가실시・채점・성적부여', '평가결과의 분석・해석・활용・의사소통', '평가의 윤리성'에 따라 개발되었다. 내용 영역별로 2~3개의 전문성 기준이 기술되었으며, 전문성 기준은 교사가 바람직한 학생평가를 계획하고 실천하기 위해 반드시 필요한 지식, 기능, 태도로 함축적이고 당위적으로 표현되어 있다. 또한 각 전문성 기준들은 교사가 학생평가의 역할과 책무를 수행하기 위해서 갖추고 있어야 하는 55개의 평가 지식과 기술로 구체화되었다.

이러한 기준은 교사의 평가 전문성을 계발하고 신장시키기 위한 실제적인 지침으로 활용될 수 있으며, 교사의 학생평가 전문성을 계발하고 신장시키기 위한 노력을 기울이기 위해서는 교사의 학생평가 전문성 수준을 진단하고, 전문성이 부족한 영역에 대한 확인과 점검이 선행되어야 할 것이다. 전문가 집단인 교사 스스로 자신의 전문성을 더욱 신장시킬 수 있도록 학생평가 전문성 기준을 내면화해야 할 것이며, 그러한 기회를 제공할 수 있는 교육평가 정책과 방안 또한 필요하다고 하겠다.

실제로 교사들은 스스로 자신의 학생평가 전문성이 낮다고 생각하지 않지만(백순근, 2006), 교육 전문가와 학생평가 전문가들은 일선 교사들의 평가 전문성이 낮다고 인식하고 있다(이인제 외, 2004). 이렇듯 학생평가 전문성에 대한 상반된 인식이 공존하는 상황에서 교사에게 요구하는 학생평가 전문성 기준을 중심으로 교사의 학생평가 전문성 수준을 진단하고 점검할 필요가 있다.

본 연구에서는 이러한 측면에서 교사 스스로 학생평가 전문성을 진단하고 전문성이 부족한 영역을 점검할 수 있도록 '교사의 학생평가 전문성 진단도구'를 개발하였다[1]. 학생평가 전문성 진단도구는 교사의 학생평가 전문성 기준에 준하여 기준에서 요구되는 평가지식과 기술 하나하나를 교사 스스로 점검할 수 있어 교사 자신이 학생평가에 대해 알고 있는 것과 할 수 있는 것을 진단할 수 있다. 또한, 전문가 수준에서 합의하여 설정된 학생평가 전문성 기준선(baseline)의 절대적인 수준과 자신이 진단한 수준을 비교함으로써 자신의 학생평가 전문성의 정도를 평가하고, 신장시켜야 할 전문성의 수준과 구체적인 영역이 무엇인지를 지각하는 기회를 제공하게 될 것이다.

1) 한국교육과정평가원에서는 3년 수행과제인 '교사의 학생평가 전문성 신장 연구'를 통해 1차년(이인제 외, 2004)에 우리나라 교사들의 학생평가 전문성 실태조사와 평가 관련 직무 분석, 2차년(김수동 외, 2005)에 교사의 학생평가 전문성 기준과 진단도구 개발, 3차년(남명호 외, 2006)에 교사의 학생평가 전문성 기준 해설과 진단도구의 타당화를 수행하였다. 본 연구는 2차년과 3차년 연구결과의 일부를 수정・보완한 것이다.

Ⅱ. 교사의 학생평가 전문성 진단도구 개발

1. 교사의 학생평가 전문성 진단도구 구성

‘교사의 학생평가 전문성 진단도구’는 ‘교사의 학생평가 전문성 기준’에서 요구하는 평가 지식과 기능을 교사 스스로 점검하고, 자신의 학생평가 전문성 수준을 절대적 기준에 비추어 평가하여 전문성 영역 중 향상시켜야 할 영역을 파악하도록 하는 것을 목적으로 하는 검사도구이다.

이 검사도구의 명칭에서도 알 수 있듯이 진단의 대상은 교사이고, 진단의 내용은 학생평가 전문성이다. 교사의 학생평가 전문성 진단도구에서 다루는 학생평가 전문성의 범위는 ‘교과내용에 대한 지식’, ‘학습자에 대한 지식’ 등과 같이 교사의 전문성 일반에 대한 것은 제외하고, 평가 고유의 영역에 한정되어 있다. 다시 말하면 교사의 학생평가 전문성 진단도구에서 측정하고자 하는 내용은 교사가 학생을 대상으로 평가활동을 수행할 때 발휘해야 하는 전문적인 능력으로서, 수업 실제의 학생평가에서 반드시 필요한 평가지식과 기능을 포괄한다.

학교 현장교사들이 자신의 학생평가 전문성의 수준을 스스로 평가할 수 있도록 하기 위한 진단도구의 개발 방향은 심층적인 논의 과정을 거쳐 다음과 같이 결정되었다.

첫째, 학생평가 전문성에 대한 진단의 초점을 교사의 ‘지식’과 ‘실천’ 중 어디에 둘 것인가의 문제이다. 교사의 학생평가 전문성을 능력의 개념으로 정의하였기 때문에 학생평가 전문성에 대한 진단은 교사가 학생평가를 수행하는 과정에서 반드시 알고 있어야 하는 지식과 할 수 있어야 하는 기능에 초점을 두는 것이 논리적이다. 따라서 진단도구를 통하여 교사가 학생평가에 대해 알고 있고 할 수 있는 것을 진단할 수 있도록 개발하였다.

둘째, 진단의 초점을 평가지식과 기능에 두었기 때문에 평가지식과 관련된 문항은 알고 있는 상태를 진단할 수 있도록 ‘~알고 있다’, 평가기능과 관련된 문항은 ‘~할 수 있다’, 평가윤리와 관련된 문항은 당위성이 강하므로 ‘~알고 있다’ 또는 ‘~하고 있다’로 진술하였다.

셋째, 학생평가 전문성 진단도구는 교사들이 진단도구를 통해 자신의 전문성 정도를 진단할 수 있도록 강한 부정에서 강한 긍정까지의 응답이 가능한 5점 평정척도(rating scale)로 구성하였다. 각 문항의 진술문마다 ‘전혀 그렇지 않다’의 1점에서부터 ‘매우 그렇다’의 5점까지 전문성 정도가 측정되도록 하였다. 또한 자기보고(self report) 형식으로 교사들로 하여금 진단도구의 각 문항에서 요구하는 평가 전문성을 스스로 점검하여 응답하게 하였다. 체크리스트 형식도 고려 대상이 되었지만 교사교육을 이수한 교사들은 이미 어느 정도의 학생평가

전문성을 갖추었을 것으로 기대되므로, 학생평가의 핵심적인 지식과 기술이 있는지 없는지를 질문하는 것보다 그 정도를 점검하는 것이 학생평가 전문성에 대한 기본 가정에 더 부합하기 때문에 평정척도로 구성하였다.

넷째, 교사의 학생평가 전문성 기준에서 기대하는 절대적인 수준과 비교할 수 있도록 평가 전문성에 대한 기준선(baseline)을 설정하여 교사 스스로 진단한 자신의 전문성을 비교할 수 있도록 하였다. 이 진단도구는 절대적인 기준선과 교사의 자기진단 결과와의 비교를 통해 교사 스스로 자신의 학생평가 전문성의 정도를 평가하고 자신이 계발하고 신장시켜야 할 전문성의 수준과 전문성의 구체적인 능력요소가 무엇인지를 지각하게 되는 기회를 제공할 것이다.

이상과 같이 설정한 방향에 따라 진단도구는 '교사의 학생평가 전문성 기준'에 명시된 평가지식과 기술 55개를 교사 스스로 점검하고 진단할 수 있는 문항으로 각각 구성하였다. 진단도구는 5개의 내용 영역으로 구성되어 있는데, 각 내용 영역은 교사가 학생을 대상으로 평가활동을 수행할 때 전문성을 발휘해야 하는 영역으로서 '평가방법의 선정', '평가도구의 개발', '평가실시・채점, 성적부여', '평가결과의 분석・해석・활용・의사소통', '평가의 윤리성'이다.

2. 교사의 학생평가 전문성 진단도구의 개발 절차

교사의 학생평가 전문성 진단도구의 각 문항은 '교사의 학생평가 전문성 기준'에서 요구하는 평가지식과 기술에 근거하여 11개의 기준과 55개의 평가지식과 기술을 다룰 수 있도록 하였다. 전문성 기준별 문항 개발과 전문가 집단의 검토・수정, 그리고 기준선 설정 등의 과정을 거쳐서 개발되었다.

좀 더 구체적으로는 첫째, '교사의 학생평가 전문성 기준'에 토대하여 각 영역별로 문항을 개발하였고, 둘째, 영역별 비중을 고려하여 문항 수를 조정하고 문항의 진술내용을 수정하였으며, 셋째, 전문성 수준을 해석하기 위한 준거로서의 기준선을 설정하는 과정을 거쳤다.

가. 문항 개발

'교사의 학생평가 전문성 기준'에 준한 진단도구 초안의 문항은 129개로 '평가방법의 선정' 32개, '평가도구의 개발' 33개, '평가실시・채점・성적부여' 21개, '평가결과의 분석・해석・활용・의사소통' 32개, '평가의 윤리성' 11개로 구성되었다. 〈표 1〉에 의하면, 진단도구 초안의 문항 구성이 각 기준에서 요구하는 평가지식과 기술에 준하여 개발된 점을 감안할 때 '평가방법의 선정'과 '평가도구의 개발' 영역에서 문항이 상대적으로 많은 경향이 있어 해당 영역에서의 문항 선제와 전체적인 조정이 필요하였다.

나. 전문가 수준의 1차 검토와 1차 수정

진단도구의 수정・보완은 전문가 수준에서 여러 차례의 상세검토를 통해 이루어졌다. 전문가에 의한 상세검토 의견은 측정・평가 전문가 8명에게 의뢰하여 수렴하였다. 문항마다 수정, 삭제 의견을 제시할 수 있도록 검토지를 구성하였고, 문항 수정과 선제를 위한 협의과정을 거쳤다. 수렴된 의견을 중심으로 수정・보완된 사항은 다음과 같다.

129개 문항의 수가 과도하게 많다는 지적이 있었다. 전문성 기준의 평가지식과 기술을 고려할 때 적정 수의 문항에 대한 전문가들의 의견은 55~100개 정도의 범위로 모아졌다. 내용 영역별로 지식과 기술의 비중을 고려하여 문항을 선제하되, 우선적으로 선제에서 제외될 문항은 문항에서 요구하는 전문성의 의미가 모호하거나 전문가들의 50% 이상이 삭제를 요구한 것이었다. 전문가 협의과정에서 100개의 문항이 선제되었으며, 내용 영역별로 조정된 문항의 수는 〈표 1〉에 제시되어 있다.

그 외 각 문항에 대한 수정 의견을 수렴하여 문항의 진술 내용을 정련하는 과정을 거쳤고, 진단도구는 100개 문항으로 수정되었다.

다. 전문가 수준의 2차 검토와 2차 수정

100개 문항으로 구성된 진단도구 1차 수정본을 국어, 사회, 수학, 과학, 영어의 교과 전문가 8명, 측정・평가 전문가 18명 총 26명에게 의뢰하여 각각의 문항에 대한 교사의 전문성 기대수준과 전체 진단도구에 대한 기대수준을 설정하였다.

100개의 문항 중에서 전문가들(동시에 패널이 됨)이 판단한 전문성의 수준이 전체 문항에 대한 전문가들의 평균 기대수준에 비해 낮거나 불필요하다는 지적을 받은 문항을 중심으로 최종 문항에서 제외될 문항 10개를 추출하였다. 또한 문항의 진술된 내용을 재검토하여 12개의 문항에서 측정하려고 하는 평가지식과 기술의 의미를 살리면서 명료성을 기하는 차원에서 문장 어미를 수정하였다. 교사의 학생평가 전문성 진단도구는 90개의 문항으로 수정・선제되었고, 그 구성은 〈표 1〉과 같다.

〈표 1〉은 진단도구의 문항 개발 초안에서부터 2차 수정된 문항까지 진단도구의 내용 영역별 문항 구성의 변화를 보여준다. 교사의 학생평가 전문성 진단도구의 2차 수정본은 '평가방법의 선정' 18개, '평가도구의 개발' 22개, '평가실시・채점・성적부여' 15개, '평가결과의 분석・해석・활용・의사소통' 27개, '평가의 윤리성' 8개의 문항으로 구성되었다.

〈표 1〉 교사의 학생평가 전문성 진단도구의 문항 구성

내용 영역	전문성 기준	문항 수		
		초안	1차 수정	2차 수정
평가방법의 선정	교사는 학습목표와 평가의 목적을 확인하고 명료화해야 한다.	5	4	3
	교사는 학습목표 및 평가의 목적에 적합한 평가방법을 선정해야 한다.	27	16	15
평가도구의 개발	교사는 평가의 목적과 내용에 적합한 평가도구를 개발하거나 선택하여 사용해야 한다.	21	15	14
	교사는 평가도구의 질을 스스로 점검하고 개선해야 한다.	12	10	8
평가실시・채점・성적부여	교사는 평가계획에 부합하도록 평가를 실시해야 한다.	8	8	7
	교사는 정확하게 채점하고 평가의 목적에 부합하도록 성적을 부여해야 한다.	13	9	8
평가결과의 분석・해석・활용・의사소통	교사는 평가결과를 정확하고 타당하게 분석, 해석해야 한다.	10	9	9
	교사는 평가결과를 수업 및 학생에 대한 교육적 의사 결정에 활용해야 한다.	13	11	10
	교사는 학생, 학부모, 교육 관련자와 평가결과에 대해 정확하게 의사소통해야 한다.	9	8	8
평가의 윤리성	교사는 학생의 인격을 존중하며 모든 평가활동 시 윤리적・법적 책임을 준수해야 한다.	8	8	6
	교사는 학생의 특성과 배경을 고려하여 공정하게 평가를 실시하고 평가의 적절성을 판단해야 한다.	3	2	2
합 계		129	100	90

라. 교사의 학생평가 전문성 진단도구의 전문성 기준선 설정

2차 전문가 검토과정에서 26명의 패널이 참여하여 각 문항과 진단도구 전체에 대해 수준을 설정하였다. 즉, 학생평가 전문성의 기대수준은 문항 단위와 검사 단위의 두 가지 방법으로 설정하였다.

우선 문항 단위의 기준선은 패널들이 문항별로 기대하는 평가 전문성의 정도를 판정하여 전체 문항에 대한 기준선을 산출하였다. 이러한 문항 단위의 기준선을 각 영역별로 합산하여 영역의 기준선을 산출하였고, 다시 영역의 기준선을 기초로 하여 전체 진단도구에 대한 기준선을 산출하였다.

검사 단위의 기준선은 패널들이 진단도구 전체에서 우리나라 교사에게 기대하는 학생평가 전문성의 수준을 어느 정도로 판단하는지를 5점 척도를 기준으로 설정하도록 하였다. 문항 단위의 기준선과 검사 단위의 기준선은 〈표 2〉와 같다.

〈표 2〉 교사의 학생평가 전문성 진단도구의 전문성 기준선

구분	내용 영역	평균	(표준편차)
문항 단위의 기준선	평가방법의 선정	3.98	(.58)
	평가도구의 개발	4.13	(.46)
	평가실시・채점・성적부여	4.11	(.52)
	평가결과의 분석・해석・활용・의사소통	3.98	(.59)
	평가의 윤리성	4.30	(.64)
	전체	4.07	(.51)
검사 단위의 기준선		3.82	(.49)

문항 단위와 검사 단위에서 설정된 기준선은 각 영역과 전체 검사에서 척도 평균 3.82~4.30점으로 나타나 4점을 기준으로 할 때 영역 간의 큰 차이는 없었다. 각 영역별로 평가 전문성의 기준선을 보면 평가의 윤리성 기준에 대한 전문가들의 기대수준이 다른 영역에 비해서 높았고, '평가방법의 선정'과 '평가결과의 분석・해석・활용・의사소통' 영역에 대한 전문성 기대수준이 좀 더 낮았다. 또한 기준선의 표준편차를 살펴볼 때 교과 전문가들과 측정・평가 전문가들로 구성된 패널들의 의견 차이는 크지 않는 것으로 나타났다. 5점 척도의 척도점 범위(1~5점)와 간격을 고려할 때 표준편차 .46~.65의 범위는 패널들의 기준선 판정이 일관적이었음을 시사한다.

패널들의 기준선 설정 결과, 전반적으로 교사들의 평가 전문성 기준선은 척도점의 범위인 1~5점에서 4점을 기준으로 ±0.5점의 범위에서 설정하는 것이 적합하였다. 즉 3.5점 이상에서 4.5점 미만, 90문항을 기준으로 할 때 전체 점수가 315점에서 404점일 때 학생평가 전문성의 기준선을 통과하였다고 할 수 있다. 교사 스스로 진단한 결과가 평균이 3.5점 미만이거나 90문항 전체의 총점이 315점 미만일 때 평가 전문성이 부족하다고 해석할 수 있다. 교사 스스로 진단한 평가 전문성을 해석할 수 있는 평가 준거는 〈표 3〉과 같다.

〈표 3〉 교사의 학생평가 전문성 진단도구의 전문성 기준선 해석 준거

척도 평균의 기준선 범위	평가 전문성 해석
4.5 이상	평가 전문성이 높다
3.5 이상 4.5 미만	평가 전문성이 있다
2.5 이상 3.5 미만	평가 전문성이 낮다
2.5 미만	평가 전문성이 매우 낮다

Ⅲ. 교사의 학생평가 전문성 진단도구 타당화

1. 교사의 학생평가 전문성 진단도구의 타당화 절차

검사도구에 대한 타당도의 검증 없이 얻은 검사 결과에 의한 해석은 잘못된 결론을 유도할 가능성이 있기 때문에 교사의 학생평가 전문성 진단도구를 타당화하는 작업은 반드시 필요하다. 여기서 말하는 타당도(validity)란 검사가 측정하고자 의도하는 것을 얼마나 충실히 측정하였느냐, 검사 점수가 검사의 사용 목적에 얼마나 부합하느냐를 확인하는 문제다(성태제, 1995). 그래서 Ⅱ장에 기술된 진단도구가 '교사의 학생평가 전문성'을 제대로 측정하는지, 아니면 이와 무관하거나 어긋나는 것을 측정하는지를 종합적으로 분석하고 평가하는 타당화 작업을 수행하였다.

가. 교사의 학생평가 전문성 진단도구 타당화 방향

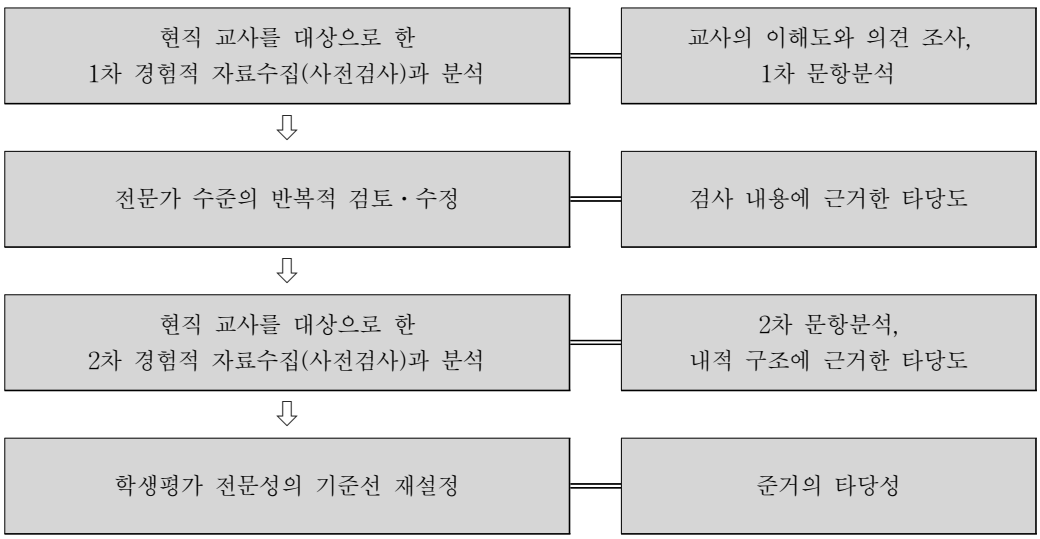
교사가 스스로 학생평가 전문성 정도를 측정하도록 개발된 진단도구는 '교사의 학생평가 전문성 기준'에서 요구하는 평가지식과 기능에 근거하고 있으며, 교사가 스스로 진단한 전문성 정도는 기준선에 따라 준거참조평가의 방식으로 해석한다. 개별 교사의 전문성 수준을 〈표 3〉에 제시되어 있는 절대적인 준거에 비추어 해석하므로 진단도구는 준거참조검사이다.

준거참조검사 타당성의 중요한 기준은 내용타당도가 된다(임인재, 1976). 주어진 내용과 목표를 잘 반영할 수 있도록 대표적으로 표집된 문항으로 검사가 구성되었는지가 준거참조검사의 관건이다. 교사의 학생평가 전문성 진단도구에서 측정하고자 하는 내용은 학생평가 전문성이고, 학생평가 전문성을 갖춘 교사가 평가활동에서 보일 수 있는 지식과 기능은 '교사의 학생평가 전문성 기준'에 진술되어 있다. 따라서 이 연구에서는 학생평가 전문성 진단도구의 문항내용에 대해 학생평가 전문성 기준의 대표성, 부합성과 그 점수를 해석하기 위

한 준거의 적절성에 초점을 두고 진단도구의 타당화를 수행하였다. 뿐만 아니라 가능한 한 다양한 접근을 시도하여 검사 내용에 근거한 내용타당도 이외에도 검사의 내적 구조에 근거한 구인타당도도 검증하였다.

나. 교사의 학생평가 전문성 진단도구 타당화 절차

교사의 학생평가 전문성 진단도구의 타당화는 진단도구의 목적, 구성, 특성, 문항 유형, 개발 과정, 실용성 등의 측면에 대한 검토에서부터 시작하였다. 이러한 검토를 통하여 요목화한 점들을 바탕으로 하여 전문가의 논리적 분석과 경험적 자료수집과 통계분석을 중심으로 타당화를 진행하였고, 그 절차를 도식화하면 〔그림 1〕과 같다.



〔그림 1〕 교사의 학생평가 전문성 진단도구 타당화 절차

우선 1차 사전검사를 실시하여 현직 교사를 대상으로 문항내용 이해도를 조사하고 응답 경향을 관찰하였으며, 교사들의 응답 자료를 수집하여 문항의 양호도를 분석하였다. 교육 측정·평가 전문가와 교과 전문가들이 문항내용을 논리적으로 분석하고 검토하여 검사 내용에 기초한 타당도, 즉 내용타당도를 검증하였으며, 교사들의 요구와 전문가의 의견을 반영하여 진단도구의 문항 수를 축소하고 문항의 진술을 정련하였다.

그리고 2차 사전검사를 실시하여 수정한 진단도구에 대한 교사들의 응답 자료를 수집하여 문항반응분포를 분석하고 검사의 내적 구조에 기초한 타당도, 즉 구인타당도를 검증하였다. 마지막으로 전문성의 수준을 해석하기 위한 준거인 기준선을 재설정하여 그 타당

성을 확보하였다. 더 구체적인 진단도구의 타당화 과정과 그 결과는 2절에 상세히 기술되어 있다.

다. 교사의 학생평가 전문성 진단도구 타당화 연구 대상

교사의 학생평가 전문성 진단도구 타당화는 전문가의 논리적 분석, 경험적 자료수집, 통계분석을 중심으로 수행되었는데, 이 절에서는 경험적 자료수집과 통계분석에 기초한 타당화 연구 대상에 대해 기술한다.

1) 1차 경험적 자료

교사의 학생평가 전문성 진단도구 2차 수정본 90문항의 양호도를 파악하기 위하여 분석한 경험적 자료는 초·중·고등학교 교사 445명의 응답 자료이다. 이 자료는 2005년 11월에 수집한 교사 198명의 응답 자료와 2006년 4~5월 수집한 교사 247명의 응답 자료를 통합한 것이다. 2005년 11월에 수집된 자료는 문항분석과 검사도구의 양호도 분석을 안정적으로 수행할 수 있을 만큼 사례 수가 충분하지 않았을 뿐만 아니라, 실제 교사들이 응답하는 것을 관찰하기 위하여 자료를 추가적으로 수집하였다. 자료를 수집한 시점의 간격은 교사들의 학생평가 전문성의 수준이 변화할 만큼 길지 않아 두 자료를 합쳐서 분석하여도 무방하다고 할 수 있다.

2) 2차 경험적 자료

문항의 수를 60문항으로 줄이고 진술내용을 정련한, 교사의 학생평가 전문성 진단도구 최종 수정본을 경험적 자료에 기초한 타당도를 검증하는 데 사용된 자료는 다음과 같다. 실제 중·고등학교에 재직 중인 교사 중 국민공통기본 교육과정에 해당하는 10개 교과인 국어, 사회, 수학, 과학, 영어, 도덕, 기술·가정, 체육, 음악, 미술을 담당하고 있는 교사 1,000명을 표집하고 우편을 통해 진단도구 최종 수정본에 대한 응답 자료를 수집하였다. 전국의 중·고등학교 학교 수를 고려하여 중학교 60개, 고등학교 40개를 표집하되, 16개 시·도별 표집학교 수는 전체 학교 수에 비례하여 배분하였고(〈표 4〉 참조), 각 학교에서 10명의 교사를 표집하였다.

표집된 교사 1,000명 중 793명의 교사 응답 자료가 회수되었고, 불성실하게 응답한 경우를 제외하고 전체 707명의 교사 응답 자료가 진단도구 최종 수정본의 타당도 검증을 위한 통계분석에 사용되었다.

〈표 4〉 진단도구의 2차 사전검사를 위한 표집학교 수

2005. 4. 1. 기준

구분	중학교		고등학교	
	전체	표집	전체	표집
서울	363	7	214	6
부산	166	3	91	3
대구	118	2	65	2
인천	114	2	70	2
광주	75	2	48	1
대전	76	2	43	1
울산	51	1	29	1
경기	472	10	212	6
강원	161	3	66	2
충북	123	3	49	2
충남	187	4	75	2
전북	201	4	74	2
전남	248	5	85	3
경북	283	6	125	3
경남	255	5	118	3
제주	42	1	18	1
총계	2,935	60	1,382	40

※ 출처 : 교육통계연감

2. 교사의 학생평가 전문성 진단도구 타당화 결과

가. 문항내용 이해도와 의견 조사

교사의 학생평가 전문성 진단도구에 응답하는 교사가 각 문항이 의미하는 바를 정확히 이해하지 못하면 그 문항에 대한 응답은 적절하지 못하고, 진단 결과는 쓸모없을 것이다. 따라서 진단도구의 타당화를 위해 우선적으로 진단도구의 실수요자, 즉 전문성 진단의 대상이 되는 중·고등학교 교사 47명에게 2차 수정본의 각 문항내용에 대한 이해도 등을 자유로운 형식으로 조사하여 진단도구가 제 역할을 하고 있는지 알아보았다. 조사 결과, 전반적으로 '문항이 어려운 편이다'고 응답하였는데, 이는 백분위점수, 준거참조평가 등 전문적인 평가용어 자체를 모르기 때문이었다. '문항 수가 너무 많다', '중복된 문항이 있다'는 응답들도 있

었으며, 90개 문항에 응답하는 데 소요되는 시간이 길어서 지루해 하는 경향을 보였다. 이러한 조사 결과는 문항 수를 축소하여 진단도구의 응답 시간을 단축함으로써 더 간편한 도구로 수정할 필요가 있음을 보여주었다.

나. 문항의 양호도 분석

교사의 학생평가 전문성 진단도구 2차 수정본 90문항에 대해 실제 현장교사들이 응답한 자료를 수집하고 문항분석을 실시하였다. 이를 위해 초·중·고등학교 교사 445명이 진단도구의 각 문항에 대한 응답 자료를 수집하였다. 교사들은 각 문항의 진술을 읽고, '전혀 그렇지 않다'에서 '매우 그렇다'로 구분된 5점 척도 중의 하나를 선택하였다. 진단도구의 문항 선별을 위하여 통계분석으로 문항별 평균, 표준편차, 영역별 문항－총점 간 상관분석을 하였다. 문항별 평균은 2.99~4.36이었고, 표준편차는 0.65~0.97이었으며, 문항－총점 간 상관계수는 0.40~0.69이었다. 표준편차의 크기를 볼 때, 대체로 반응은 다양하지 않았다.

이 같은 결과는 진단도구가 개별 교사의 점수를 상대비교가 주요 목적인 규준참조검사가 아니라 절대기준 도달 여부를 판정하는 것이 주요 목적인 준거참조검사의 성격을 지니기 때문이다. 문항분석 결과 평균을 볼 때 교사들의 반응이 지나치게 편포되어 있는 문항은 없는 것으로 나타났다.

다. 문항내용에 근거한 타당도 검증 : 내용타당도

교사의 학생평가 전문성 진단도구가 측정하고자 하는 내용, 즉 전문성 기준에 기술되어 있는 바에 대해 충분히 알고 있는 교육 측정·평가 전문가 5명에게 진단도구가 측정하고자 하는 속성, 즉 교사의 학생평가 전문성의 5개 내용 영역을 제대로 측정하고 있는지에 대해 면밀히 검토하도록 하였다.

전문가에게 전문성 기준과 진단도구의 목적, 대상, 내용 등에 대하여 설명하고 진단도구와 문항 검토지를 배부하여 각자 독립적으로 진단도구의 각 문항과 전문성 기준과의 관련 여부를 분석하도록 의뢰하였다. 또한 문항의 수를 축소한다면 몇 개로 줄이며 어떤 문항을 삭제할 것인지, 문항의 진술에 사용된 단어는 적절한지, 문법적으로 오류는 없는지, 애매모호한 점은 없는지 등에 대한 검토도 함께 이루어졌다. 문항마다 수정과 삭제 의견을 제시할 수 있도록 배부한 검토지에 기록한 상세검토 의견을 토대로 문항 수정, 선제를 위한 전문가 협의회를 거쳤다. 전문가들에 의한 문항내용의 논리적 분석·검토 의견을 수렴하고 종합한 결과는 다음과 같다.

첫째, 교사의 학생평가 전문성 진단도구를 구성하고 있는 90문항 중 전문성 기준에서

벗어나는 문항은 전혀 없었고, 학생평가 전문성의 5개 내용 영역별 비중을 볼 때 특정한 영역에 국한되거나 치중되지 않다고 분석되었다. 따라서 진단도구의 문항들은 측정하고자 의도하는 내용, 즉 학생평가 전문성을 충분히 대표하고 있으므로 내용타당도가 있다고 평가되었다.

둘째, 교사의 학생평가 전문성 진단도구의 문항 수에 대한 전문가들의 의견은 55~75개 정도의 범위로 모아졌다. 전문성 영역별로 지식과 기능의 비중을 고려하여 문항 선제를 하되, 우선적으로 선제에서 제외되거나 통합될 문항은 다른 문항의 진술과 중복되는 부분이 있는 문항이었다. 문항을 통합할 경우 전문가들은 문항이 다루는 기준의 성격에 따라 지식과 기술을 구분하여 진술하거나 지식과 기술을 통합하여 진술하는 것이 적절하다고 합의하였다.

라. 진단도구의 수정

문항내용에 대한 교사들의 이해도, 전문가들의 검토 의견, 전문성 영역별 비중, 문항의 중복성이나 상대적 중요도 등을 고려하여 문항을 제거 혹은 통합하여 55개 문항을 선제하고, 각 문항별 수정 의견을 반영하여 문항의 진술을 정련함으로써 문항에서 측정하려고 하는 전문성 기준의 의미를 살리면서 명료하게 진술하였다. 문항의 수를 55개로 선제한 것은 가급적 최소한의 문항으로 축소하여 실제적 유용성을 높이고자 하기 위함이었다.

55개 문항으로 구성된 진단도구의 수정안에 대한 검토를 교육측정·평가와 교과교육 분야의 전문가 10명에게 의뢰하였다. 각 문항마다 삭제 혹은 잔류에 대한 찬반 의견을 제시할 수 있도록 검토지를 배부하여 각자 독립적으로 심층 검토하도록 하였다. 전문가들의 검토 내용은 진단도구의 길이(문항 수), 기준과 문항의 관련성, 선제의 적절성, 문항 진술의 명료성 등이다.

전문가들의 각 문항별 검토 의견과 전문성 기준의 구체화된 지식과 기능을 고려하여 문항을 재선제하고 재수정·보완하는 과정을 거쳐 최종 60개 문항을 확정하였다. 〈표 5〉에 제시된 바와 같이 최종 수정·보완된 진단도구는 '평가방법의 선정' 영역에 대한 문항 13개, '평가도구의 개발' 영역에 대한 문항 13개, '평가실시·채점·성적부여' 영역에 대한 문항 9개, '평가결과의 분석·해석·활용·의사소통' 영역에 대한 문항 18개, '평가의 윤리성' 영역에 대한 문항 7개로 총 60개의 문항으로 구성되어 있다. 각 문항의 내용은 [부록]에 제시되어 있다.

〈표 5〉 교사의 학생평가 전문성 진단도구 최종 수정본의 문항 구성

내용 영역	전문성 기준	문항 수
평가방법의 선정	교사는 학습목표와 평가의 목적을 확인하고 명료화해야 한다.	2
	교사는 학습목표와 평가의 목적에 적합한 평가방법을 선정해야 한다.	11
평가도구의 개발	교사는 평가의 목적과 내용에 적합한 평가도구를 개발하거나 선택하여 사용해야 한다.	7
	교사는 평가도구의 질을 스스로 점검하고 개선해야 한다.	6
평가실시・채점・성적부여	교사는 평가계획에 부합하도록 평가를 실시해야 한다.	3
	교사는 정확하게 채점하고 평가의 목적에 부합하도록 성적을 부여해야 한다.	6
평가결과의 분석・해석・활용・의사소통	교사는 평가결과를 정확하고 타당하게 분석・해석해야 한다.	7
	교사는 평가결과를 수업과 학생에 대한 교육적 의사 결정에 활용해야 한다.	6
	교사는 학생, 학부모, 교육 관련자와 평가결과에 대해 정확하게 의사소통해야 한다.	5
평가의 윤리성	교사는 학생의 인격을 존중하며 모든 평가활동 시 윤리적・법적 책임을 준수해야 한다.	4
	교사는 학생의 특성, 배경을 고려하여 공정하게 평가를 실시하고 평가의 적절성을 판단해야 한다.	3
합 계		60

마. 문항 분석

반복적인 수정・검토 과정을 통해 문항의 수를 줄이고 진술 내용을 정련한 교사의 학생평가 전문성 진단도구 최종 수정본의 내용타당도는 앞에서 기술한 바와 같이 전문가의 논리적 분석과 판단으로 입증되었다. 내용타당도 이외에 추가적으로 경험적 자료에 기초한 타당도를 검증하기 위해서 실제 현장 교사 707명의 학생평가 전문성 진단도구의 응답 자료를 분석하였다. 현장 교사에게 학생평가 전문성 진단도구를 적용하여 '매우 그렇다'에 5점, '전혀 그렇지 않다'에 1점으로 평정하게 하였다. 교사의 학생평가 전문성 진단도구에 대한 문항반응 분포는 〈표 6〉과 같다.

〈표 6〉 교사의 학생평가 전문성 진단도구 최종 수정본의 문항 반응분포

내용 영역	문항번호	빈도 (백분율)					평균	표준편차
		전혀 그렇지 않다	그렇지 않다	보통이다	그렇다	매우 그렇다		
평가방법의 선정	1	3 (0.42)	12 (1.70)	114 (16.12)	337 (47.67)	241 (34.09)	4.13	0.59
	2	3 (0.42)	13 (1.84)	113 (15.98)	345 (48.80)	233 (32.96)	4.12	0.59
	3	54 (7.64)	157 (22.21)	252 (35.64)	169 (23.90)	75 (10.61)	3.08	1.19
	4	14 (1.98)	78 (11.03)	218 (30.83)	294 (41.58)	103 (14.57)	3.56	0.88
	5	5 (0.71)	34 (4.81)	209 (29.56)	324 (45.83)	135 (19.09)	3.78	0.69
	6	6 (0.85)	54 (7.64)	248 (35.08)	302 (42.72)	97 (13.72)	3.61	0.72
	7	5 (0.71)	52 (7.36)	216 (30.55)	333 (47.10)	101 (14.29)	3.67	0.70
	8	5 (0.71)	79 (11.17)	282 (39.89)	281 (39.75)	60 (8.49)	3.44	0.68
	9	1 (0.14)	26 (3.68)	187 (26.45)	389 (55.02)	104 (14.71)	3.80	0.53
	10	7 (0.99)	96 (13.58)	283 (40.03)	266 (37.62)	55 (7.78)	3.38	0.72
	11	19 (2.69)	97 (13.72)	275 (38.90)	254 (35.93)	62 (8.77)	3.34	0.84
	12	6 (0.85)	76 (10.75)	292 (41.30)	277 (39.18)	56 (7.92)	3.43	0.67
	13	5 (0.71)	44 (6.22)	249 (35.22)	333 (47.10)	76 (10.75)	3.61	0.62
평가도구의 개발	14	4 (0.57)	32 (4.53)	225 (31.82)	347 (49.08)	99 (14.00)	3.71	0.61
	15	19 (2.69)	141 (19.94)	276 (39.04)	200 (28.29)	71 (10.04)	3.23	0.94
	16	21 (2.97)	161 (22.77)	316 (44.70)	184 (26.03)	25 (3.54)	3.04	0.75
	17	8 (1.13)	92 (13.01)	272 (38.47)	273 (38.61)	62 (8.77)	3.41	0.75
	18	12 (1.70)	96 (13.58)	293 (41.44)	262 (37.06)	44 (6.22)	3.33	0.72
	19	5 (0.71)	63 (8.91)	236 (33.38)	313 (44.27)	90 (12.73)	3.59	0.72
	20	16 (2.26)	114 (16.12)	336 (47.52)	217 (30.69)	24 (3.39)	3.17	0.67
	21	12 (1.70)	57 (8.06)	263 (37.20)	300 (42.43)	75 (10.61)	3.52	0.72
	22	8 (1.13)	68 (9.62)	250 (35.36)	300 (42.43)	81 (11.46)	3.53	0.74
	23	9 (1.27)	91 (12.87)	262 (37.06)	276 (39.04)	69 (9.76)	3.43	0.77
	24	5 (0.71)	28 (3.96)	169 (23.90)	393 (55.59)	112 (15.84)	3.82	0.59
	25	4 (0.57)	39 (5.52)	208 (29.42)	369 (52.19)	87 (12.31)	3.70	0.60
	26	23 (3.25)	118 (16.69)	266 (37.62)	230 (32.53)	70 (9.90)	3.29	0.93

평가실시·채점·성적부여	27	4 (0.57)	30 (4.24)	114 (16.12)	287 (40.59)	272 (38.47)	4.12	0.75
	28	7 (0.99)	52 (7.36)	229 (32.39)	307 (43.42)	112 (15.84)	3.66	0.75
	29	8 (1.13)	9 (1.27)	62 (8.77)	243 (34.37)	385 (54.46)	4.40	0.63
	30	6 (0.85)	8 (1.13)	63 (8.91)	277 (39.18)	353 (49.93)	4.36	0.58
	31	8 (1.13)	16 (2.26)	128 (18.10)	358 (50.64)	197 (27.86)	4.02	0.65
	32	2 (0.28)	20 (2.83)	178 (25.18)	391 (55.30)	116 (16.41)	3.85	0.53
	33	30 (4.24)	114 (16.12)	269 (38.05)	243 (34.37)	51 (7.21)	3.24	0.90
	34	31 (4.38)	134 (18.95)	267 (37.77)	229 (32.39)	46 (6.51)	3.18	0.92
	35	12 (1.70)	67 (9.48)	226 (31.97)	317 (44.84)	85 (12.02)	3.56	0.78
평가결과의 분석·해석·활용·의사소통	36	31 (4.38)	121 (17.11)	256 (36.21)	218 (30.83)	81 (11.46)	3.28	1.04
	37	6 (0.85)	67 (9.48)	273 (38.61)	285 (40.31)	76 (10.75)	3.51	0.71
	38	8 (1.13)	91 (12.87)	255 (36.07)	293 (41.44)	60 (8.49)	3.43	0.74
	39	2 (0.28)	25 (3.54)	226 (31.97)	355 (50.21)	99 (14.00)	3.74	0.56
	40	8 (1.13)	93 (13.15)	290 (41.02)	265 (37.48)	51 (7.21)	3.36	0.71
	41	14 (1.98)	92 (13.01)	265 (37.48)	268 (37.91)	68 (9.62)	3.40	0.81
	42	4 (0.57)	44 (6.22)	223 (31.54)	361 (51.06)	75 (10.61)	3.65	0.60
	43	4 (0.57)	49 (6.93)	277 (39.18)	304 (43.00)	73 (10.33)	3.56	0.63
	44	5 (0.71)	58 (8.20)	265 (37.48)	304 (43.00)	75 (10.61)	3.55	0.67
	45	3 (0.42)	30 (4.24)	219 (30.98)	356 (50.35)	99 (14.00)	3.73	0.59
	46	1 (0.14)	42 (5.94)	210 (29.70)	371 (52.48)	83 (11.74)	3.70	0.57
	47	1 (0.14)	37 (5.23)	201 (28.43)	386 (54.60)	82 (11.60)	3.72	0.55
	48	13 (1.84)	63 (8.91)	268 (37.91)	288 (40.74)	75 (10.61)	3.49	0.75
	49	4 (0.57)	50 (7.07)	242 (34.23)	311 (43.99)	100 (14.14)	3.64	0.69
	50	5 (0.71)	54 (7.64)	254 (35.93)	306 (43.28)	88 (12.45)	3.59	0.69
	51	6 (0.85)	78 (11.03)	267 (37.77)	283 (40.03)	73 (10.33)	3.48	0.73
	52	5 (0.71)	58 (8.20)	271 (38.33)	311 (43.99)	62 (8.77)	3.52	0.63
	53	4 (0.57)	41 (5.80)	234 (33.10)	348 (49.22)	80 (11.32)	3.65	0.60
평가의 윤리성	54	4 (0.57)	17 (2.40)	133 (18.81)	329 (46.53)	224 (31.68)	4.06	0.65
	55	4 (0.57)	14 (1.98)	88 (12.45)	264 (37.34)	337 (47.67)	4.30	0.65
	56	8 (1.13)	25 (3.54)	128 (18.10)	287 (40.59)	259 (36.63)	4.08	0.64
	57	6 (0.85)	31 (4.38)	185 (26.17)	318 (44.98)	167 (23.62)	3.86	0.78
	58	3 (0.42)	18 (2.55)	113 (15.98)	308 (43.56)	265 (37.48)	4.15	0.73
	59	8 (1.13)	54 (7.64)	230 (32.53)	288 (40.74)	127 (17.96)	3.67	0.80
	60	6 (0.85)	54 (7.64)	246 (34.79)	296 (41.87)	105 (14.85)	3.62	0.74

〈표 6〉에서 보는 바와 같이 교사의 학생평가 전문성 진단도구 60문항에 대한 반응분포는 대다수 문항에서 부적편포를 나타냈다. 거의 모든 문항에 대해 '전혀 그렇지 않다'와 '그렇지 않다'에 응답한 교사의 비율은 낮았다. 문항점수의 평균이 3.04~4.40, 표준편차는 0.53~1.19로 극단의 평균점수를 보이거나 지나치게 적은 표준편차를 갖는 문항은 없었다.

바. 내적구조에 근거한 타당도 검증 : 구인타당도

일반적으로 검사도구의 내적구조에 근거한 타당도를 검증하기 위하여 사용되는 통계적 방법은 요인분석, 집단비교법, 상관계수법 등이 있다. 본 연구에서 타당도를 평가하고자 하는 '교사의 학생평가 전문성 진단도구'는 5개의 내용 영역으로 구성되어 있다. 진단도구에서 측정하려는 잠재적 특성은 교사가 평가 활동 전반에서 발휘해야 하는 전문성으로, 진단도구의 내용 영역은 교사의 학생평가 활동의 영역이지 전문성의 하위 구인은 아니다. 5개의 내용 영역 중 평가방법의 선정, 평가도구의 개발, 평가실시・채점・성적부여, 평가결과의 분석・해석・활용・의사소통 영역은 시계열성을 가지고 있으며, 평가의 윤리성 영역은 나머지 4개 영역의 바탕이 되는 태도에 대한 것이다.

이와 같이 학생평가 전문성의 내용 영역이 서로 상당히 밀접하게 관련되어 있으므로 진단도구는 단일 요인으로 구성되는 것이 논리적이다. 진단도구의 구인을 확인하기 위하여 공통요인모형으로 단일주축분해를 사용하여 요인분석을 수행하였다. 초기추정치는 다중상관제곱치(SMC; Squared Multiple Correlation)였으며, 요인구조의 회전은 사교회전(Quartimax with Kaiser Normalization)의 방법을 사용하였다. 요인분석 결과 제1요인이 설명하는 분산의 비율은 35.5%, 제2요인이 설명하는 분산의 비율은 5.0%로 단일 요인 구조의 척도로 해석할 수 있는 결과가 나타났다. 따라서 본 연구에서는 단일한 특성을 측정하기 위한 진단도구가 의도한 대로 학생평가 전문성을 측정하고 있는지를 검증하기 위하여 전문성 수준에 따른 집단비교법을 사용하는 것이 적절하다고 판단하였다.

'교사의 학생평가 전문성 진단도구'가 타당하다면, 평가관련 연수에 참여할수록 교사들의 평가 전문성이 높아질 것이라 예상할 수 있으며, 진단도구에 의해 진단된 전문성 수준이 평가관련 연수의 참여여부에 따라 차이가 있을 것이다. 평가관련 연수의 참여여부에 따라 교사를 두 집단으로 구분하고, 각 집단의 학생평가 전문성 수준을 비교한 결과, 〈표 7〉에서 보듯이 유의한 차이가 있었다.

〈표 7〉 평가관련 연수 경험 유무에 따른 교사의 학생평가 전문성 수준 차이

평가관련 연수 경험 / 내용 영역	경험 없음(367명)		경험 있음(340명)		평균 차이
	평균	표준편차	평균	표준편차	
평가방법의 선정	45.99	6.85	47.98	7.55	1.97*
평가도구의 개발	43.47	7.67	46.20	7.77	2.73*
평가실시・채점・성적부여	34.01	5.07	34.79	4.96	0.78*
평가결과의 분석・해석・활용・의사소통	62.69	10.63	65.42	10.59	2.74*
평가의 윤리성	27.49	4.61	28.01	4.22	0.52
전체	213.65	29.61	222.39	30.53	8.74*

* $p<.01$

학생평가 전문성 영역별로 교사의 수준을 살펴보면 연수 경험이 있는 교사 집단의 전문성 수준이 연수 경험이 없는 교사 집단의 전문성 수준보다 높은 것으로 나타났으며, '평가의 윤리성' 영역을 제외한 4개 내용 영역의 집단별 평가 전문성 점수의 차이는 통계적으로 유의하였다. 이와 같이 '교사의 학생평가 전문성 진단도구'의 점수는 학생평가와 관련된 지식과 기술에 차이가 있을 것으로 예상되는 두 집단의 학생평가 전문성 수준을 변별하는 것으로 나타나 진단도구의 내적구조에 의한 타당도가 입증되었다.

사. 전문성 기준선의 적절성

교사의 학생평가 전문성 진단도구 최종 수정본 60문항을 국어, 사회, 수학, 과학, 영어, 도덕, 기술・가정, 체육, 음악, 미술의 교과 전문가와 교육 측정・평가 전문가에게 의뢰하여 각각의 문항에 대한 기준선을 재설정한 다음 그 적절성을 검토하였다. 우리나라 교사에게 기대하는 학생평가 전문성의 수준을 어느 정도로 판단하는지를 5점 척도를 기준으로 설정하도록 하였다.

학생평가 전문성에 대한 문항 단위의 기준선은 전문가 패널들이 문항별로 기대하는 평가 전문성의 정도를 판정하여 60개의 문항에 대한 기준선을 산출하였다. 이러한 문항 단위의 기준선을 각 영역별로 합산하여 영역의 기준선을 산출하였고, 다시 영역의 기준선을 기초로 하여 전체 진단도구에 대한 기준선을 산출하였다. 최종 수정된 진단도구 60문항에서 설정된 기준선과 2차 수정본 90개 문항에서의 기준선은 〈표 8〉과 같다.

90개 문항에 대한 전문성 기준선과 60개 문항일 때의 전문성 기준선은 5개 내용 영역과 전체 검사에서 별 차이가 없이 유사하였다. 문항 단위에서 설정된 기준선은 각 내용 영역과 전체 검사에서 척도 평균 3.95~4.31점으로 나타나 4점을 기준으로 할 때 내용 영역 간의 큰 차이는 없었다. 또한 기준선의 표준편차를 살펴볼 때 전문가 패널들 간의 의견 차이는 크지

않은 것으로 나타났다. 5점 척도의 점수 범위(1~5점)와 간격을 고려할 때 표준편차 .48~.59의 범위는 패널들의 기준선 판정이 일관적이었음을 시사한다.

〈표 8〉 '교사의 학생평가 전문성 진단도구'의 전문성 기준선 검토

내용 영역	90문항	60문항
	평균(표준편차)	평균(표준편차)
평가방법의 선정	3.98 (.58)	4.02 (.51)
평가도구의 개발	4.13 (.46)	4.21 (.49)
평가실시・채점・성적부여	4.11 (.52)	4.21 (.51)
평가결과의 분석・해석・활용・의사소통	3.98 (.59)	4.00 (.56)
평가의 윤리성	4.30 (.64)	4.31 (.59)
전체	4.07 (.51)	4.12 (.48)

패널들의 기준선 재설정 결과, 2차 수정본의 기준선과 큰 차이가 없었으므로 척도점의 범위인 1~5점에서 4점을 기준으로 ±0.5점의 범위에서 설정한 교사들의 평가 전문성 기준선(〈표 3〉 참조)은 적절하며, 이는 60문항으로 축소된 진단도구 최종 수정본에서도 수정 없이 공통적으로 사용해도 무방하다고 하겠다. 교사 스스로 진단한 결과가 평균이 3.5점 이상, 60문항의 전체 점수가 210점 이상일 때 학생평가 전문성의 기준선을 통과하였다고 할 수 있고, 평균이 3.5점 미만이거나 총점이 210점 미만일 때 평가 전문성이 부족하다고 해석할 수 있다.

아. 신뢰도 분석

교사의 학생평가 전문성 진단도구의 신뢰도를 분석하기 위하여 각 내용 영역과 전체 문항의 내적 일관성 신뢰도 계수, Cronbach α를 추정하였고, 구체적인 결과는 〈표 9〉와 같다.

〈표 9〉 교사의 학생평가 전문성 진단도구의 신뢰도 계수

내용 영역	문항 수	Cronbach α
평가방법의 선정	13	.89
평가도구의 개발	13	.92
평가실시・채점・성적부여	9	.84
평가결과의 분석・해석・활용・의사소통	18	.94
평가의 윤리성	7	.87
전체	60	.97

〈표 9〉에 제시된 바와 같이 내용 영역별 신뢰도 계수는 '평가방법의 선정' 영역이 .89, '평가도구의 개발' 영역이 .92, '평가실시・채점・성적부여' 영역이 .84, '평가결과의 분석・해석・활용・의사소통' 영역이 .94, '평가의 윤리성' 영역이 .87이며, 전체 문항의 신뢰도 계수는 .97로 나타났다. 교사의 학생평가 전문성 진단도구 전체와 내용 영역의 신뢰도를 살펴본 결과, 진단도구는 상당히 일관된 내용을 평가하는 것으로 나타났다.

Ⅳ. 교사의 학생평가 전문성 진단도구 활용

실제적으로 교사들은 업무 시간의 1/3~1/2의 시간을 평가와 관련된 활동에 할애하고 있으며, 이러한 전문적인 업무의 중요한 부분으로 평가 관련 능력이 필요하지만(Stiggins & Conklin, 1992; Nitko, 2001) 교사들은 평가와 관련된 전문적인 훈련을 받은 경험이 없거나 적으며, 평가활동을 수행하는 데 준비가 덜 되어 있다고 스스로 느끼기도 한다(Schafer, 1993; 김신영, 2002; 이정애, 2003). 또한 교원 양성기관인 대학에서 교육평가 관련과목을 전혀 이수하지 않은 채 현장에 나온 교사들도 있고, 이수했다고 해도 현장에서의 평가관련 업무에 도움이 된다고 생각하는지 의심스럽다(2007, 김신영). 교사연수 프로그램에서 제공하는 평가 관련 연수 내용들이 학생평가에서 발생하는 문제들을 해결하는 데 필요한 지식이나 기술과는 거리가 있다는 지적도 있다(Stiggins, 1991; 김신영, 2002). 따라서 현장 교사와 예비 교사들에게 학생평가력을 제고할 수 있는 실질적인 기회들이 제공될 필요가 있다. 이러한 기회들은 실효성 있는 교사 대상의 교육 프로그램의 개발과 지속적인 운영을 의미하며, 아울러 교직 전문가로서 교사 스스로의 노력에서도 제공될 수 있다.

전문가 집단은 장기간의 훈련과정을 거쳐 전문적 직무수행에 요구되는 지식과 기술을 갖춘 후 사회적으로 공인된 자격을 획득하며 직업윤리를 준수하고, 직무수행과 관련된 의사결정을 위해서는 자율적으로 판단하고 그 판단에 대한 책임도 스스로 진다. 또한 전문성의 내면화에도 상당한 시간과 노력이 필요하며 일시적으로도 성취될 수 없다. 교사의 학생평가 전문성을 갖추고 신장시키기 위해서는 학생평가와 관련된 지식, 기술과 이의 적용에 대한 지속적인 내면화 과정이 요구된다. 이러한 의미에서 교사의 학생평가 전문성 진단도구는 교사들의 학생평가 전문성을 진단하고, 필요한 지식과 기술을 내면화하기 위한 중요한 연장으로 활용될 수 있다.

학생평가 전문성 진단도구는 전문가 집단의 공적인 합의 수준을 거쳐 타당화된 도구이기 때문에 교사의 학생평가 전문성을 갖추고 신장시키기 위한 체계적인 훈련과정으로서 제공되는 전・현직 교육 프로그램에 대하여 학생평가에 관련된 프로그램의 실효성을 증진시키기

위해 활용될 수 있다. 구체적으로 학생평가 전문성 진단도구의 실제 내용과 기준선은 교육 프로그램의 내용과 투입 수준을 결정할 수 있으며, 프로그램의 효과를 검증하기 위한 타당한 도구로 사용될 수 있다.

또한 전문가 집단인 교사들의 자율적인 노력을 위해 진단도구는 학생평가 전문성을 스스로 점검하고 확인할 수 있는 중요한 도구가 될 것이다. 학생평가 전문성에 대한 교사의 자기 진단은 알고 있어야 할 '지식'과 할 수 있는 '기능'에 초점을 두고 있기 때문에, 교사가 스스로 학생평가에 대해 무엇을 알고 할 수 있는지를 스스로 꼼꼼하게 점검하고 부족한 영역을 확인함으로써 자기장학을 할 수 있는 기회를 제공할 것이다.

아울러 교사의 학생평가 전문성 진단도구는 교사연수 기회를 제공하는 중요한 준거로 활용될 수 있다. 학생평가 전문성 진단을 통해 절대적인 기준선에 도달하지 못한 교사에게는 좀 더 많은 시·도 교육청 차원의 연수 기회를 제공하거나 스스로 자기연수를 계획하도록 할 수 있어서 차별화되고 효율적인 연수 프로그램의 개발, 운영이 가능하다.

참 고 문 헌

교육인적자원부(2005). **학업성적 신뢰제고 조치계획**. 서울: 교육인적자원부.

김경희, 김신영, 김성숙, 지은림, 반재천, 김수동(2006). 교사의 학생평가 전문성 기준 개발. **교육평가연구, 19**(2), 89-112.

김수동, 이의갑, 김경희, 김선희, 박은아, 신명선, 김수진, 박가나, 서수현, 전영석 (2005). **교사의 학생평가 전문성 신장 연구(Ⅱ)**. 한국교육과정평가원 연구보고 RRE 2005-3.

김신영(2002). 교실내 평가의 특성과 교사의 평가 전문성. **한국외국어 대학교 논문집**. 34, 541-558.

김신영(2007). 교사의 학생평가전문성과 중등교사 양성과정. **교육평가연구, 20**(1), 1-16.

남명호, 박소영, 송미영, 김국현, 김수동, 조일수, 임완성, 이경애, 오수학, 강민선, 강진호(2006). **교사의 학생평가 전문성 신장 연구(Ⅲ)**. 한국교육과정평가원 연구보고 RRE 2006-5.

백순근(2006). 학생평가에 대한 교사들의 학생부 신뢰도 제고 방안. 교육혁신위원회, **교사의 학생평가 전문성과 학생부 신뢰도 제고 방안 탐색**. 한국교육평가학회 공동 개최 학술세미나 자료집. 61-80.

서울시 교육청(2007). **2007학년도 중등 장학계획**. 서울: 서울시 교육청.

성태제(1995). **타당도와 신뢰도**. 서울: 양서원.

이인제, 이범홍, 박정, 진재관, 김옥남, 서수현, 김신영(2004). **교사의 학생평가 전문성 신장 모형과 기준**. 한국교육과정평가원 연구보고 RRE 2004-5-2.

이정애(2003). **중학교 교사의 학생평가에 대한 인식과 평가 전문성 연구**. 석사학위논문. 이화여자대학교 교육대학원. 미간행.

임인재(1976). **절대기준평가의 원리와 실제**. 서울: 배영사

American Federation of Teacher, National Council on Measurement in Education, & National Education Association(AFT, NCME, & NEA) (1990). Standards for teacher competence in educational assessment of students. *Educational Measurement: Issues and practice. 9*(4), 30-32.

McMillan (2004). *Classroom assessment principle and practice for effective instruction*. Boston: Allyn & Bacon.

Nitko, A. J. (2001). *Educational assessment of students* (3rd). Merrill Prentice Hall.

Schafer, W. D. (1993). Assessment literacy for teachers. *Theory into Practice, 32*, 118-126.

Stiggins, R. J. (1991). Assessment literacy. *Phi Delta Kappan, 72*, 534-239.

Stiggins, R. J. & Conklin, N. F. (1992). *In teachers' hands: Investigating the practices of classroom assessment*. Albany, NY: SUNY Press.

• *논문 접수 : 2007년 4월 15일 / 수정본 접수 : 2007년 5월 15일 / 게재 승인 : 2007년 5월 25일*

ABSTRACT

Development & Validation of Teacher's Assessment Professional Competency Test

Mi-Young Song
(Associate Research Fellow, Korea Institute of Curriculum & Evaluation)
Kyung-Hee Kim
(Research Fellow, Korea Institute of Curriculum & Evaluation)

Teacher's competence in student assessment can be defined as 'the competence or capability of collecting, interpreting as well as applying the students' assessment information about students' learning and achievement'. This competence is the assessment literacy that should be required before actually assessing the students. A teacher's competence test in student assessment was developed. This test are designed for teachers to measure and diagnose the competence in student assessment by themselves based on 'the standards of teacher's competence in student assessment'.

The number of items were reduced 90 to 60 in order to be more useful in practice. Validity evidence was obtained from an logical analysis of the relationship between the construct and wording and format of the items, and an empirical analysis of the internal structure. A baseline of essential competence was also validated for teachers to compare their diagnosed competency with it for each domains of the standards such as 'Choosing assessment methods', 'Developing assessment instrument', 'Administering, scoring, and grading', 'Analizing, interpreting, using, and communicating assessment results', and 'Recognizing ethics in assessment'.

The ways to apply 'the standards of teacher's competence in student assessment' and 'the teacher's assessment competence test' at schools and districts were showed in various aspects.

Key Words : student assessment, teacher's competence, test, validation

[부록] 교사의 학생평가 전문성 진단도구 문항

1. 학생이 달성해야 할 학습목표를 확인하고, 평가할 학습목표를 구체적으로 진술할 수 있다.
2. 진단・형성・총합평가의 목적과 특징을 알고 있다.
3. 규준참조평가와 준거참조평가의 특징과 장・단점을 알고 있다.
4. 양적평가와 질적평가의 특징과 장・단점을 알고 있다.
5. 다양한 학습목표와 학습활동에 따라 적합한 평가방법을 선택할 수 있다.
6. 학생들의 특성과 수준(사회문화적 배경, 흥미, 학습능력 등)에 적합한 평가방법을 선택할 수 있다.
7. 학생의 선행학습 수준이나 기초 능력을 진단하기에 적합한 평가방법을 선택할 수 있다.
8. 교수・학습 과정에서 학생이 겪는 어려움을 진단하기에 적합한 평가방법을 선택할 수 있다.
9. 학생의 성취 수준을 파악하기에 적합한 평가방법을 선택할 수 있다.
10. 학생의 성장이나 변화 정도를 파악하기에 적합한 평가방법을 선택할 수 있다.
11. 특별한 지도를 필요로 하는 학생을 변별하기에 적합한 평가방법을 선택할 수 있다.
12. 교수・학습 과정에서 학생 자신에 의한 평가나 학생 상호 간에 의한 평가가 필요한 상황을 판단하고 적용할 수 있다.
13. 각 평가방법이 교수・학습 활동에 미치는 긍정적인 영향과 부정적인 영향을 파악할 수 있다.
14. 평가의 목적과 내용에 적합한 평가도구를 선택할 수 있다.
15. 평가의 목적과 내용에 부합하는 평가도구가 없을 경우 평가도구를 직접 개발할 수 있다.
16. 검사제작의 일반적 절차에 따라 평가도구를 개발하기 위한 계획을 구체적으로 세울 수 있다.
17. 선택형 문항제작원리를 충실히 반영하여 평가도구를 개발할 수 있다.
18. 서답형 문항제작원리를 충실히 반영하여 평가도구를 개발할 수 있다.
19. 수행평가 과제 및 채점기준 개발 원리를 충실히 반영하여 평가도구를 개발할 수 있다.
20. 표준화 검사의 개념과 특징 및 종류를 알고, 목적에 맞게 선택하여 사용할 수 있다.
21. 타당도의 개념을 알고, 타당도를 평가할 수 있다.
22. 신뢰도의 개념을 알고, 신뢰도를 평가할 수 있다.
23. 채점자 신뢰도의 개념을 알고, 채점자 신뢰도를 평가할 수 있다.

24. 문항 난이도의 개념을 알고, 문항 난이도를 평가할 수 있다.
25. 문항 변별도의 개념을 알고, 문항 변별도를 평가할 수 있다.
26. 오답지 매력도의 개념을 알고, 오답지 매력도를 평가할 수 있다.
27. 학생에게 평가계획(평가목적, 평가영역, 평가일시, 평가도구의 성격, 평가기준 등)에 대해 사전에 공지하고 있다.
28. 평가실시에 적합한 물리적・심리적 환경을 조성・관리하고 있다.
29. 평가과정에서 부정행위가 초래되지 않도록 학생을 지도・감독하고 있다.
30. 채점기준을 정확하게 숙지하여 채점할 수 있다.
31. 학생의 답안을 점검하여 채점기준을 수정・보완할 수 있다.
32. 학생에게 부여한 점수의 범위를 확인하여 학생의 능력을 제대로 변별하고 있는지 확인할 수 있다.
33. 규준참조평가에서의 검사점수의 의미를 이해하고 점수를 부여할 수 있다.
34. 준거참조평가에서의 검사점수의 의미를 이해하고 점수를 부여할 수 있다.
35. 다양하게 수집된 평가자료의 중요도를 고려하여 총합점수를 산출할 수 있다.
36. 학생들의 점수 분포나 문항 특성에 대한 기초 통계자료(빈도, 평균, 분산, 표준편차 등)를 산출할 수 있다.
37. 관찰이나 면담을 통해 수집된 질적 평가자료를 객관적인 관점에서 이해하고 해석할 수 있다.
38. 평가방법에 따라 동일한 학생에 대한 평가결과가 일치하지 않을 경우 그 원인을 분석하여 해석할 수 있다.
39. 평가의 목적 및 내용에 기초하여 평가결과를 해석할 수 있다.
40. 학생, 평가도구, 평가환경, 평가자 등에 의한 오차를 고려하여 평가결과를 해석할 수 있다.
41. 표준화 검사에서 보고되는 여러 가지 검사 점수(백분위, 표준점수, 등급점수 등)를 정확하고 타당하게 해석할 수 있다.
42. 평가결과에 기초하여 학생의 성취 수준, 강점, 약점 등을 판단할 수 있다.
43. 다양한 측면의 양적・질적 평가결과를 종합하여 학생에 대한 의사결정 자료로 활용할 수 있다.
44. 다양한 측면의 양적・질적 평가결과를 종합하여 교수・학습에 대한 의사결정 자료로 활용할 수 있다.
45. 진단평가결과에 기초하여 학생들의 수준과 요구에 적합한 수업계획을 수립할 수 있다.

46. 형성평가결과에 기초하여 학생들의 학습 곤란 정도와 문제점을 진단하고 해결방안을 제공할 수 있다.
47. 총합평가결과에 기초하여 학생의 성취정도와 변화정도를 파악하고 차기 수업계획을 위한 정보로 활용할 수 있다.
48. 시·도 교육청이나 국가수준의 성취도 평가결과를 소속 학급 또는 학교의 교육내용의 적절성과 수준을 점검하기 위한 자료로 활용할 수 있다.
49. 학생, 학부모, 교육 관련자들에게 평가목적, 평가결과, 해석방법에 대해 설명할 수 있다.
50. 학생의 학업수행 배경과 수준을 고려하여 평가결과를 해석하고 설명할 수 있다.
51. 검사의 규준에 근거하여 학생들의 상대적 위치에 대해 정확하게 해석하고 설명할 수 있다.
52. 검사의 준거에 기초하여 학생들이 무엇을 알고 할 수 있는지를 정확하게 해석하고 설명할 수 있다.
53. 평가결과에 기초하여 학생과 학부모에게 학생의 강점과 약점에 대해 설명하고 개선 방향을 제시할 수 있다.
54. 학급·학교·교육청·국가수준의 평가 관련 지침 및 규정을 준수하고 있다.
55. 학생의 평가결과에 부당한 영향을 미치지 않도록 평가도구의 보안에 주의를 기울이고 있다.
56. 학생의 권리를 보호하기 위해 개인의 평가결과를 엄격하게 보안·유지하고 있다.
57. 평가권한의 오남용이 학생의 정의적 특성에 미치는 영향에 대해 알고 있다.
58. 합리적이고 공정한 절차에 의해 평가를 실시하고 있다.
59. 평가도구에 성별이나 사회경제적 배경에 따라 불리하게 작용할 수 있는 내용이나 방법이 선정되지 않았는지 확인하고 있다.
60. 평가계획, 평가도구의 개발, 시행, 해석·활용·의사소통 등 평가활동의 전 과정을 반성적으로 평가하고 있다.