

프로그램 학습성과 평가에 대한 수요지향 평가자 간 신뢰도 분석 및 개선 방안

Reliability Analysis and Improvement Plan for Evaluation of Program Outcomes among Demand-driven Raters

이영호, 신영학, 김종화
국립목포대학교 컴퓨터공학과

Youngho Lee(youngho@ce.mokpo.ac.kr), Younghak Shin(youngihak@mokpo.ac.kr),
Jonghwa Kim(kimjh@mokpo.ac.kr)

요약

공학교육인증제를 운용하는 공학교육 프로그램에서 프로그램 학습성과는 학생이 졸업하는 시점까지 갖추어야 할 지식, 기술 그리고 태도를 말한다. 일반적으로, 프로그램 학습성과 성취도를 측정하는 평가도구로 캡스톤디자인을 사용한다. 본 논문에서는 프로그램 학습성과 평가에 참여한 수요지향 평가자들의 평가 결과의 신뢰도를 측정하기 위해 급내 상관계수(ICC)를 계산한다. 즉, 같은 평가도구를 이용하여, 여러 명의 평가자가 학습성과를 평가하게 하고, 그 결과를 이용하여 평가자의 ICC를 구한다. ICC는 단체로 수집되거나 단체로 분류된 데이터에 대한 신뢰도의 등급을 측정하는 데 이용된다. ICC가 1에 가까우면 평가자 간 신뢰도가 높다고 판단할 수 있다. 이러한 방법을 지난 2년간의 4년제 대학 컴퓨터공학과 PO 평가 결과에 적용하여 평가도구의 신뢰성을 측정하고 분석하였다. 그리고 이 결과를 이용하여 평가자 간 ICC 측정값을 이용하면, 객관성이 부족한 평가도구를 찾아 보완할 수 있음을 알 수 있었다. 이러한 결과를 바탕으로, PO 평가체계에 대한 개선 방안과 PO 수행 지침을 도출한다.

■ 중심어 : | 공학교육 | 프로그램 학습성과 | 급내상관계수 | 수행준거 | 수요지향 평가자 |

Abstract

In a program that runs an engineering education certification, program outcomes refer to the knowledge, skills, and attitudes a student must have until graduation. In general, capstone design is used as a tool for evaluating program outcomes. This paper applies the intraclass correlation coefficient (ICC) to measure the raters' reliability in assessing program outcomes.

Several raters evaluate program outcomes, and the result is used to obtain the raters' ICC. ICC measures the reliability of ratings or measurements for clusters — data that has been collected as groups or sorted into groups. If the ICC is close to 1, it means that the reliability among the raters is high. We evaluated the proposed method's usefulness through case analysis. As a method for assessing an evaluation tool's objectivity, multiple raters measure the same evaluation tool. As a result, we measured the ICC values for all POs, and analyzed the cause for the low measured POs. We applied this method to evaluate program outcomes of the Department of Computer Engineering in the past two years. As a result, we derived guidelines for improvement and program outcomes.

■ keyword : | Engineering Education | Program Outcomes | Intraclass Correlation Coefficient | Assessment | Rubrics | Demand-driven Evaluator |

I. 서론

공학교육인증제를 운영하는 공학교육 프로그램에서, 프로그램 학습성과는 학생이 졸업하는 시점까지 갖추어야 할 지식, 기술 그리고 태도를 말한다. 프로그램에서는 프로그램 학습성과(Program Outcomes: PO)별 성취도를 측정, 분석하는 일련의 절차와 방법을 수행준거(Performance Criteria, PC), 성취 수준, 평가도구, 채점 기준, 달성목표 등을 포함하여 정량적인 평가가 가능하도록 평가체계를 수립한다(한국공학교육인증원 2020년 적용 컴퓨터·정보공학교육인증기준 KCC2015, KCC2015 판정가이드)[1][2]. 그 평가 결과를 프로그램 개선을 위한 분석자료로 활용하고 있다.

이러한 프로그램 학습성과 평가를 위한 평가진은 프로그램 교수진으로 구성하는 것이 일반적이다. 하지만, 산학 간 인재상의 불일치를 최소화하는 수요 지향적 교육과정의 운영을 위한 산업체 관점의 평가 자료를 얻기 위하여, 고용주를 포함한 관련 산업 분야의 전문가로 구성된 수요지향 평가자와 프로그램 교수진으로 평가진을 구성하기도 한다. 본 논문에서는 프로그램 학습성과 평가에 참여한 고용주 입장의 산업체 평가자를 수요지향 평가자라고 정의한다.

하지만, 프로그램의 교육내용과 PO 평가체계에 대한 이해가 상대적으로 낮은 수요지향 평가자의 관점에서 프로그램 자체적으로 수립한 평가도구와 PO별 성취 수준을 짧은 시간에 이해하고 객관적으로 평가한다는 것은 어려운 일이다. 그렇다고 이러한 애로사항을 개선하기 위하여 평가도구를 정량적인 수치 계산으로 만든다면, 수요지향 평가자를 초대하여 평가하는 의미가 없다. 정량적 방법보다는 평가자의 전문성을 믿고 주관적인 평가를 진행하는 것도 필요하다. 그러므로 단순히 수치를 입력하는 정량적 평가도구 보다 정성적인 평가가 가능하며, 평가자의 정성적 판단을 신뢰할 수 있는 평가체계를 수립하여, 평가진의 구성과 무관하게 평가 결과의 편차를 최소화할 수 있는 PO 성취도 평가체계를 구축할 수 있어야 한다. 이러한 PO 성취도 평가체계 구축을 위하여서는 평가자의 평가 결과를 이용하여 평가자 간의 편차를 개선하는 방법에 관한 연구가 선행되어야 한다.

본 논문에서는 프로그램 학습성과 평가체계의 개선 방안을 도출하기 위해, 프로그램 학습성과 평가에 참여한 수요지향 평가자들의 급내 상관계수(ICC)를 이용하여 수요지향 평가자 간 신뢰도를 측정하고 그 결과를 분석한다. 프로그램 학습성과 평가도구로 캡스톤디자인을 사용한다. 우리는 평가체계의 객관성을 평가하는 방법으로 같은 평가도구를 이용하여, 여러 명의 평가자가 PO를 평가하게 하고, 그 결과를 이용하여 평가자의 ICC를 구한다. ICC가 1에 가까우면 평가자 간 신뢰도가 높다고 판단할 수 있다. 이러한 방법을 지난 2년간의 4년제 대학 컴퓨터공학과 PO 평가 결과에 적용하여 평가도구의 신뢰성을 측정하고 분석하였다.

본 논문의 목적은 다음과 같다. 먼저, 공학교육인증제에서 프로그램 학습성과의 평가에 관한 연구를 살펴본다. 두 번째 캡스톤디자인을 평가도구로 하는 프로그램 학습성과 평가의 수요지향 평가자 신뢰도를 측정하고 분석한다. 지난 2년간의 4년제 대학 컴퓨터공학과 PO 평가 결과에 적용하여 평가도구의 신뢰성을 측정하고 분석하였다. 그리고 이 결과를 이용하여 수요지향 평가자 간 ICC 측정값을 이용하면, 객관성이 부족한 평가도구를 찾아 보완할 수 있음을 알 수 있었다. 이러한 경험을 바탕으로, PO 평가체계에 대한 개선 방안과 PO 수행 지침을 도출한다.

II. 프로그램 학습성과 평가에 관한 관련 연구

1. 공학교육 인증제에서 프로그램 학습성과

이희원, 김성환, 박근, 김정엽(2010)의 연구에서는 서울과학기술대학교의 2010년 캡스톤디자인 교과목에서 학습성과를 측정하면서 평가의 내실 성에 대해 논의하였다[3]. 이 논문에서는 프로그램 학습성과 평가가 얼마나 믿을 수 있는 결과인가에 대한 의문을 제기하면서, 참가하는 학생들의 진정성, 평가하는 교수들의 신중성, 마지막으로 참여 학생의 범위에 대해 고민해야 학습성과 평가의 내실 성을 높일 수 있다고 주장하였다.

박윤국의 연구에서는 홍익대학교의 2010년 졸업자를 대상으로 2009년 1학기와 2학기의 교과목을 수강한 학생들의 학습성과 차이를 비교하였으나 큰 차이를

보이지 않았다고 보고하였다[4]. 이 대학에서는 교과목 학습성과를 이용하여 프로그램 학습성과를 평가하는 방안을 이용하였다. 이 논문의 마지막에는 본 연구가 실질적이면서 덜 수고스러운 방법으로 프로그램 학습성과를 평가하는 방안으로 활용되기를 바란다고 언급하고 있다.

2013년 서울대학교에서 성찰 저널을 활용한 프로그램 학습성과 평가체계를 개발하였다[5]. 이 연구에서는 프로그램 학습성과 평가체계가 중요하지만, 평가자나 교육실행자 모두에게 어려운 점을 지적하였다. 그 이유는 프로그램 학습성과 평가체계가 표준화되고 정형화된 체계가 있는 것이 아니라 각 프로그램의 교육목표와 상황에 적합하게 자체적으로 수립하고 운영해야 하기 때문이다. 평가체계 구축이 어려운 이유는 프로그램 학습성과 의미에 대한 이해 부족과 평가도구 개발에 대한 이론적 고찰 및 평가도구 타당성에 대한 분석이 없기 때문이라고 지적하였다.

고현선, 오준석, 최금진, 박길문(2013)의 연구에서는 조선대학교에서 운영 중인 13개의 프로그램 중 기계공학과 프로그램의 학습성과 평가체계에 대해 정리하였다[6]. 이 논문의 결론을 살펴보면, 프로그램 학습성과에 대한 평가 분석은 평가체계 및 평가도구 개발이라는 관점에서 지속적인 연구가 필요한 분야이며, 타당성과 신뢰도 향상이 필요하므로 대학의 지속적인 연구와 지원이 필요하다고 언급하였다. 즉 한 번에 완벽한 평가체계를 구축하고 변경 없이 수행하기 어렵다는 점을 지적하였다.

김해진, 천의영, 김은경(2020)의 연구는 간호학과 학생을 대상으로 기본 간호학실습 교육에서 의사소통 능력 평가에 사용할 수 있는 평가도구를 제작하고, 적용 결과를 기술한 사례 연구이다[7]. 이 연구에서는 기본 간호학실습 교과목을 수강 중인 학생을 대상으로 의사소통 능력 항목 5가지를 측정하였다. 평가 결과 본 연구에서 도출된 의사소통 능력 평가도구의 타당성과 신뢰성 확보를 위한 추가연구와 다양한 평가도구와의 비교 연구가 필요하다고 언급하였다.

전술한 PO 평가에 관한 연구 결과를 살펴보면, 많은 공학교육 인증제 운영 프로그램에서 PO별 성취도 평가 방법의 타당성 확보를 위하여 큰 노력과 시간을 투자하

여 프로그램의 개선 방안을 도출하였다. 하지만, 학습성과 평가 방법은 각 대학의 교육목표와 상황이 상이하여, 표준화된 절차를 개발하여 모든 대학이나 학과에서 같은 평가 방법을 적용하는 것은 불가능하다. 그리고, 평가가 진행된 후에 그 평가의 신뢰도에 대한 논의가 더 중요하다. 따라서, 각 프로그램의 특색에 맞는 평가체계를 갖추기 위해, 측정 결과의 신뢰도를 분석하는 과정이 필요하다.

2. 평가도구 분석 방법 - 신뢰도 측정

표 1. Koo and Li(2016)가 제안한 ICC 값 해석 방법[8]

ICC	평가	
$.9 \leq ICC$	Excellent	분석에 활용하기 좋은 수치
$.75 \leq ICC < .9$	Good	
$.5 \leq ICC < .75$	Moderate	
$ICC < .50$	Poor	아쉽지만 분석에 활용할 수 없는 수준

신뢰도(reliability)는 구성개념의 측정이 일관적이거나 의존적인 정도를 나타내는 것이다. 다시 말해서, 연구 환경이 변하지 않는 가정하에서 특정한 구성개념을 여러 번에 걸쳐 측정하였을 때, 매번 같은 결과를 도출할 수 있는 정도를 말한다. 관찰자 간의 신뢰도(inter-observer reliability)는 같은 구성개념에 대해 측정하고자 할 때, 두 명 이상의 독립적인 평가자(관찰자) 사이에서 일관성 있는 결과를 도출했는지 평가하는 것이다. 즉, 여러 측정자 사이의 신뢰도를 분석하기 위해 ICC (Intraclass correlation coefficient)를 사용한다[9][10]. 급내상관 계수(ICC)의 값이 [표 1]과 같이 0.9 이상은 Excellent, 0.75 이상은 Good, 0.5~0.75는 Moderate, 0.5 이하는 분석에 사용할 수 없는 수준이다.

III. 프로그램 학습성과 평가자의 신뢰도 분석

1. 컴퓨터공학과 프로그램 학습성과 수행 준거

본 프로그램에서는 공학교육 인증기준 KCC2015의 프로그램 학습성과를 반영하여 10개의 PO를 [표 2]와

같이 수립하였다. 이중 수요지향 평가자가 졸업예정자의 발표를 보고 평가하는 항목인 PO2~PO7은 학생들의 졸업 역량을 측정할 수 있도록 캡스톤디자인 산출물들을 성취도 평가도구로 설정하였다. 산업체와 대학 교육의 인재상 불일치를 최소화하기 위해, 고용주를 포함한 전공 관련 산업 분야의 전문가로 평가자를 구성한다. 또한, 각 PO는 PO별 각 항목의 내용을 기반으로 평가하기 위하여 PO9은 한 개, PO2, PO5, PO6, PO8, PO10은 두 개, PO1, PO3, PO4와 PO7은 3개의 수행 준거가 설정되었다.

수요지향 평가자가 이 수행 준거를 평가할 때 모든 평가자가 신뢰성과 타당성을 유지하며 평가하는 것이 중요하다. 지나치게 주관적인 평가를 하게 되면 평가도구로서의 신뢰도가 떨어져 졸업예정자의 학습성과 평가가 무의미해진다. 따라서 신뢰성과 타당성 있는 평가를 위해 수행 수준을 구체적으로 제시하여 수요지향 평가자가 '상', '중', '하' 중 하나를 선택할 수 있도록 한다. 여기서 '상'은 5점, '중'은 3점, 마지막으로 '하'는 1점으로 정한다.

표 2. KCC2015에 제시된 학습성과와 수행준거

컴퓨터공학심화 프로그램 학습성과		수행 준거(PC)
PO1	수학, 기초과학, 인문 소양 및 컴퓨터공학 지식을 컴퓨팅 분야의 문제 해결에 응용할 수 있는 능력	수학 지식 응용 능력
		기초과학 지식 응용 능력
PO2	이론이나 알고리즘을 수식 또는 프로그래밍 등을 통해 검증할 수 있는 능력	컴퓨터 정보(공)학 지식 응용 능력
		효율적 알고리즘 설계
PO3	컴퓨팅 분야의 문제를 정의하고 모델링할 수 있는 능력	구현 및 테스트
		프로젝트 개발목적과 최종목표
PO4	컴퓨팅 분야의 문제를 해결하기 위해 최신 정보, 연구 결과, 프로그래밍 언어를 포함한 적절한 도구를 활용할 수 있는 능력	프로젝트 문제 정의
		설계모델링
PO5	사용자 요구사항과 현실적 제한조건을 고려하여 하드웨어 또는 소프트웨어 시스템을 설계할 수 있는 능력	최신정보(주제 지식)
		연구결과(자료 지원)
PO6	컴퓨팅 분야의 문제를 해결하는 과정에서 팀의 구성원으로서 팀 성과에 기여할 수 있는 능력	프로그래밍 언어(구성)
		요구사항 분석을 통한 시스템 설계(기어)
PO7	다양한 환경에서 효과적으로 의사소통할 수 있는 능력	현실적 제한조건을 고려한 시스템 설계(자료 지원)
		팀의 구성원으로서 팀 성과에 기여(팀워크)
PO8	컴퓨팅 분야의 해결방안이 안전, 경제, 사회, 환경 등에 미치는 영향을 이해할 수 있는 능력	발표능력(콘텐츠와 창의성)
		프레젠테이션 작성능력(일관성과 조직)
		말하기 능력과 참여(말하기 능력과 참여)
		컴퓨팅 분야의 해결방안을 제시(사례 수)
		안전, 경제, 사회, 환경에 미치는 영향 이해(기대효과에 대한 타당성)

PO9	컴퓨터공학인으로서의 직업윤리와 사회적 책임을 이해할 수 있는 능력	직업윤리와 사회적 책임을 이해할 수 있는 능력(윤리관에 대한 논리성)
P10	기술환경 변화에 따른 자기개발의 필요성을 인식하고 지속적으로 자기주도적으로 학습할 수 있는 능력	기술환경 변화에 따른 자기개발의 필요성
		자기주도적으로 학습할 수 있는 능력

2. 프로그램 학습성과 성취도 측정 일정 및 방법

본 절에서는 4년제 대학 컴퓨터공학과에서의 프로그램 학습성과 평가 일정과 방법, 그리고 신뢰도 측정값을 정리한다. PO 성취도 측정은 2018년과 2019년 11월에 진행되었다. 캡스톤디자인은 4~5명으로 팀을 구성한다. 2018년에는 24명 6개 팀, 2019년에는 31명 7개 팀이 참여하였다. 평가자는 팀당 약 20분 정도의 발표와 질의응답 시간을 통하여 PO2~PO7을 평가한다. 평가장에는 학과 교수진, 수요지향 평가자와 관심 있는 재학생이 참관할 수 있다. 누구나 질문할 수 있으나 대체로 수요지향 평가자가 질문의 우선권이 있다.

표 3. 2018년과 2019년의 평가자의 전문성에 관한 정보

평가자	회사 분야	담당업무	캡스톤디자인 멘토	
'18	A	시스템통합	시스템설계	O
	B	시스템통합	시스템설계	X
	C	시스템통합	기획	O
'19	A	스마트공장	시스템설계	O
	B	가상현실/콘텐츠	소프트웨어개발	O
	C	시스템통합	기획	O
	D	증강현실	기획/소프트웨어개발	O

평가에 참여한 수요지향 평가자의 수는 2018년에 3명, 2019년에 4명이다. 2018년에는 학생의 캡스톤디자인 주제가 시스템통합, 즉 회사, 관공서, 병원 등의 인터넷이나 서비스를 개발하는 주제를 주로 다루었고, 2019년에는 스마트공장과 가상현실도 추가되었다. 평가자 대부분은 신규인력을 고용할 수 있는 권한을 가진 위치에 있는 대표이며, 현장 업무에 대한 지식도 갖추고 있다. [표 3]은 평가자의 회사 분야, 담당업무, 캡스톤디자인 멘토 활동 여부에 대해 정리한 표이다.

일반적으로 평가자가 평가대상과 분리하는 것이 원칙이나, 프로그램 학습성과 성취도 측정의 경우 학생 개인을 평가하는 것이 아니라 심화 프로그램을 평가하는 것이고, 심화 프로그램의 교육내용에 대해 깊이 있는 평가를 하기 위해서는 학과의 교육과정과 졸업생의 수준을 잘 이해하는 평가자가 도움이 될 것이다.

학생들은 평가 당일 지난 1년간 준비한 캡스톤디자인의 최종보고서, 제안서, 회의록 등 각종 자료와 발표 자료를 평가자에게 제공한다. 수요지향 평가자는 프로그램 학습성과, 수행 준거, 그리고 수행 수준(루브리)이 표시된 평가지에 학생들의 발표와 제공된 최종보고서 등의 자료를 보고 평가한다.

3. 평가자 간 급내 상관계수 측정

신뢰도 측정값을 분석하기 위해 SPSS를 사용하였다. SPSS를 이용하여 수요지향 평가자의 평가 결과에 대해 각 학습성과 수행 준거의 ICC를 구하여 표로 작성하였다. 그 결과, [표 4]에는 3명의 평가자, [표 5]에는 4명의 평가자의 ICC가 정리되었다.

본 평가에서는 비슷한 성향의 모집단에서 수요지향 평가자를 선택하였기 때문에 two-way random effects 모델을 적용하였다. 또한, 모든 측정 단위가 같아서 single 타입을 선택, 평가자의 평가 일치도를 보기 위해 absolute agreement를 옵션으로 선택하였다.

표 4. 2018년 평가자 3명의 학습성과에 따른 급내상관계수 (ICC)

PO	수행 준거	ICC	95%CI.L B	95%CI.U B	p-value
2	알고리즘 설계	0.270	-0.338	0.646	0.159
	구현 및 테스트	0.559	0.163	0.790	0.006
3	프로젝트계획서	0.674	0.363	0.847	0.000
	문제정의	-0.084	-0.682	0.407	0.605
	설계모델링	0.526	0.047	0.782	0.018
4	최신정보	0.229	-0.102	0.545	0.034
	연구결과	0.299	-0.093	0.617	0.024
	프로그램밍언어	0.400	-0.055	0.703	0.001
5	요구사항 분석을 통한 시스템설계	0.276	-0.081	0.597	0.002
	현실적 제한조건을 고려한 시스템설계	0.236	-0.122	0.561	0.059
6	팀워크1	0.438	-0.042	0.733	0.000
	팀워크2	0.391	-0.060	0.697	0.001
7	콘텐츠와 창의성	0.612	0.258	0.817	0.002
	일관성과조직	0.371	-0.059	0.676	0.024
	말하기 능력과 참여	0.345	-0.114	0.666	0.056

표 5. 2019년 평가자 4명의 학습성과에 따른 급내상관계수 (ICC)

PO	수행 준거	ICC	95%CI.L B	95%CI.U B	p-value
2	알고리즘 설계	0.482	-0.155	0.714	0.002
	구현 및 테스트	0.659	0.388	0.822	0.000
3	프로젝트계획서	0.507	0.188	0.730	0.001
	문제정의	0.865	0.768	0.929	0.000
	설계모델링	0.783	0.551	0.896	0.000

4	최신정보	0.639	0.340	0.814	0.000
	연구결과	0.471	0.101	0.717	0.009
	프로그램밍언어	0.637	0.376	0.807	0.000
5	요구사항 분석을 통한 시스템설계	0.738	0.458	0.874	0.000
	현실적 제한조건을 고려한 시스템설계	0.471	0.101	.0717	0.009
6	팀워크1	0.712	0.434	0.857	0.000
	팀워크2	0.712	0.434	0.857	0.000
7	콘텐츠와 창의성	0.470	0.117	0.710	0.000
	일관성과조직	0.588	0.290	0.780	0.000
	말하기 능력과 참여	0.546	0.245	0.753	0.000

IV. 분석 및 개선사항 도출

1. 평가도구와 평가자에 대한 분석

1.1 2018년 프로그램 학습성과에 대한 분석

2018년 PO의 급내 상관계수(ICC) 계수를 살펴보면, PO3의 '프로젝트계획서'와 PO7의 '콘텐츠 창의성' 각각 0.674, 0.612로 신뢰도가 높은 문항임을 알 수 있다. 급내 상관계수(ICC)가 0.5~0.59 사이인 수행 준거는 PO2 '구현 및 테스트', PO3 '설계모델링' 두 개이다. 나머지는 계수의 값이 낮아 의미가 없거나 분석할 수 없는 상태이다.

아쉽게도, 2018년 평가 결과는 대체로 급내 상관계수(ICC)가 낮아 수행 준거(설명 불충분 포함)가 문제인지 평가자가 문제인지 알기 어렵다. 이런 이유로 2018년 평가 결과는 신뢰도가 낮아 개선 방안 도출에 이용하는 것은 무리가 있다. 평가 결과의 모호한 점을 개선하기 위해 평가자에게 충분한 설명과 자료를 제공하여 심층 평가가 이루어지게 할 필요가 있다.

1.2 2019년 프로그램 학습성과에 대한 분석

2019년 프로그램 학습성과 평가 결과의 급내 상관계수를 살펴보면, '문제정의', '설계모델링', '시스템설계', '팀워크'에서 0.7 이상의 신뢰도로 평가되었다. 특히, PO3의 '문제정의'는 ICC가 0.865가 나와 평가자 간의 신뢰도가 높게 측정되었다. 2018년과 다른 점은 평가 1주일 전에 제안서, 최종보고서와 평가 기준 등의 자료를 평가자에게 배포하여 미리 자료를 검토하고 평가 기준을 확인할 수 있는 시간이 있었다는 점이다. 또한, 2018년에는 3명의 평가자를 선정하였지만, 2019년에

는 평가의 객관성을 높이기 위해 4명의 평가자를 선정하였다.

2019년의 학습성과 평가에서 PO2의 '알고리즘 설계', PO4의 '연구 결과', PO5의 '현실적 제한조건을 고려한 시스템 설계', 그리고 PO5의 '콘텐츠와 창의성'은 ICC가 0.4로 낮게 측정되어 수행 준거(루브릭 포함)가 객관적인지 검토할 필요가 있다.

PO2의 '알고리즘 설계'의 경우 평가자들이 최종보고서와 발표를 보고 문제를 해결하기 위해 효율적인 알고리즘을 적용하였는지 평가하는 항목이다. 이 수행 준거의 경우 객관적인 평가가 어려운 주관적인 평가를 해야 하므로 신뢰성이 낮은 것으로 보인다. 수행 수준에서 알고리즘의 정확성과 효율성을 퍼센트(%)로 평가하게 되어 있는데, 알고리즘을 객관적으로 평가하기에 어려운 수행 수준이므로 개선할 필요가 있다.

PO5의 '현실적 제한조건을 고려한 시스템 설계'에 대한 평가 결과의 경우 0.471로 낮은 ICC를 보인다. PO5의 경우 요구사항 분석과 현실적 제한조건을 최종보고서와 발표에서 다루고 있는지를 평가하는 것으로 수행 준거에는 문제가 없어 보이나, 평가자들의 이해를 도울 수 있도록 수행 준거의 수정 보완이 필요하다.

제일 낮은 수치를 보인 수행 준거는 0.470인 '발표 능력(콘텐츠와 창의성)'이다. 이 수행 준거는 학생들의 발표 자료의 논리적 흐름에 대한 평가인데, [표 6]을 보면 평가 기준을 설명하는 문장의 길이가 길고, '주요 주장과 관련된 풍부한 자료 포함 여부', '합리적 주장', '창의적 방법' 등 모호하고 주관적으로 판단할 수밖에 없는 단어로 설명되어 있다. 따라서 이 수행 준거의 경우 평가자가 객관적으로 평가하기 어려워져 발생한 것이라 판단된다.

표 6. 2019년 측정된 평가자 급내 상관계수(ICC) 중 낮은 ICC 가 측정된 학습성과 수행 준거와 수행 수준

수행 준거	수행수준(루브릭)		
	상	중	하
효율적 알고리즘 설계	알고리즘의 논리구조의 정확성 및 효율성이 90% 이상이다.	알고리즘의 논리구조의 정확성 및 효율성이 50-89% 정도이다.	알고리즘의 논리구조의 정확성 및 효율성이 50% 미만이다.
연구결과(자료지원)	모든 관련 정보를 수집했고 정보의 출처는 유효했다. 정보를 이용하여 분석과 설계 고려사항(현실적 제한조건)	충분한 정보를 수집했고 대부분 정보출처가 유효했다. 정보를 이용하여 분석과 설계 고려사항을 대부분 지원했다.	정보 수집이 부족하고 출처는 유효성에 결함이 있다. 수집된 정보를 이용하여 분석과 설계 고려사항

	설계 구성요소)을 지원했다.	원했다.	을 지원하지 않았다.
현실적 제한조건을 고려한 시스템 설계(자료지원)	모든 관련 정보를 수집했고 정보의 출처는 유효했다. 정보를 이용하여 분석과 설계 고려사항(현실적 제한조건, 설계구성요소)을 지원했다.	충분한 정보를 수집했고 대부분 정보출처가 유효했다. 정보를 이용하여 분석과 설계 고려사항을 대부분 지원했다.	정보 수집이 부족하고 출처는 유효성에 결함이 있다. 수집된 정보를 이용하여 분석과 설계 고려사항을 지원하지 않았다.
발표능력(콘텐츠와 창의성)	발표는 주요한 주장과 명백히 관련된 풍부한 자료를 포함하고 있다. 외부 연구를 이용하여 합리적인 주장과 해결책을 제시했다. 발표 자료는 원본이고 청중의 주목을 유지하기 위해 창의적인 방법을 사용하였다.	발표는 주요한 주장을 뒷받침하는 자료를 포함하고 있으나 1)모든 자료가 주요 주장과 명백히 관련되지 않고 2)주장 또는 해결책을 정당화하기 위해 사용한 외부연구가 제한적이고 3)발표 자료는 적절하나 청중의 주목이 다소 낮다.	청중은 주요한 논리와 아이디어의 흐름을 이해하기 위해 상당한 노력을 기울여야 했다. 분석과 권고의 주요한 측면이 없다. 주장과 해결책을 정당화하기 위해 외부 연구를 사용하지 않았다. 발표는 창의성이 부족하고 청중의 주목을 이끌지 못한다.

2. 개선사항

프로그램 학습성과 성취도 평가에서 평가자 간의 신뢰성을 확보하는 방안으로는 PO 평가체계에 대한 객관성과 일관성 유지를 위한 수행 준거와 수행 수준의 개정, 평가자의 눈높이 조정 및 프로그램의 교육내용과 PO 평가체계에 대한 이해도 향상 등이 있다.

이들 중 가장 우선으로 고려되어야 하는 체계적인 측면인 PO 평가체계의 수행 준거와 수행 수준에 대하여 4.1절의 분석자료를 활용하여 분석하면 다음과 같다. 2019년 급내 상관계수(ICC)가 0.6보다 낮은 PO2의 '알고리즘 설계', PO3의 '프로젝트계획서', PO 4의 '연구 결과', PO 5의 '현실적 제한조건을 고려한 시스템 설계', 그리고 PO5의 '콘텐츠와 창의성', PO7의 '콘텐츠와 창의성', '일관성과 조직성', '말하기 능력과 참여'의 수행 준거나 수행 수준을 개선할 필요가 있다. [표 6]은 급내 상관계수(ICC)가 0.5보다 낮은 프로그램 학습성과 수행 준거를 정리한 표이다.

[표 6]의 수행 수준을 살펴보면, 대체로 평가자에 따라 의미 해석의 일관성을 유지할 수 없어 개선이 필요하다. '알고리즘 설계'의 경우 프로젝트에서 사용된 알고리즘의 정확성과 효율성을 수치로 평가해야 하는 어려움이 있다. 대부분의 캡스톤디자인이 시스템 제작을 결과물로 제출하기 때문에 개별적인 알고리즘을 살펴보기 어려운 점도 평가를 어렵게 하는 요소이다. 따라서 '알고리즘 설계'를 평가하기 위해서는 학생들의 보고서에 명시적으로 이 내용을 추가하게 하고, 발표 시에

도 언급하게 해야 하며, 수행 수준을 평가할 수 있게 수정할 필요가 있다.

‘현실적 제한조건을 고려한 시스템 설계’의 경우, 이 수행 준거 앞에 ‘요구사항분석을 통한 시스템 설계’의 신뢰도가 높은 데 반해 낮은 ICC를 나타내고 있다. 그 이유는 학생들의 발표와 보고서에서 요구사항 분석과 현실적 제한조건을 분석한 내용이 겹쳐있어 구별하기 어렵기 때문이라 여겨진다. 이 수행 준거는 학생들의 보고서에서 명확히 분리하여 작성하는 것을 권고할 필요가 있다.

발표와 관련 있는 ‘발표 능력’, ‘프레젠테이션 작성 능력’, ‘말하기 능력과 참여’는 객관적으로 분석하거나 보고서 내용으로 판단할 수 없다. [표 6]의 해당 수행 준거를 살펴보면, 설명이 길고 내용이 복잡해서 단순하게 ‘상’, ‘중’, ‘하’로 평가하기 어렵다. 이 3가지 수행 준거를 객관화할 수는 없지만, 이 3가지 수행 준거가 명확히 구별되지 않아 서로 독립적인 평가 기준이 될 수 있도록 수정할 필요가 있다. 이상의 개선사항을 고려하여 프로그램 규정에서 정하고 있는 일정(3년)에 따라 PO 평가체계의 수행 준거와 수행 수준에 대한 일관성과 객관성을 확보할 수 있도록 개선하고자 한다.

2018년 대비 2019년 ICC 분석 결과를 살펴보면, ‘콘텐츠와 창의성’에 대해서는 2018년 0.612에서 2019년 0.470으로 소폭 낮아졌다. 2018년과 2019년의 심사위원이 서로 다르므로 그 원인을 정확히 추측하는 것은 어려운 일이다. 원인을 추론해 보면, 2018년 평가자는 시스템통합(SI) 관련 기업의 3명의 평가자였고, 2019년의 심사위원 4명 중 2명은 가상현실/증강현실 관련 콘텐츠를 개발하는 기업의 평가자였다. SI 분야보다 가상현실/증강현실 분야는 발표 자료에 그래픽스와 동영상 삽입되는 등 상대적으로 화려한 발표가 요구되기 때문이 아닌가 추측된다. 즉, SI 분야와 가상/증강현실 분야의 두 그룹의 평가자 간에 평가 기준의 차이가 있어 결과가 일치하지 않는 것이다. 하지만 현재의 분석자료로는 추측을 검증할 방법이 없어 추후 연구로 수행해야 한다.

V. 토론

1. 수행 준거의 객관성 유지 노력

PO의 수행 준거를 객관적으로 평가할 내용과 주관적으로 평가하여 수요지향 평가자의 전문적 의견을 반영해야 할 두 부류로 분류하여 신뢰성을 높이는 방향으로 수행 준거를 제시해야 한다. 양적으로 측정 가능한 수행 준거라면, 정확한 숫자 범위를 제시하여 평가자가 일관성 있게 평가할 수 있게 해야 한다. 수요지향 평가자가 자신의 전문적 역량을 발휘하여 평가해야 하는 수행 준거라면, 그 분야에 맞는 적절한 용어를 사용하여 수행 준거를 제시해야 한다. 예를 들어 [표 6]의 수행 준거 중 ‘알고리즘의 논리구조의 정확성 및 효율성이 90% 이상이다.’와 같은 경우 정확성과 효율성을 90%의 수치로 표현하는 방법을 제시하지 않아 평가자가 평가할 수 없는 수행 준거를 제시하였다. 이 수행 준거는 논리적으로 문제가 있어 수요지향 평가자가 제대로 된 평가를 할 수 없으므로 수정이 필요하다.

2. 전문성을 갖춘 수요지향 평가위원의 중요성

PO 평가를 위한 수요지향 평가자 선정에 주의를 기울여야 한다. 수요지향 평가자를 선정할 때는 친분이나 초청의 편의성보다, 평가에 적합한 평가자의 기준을 마련하는 것이 좋다. 수요지향 평가자의 회사업종, 평가자의 담당업무, 캡스톤디자인 멘토 활동 여부, 해당 회사에 최근 2~3년간 졸업생 취업 여부도 기준이 될 수 있다. 예를 들면, 컴퓨터공학전공 관련된 IT 기업 중 과제나 직원 평가 경험이 있는 CEO 혹은 CTO가 적합하다.

수요지향 평가자가 프로젝트를 평가할 때 평가자 스스로 전문성의 단계를 표시하여 가중치로 사용하는 방법도 시도해 볼 수 있다. 컴퓨터공학 관련 학술대회에서 논문을 평가할 때, 평가자 스스로 자신의 전문성을 1~4중 선택하고 최종 점수에 가중치로 사용된다. 이렇게 하는 이유는 평가자가 모든 컴퓨터공학 분야의 논문에 전문성을 갖추기 어렵기 때문이다.

3. 프로그램 학습성과 평가 지침도출

본 장에서는 앞서 이루어진 평가 결과를 분석하는 과정에서 올바른 PO 성취도 평가를 수행하기 위한 평가 지침을 도출하였다. 수요지향 평가자가 참여하는 PO 평가를 진행하기 위해서는 고려해야 할 사항들을 지침

으로 정리하였다.

GL1. 각 PO 평가에 적합한 수요지향 평가자의 수준과 자격에 대한 기준을 마련해야 한다. 예를 들면, 수요지향 평가자가 소속된 기업 분야, 수요지향 평가자의 전문분야, 학생의 프로젝트와의 관련성 등을 기준으로 삼아 초청할 수 있다.

GL2. 수요지향 평가자들에게 미리 심사에 필요한 제안서, 최종보고서, 동영상, 평가표 등의 자료를 배포하여 평가 당일 신속하고, 정확하게 평가할 수 있도록 한다. 평가의 공정성을 위해 학생들이 작성한 프로젝트 보고서와 각종 자료가 일관성이 있게 작성되어야 한다.

GL3. 수요지향 평가자가 학생들의 캡스톤디자인 과제 내용을 파악할 수 있는지 전문성에 대한 자가 평가를 반영할 수 있다. 수요지향 평가자가 A 분야의 전문가라 하더라도 B 분야와 관련된 프로젝트를 평가하기는 쉬운 일이 아니다. 하지만 각 프로젝트에 맞춰서 평가자를 초청하는 것은 어려운 일이므로, 평가자가 자신의 전문성을 표시하도록 하는 것도 한 방법이다.

GL4. 학생들이 최종보고서와 발표 자료를 제작할 때, 수행 준거에 따라 제작하게 하면, 평가자가 평가하기 쉬워진다. 학생들이 보고서를 작성할 때 수행 준거에 맞춰 목차를 정해준다면 평가자가 평가를 수행하기 수월해질 것이다.

GL5. PO 평가 결과를 학생들에게 공개하여 자기 주도적으로 개선할 수 있도록 자료를 제공해야 한다. PO 평가가 종료된 후, 성적에 반영되는 평가가 아니므로 학생들의 관심은 멀어진다. 하지만, 이왕 수집된 전문가의 평가의견을 학생들에게 전달하여 실력향상에 도움이 되도록 할 수 있다.

GL6. 평가 결과를 분석하여 프로그램 위원회에 제출하여야 한다. PO 평가는 프로그램을 개선하기 위한 도구일 뿐 절대적인 것은 아니며, 교수자가 이 결과를 이용하여 분석하고 토론하여 개선사항을 도출해야 한다. 위의 평가지침에 열거된 것뿐 아니라 각 프로그램에서 중요하게 고려되어야 할 사항을 발견할 수도 있다.

VI. 결론 및 추후 연구

공학교육인증제에 참여하고 있는 각 대학의 프로그램에서 프로그램 학습성과를 평가하고 분석하는 것은 중요한 일이다. 하지만, 각 대학의 교육목표와 상황이 달라 표준화된 학습성과 평가 절차를 개발하는 것은 불가능하다. 따라서, 학습성과 평가가 제대로 이루어졌는지 신뢰도 분석 방법을 이용하여 평가하는 시도가 필요하다.

본 논문에서는 PO 성취도 평가 시 평가자와 수행 준거의 신뢰도를 분석하기 위한 통계 도구로 급내 상관계수(ICC)를 이용하였다. 지난 2년간의 PO 평가 결과에 급내 상관계수(ICC)를 측정함으로써 평가도구의 신뢰성을 분석하였고 그에 따른 개선 방안과 PO 평가지침을 도출하였다. PO는 프로그램 교육목표에 기반한 인재상을 구현하기 위한 구체적인 실행방안으로, 이에 대한 평가는 프로그램에서 정한 졸업예정자가 갖추어야 할 졸업 역량에 대한 성취도 평가를 통한 프로그램 교육과정의 자율순환적인 개선을 위한 중요한 측정 도구이다. 따라서 PO를 측정하고 분석하는 것은 중요한 일이다.

도출된 지침은 학습성과 평가를 수행할 때, 고려해볼 사항이며, 대학마다 조건과 상황이 달라서 일률적으로 적용할 수는 없다. 또한 도출된 지침은 학습성과 평가 전반적인 것에 관한 내용을 다루고 있지 않고 평가자에 관한 내용을 중점적으로 다루고 있다. 많은 대학에서 이 지침을 참고하여 프로그램 학습성과 평가를 수행하고, 그 결과를 논문으로 학회에 보고하여 개선 보완되길 바란다.

본 연구를 수행하면서, ICC 분석 결과를 통해 매년 이루어지는 프로그램 학습성과 성취도 측정을 연차별로 비교하기 어렵다는 한계점을 발견하였다. 그 이유는 평가도구를 제외한 수요지향 평가자, 평가에 참여하는 학생, 학생들의 프로젝트 주제를 일관성 있게 일치시키기 어렵기 때문이다. 그러한 이유로, 2018년과 2019년의 ICC의 변화에 따른 근본적인 이유를 분석해 내기 어려웠다.

PO 평가는 학생의 성취도를 측정하는 방안으로써 객관적 수치로만 나타내는 것은 부적절하다. 즉, 이러한 평가 결과와 통계치는 참고자료로 이용되어야 하며 모든 판단과 결정은 교수자를 포함하여 프로그램에서 직

접 하는 것이 중요하다. 그리고 이러한 수요 지향적인 PO 평가를 하는 가장 중요한 이유는 자율순환적 개선을 위한 것임을 잊지 말고 시행해야 한다. 다음에는 코로나19로 인해 온라인 PO 평가가 수행될 때 기존의 PO 평가와 차이가 없이 진행될 수 있는 PO 평가 방안에 관해 연구할 계획이다.

참 고 문 헌

[1] 한국공학교육인증원 2020년 적용 컴퓨터·정보공학교육인증기준 KCC2015, <http://www.abEEK.or.kr/intro/standard>, 2020.10.4.

[2] 한국공학교육인증원 2020년 적용 컴퓨터·정보공학교육인증기준 KCC2015 판정가이드, <http://www.abEEK.or.kr/intro/standard>, 2020.10.4.

[3] 이희원, 김성환, 박근, 김정엽, “캡스톤디자인 교과목 기반의 프로그램 학습성과 평가 연구,” 공학교육연구, 제13권, 제6호, pp.143-151, 2010.

[4] 박윤국, “공학교육인증에서 교과목 학습성과와 프로그램 학습성과의 관계 - 실질적인 프로그램 습성과 평가를 지향하며,” 공학교육연구, 제13권, 제4호, pp.87-92, 2010.

[5] 이영태, 임철일, “성찰저널을 활용한 프로그램 학습성과 평가체계 개발,” 공학교육연구, 제16권, 제3호, pp.42-50, 2013.

[6] 고현선, 오준석, 최금진, 박길문, “조선대학교 기계공학프로그램 학습성과 평가 사례 연구,” 공학교육연구, 제16권, 제6호, pp.65-77, 2013.

[7] 김해진, 천의영, 김은경, “기본간호학실습에서 학습성과 평가도구의 적용 사례: 의사소통능력 평가를 중심으로,” 한국산학기술학회 논문지, 제21권, 제4호, pp.173-180, 2020.

[8] T. K. Koo and M. Y. Li, "A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research," Journal of Chiropractic Medicine, Vol.15, No.2, pp.155-163, 2016. doi:10.1016/j.jcm.2016.02.012

[9] J. J. Bartko, "The intraclass correlation coefficient as a measure of reliability," Psychol Rep, Vol.19, No.1, pp.3-11, 1966. doi:10.2466/pr0.1966.19.1.3

[10] Jin-hong Kim and Bo-ram Choi, "Intra- and Inter-rater Reliability of BioRescue," 한국콘텐츠학회논문지, 제18권, 제11호, pp.348-352, 2018.

저 자 소 개

이 영 호(Youngho Lee)

정회원



- 1999년 : KAIST 수학과 졸업
- 2001년 : GIST 정보통신공학과 석사
- 2008년 : GIST 정보통신공학과 박사
- 2009년 ~ 현재 : 목포대학교 교수

〈관심분야〉 : 공학교육인증, HCI

신 영 학(Younghak Shin)

정회원



- 2009년 : 광운대학교 전자통신공학과 졸업
- 2011년 : GIST 정보통신공학과 석사
- 2016년 : GIST 정보통신공학과 박사
- 2016년 ~ 2018년 : 노르웨이과학기술대학교(NTNU) 박사후연구원
- 2018년 ~ 2020년 : LGCNS 책임

연구원

- 2020년 ~ 현재 : 목포대학교 조교수

〈관심분야〉 : 인공지능, 신호처리, 공학교육인증

김 중 화(Jonghwa Kim)

정회원



- 1991년 ~ 현재 : 목포대학교 교수
- 2004년 ~ 현재 : 한국공학교육인증 평가단장/평가위원
- 2016년 ~ 2019년 : 한국공학교육인증원 부원장

〈관심분야〉 : 공학교육, 교육인증, 공학설계교육, 센서공학, Digital System Design, IoT System, Embedded System, EMS