Edith J. CISNEROS-COHERNOUR

Yucatan autonominis universitetas • Universidad Autónoma de Yucatán

DĖSTYMO AUKŠTOJOJE MOKYKLOJE PAGRĮSTUMAS IR ĮVERTINIMAI PAGAL POZITYVISTINĘ PARADIGMĄ

VALIDITY AND EVALUATIONS OF TEACHING IN HIGHER EDUCATION INSTITUTIONS UNDER POSITIVISTIC PARADIGM

SANTRAUKA

Straipsnyje nagrinėjamas tyrimų, atliktų pagal vyraujančią dėstymo aukštojoje mokykloje vertinimo paradigmą, pagrįstumas (validumas). Siekiant nustatyti tyrimų, kuriuose analizuojami studentų sudaryti dėstymo reitingai, privalumus ir trūkumus, taikoma Messick pagrįstumo sistema. Taip pat keliamos problemos, į kurias reikėtų atkreipti dėmesį atliekant tokius tyrimus ateityje.

PAGRINDINIŲ SĄVOKŲ APIBRĖŽIMAI

• *Pagrįstumas (tinkamumas)* – siejasi su interpretacijos prasmingumu, verte ir tinkamumu. Jis parodo, kokiu mastu empiriniai duomenys ir teorija atspindi vertinimo arba įvertinimo adekvatumą ir tinkamumą (Messick, 1995, p. 741).

• *Fakulteto darbuotojų įvertinimas* – aukštųjų mokyklų (kolegijų ir universitetų) administracinio, mokomojo ar kito akademinio personalo kompetencijos įvertinimas, remiantis apibrėžtais kriterijais.

• *Dėstymas kolegijose* – procesas, kurio metu sąmoningai perteikiamos žinios, požiūriai ar įgūdžiai. Sąvoka apima visą aukštojo mokslo institucijų studijų procesą nuo planavimo iki įgyvendinimo atsižvelgiant ir į grįžtamąjį ryšį.

• *Aukštasis mokslas* – bet koks aukštesnis už vidurinį (12 klasių) išsilavinimas, kurį baigus įgyjamas kvalifikacinis laipsnis.

• *Įvertinimas* – kieno nors privalumų, gerųjų savybių ar vertingumo nustatymas arba šio proceso rezultatas.

• *Vertinimas* – dažnai vartojamas kaip sinonimas įvertinimo sąvokai apibūdinti, bet kai kuriais atvejais gali apibūdinti procesą, kuriame labiau akcentuojami kiekybiniai ir/ar tyrimo tikslai.

• *Dėstymas* – instruktavimo sinonimas.

• *Instruktavimas* – tikslingas žinių, požiūrių ar įgūdžių perteikimo procesas; apima visą instruktavimo procesą pradedant planavimu ir įgyvendinimu bei įvertinimu ir baigiant grįžtamuoju ryšiu.

• *Patikimumas (tikslumas)* – tam tikrų matavimų ar bandymų rezultatų pastovumas arba stabilumas. Kai pakartotiniai to paties reiškinio matavimai duoda tuos pačius arba panašius rezultatus, teigiama, kad matavimo priemonė yra patikima.

• *Įvertinimo instrumentas* – priemonė, skirta nustatyti kieno nors privalumus, jo gerąsias savybes arba vertingumą.

• *Vertinimo instrumentas* – apibūdina priemonę, naudojama įvertinimui.

• *Apibendrinamumas* – išvadų, kurias galima padaryti apie populiaciją remiantis turima informacija apie bandinį, apimties mastas.

ABSTRACT

This paper focuses on the validity of the research conducted under the leading paradigm in the evaluation of teaching in higher education. Messick's framework on validity is used to identify the strengths and limitations of the research, mostly centered on the study of student ratings of instruction. Critical issues that need to be addressed by future studies in this area are also identified.

DEFINITIONS OF KEY WORDS

• *Validity* – linked to the meaning, value and the appropriateness of interpretation. It is the overall judgment of the extent of which empirical evidence and theory support the adequacy and appropriateness of the interpretations based on the assessment or evaluation (Messick, 1995, p. 741).

• *Faculty evaluation* – judging the value or competence of administrative, instructional, or other academic staff in higher schools (colleges, universities) based on established criteria.

• *College teaching* – refers to the process by which knowledge, attitudes, or skills are deliberately conveyed – includes the total instructional process, from planning and implementation through evaluation and feedback that takes place in higher education institutions.

• *Higher education* - all education beyond the secondary level (12th grade), leading to a formal degree.

• *Evaluation* – the process of determining the merit, worth or value of something; or the product of that process (Scriven, 1991).

• *Assessment* – offer used as a synonym for evaluation, but sometimes used to refer to a process that is more focussed on quantitative and /or testing approaches. (Scriven, 1991).

• *Teaching* – synonym for instruction.

• *Instruction* – process by which knowledge, attitudes, or skills are deliberately conveyed — includes the total instructional process, from planning and implementation through evaluation and feedback.

• *Reliability* – the consistency or stability of a measure or test from one use to the next. When repeated measures of the same thing give identical or very similar results, the measurement instrument is said to be reliable. (Vogt, 1993)

• *Evaluation instrument* – intrument used for determining the merit, worth or value of something.

• *Assessment instrument* – it refers to the instrument used for evaluation.

• *Generalizibility* – the extent to which we can come to conclusions about population based on information about a sample.

ĮVADAS

Tradiciniam pozityvistiniam dėstymo aukštoje mokykloje įvertinimo požiūriui būdingas praktika pagrįstų matavimų objektyvumo pabrėžimas. Tyrėjas arba įvertintojas yra „objektyvus“ informacijos rinkėjas, kuris remiasi kiekybinės analizės metodais. Erickson (1986) teigia, kad vyraujanti dėstymo tyrimų paradigma yra kilusi iš tradicinio gamtos mokslų modelio:

„Pastaruosius 20 metų pozityvistiniai tyrimai daugiau dėmesio skyrė analitiniams procesams, o ne teorinių modelių tobulinimui. Manyta, kad bendrumai tarp skirtingų auditorijų išryškės atliekant tyrimus, o nežymius skirtumus galima atmesti kaip nereikšmingus.“ (1986, p. 131).

Šiam modeliui pritariantys mokslininkai dėstymą sieja su elgsena, o įvertinimą su efektyvumu. Taigi, dėstymo efektyvumas yra „nustatomas pagal galutinius įvertinimus, standartizuotus pasiekimų testus ir konkrečią dėstymo praktiką“ (Erickson 1986, p. 131).

Pozityvizmo modelio pavyzdys yra proceso kaip produkto tyrimai, kurių metu pabrėžiamas „tiesioginis“ dėstymas, siekiamų žinių ir elgsenos pateikimas ar atkartojimas. Tokiuose tyrimuose dėstymo efektyvumu laikomi „atskiri stebimi paties dėstymo proceso deriniai, veiksmingi nepriklausomai nuo laiko ar vietos“ (Shulman 1986, p. 10).

Kaip teigia Dunkin ir Barnes (1986), proceso kaip produkto tyrimai, atlikti septintame dešimtmetyje ir aštuntojo dešimtmečio pradžioje, yra dabartinio dėstymo aukštosiose mokyklose pagrindas. Tačiau, skirtingai nei daugumа tokio pobūdžio tyrimų, atliktų kitose švietimo pakopose, aukštojo mokslo lygmenyje „proceso analizė buvo apibrėžiama preskriptyviai arba vertinama nepakankamai parengtų stebėtojų, tačiau nebuvo grindžiama atidžiu stebėjimu “ (Dunkin, Barnes, 1986, p. 774).

Kiti tyrinėtojai pažymi, kad formali gero dėstymo samprata aukštojoje mokykloje atsirado ne iš proceso kaip produkto tyrimų, o iš gerą dėstymą apibrėžiančių charakteristikų ir savybių sąrašo. Tokie sąrašai sudaryti remiantis dėstytojų ir studentų apklausomis, kurių metu respondentai turėjo apibūdinti, kas apima „gero dėstymo“ sampratą. Feldman (1988), Frey (1979) ir Marsh (1997) pritaria tokiam charakteristikų ar elgesio apibūdinimo panaudojimui kuriant dėstymo kokybės nustatymo metodiką. Jie teigia, kad apklausos gali garantuoti dėstytojams grįžtamąjį ryšį, galimybę išsiaiškinti, kaip vertinamas jų dėstymas, nustatyti poreikius ir trūkumus.

Dar viena mokslininkų grupė teigia, kad dėstymas turėtų būti vertinamas ne pagal tam tikras jo savybes ar dimensijas, o „globaliai“. Cashin ir Downey (1992), Cohen (1986) ir Abrami (1990; 1993) pabrėžia globalius elementus arba „atidžiai nustatomą faktoriaus taškų vidurkį“, kai dėstytojų reitingavimas sąlygoja administracinius sprendimus (Abrami 1990, p. 98). Abrami nuomone, nors „geras dėstymas“ susideda iš daugybės komponentų, jį reikėtų vertinti „globaliai“, lyginant įvairių kursų, fakultetų ir įvairiomis aplinkybėmis dirbančius dėstytojus. Mokslininkas abejoja konstrukto, kuris neapima visų gero dėstymo dimensijų ir charakteristikų įvertinimo, validumu.

Tradiciniame įvertinime, kuris grindžiamas pozityvizmo modeliu, pabrėžiama, kad labai svarbu nustatyti ir apiben-

INTRODUCTION

The traditional positivistic approach for evaluating teaching in higher education has been characterized by a strong emphasis on objectivity in measurement that excludes attention to values behind the practice. The researcher or evaluator is an “objective” data gatherer who strongly relies on quantitative methods. According to Erickson (1986), the mainstream paradigm for research on teaching has its roots in the traditional model of the natural sciences:

“The history of the positivistic research on teaching for the past 20 years is one of analytical bootstrapping with very partial theoretical models of the teaching process, on the assumptions that what was generic across classrooms would emerge across studies and that the subtle variations across classrooms were trivial and could be washed out of the analysis as error variance.” (p. 131).

Researchers following this paradigm tend to link the idea of teaching to the idea of treatment, and evaluation to the idea of effectiveness. Teaching effectiveness, then, is “measured by looking at end-of-the-year scores or standardized achievement tests, and to particular teaching practices.” (Erickson 1986, p. 131).

A clear example of this paradigm is the process-product research that strongly supports “direct” instruction, the presentation and recitation of desired knowledge and behaviors. In this research, the effectiveness of teaching is “attributable to combinations of discrete and observable teaching performances per se, operating relatively independent of time and space.” (Shulman 1986, p. 10).

According to Dunkin and Barnes (1986), process-product research of the 60’s and early 70’s is the underlying rationale for teaching in higher education today. But unlike most of this research (conducted at other levels of education), in higher education “the process part has been assumed on the basis of prescriptive definitions, or rated by untrained observers, rather than documented through careful observation.” (Dunkin & Barnes, p. 774).

Other researchers point out that the formal conception of good teaching in higher education has not resulted from process-product research but from lists of characteristics or qualities that are used as descriptors of good teaching. Some of these lists are the result of surveys to faculty members and students who have been asked to describe what constitutes “good teaching.” Feldman (1988), Frey (1979) and Marsh (1997) support the use of these characteristics or behaviors for designing instruments for assessing teaching quality. They say that including multiple dimensions can produce useful information as feedback for faculty about their teaching, and for identifying faculty needs for improvement of instruction.

Another group of researchers supports a different point of view. These researchers claim that teaching should be evaluated “globally” rather than paying attention to particular characteristics or dimensions of instruction. Cashin and Downey (1992), Cohen (1986) and Abrami et al (1990; 1993), are among the researchers who support the use of global items or a “carefully weighted average of the factor scores” when the ratings are used for making administrative decisions (Abrami et al. 1990, p. 98). According to Abrami, even though “good teaching” is a construct of multiple components, it is more appropriate to evaluate teaching “globally” when comparing instructors across courses, departments and settings. He expresses concern about

drinti priežasties ir pasekmės sąsajas. Dažniausiai įvertinimui naudojamos reitingo skalės, pusiau struktūrinės apklausos, klausimynai ir testai (Feldman, 1986, 1986; Falk, 1971). Vis dėlto dažniausios dėstymo įvertinimo priemonės kolegijose yra klausimynai. Paprastai į klausimyną įtraukiami globalūs ir/ arba standartizuoti klausimai apie dėstymo charakteristikas ir dimensijas. Įvertinimo formų administravimas taip pat standartizuojamas. Analizuojant tokių apklausų rezultatus, tyrimų duomenys verčiami taškais ar skaičiais. Po to rezultatai gauti iš kitų įvertinimų lyginami su duomenimis, gautais iš kitų fakulteto atstovų, arba sugretinami su iš anksto nustatytais kriterijais ir standartais. Atliekant netgi darbo auditorijoje stebėjimus remiamasi kiekybiniais vertinimais.

Dėstymo įvertinimo pagrįstumo tyrimuose aukštosiose mokyklose taip pat analizuojamas studentų atsakymų pagrįstumas. Nors tyrimų apie tai, kaip studentai vertina dėstymą, gana nemažai, tačiau jie dažniausiai daugiasekcijiniai ar daugybiniai, tad reikalinga tikslesnė (tyrimų) pagrįstumo analizė pagal naują Messik (1989) pagrįstumo modelį. Messik (1989) modelis ne tik lėmė naujas tyrinėjimų kryptis, bet ir sąlygojo naujų studentų vertinimo standartų sukūrimą (AERA, APA ir NCME, 1999).

1 PAGRĮSTUMAS IR PAPLITĘ DĖSTYMO ĮVERTINIMAI

Pagrįstumo sąvokoje slypi du klausimai: „Ar matuojame tikrai tai, ką norime išmatuoti? Ar įžvalgos ir veiksmai apie įvertinamuosius[1] yra pagrįsti faktais?“ Kadangi pagrįstumas apima interpretacijos prasmingumo, reikšmingumo ir tinkamumo sąvokas, jis yra svarbiausias dėstymo įvertinimo faktorius. Tai sutampa ir su naujais pagrįstumo teorijos atradimais.

Devintojo dešimtmečio pabaigoje ir dešimtojo pradžioje, vertinimo literatūra pasipildė naujais tyrimais, kurie lėmė naujos pagrįstumo sampratos atsiradimą. Šių naujoviškų tyrimų pradininkas buvo Samuelis Messick (1989), parašęs straipsnį apie pagrįstumą. Shepard (1993), Lane, Park ir Stone (1998); Moss (1992; 1998), Reckase (1998); Yen (1998) ir Cronbach (1989) tęsė Messick tyrinėjimų kryptį. Naujasis modelis atmeta fragmentišką ir pateikia vientisą pagrįstumo sampratą. Pagal šį modelį pagrįstumas yra konstruktyvus reiškinys. Kaip teigia Messick, naujasis modelis „jungia turinį, kriterijus ir pasekmes į konstruktyvią visumą, leidžiančią empiriškai patikrinti racionalias hipotezes apie skaičių reikšmę ir teorinius taikomojo ir mokslinio pobūdžio santykius“ (Messick 1995, p. 751).

Be to, pagrįstumas yra ne testo ypatumas, o „visapusis sprendimas, kuriame nustatoma, kaip, kokiais mastais įvertinimo interpretacijų adekvatumas ir tinkamumas paremtas empiriniais duomenimis ir teorija“ (Messick 1995, p. 741). Pagrįstumas apima ne tik vertinimo rezultatų prasmingumą ir interpretacijas, bet ir išvadas bei socialines įvertinimo pasekmes. Taigi galima teigti, kad pagrįstumas remiasi prasmingumu ir rezultatais (Messick, 1989, 1995).

[1] Įvertinimo objektas.

construct validity problems that could result if the evaluation is unable to include all relevant dimensions and characteristics of good teaching.

Traditional evaluation under a positivistic paradigm puts strong emphasis on generalization, and on the establishment of cause and effect linkages. In most cases, the evaluation is conducted by using rating scales, semi-structured interviews, personality tests or questionnaires (Feldman, 1986, 1989; Falk, 1971). Questionnaires, however, are the most used instruments in the evaluation of college instruction by the researchers. In most cases, the questionnaires include global items and/or pre-ordinate standardized sets of items about teaching characteristics and dimensions. The evaluation forms are administered in standardized way. Findings of these surveys are commonly analyzed in a way that reduces the results to a rating or score. Then, results obtained from the evaluation are compared with those obtained by other faculty members or against a predetermined criterion or standard. When classroom observations are conducted, the tendency again is towards quantification.

Studies on the validity of the evaluation of teaching in higher education have also centered on the validity of student ratings. Although there is a broad number of studies about the validity of student ratings of instruction, mainly as multi-section and multitrait studies, there is a need for examining the validity of the research under the new validity framework developed by Messik (1989), that has resulted in a shift in the validity literature and influenced the creation of new standards for student assessment (AERA, APA, and NCME , 1999).

1 VALIDITY AND CURRENT EVALUATIONS OF TEACHING

Validity is concerned with the questions: Are we measuring what we think we measure? Are our inferences and actions about the evaluand[1] supported by evidence? Because validity is linked to the meaning, value and the appropriateness of interpretation, validity is the most critical consideration in evaluation. Issues of validity in the evaluation of teaching are important given the new developments in validity theory.

In the late 80's and early 90's, a shift took place in the assessment literature that resulted in a new conceptualization of validity. Samuel Messick was responsible for this shift with his famous chapter on validity (1989), followed by Shepard (1993), and other authors, such as Lane, Park, and Stone, (1998); Moss, (1992, 1998); Reckase, (1998); Yen, (1998); and Cronbach, (1989). The new framework moves away from a fragmented to a unified concept of validity. Under this new framework, all validity is about construct validity. As Messick states, the new framework "integrates considerations of content, criteria, and consequences into a construct framework for the empirical testing of rational hypotheses about score meaning and theoretically relevant relationships including those of an applied and a scientific nature" (Messick 1995, p. 751).

In addition, validity is not a property of a test but "an overall judgment of the extent of which empirical evidence and theory support the adequacy and appropriateness of the interpretations based on the assessment" (Messick 1995, p. 741). Moreover, validity refers not only to meanings and interpretation of assessment scores, but also to the inferences

[1] Evaluation object.

1.1. KONSTRUKTO PAGRĮSTUMO ASPEKTAI

Messick (1989, 1995) išskiria šešis konstrukto pagrįstumo aspektus, kuriais reikėtų vadovautis atliekant švietimo vertinimą ir siekiant nustatyti, ar tyrimai patikimi: turinio, esminį (substancialųjį), struktūrinį, išorinį, apibendrinamąjį ir pasekmių. Ory ir Ryan (2001) tyrinėdami, kiek pagrįsti yra studentų pateikti dėstymo reitingai nustatė, kad kai kuriuose moksliniuose darbuose analizuojamas vienas ar kitas pagrįstumo aspektas, tačiau daug svarbių aspektų dar netyrinėti. Taip pat įdomu būtų išsiaiškinti, kaip įvertinimo kontekstas sąlygoja studentų tyrimų, skirtų studentų pateikiamiems reitingams, pagrįstumą.

1.1.1. TURINIO PAGRĮSTUMAS

Vieną svarbiausių pagrįstumo aspektų sudaro įvertinimo galimybė atspindėti matuojamo konstrukto turinį. Svarbiausias su pagrįstumu susijęs klausimas: „Ar egzistuoja ryšys tarp įvertinimo turinio ir matuojamo konstrukto?“ Atliekant dėstymo įvertinimą aukštojoje mokykloje turinio pagrįstumas siejasi su dėstymo kokybės įvertinimu. Konstruktas gali būti neefektyvus, kai įvertinimo procese nepajėgiama nustatyti visų gero dėstymo komponentų. Netikslumų gali atsirasti, kai vertinami kintamieji, nesusiję su dėstymo kokybe.

Pagal Ory ir Ryan (2001), kadangi dauguma įvertinimo formų „yra sukurtos nesiremiant teorija ar konstruktų sfera“ (p. 11), kyla klausimų dėl įvertinimo duomenų interpretacijos pagrįstumo. Be to, standartizuotų procedūrų, skirtų kolegijų įvertinimui, taikymas gali būti probleminis, nes, pavyzdžiui, įvertinimo procese gali būti nepajėgiama pilnai pateikti vertinamą konstruktą, kadangi neturima visų duomenų arba gali būti analizuojami su konstruktu nesusiję kintamieji (Stake, Cisneros-Cohernour, 2000).

Kitos pagrįstumo problemos kyla dėl to, kad šiuolaikiniuose įvertinimuose laikomasi siaurų dėstymo apibrėžimų, nebeatitinkančių šiuolaikinių dėstymo ir studijavimo teorijų reikalavimų. Nors dėstymo teorijos yra pažengusios nuo paprastų iki sudėtingų koncepcijų, įvertinime tokie pokyčiai dar neįvykę. Jeigu duomenų reikšmės nėra aiškios, abejotina, ar įvertinimo rezultatai tinkamai atspindi dėstymo kokybę.

Vadinasi, nevykęs konstrukto vertinimas ir duomenų nepagrįstumas atsiranda, kai vertinimas nėra plačiai apibrėžtas, ne visos gero dėstymo dimensijos ir ne visi svarbūs konstrukto elementai pateikti. Pagrįstumo ir įvertinimo neatitikimus gali lemti skirtingos studijavimo sąlygos įvairiose auditorijose ar skirtingas kurso išdėstymas. Svarbu išsiaiškinti, ar vertinimas nešališkas, t.y. ar nepalaikomas vienas požiūris į dėstymą ir studijavimą; ar nekritikuojami alternatyvūs, netradiciniai požiūriai.

1.1.2. ESMINIS PAGRĮSTUMAS

Šis konstrukto pagrįstumo aspektas svarbus analizuojant respondentų atsakinėjimo procesus testo metu ir pildant įvertinimo formas siekiant nustatyti, ar yra atitikimas tarp to, ką atsako respondentai, ir tos informacijos, kurią vertinimo priemone norėta surinkti. Esminis pagrįstumas atskleidžiamas tuo atveju, kai matuojamas konstruktas ir apklausos turinys koreliuoja tarpusavyje. Iš Ory ir Ryan (2001, p. 14) pavyzdžio galime matyti, kad, „kai respondentas, atsakinėdamas į kritinio mąstymo testo klausi-

and social consequences that result from the evaluation. Indeed, meaning and consequences are essential to validity (Messick, 1989, 1995).

1.1. ASPECTS OF CONSTRUCT VALIDITY

Messick (1989, 1995) identified six important aspects of construct validity to be used for all educational assessments to identify sources of invalidity: construct, substantive, structural, external, generalizability, and consequential. In their research on the validity of student ratings of instruction, Ory and Ryan (2001) found that some studies have been conducted on some aspects of these aspects of validity, but that other important aspects have not been addressed by the research. There is also a need for examining how the evaluation context could raise issues related to the validity of the student ratings research.

1.1.1. CONTENT VALIDITY

One of the most important aspects of validity is the capacity of the evaluation to reflect the content of the construct that it is intended to measure. This aspect of validity addresses the question: Is there a relationship between the content of the evaluation and the construct intended to be measured? In the evaluation of teaching in higher education, content validity refers to the capacity of the evaluation to measure teaching quality. Consequently, there is construct under-representation when the evaluation is not broad enough broad to measure all the components of good teaching. There is construct irrelevant variance when the assessment includes variables other than teaching quality.

According to Ory and Ryan (2001), because most evaluation forms “are developed without too much thought of theory or construct domains” (p. 11), there is a raising concern about the validity of the interpretations based on evaluation scores. In addition, the use of standardized procedures for evaluating college raises issues of construct under-representation if the assessment fails in representing the construct been measured either because it lacks important elements of the construct or because the measurement includes variables non relevant to this construct (Stake & Cisneros-Cohernour, 2000).

Moreover, validity problems increase because current evaluations of teaching remain focus on narrow definitions of teaching that are not consistent with current theories of teaching and learning. While the research on teaching has evolved from simplistic to more complex conceptualizations, these changes have not taken place in the evaluation. Unless there is certainty about the meaning of the scores, it can not be said for certain that the evaluation results are valid representations of instructional quality.

Since construct under-representation could occur when if the assessment is not defined broadly enough to include critical dimensions of the construct good teaching, the scores can not be interpreted as good teaching if they have failed to include all the important elements of the construct. Having different conditions of learning taking place in different classrooms and implementing course content in different ways can also result in a threat to validity if as a result of this a subgroup of instructors receives an unfair advantage in the evaluation. It is also important to determine if the assessment is encouraging a particular type of teaching and learning, and also if the assessment results in punishing alternative approaches that stress non-traditional views of teaching and learning.

mus, mąsto kritiškai, testo rezultatai laikomi iš esmės pagrįstais“.

Konstrukto esminio pagrįstumo tyrimuose siekiama atsakyti į keletą klausimų: „Kas lemia rezultatų skirtumus? Kas yra žinoma apie atsakinėjimo procesus skirtingose situacijose? Jei studentų atsakymai duotoje situacijoje teigiamesni, ar jie teisingi? Ar įvertinimo proceso pobūdis atitinka matuojamą konstruktą?“

Svarbu suvokti ne tik tai, kad rezultatai gali varijuoti skirtingose situacijose, bet ir kodėl jie varijuoja. Taip pat reikėtų išsiaiškinti, kaip studentai suvokia rangavimo skalę ir ar jų skalė sutampa su testo skale; ar visi studentai atsako į klausimus ta pačia veiksmų seka, ar tam tikra grupė atsako kitaip nei likusieji; ar vertinimas objektyvus, jei studentai yra iš skirtingų etninių, kultūrinių grupių.

Buvo atlikta keletas tyrimų (Marlin, 1987; Dwinell ir Higben, 1993; Ballantyne, 1998), kur bandyta išsiaiškinti studentų požiūrį dėl įvertinimo, tačiau vis dar mažai žinoma „kokie procesai sąlygoja studentų atsakymus į rangavimo klausimus“ (Ory ir Ryan, 2001, p. 26). Mokslininkai pažymi, kad „atlikti tyrimai tik parodė kaip kinta vertinimai skirtingose situacijose, tačiau nepaaiškino, kodėl jie kinta“ (2001, p. 15). Norint nustatyti, kaip studentai suvokia rangavimo skalę atsakinėdami į klausimus, reikia naujų, tikslesnių tyrimų. Ory ir Ryan (2001, p. 15) kelia klausimą, „kaip studentai supranta vidutinį įvertinimą, jei pasirenkama penkiabalė Likerto skalė? Ar pažymėtas trejetas reiškia neigiamą, vidutinį ar atsainų įvertinimą? Kaip reaguojama į skalę, kurioje pažymėti tik kraštutiniai įvertinimai? Ar vieni studentai labiau linkę rinktis kraštutinius įvertinimus nei kiti? Ar kai kurie studentai mano, kad „idealus penketas“ yra neįmanomas? Norint padaryti pagrįstas išvadas, reikia išsiaiškinti, ar studentai ir vertintojai vienodai suvokia vertinimo skalę.“

Kritikai iškelia dar vieną problemą, kuri standartizuotuose dėstymo įvertinimuose apeinama – gauti rezultatai nebūtinai atspindi realius skirtumus tarp žmonių ir neeliminuoja kultūrinių skirtumų. Akivaizdu, kad būtina ištirti dėstymo aukštojoje mokykloje įvertinimo esminį pagrįstumą. Rezultatų įvertinimas ir jų interpretavimas gali būti patikimas tik tuo atveju, kai studentų atsakymų sistema ir atsakymų skirtumų priežastys yra išsiaiškinamos. EL-Hassan (1995) teigia, kad taip pat reikėtų išsamiau paanalizuoti, kaip studentų atsakymus veikia įvairūs veiksniai: dėstytojo užimamos pareigos, lytis, dėstomas kursas (privalomas ar pasirenkamas, disciplinos pobūdis, grupės dydis, užduočių sunkumas), studento motyvacija, gaunami pažymiai.

1.1.3. STRUKTŪRINIS PAGRĮSTUMO ASPEKTAS

Šis konstrukto pagrįstumo aspektas reikalauja, kad „pasirinkto konstrukto teorija sąlygotų ne tik atitinkamų vertinimo užduočių pasirinkimą ar konstravimą, bet ir racionalų konstruktu pagrįsto vertinimo kriterijų ir jų dalių sukūrimą“ (Messick, 1994, p. 15). Struktūrinis pagrįstumo aspektas atsako į klausimą: „Koks ryšys tarp skirtingų įvertinimo procedūros komponentų ir įvertinamo konstrukto?“ Įvertinimas turi pagrįsti ryšį tarp atskirų įvertinimo mechanizmo komponentų bei konstrukto struktūros. Taip pat svarbu žinoti, ar rezultatų sumavimo būdas atitinka konstrukto ribas.

1.1.2. SUBSTANTITIVE VALIDITY

For this aspect of construct validity it is important to analyze response processes of those taking the test and completing the evaluation forms in order to see if there is a fit between the process used to answer and the process for which the assessment was developed. Evidence of substantial validity can be found when there is a fit between what is been tested and the construct measured. As Ory and Ryan (2001) illustrate, “When an examinee uses critical thinking to answer items on a test of critical thinking there is evidence for the substantial validity of the test scores.” (p. 14).

Studies on the substantive validity aspects of construct validity focus on questions such as: What accounts for score differences? What do we know about the response processes in different situations? If students respond more positively in a given situation, are they responding more or less truthfully? Does the nature of the evaluation process match the construct being measured?

It is not enough to know that the scores change in different situations, it is necessary to know why the change takes place. We also need to understand how students use the rating scales to respond, and if there is a fit between the intended meaning of the scale and the meaning of the scale for students. It is important to determine if all students follow similar processes when responding to the tests. Do some subgroups of students respond differently than others? Is the assessment appropriate for different groups of students of diverse ethnic and cultural backgrounds?

Several studies (Marlin, 1987; Dwinell and Higben, 1993; Ballantyne, 1998) have been conducted about student attitudes about the evaluation, specifically towards the student ratings. But, little is still known about “the actual process followed by students to respond to rating forms.” (Ory and Ryan, p. 26) According to these authors “past research efforts have indicated how ratings change in different situations but they do little to help us understand why the change occurs” (p. 15). More research is also needed to understand how students use the rating scales to respond. As Ory and Ryan (2001) state, “If items are presented with a five point Likert scale, how do students interpret and use the middle category? Do students mark a “3” to indicate an inability to respond, a middle response, or a lack of interest? If only the endpoints are labeled, how do students interpret and use the other scale points? Are some students more reluctant than others to use the extreme ends of the scale? Do some students believe that a “perfect five” is unobtainable? To make valid inferences from student ratings we need to determine if there is proper fit between what the meaning of the scale was for students, and the intended meaning of the scale.” (p. 15)

There is also an important problem not addressed with standardized evaluations of teaching on campus, as identified by critical scholars, is that the scores do not necessarily reflect real differences among people, and they often do not adequately eliminate underlying biased cultural assumptions built into the test as a whole. Consequently, there is a need for conducting research on the substantive validity of the evaluation of teaching in higher education. The interpretation and use of evaluation results can be improved if the student response pattern and the differences in response patterns among different students are understood. As El-Hassan (1995) states, there is also a need for examining more deeply how variables related to the instructor (faculty rank and gender) or the course (required versus elective, academic discipline, class size, workload-difficulty), or the student (motivation towards the course, expected grades) could influence student response patterns.

Daugeliu tyrimų bandyta nustatyti gero dėstymo charakteristikas ir procesus remiantis studentų, dėstytojų ar administracijos pateiktais duomenimis. Didžioji dalis įvertinimo priemonių yra sukurtos būtent tokių tyrimų pagrindu. Mokslininkai nustatė koreliacijas tarp įvairių savybių, charakteristikų ir dėstymo reitingavimo. Pavyzdžiui, Centra (1993) ir Feldman (1976) išanalizavę keletą įvertinimo formų, rado tam tikrų bendrumų. Kita vertus, reikia atkreipti dėmesį į tai, kad studentai panašiai atsako į tam tikrus klausimus ne todėl, kad jie priklauso žinomai sričiai. Tarsi būtų analizuojami studentų atsakymai į šimtus matematinių klausimų, sugrupuojant juos į atsakymais pagrįstus klasterius, identifikuojant kaip esminius gebėjimus, kurie padeda išspręsti matematinius klausimus (Ory ir Ryan 2001, p. 18). Galima paminėti Kulik ir McKeachei (1975) bei Feldman (1987) metaanalizės studijas, t.y. nors mokslininkai kruopščiai išanalizavo daugybę skirtingų duomenų, tačiau nesugebėjo nustatyti dėstymo kokybę apibūdinančių esminių charakteristikų ir elgsenos. Ory ir Ryan (2001, p. 11) teigia, kad „būtina apibrėžti tikslias efektyvaus dėstymo charakteristikos ribas. Tų ribų neapibrėžus institucijos kelia klausimą, kokiu pagrindu buvo sudarytos įvertinimo formos ir, kas dar svarbiau, kaip analizuojami ir interpretuojami gauti duomenys.“ Kadangi neturima empirinių duomenų, kad pasirinkti elementai iš tiesų atspindi gerą dėstymą, būtini tolesni struktūrinio pagrįstumo aspekto tyrimai. Be to, svarbu išsiaiškinti įvertinimo formų konstrukto ribas: kiek jo ribos lemia dėstymo kokybę ir kaip skirtingi asmenys suvokia dėstymo reitingavimą?

1.1.4. IŠORINIO PAGRĮSTUMO ASPEKTAS

Šis aspektas nurodo įvertinimo ryšį su kitais kintamaisiais, kurie yra išoriniai atliekamam vertinimui, siekiant nustatyti informacijos pagrįstumo akivaizdumą. Taigi „taškų reikšmė patvirtinama išoriškai įvertinus empirinį duomenų atitikimo laipsnį kitų matavimų duomenims, o esant skirtumams, ar tie įverčiai atspindi jų tikrąją reikšmę“ (Messick 1994, p. 16). Be to „svarbūs išorinių santykių ryšiai susiformuoja tarp vertinimo taškų ir kriterijų, susijusių su atranka, įdarbinimu, licencijavimu, programos įvertinimu bei kitų su atsiskaitomybės procesu siejamų analizės kriterijų“ (Messick 1994, p. 17).

Ankstesniuose studentų pateikto dėstymo reitingavimo pagrįstumo tyrimuose buvo analizuojami svarbūs pagrįstumo aspektai, bandyta nustatyti ar egzistuoja ryšys tarp studentų vertinimo ir jų pasiekimų[2]. Dažniausiai per tokius tyrimus nustatomas įvertinimo pagrįstumas analizuojant vieno kurso, kurį dėsto skirtingi dėstytojai, įvertinimo rezultatus ir jų santykį su studentų pasiekimais (Ory ir Ryan, 2001). Pavyzdžiui, Cohen (1981) tyrinėjo koreliaciją tarp studentų pateikto reitingavimo ir jų pasiekimų; Murray (1983) įtraukė į tyrimus parengtus stebėtojus, kurie turėjo nustatyti skirtumus tarp geriausiai ir prasčiausiai įvertintų dėstytojų.

Peržiūrėjus didelį kiekį daugiaplanių tyrimų, atliktų Abrami, D'Apollonia ir Cohen (1990), paaiškėjo, kad būtina tikslesnė analizė, jei siekiama suprasti, kokios yra apibendrinamumų ribos, kriterijų efektyvumas, reitingavimo dimensijos ir dėstymo sąlygos. Kadangi grupių ho-

[2] Parodomų pažymiais

1.1.3. STRUCTURAL ASPECT OF VALIDITY

This aspect of construct validity stresses that “the theory of the construct domain should guide not only the selection or construction of relevant assessment tasks, but also the rational development of construct-based scoring criteria and rubrics.” (Messick 1994, p. 15). This aspect of validity addresses the question: To what extent does the relationship among different components of the evaluation procedures correspond with the construct being evaluated? In this way, the evaluation needs to provide evidence that the relationship among the different components of the assessment instrument correspond with the structure of the construct domain. It is also important to know how well the scoring structure is consistent with the construct domain.

A large number of studies have been conducted in order to determine the characteristics or behaviors that constitute good teaching, based mostly on the perceptions of students, teachers or administrators. Most evaluation instruments are based on those characteristics. Some researchers have also found correlations between these and other sets of characteristics and behaviors and the ratings of instruction. For example, Centra (1993) and Feldman (1976) found common dimensions after analyzing several evaluation forms. However, it is important to consider that items are included on many forms because students appear to respond similarly to particular ones not because they come from a known domain of targeted characteristics. It is somewhat like analyzing student responses to hundreds of math items, grouping the items into response-based clusters, and then identifying the clusters as essential skills necessary to solve math problems. (Ory & Ryan 2001, p. 18). In addition, some studies such as the work of Kulik and McKeachie (1975), and Feldman (1987) were meta-analyses. Although there is consistency across the different data analyzed, the researchers have not been able to identify a single set of characteristics and behaviors, as those essential for defining the construct teaching quality. As Ory and Ryan (2001) state, “Without a clearly defined target domain of effective instructional characteristics, it is unclear how institutions select the content of their evaluation forms, and more importantly, what do these institutions infer as the meaning of their ratings” (p. 11). Since there is no empirical evidence that the items selected are indeed elements of good teaching, there is a need of more research on this aspect of construct validity. Consequently, an important area of research needs to focus on what is the construct domain in the evaluation forms? How this domain relates to instructional quality? What is the meaning of the ratings for different stakeholders?

1.1.4. EXTERNAL ASPECT OF VALIDITY

It refers to the relationship of the evaluation to other variables, external to the assessment in order to provide source of validity evidence. In this way, “the meaning of the scores is substantiated externally by appraising the degree to which empirical relationships with other measures, or the lack thereof, is consistent with that meaning” (Messick, 1994, p. 16). In addition, “special importance among external relationships are those between the assessment scores and criterion measures pertinent to selection, placement, licensure, program evaluation, or other accountability purposes in applied settings” (Messick, 1994, p. 17).

Prior research on the validity of student ratings has been conducted to address this important aspect of validity. Some of these studies have been conducted to determine if there is a relationship between student ratings and student achievement.[2] The multisection studies are an example of this kind of research that determines the validity of the evaluation by analyzing the correla-

[2] Defined as grades.

mogeniškumas ir kiti veiksniai kinta, koreliacinių tyrimų rezultatai turėtų būti traktuojami labai atsargiai. Be to, minėti tyrimai dažniausiai buvo atlikti pirmakursiams ir antrakursiams skirtuose įvadiniuose kursuose.

Daugiaplaniuose tyrimuose studentų pateikti reitingavimai lyginti su kitokiais duomenų šaltiniais, pavyzdžiui, bendraamžių ir absolventų pateiktais reitingavimais, savianalize ir t.t. Taip buvo siekiama patikrinti, ar rezultatai, gauti iš skirtingų šaltinių, neprieštarauja vieni kitiems. Paaiškėjo, kad studentų ir absolventų atsakymai akivaizdžiai sutampa, buvo nustatytos stiprios teigiamos koreliacijos tarp šių kintamųjų. Kituose tyrimuose buvo lygintos skirtingos duomenų rinkimo formos, pavyzdžiui, ar skiriasi studentų atsakymai į uždarus ar atvirus klausimus, grupės interviu ir t.t. (Ory, Braskamp ir Pieper 1980; Ory ir Ryan 2001).

1.1.5. PAGRĮSTUMO APIBENDRINAMUMO ASPEKTAS

Vertinimo pagrįstumo apibendrinamumo aspektu siekiama nustatyti, ar egzistuoja ryšys tarp „jau įvertintų užduočių ir kitų užduočių, kurios reprezentuoja konstruktą ar atskirus jo aspektus“ (Messick 1994, p. 15). Apibendrinamumo aspektas rodo „įverčių reikšmių ribas“ (Messick 1994, p. 15).

Pagrįstumo apibendrinamumo aspektas kelia tokius klausimus: „Ar galima lyginti skirtingų dalykų aplinkos ir skirtingų laikotarpių įvertinimų reikšmes? Ar galima daryti tas pačias išvadas apie įvertinimus, atliktus skirtingoje aplinkoje? Ar pagrįsta lyginti įverčius, naudotus skirtingiems tikslams? Ar įverčiai, gauti skirtingoje aplinkoje gali būti lyginami?“ Apibendrinamumo tyrimais bandoma nustatyti ir atskleisti skirtumus, suprasti, kodėl jie atsiranda, išmokti, kaip juos apibendrinti pristatant vertinimo rezultatus ir siekiant sustiprinti vertinamo proceso pagrįstumą.

Nors kai kurie mokslininkai teigia, kad skirtingų grupių studentų vertinimus galima apibendrinti, Abrami, d'Apollonia ir Cohen (1990) tuo abejoja, sakydami jog „didžioji dalis tyrimų buvo atlikti įvadiniuose pirmakursiams ir antrakursiams skirtuose kursuose“ (Ory ir Ryan 2001, p. 20). Taigi reikalingi detalesni šio konstrukto pagrįstumo tyrimai.

1.1.6. PASEKMĖS PAGRĮSTUMAS

Vertinimo pasekmių pagrįstumo aspektas rodo trumpalaikes ar ilgalaikes vertinimo ir jo rezultatų interpretavimo pasekmes (Wilson, 1999). Svarbios tiek numatomos, tiek nenumatomos pasekmės. Pasekmių pagrįstumas skatina tyrinėti įvertinimo teorijų vertybes, išvadas ir principus, taip pat ideologiją, kurioje teorija yra taikoma. Didelis dėmesys skiriamas pasekmėms, „sietinoms su vertinimo šališkumu, neteisinga interpretacija ar nesąžiningu testo taikymu“ (Messick, 1994, p. 17).

Įvertinant dėstymą aukštojoje mokykloje, pasekmių pagrįstumas kol kas nesulaukė pakankamo tyrinėtojų dėmesio; dėl to reikalingi tikslesni įvertinimo šališkumo, numatomų ir nenumatomų pasekmių tyrimai. Svarbu įvertinti „teorijos principus ir potencialias ar esamas problemas, su kuriomis gali susidurti institucija“ (Ory ir Ryan 2001, p. 26); nustatyti potencialias neigiamas įvertinimo pasekmes, galinčias kilti dėl nepakankamo reprezentatyvumo, šališkumo ar nesąžiningumo.

tion of evaluation results of a single course that is taught by different instructors with the section mean of student achievement (Ory and Ryan, 2001). In addition, Cohen conducted correlation studies (1981) on the relationship between student ratings and student achievement, and other researchers such as Murray (1983) used trained observers to determine teaching differences among instructors who obtained high and low ratings.

A review of several dozen multi-section studies conducted by Abrami, D' Apollonia, and Cohen (1990) showed that although consistent, more research needs to be conducted to understand the limits on generalizability of rating validity across rating dimensions, effectiveness criteria, and conditions of instruction. As is well known, correlation research findings need to be taken carefully because group homogeneity and other factors vary so. In addition, many of these studies were conducted in low learning, introductory courses taught primarily to freshmen and sophomores.

Multi-trait studies have also compared the results obtained from student ratings with other data sources, such as peers, alumni, self-ratings, etc. Researchers have studied different evaluation sources to determine the consistency among different data sources in evaluating teaching. Researchers have found high positive correlations between student ratings and alumni ratings. In addition, another group of studies have studied the correlation between different forms of data collection, such as "student overall ratings of instructor competence as measured rating items, written comments to open-ended items, and group interviews (Ory, Braskamp and Pieper, 1980; Ory and Ryan, 2001).

1.1.5. GENERALIZABILITY ASPECT OF VALIDITY

It examines if there is correlation "of the assessed tasks with other tasks representing the construct or aspects of the construct" (Messick 1994, p. 15). Generalizability refers to the "boundaries of score meaning" (Messick 194, 15).

Generalizability as an aspect of construct validity addresses questions such as: Can we make comparable inferences about the meaning of the scores across subjects, settings, and time? Can we make the same inference about ratings collected in different settings? Can we make valid comparisons between scores used for one purpose versus another purpose? Are assessment scores collected in different settings comparable? Generalizability studies focus on determining differences, understanding why they occur, and learning how to account for them in reporting assessment results to enhance the validity of the assessment process.

Although some researchers support the generalizability of student ratings across different sections, other researchers such as Abrami, d'Apollonia, and Cohen (1990) have questioned the generalizability of the evaluations of teaching using student ratings "because so many of the studies were conducted in lower-learning, introductory courses taught primarily to freshmen and sophomores" (Ory and Ryan, 2001, p. 20). More research is needed on this important aspect of construct validity.

1.1.6. CONSEQUENTIAL VALIDITY

This refers to the short and long-term consequences of evaluation use and the consequences associated with interpretations of evaluation scores (Wilson, 1999). Both intended and unintended consequences are important. Consequential validity also implies the need for appraising the value implications of the theory underlying evaluation scores, as well as the ideology in which the theory is embedded. When collecting evidence about the consequential aspect of validity, especial emphasis is given to consequences "associated with bias in scoring and interpretation or with unfairness in test use" (Messick, 1994, 17).

2
POZITYVISTINĖS PARADIGMOS RIBOTUMAI

Pozityvistinio dėstymo įvertinimo privalumai yra organizuotumas ir paprastumas, tačiau, daugumos autorių nuomone, pastebėti ribotumai sumenkina šiuos privalumus. Standartizuotas studentų vertinimas gali turėti įvairių neigiamų pasekmių.

Pirma, išskirtinis dėmesys charakteristikoms ar elgsenos savybėms „riboja žinias apie dėstymą ir studijavimą“ (Dunkin ir Barnes, 1986, p. 774). Toks įvertinimas dažniausiai yra orientuotas į dėstytoją, t.y. dėstytojas turi „nuosekliai ir aiškiai perteikti tam tikrą kurso medžiagą, o studentas įsisavinti kursą atlikdamas tradicines užduotis tradiciniais studijavimo metodais“ (Centra ir Boneshell, 1990). Iš to išplaukia, kad alternatyvius dėstymo būdus taikantys dėstytojai yra kritikuojami. Galimas neatitikimas tarp įvertinimo ir bet kokio dėstymo, besiremiančio konstruktyvistine mokymosi teorija ar asmenybės bei kognityvinio vystymosi teorijomis (Mabry, 1999).

Nors charakteristikų ir elgsenos apibendrinimai gauti proceso ir rezultato tyrimų metu koreliuoja su studentų elgesiu per egzaminus ir testus, kyla abejonių dėl pagrįstumo. Pagrįstumui įrodyti reikalingi tikslesni empiriniai tyrimai. Nustatyti elgesio ypatumai ir charakteristikos remiasi įvairių duomenų analize, tačiau „turima mažai įrodymų, kad dėstytojų elgesys auditorijoje visiškai atitinka apibrėžtą elgesio standartą“ (Shulman 1986, p. 12). Problemiškas ir apklausų metu nustatytų dėstytojų ir studentų elgesio charakteristikų, būdų, bruožų ir elgesio ypatumų panaudojimas tyrimui. Empirinių duomenų nepakanka norint tvirtinti, kad šie bruožai atspindi gerą dėstymą ar kad jie sąlygoja studijavimą (Miller, 1974; Genova, 1986).

Turinio pagrįstumo problemų atsiranda tuomet, kai vertinimo kriterijus sudaro ne visos būdingiausios charakteristikos ir elgesio ypatumai. Kaip jau minėta, per platus ar per siauras dėstymo apibrėžimas gali neatspindėti visų galimų dėstymo situacijų (Stake ir Cisneros-Cohernour, 2000). Remiantis Doyle (1982, p. 27) „...neįmanoma, kad tas pats charakteristikų rinkinys gali vienodai tikti dėstant skirtingus dalykus skirtingiems studentams, skirtingomis aplinkybėmis [...]. Tokio sąrašo sudarymas labai rizikingas“. Kai dėstymo kokybei nustatyti taikomos įvairios charakteristikos, jų preskriptyvinis panaudojimas gali riboti dėstymo kūrybingumą ir stabdyti profesinį tobulėjimą. Dėstymo būdų ir/ar bruožų kaip kriterijų taikymas dėstymo kokybės įvertinime, varžo dėstytojų darbo įvairovę, gali nukentėti dėstytojai „nepatenkantys į nustatytus rėmus“ (Stake ir Cisneros-Cohenour, 2000).

Tendencija dėstymo kokybę apibendrinti skaitmeniniais rodikliais gali išprovokuoti pastangas gerinti tik rezultatus, o ne darbo kokybę (Cisneros-Cohernour, 1997). Lyginimo rezultatai netaikant griežtos dėstymo ir studijavimo procesą lemiančių kintamųjų kontrolės gali būti neteisingi (Stake ir Cisneros-Cohenour, 2000).

Taip pat svarbu atsižvelgti į ekspertų, įvertinančių dėstymo kokybę, objektyvumą. JAV daugelis ekspertų, tikrinančių dėstymo įvertinimo pagrįstumą aukštojo

The consequential validity of the evaluation of teaching in higher education is an area that has received little attention by the researchers. More research is needed on the value implications of the evaluation results, the intended and unintended consequences of using certain criteria for defining and assessing good teaching, “the ideology within which the theory is imbedded, (and) the potential or actual problems that could result for the institution as a result of the consequences” (Ory and Ryan 2001, p. 26). More studies are needed to determine the potential negative consequences of the evaluation, especially in regard to issues of bias, fairness, primarily in relation to minority and other underrepresented groups among the faculty.

2
LIMITATIONS OF THE POSITIVISTIC PARADIGM

A positivistic orientation to the evaluation of teaching has the benefit of organization and simplicity. But many see the limitations and problems surpassing its benefits. As with standardized student assessment, this orientation could result in some serious negative consequences.

First, exclusive use of characteristics or behavioral attributes is “limiting to a certain kind of knowledge about teaching and learning.” (Dunkin and Barnes, 1986, p. 774). The evaluation usually centers on a kind of teaching that is teacher–centered. In other words, a kind of teaching in which the instructor’s task “is to cover a well defined set of topics for a course systematically and precisely, while the student’s task is to master the course content through traditional assignments and study methods,” (Centra & Boneshell, 1990). Instructors using a different teaching approach or style may be at a disadvantage. There can be a mismatch between the evaluation and any teaching consistent with constructivist learning theory, as well as with theories of human and cognitive development (Mabry, 1999).

The use of a list of characteristics and behaviors from the process-product research, although correlated with student performance in exams or tests, raises validity questions. Their validity has not been determined empirically. These behaviors and characteristics are the outcome of synthesis from aggregate data, but there is “little evidence that any observed teacher had ever performed in the classroom congruent with the collective pattern of the composite” (Shulman, p. 12). The use of characteristics, styles, traits and behaviors identified from surveys to faculty and students also presents a problem. There is little empirical evidence that any of them constitute good teaching, or that they are related to student learning (Miller, 1974; Genova et al, 1986).

There are problems when not all relevant characteristics and behaviors used as criteria are included in the assessment, problems of content validity. As said earlier, using a general and narrow definition of teaching is problematic because it may not be appropriate for all teaching situations (Stake and Cisneros-Cohernour, 2000). Doyle (1982) says, “… it seems most unlikely that any one set of characteristics will apply with equal force to teaching all kinds of materials to all kinds of students under all kinds of circumstances… To prepare such a list entails a substantial risk” (p. 27). When a number of characteristics are adopted as indicators of teaching quality, their prescribed use can result in limiting instructional creativity, and can become a barrier for professional development. The use of traits and/or teaching styles as criteria for evaluating teaching constrain diversity in instructors, penalizing those who do not “fall within the norm” (Stake and Cisneros-Cohernour, 2000).

siose mokyklose atsiduria dviguboje padėtyje kaip mokslininkai ir tie, kurie organizuoja ir atlieka įvertinimą. Jų mokslinė veikla tiesiogiai arba netiesiogiai dažniau mažina jų kaip institucijos administratorių veiklos pagrįstumą.

JAV paskelbtose publikacijose apie dėstymo tyrimus laikomasi gana vienodos nuomonės, tik kai kuriuose darbuose prieštaraujama šios tyrėjų grupės gautiems tyrimų rezultatams. Išimtis – Kanados mokslininko Brodie darbas. 1998 metais Brodie atliktų tyrimų rezultatai parodė, kad ankstesniuose darbuose, kuriuose buvo nustatytas ryšys tarp aukštesnio pažymio ir studentų pateiktų reitingavimų, „nepakankamai įvertinta tendencija rašyti aukštesnius įvertinimus dėl griežtumo stokos“ (1998, p. 17). Brodie išsiaiškino, kad „kai to paties dalyko įvertinimai atskirose grupėse smarkiai skyrėsi, dėstytojai, rašantys geresnius pažymius ir užduodantys mažiau darbų, gaudavo aukštesnius įvertinimus“ (1998, p. 17). Be to, Brodie (1999), analizuodamas santykį tarp studentų pateiktų reitingavimų ir jų pasiekimų, rado naujų nesutapimų. Nesutapimų pasitaikė ataskaitose apie tyrimų rezultatus ir ta pačia tema skelbtų straipsnių. Mokslininkas teigia, kad nesutapimai tarp dėstymo charakteristikos ir reitingavimų, skelbtų keliuose leidiniuose, atsirado dėl to, kad „kai kurie mokslininkai tiesiog ištrynė neigiamas ar žemas koreliacijas ir sufalsifikavo vertinimo skalę paversdami ją teigiama“ (1999, p. 1). Nors niekas nesidomėjo, kiek Brodie teiginiai teisingi, vis dėlto svarbu atsižvelgti į dvejopą tyrinėtojo – mokslininko ir įvertinimo sistemos administratoriaus – vaidmenį.

Kiti pozityvistinės paradigmos kritikai teigia, kad mokslininkai per daug dėmesio skiria skaičiavimams ir paverčia mokslą technika. Horkheimer ir Adorno (1948, p. 11) nuomone, „jie pakeitė koncepciją formulėmis, o priežastingumą taisyklėmis ir tikimybėmis.“ Magunsson (2000), diskutuodama dėl skirtingos kultūrinės ir etninės kilmės akademinio personalo dėstymo kokybės nustatymo pagrįstumo ir tinkamumo, yra pasakiusi, kad „tyrimai, kuriuose „mažumos“ priskiriamos nežymiems sistemos nukrypimams ar panašioms su skaičiavimų sistema susijusioms sampratoms, iš esmės neatsisako techninio psichometrijos diskurso problemų tyrimo. Kitaip tariant, jeigu rasizmas egzistuoja, jis būdingas visai organizacijai, negali būti tik nežymus skaičiavimo sistemos nukrypimas“ (Magunsson 2000, p. 45).

Menges (1998) teigia, kad reikia tirti, kaip interpretuojama ir naudojama informacija apie įvertinimus, kaip mokytojai „pritaiko įvertinimo rezultatus planuodami, dirbdami auditorinį darbą ir vertindami savo dėstymo lygį“ (p. 3). Po to Menges priduria, kad pagrindinis tyrimų trūkumas yra „dėstymo konteksto nepaisymas [...], skirtingo dalyvių požiūrio ir jų asmeninių, organizacinių ir politinių ypatumų ignoravimas“ (p. 4).

Teigiamo pozityvistinio modelio aspektu laikytina tai, kad vis labiau domimasi tyrimų prielaidų tikrinimu ir įvertinamo konstrukto pagrįstumu (Menges, 1998; Theall ir Franklin 1999, 2000; Ryan ir Johnson 1998; Ory ir Ryan 2001).

The tendency to summarize teaching quality in a numerical index can lead to the unintended consequence of people focusing more on improving the scores than on improving their teaching (Cisneros-Cohernour, 1997). Comparisons made without a rigorous control of variables influencing the teaching and learning process, can lead to unfairness (Stake and Cisneros-Cohernour, 2000).

It is also important to review the claims of objectivity made by those conducting research on the evaluation of teaching. In the U.S., most of those conducting research on the validity of the evaluations of teaching in higher education have a dual role, as scholars and as those who develop and implement the evaluation. Their scholarly work directly or indirectly more often than not supports the validity of their work as administrators in the institution.

The publications of teaching research in the US contain few studies that contradict the main findings of this research community. An exception is the work of Brodie, a Canadian researcher, who in 1988 found that prior studies on the relationship between grade inflation and student ratings “have underestimated the biasing effect of grading leniency” (p. 17). In that study, Brodie found evidence that “when grades varied markedly across sections of the same course, the professors assigning highest grades with less studying received highest evaluations” (p. 17). In addition, Brodie’s (1999) review of the research on the correlation between student evaluations of teaching and student learning raised new issues. He found discrepancy between the results in research reports and the published articles of the same study. He encountered evidence that correlations between certain teaching characteristics and the ratings as reported in several journals have been inflated and that “some researchers have deleted low and/or negative correlations, but also created positive correlations by reversing the rating scale” (p. 1). Although no research has been conducted to confirm the findings of Brodie’s research, or about the influence of the dual role of the researcher as scholar and as administrator of the evaluation system, these important questions deserve more attention.

Other critics of the positivistic paradigm perceive that the emphasis put on measurement by the researchers has been so strong that has replaced science with technique. As Horkheimer and Adorno, state, “they have replaced the concept with the formula, and causation with rule and probability” (1948, p.11). Magunsson (2000), in her discussion of the appropriateness and validity of the evaluation for assessing the quality of teaching of instructors of diverse cultural or ethnical background, adds:

“The problem with an analysis that equates ‘minority’ with small systemic variance, or other such measurement concepts, is that it constructs the issue once again within the technical discourse of psychometrics. The problem is that if there is racism, this is systemic to the entire organization and can’t be reflected merely as systemic variance related to measurement” (Magunsson 2000, p. 89)

Menges (1998), also claims that more research is needed with still little known about how the information from the evaluation is interpreted and used, and about how teachers “use the evaluation in planning, implementing, and appraising their own teaching.” (p. 3). He adds that the main shortcoming of the research is “its lack of recognition of the context of teaching ... (ignoring) the perspectives of different participants, and their personal, organizational, and political contexts” (p. 4).

What is promising is that among the researchers supporting the positivistic paradigm is a growing interest for testing the assumptions held by the research, and for examining the validity of the construct being evaluated (Menges (1998); Theall & Franklin, (1990, 2000); Ryan & Johnson (1998); and Ory and Ryan (2001).

IŠVADOS

Išaugus atsiskaitomybės poreikiui įvertinti dėstymą aukštojoje mokykloje, neformalūs požiūriai tapo sisteminiais. Be to, administracijai ėmus domėtis išmatuojamais rezultatais, dėstytojai ėmė rūpintis įvertinimo objektyvumu ir tuo, kaip jis gali sąlygoti jų darbo sutartis, paaukštinimo ar algos pakėlimo galimybes.

Kadangi tyrimai apie dėstymą ir studijavimą tapo kompleksiškesni ir išsamesni, kilo daug naujų klausimų dėl tradicinio pozityvistinio požiūrio pagrįstumo. Pagal šį požiūrį dėstymas yra susijęs su efektyvumo įvertinimu. Geras dėstymas apibrėžiamas kaip idealių charakteristikų ar elgsenos normų rinkinys, kurio dėstytojas turi laikytis. Tačiau, kai kurie tyrinėtojai gerą dėstymą laiko globaliu konstruktu, priešingu savybių, elgsenos ar dimensijų rinkiniui. Pozityvistinio požiūrio tyrimuose pabrėžiama skaičiavimų problema; didelis dėmesys skiriamas studentų pateiktų reitingavimų patikimumui ir stabilumui, veiksnių, galinčių padaryti neigiamą įtaką įvertinimui, nustatymui. Atlikti tyrimai apie iš skirtingų šaltinių gautus įvertinimo privalumus ir trūkumus (studentų, bendraamžių, išorinių stebėtojų, administratorių vertinimus, savianalizę, ir t.t.), studentų pateiktų reitingavimų ir įvairių kintamųjų, pavyzdžiui, jų pasiekimų, santykį. Taip pat analizuotas įvertinimo formų vidinis nuoseklumas (punktų analizė) ir reitingavimų stabilumas laiko skalėje.

Pozityvistinio požiūrio šalininkai įsitikinę, kad studentų pateikti dėstymo kokybės reitingai yra patikimas šaltinis. Mokslininkai nustatė koreliacijas tarp studentų pateiktų reitingų ir jų pasiekimų[3] bei to paties dėstytojo reitingų pastovumo skirtingose grupėse. Daugiaplaniuose tyrimuose buvo įrodymų apie diskriminantinį ir konvergentinį reitingų pagrįstumą. Be to, nagrinėjant galimus kintamuosius, kurie neigiamai įtakoja reitingavimą paaiškėjo, kad neigiamos įtakos pagrįstumui gali turėti kurso pobūdis (pasirenkamas ar privalomas) ir dėstomas dalykas.

Pozityvistinio požiūrio kritikai teigia, kad pervertinama apibendrinimo ir priežasties bei pasekmės nustatymo svarba. Studentų pateiktų reitingų tyrėjai nepajėgė nustatyti pagrindinių gero dėstymo konstrukto elementų. Taip pat kritikuojamas per didelis dėmesys studentų reitingų metaanalizei, ypač kai ji sąlygoja administracinius sprendimus, kurie gali paveikti darbuotojų karjerą. Be to, ankstesni tyrimai apie dėstymo aukštosiose mokyklose įvertinimo pagrįstumą remiasi tradiciniu požiūriu.

Pagrindinis tyrimų trūkumas susijęs su įvertinamo konstrukto apibūdinimu. Nors buvo keletas bandymų taikyti apibendrinamumo ir išorinį aspektus vertinant studentų atsakymų pagrįstumą, konceptualusis, esminis ir pasekmių pagrįstumo aspektai dar nėra tyrinėti. Taip pat reikėtų išsiaiškinti ar įvertinimas objektyviai atspindi dėstymo kokybę, kaip sprendimų priėmėjai ir dėstytojai pritaiko gautas išvadas profesiniam tobulėjimui ar administraciniams sprendimams.

[3] Parodomų pažymiais

CONCLUSIONS

The evaluation of teaching in higher education has evolved from informal to systematic approaches as pressures for accountability increased in this level of education. In addition, as administrators began to worry about measuring outcomes, concerns have increased among the faculty about the fairness and use of evaluation results for making administrative decisions, such as tenure, promotion and salary increases.

As the research on teaching and learning have evolved to more complex understanding of these the teaching and learning processes, new questions are raised about the validity of the traditional positivistic approach for evaluating teaching in higher education. Under this positivistic approach, teaching is linked to the idea of treatment and evaluation to the idea of effectiveness. Good teaching is defined as a set of ideal characteristics or behaviors expected from the instructor. Although, some researchers define teaching as a global construct as opposite to a set of characteristics, behaviors or dimensions. Studies under the positivistic approach have been conducted with an emphasis on measurement issues, looking especially to the reliability and stability of student ratings of instruction, as well as on the study of several variables that could negatively influence the evaluation. Other studies have been conducted about the strengths and limitations of different evaluative sources (i.e. student ratings, peers, external observers, administrators, self-evaluation, etc.), and on the relationship between student ratings and some variables, such as student achievement. Reliability studies have been conducted about the internal consistency of the evaluation forms (item analysis), and stability of the ratings over time.

Supporters of the positivistic approach for evaluating teaching claim that student rating of instruction are reliable sources for evaluating teaching. The researchers have found correlations between student ratings and student achievement,[3] and stability of the ratings when comparing the ratings of the same instructor in different sections. Multi-trait studies have also found some evidence of discriminant and convergent validity of the ratings. In addition, researchers studying the possible variables that could negatively influence the ratings have found evidence of some biasing influence, primarily by course type (required versus elective) and course discipline as biasing factors influencing the ratings.

Critics of the positivistic approach state that the strong emphasis on generalization and the establishment of causal and effect linkages have been overstressed by the research. Studies on the dimensionality of the student ratings have failed to identify the essential elements of the construct “good teaching.” The over reliance on meta analyses of students on student ratings of instruction has also been questioned, especially when evaluation data is used for making administrative decisions that can affect faculty careers. In addition, prior research on the validity of the evaluations of teaching in higher education has been conducted using a traditional approach to validity.

But the main limitation of the research is about the validity of the evaluation in representing the construct being evaluated. Although some research on the validity of student ratings have been conducted on the generalizability of the ratings and their external validity. No research has been conducted on the conceptual, substantive and consequential validity aspects of the evaluation. There is also a need for understanding if the evaluation fairly represents the quality of teaching within its context, and how decision makers and teachers use evaluation results for professional development and for making administrative decisions.

[3] Defined as grades.

LITERATŪRA / REFERENCES

American Educational Research Association, American Psychological Association,& National Council on Measurement in Education. (1985). *Standards for Educational and Psychological Tests*. Washington, DC: American Psychological Association.

Abrami P. C. (1989). How Should We Use Student Ratings to Evaluate Teaching? // *Research in Higher Education*, 30(2), p. 221–27.

Abrami P. C., D' Apollonia S., & Cohen P. A. (1990). The validity of student ratings of instruction: What we know and what we don't // *Journal of Educational Psychology*, 82, p. 219–231.

Ballantyne C. (1998). *What students think: An innovative look at student evaluations of teaching*. Paper presented at the annual meeting of the American Evaluation Association, Chicago.

Brodie D. A. (1998). Do students report that easy professors are excellent teachers? // *The Canadian Journal of Higher Education*, 23(1), p. 1–20.

Brodie D. A. (1999). *Has publication bias inflated the reported correlation between student achievement and ratings of instructors*? Paper presented at the Annual Meeting of the American Educational Research Association. Montreal, CA, April.

Cashin W. E., Downey R. G. (1992). Using Global Student Rating Items for Summative Evaluation // *Journal of Educational Psychology*, 84(4), p. 563–572.

Centra J. A. (1993). *Reflective faculty evaluation: Enhancing teaching and determining faculty effectiveness*. San Francisco, CA: Jossey-Bass Inc.

Centra J. A., Bonesteel P. (1990). College teaching: An art or a science? In M. Theall and Franklin, Student ratings of instruction: Issues for improving practice // *New Directions for Teaching and Learning*, 43, p. 7–15.

Cisneros-Cohernour E. (1998). *Trade-offs: Using the feedback of student ratings for instructional improvement*. Champaign, IL: University of Illinois at Urbana-Champaign.

Cohen P. A. (1981). Student ratings of instruction and student achievement: A meta-analysis of multisection validity studies // *Review of Educational Research*, 51, p. 281–309.

Cohen P. A. (1983). Comment on selective review of the validity of student ratings of teaching // *Journal of Higher Education*, 54, p. 448–458.

Cohen P. A. (1986). *An updated and expanded meta-analysis of multisection student rating validity studies*. Paper presented at the Annual Meeting of the American Educational Research Association. San Francisco, CA, April.

Crombach L. J. (1989). Construct validation after thirty years. In Linn R. L. (Eds.), *Intelligence: Measurement, theory, and public policy*. Chicago: University of Illinois Press, p.147–171.

Doyle K. O., Jr. (1982). *Evaluating teaching.* Lexington, Mass: Lexington Books.

Dunkin M. J., Barnes (1986). Research on teaching in higher education. In. Wittrock Merlin C. (Ed.), *Handbook of research on teaching.* American Educational Research Association, New York: NY.

Dwinell P. L., Higbee J. L. (1993). Students' perceptions of the value of teaching evaluations // *Perceptual and Motor Skills,* 76, p. 995–100.

El Hassan K. (1995). Students' ratings of instruction: Generalizability of findings // *Studies in Educational Evaluation*, 21, p. 411–429.

Embreston (Whitely) S. (1983). Construct representation versus nomothetic span // *Psychological Bulletin*, 93, p. 179–197.

Erickson F. (1986). Qualitative methods in research on teaching. In Wittrock Merlin C. (Ed.), *Handbook of research on teaching.* American Educational Research Association, New York: NY.

Falk B., Dow K. L. (1971). *The assessment of university teaching*. London: Society for Research into Higher Education Ltd.

Feldman K. A. (1976). The superior college teacher from the student's view // *Research in Higher Education*, 5, p. 243–288.

Feldman K. A. (1986). The perceived instructional effectiveness of college teachers as related to their personality and attitudinal characteristics // *Research in Higher Education*, 24, p. 139–170.

Feldman K. A. (1988). Effective college teaching from the students' and faculty's view: Matched or mismatched priorities? // *Research in Higher Education*, 28(4), p. 291–344.

Feldman K. A. (1989). *The association between student ratings of specific instructional dimensions and student achievement: Refining and extending the synthesis of data from multisection validity studies*.

Frey P. W. (1976). Validity of student instructional ratings: Does timing matter? // *Journal of Higher Education*, 42(3), p. 327–336.

Genova W. J., Madoff M. J., Chin R., Thomas G. B. (1976). *Mutual benefit evaluation of faculty and administrators in higher education.* Newton, MA: Ballinguer Publishing Co.

Lane S., Parke C.S., Stone C. A. (1998). A framework for evaluating the consequences of assessment programs // *Educational Measurement: Issues and Practice,* 17(2), p. 24–28.

Mabry L. (1999). *Portfolios plus: A critical guide to alternative assessment*. Thousand Oaks, CA: Corwin Press, Inc.

Magusson K. (2000). Personal interview. July.

Marlin J. W. J. (1987). Student perception of end-of-course evaluations // *Journal of Higher Education,* 58(6), p. 704–716.

Marsh H. W. (1987). Student's evaluations of university teaching: Research findings, methodological issues, and directions for future research // *International Journal of Educational Research*, 11, p. 253–388.

Menges R. J. (1998). *Shortcomings of research on evaluation and improving teaching in higher education*. Paper presented at the Annual Meeting of the American Educational Research Association, San Diego: CA, April.

Messick S. (1989). Validity. In Linn R. L. (Ed.), *Educational Measurement* (3rd ed.). New York: American Council on Education and Macmillan.

Messick S. (1995). Validity of psychological assessment: Validation of inferences from persons responses and performances as scientific inquiry into score meaning // *American Psychologist*, 50(9), p. 741–749.

Messick S. (1994). *Alternative modes of assessment, uniform standards of validity*. Research Report.

Messick S. L. (1989). Validity. In Linn R. L. (Ed.), *Educational Measurement* (3rd Edition). New York: American Council on Education & MacMillan.
Miller R. I. (1974). *Developing programs for faculty evaluation*. San Francisco, CA: Jossey-Bass Inc.
Moss P. A. (1992). Shifting conceptions of validity in educational measurement: Implications for performance assessment // *Review of Educational Research*, 62, p. 229–258.
Moss P. A. (1996). Enlarging the dialogue in educational measurement: Voices from interpretive research traditions // *Educational Researcher*, 25(1), p. 20–28.
Ory J. C., Braskamp L. A., Pieper D. (1980). The congruency of student evaluative information collected by three methods // *Journal of Educational Psychology,* 72, p. 181–185.
Ory J., Ryan K. (2001). How do student ratings measure up to a new validity framework? *New Directions in Institutional Research*, 109. Jossey-Bass Inc., Publishers, San Francisco: CA.
Reckase M. D. (1998) Consequential validity from the test developer's perspective // *Educational Measurement: Issues and Practice,* 17(2), p. 13–16.
Ryan K., Johnson T. (1998*). Democratizing evalution: Meanngs and methods from practice*. Paper presented at the Annual Meeting of the American Evaluation Association, Chicago, IL, November.
Shepard L. Evaluating test validity // *Review of Educational Research*, 19, p. 405–450.
Shulman L. S. (1986). Paradigms and research programs in the study of teaching: A contemporary perspective. In Wittrock M. C. (Ed.), *Handbook of research on teaching.* American Educational Research Association, New York: NY.
Stake R. E., Cisneros-Cohernour E. J. (2000). *Situational evaluation of teaching on campus*. New Directions for Teaching and Learning. Jossey-Bass Inc., Publishers, San Francisco: CA.
Theall M. (1997). On drawing reasonable conclusions about student ratings of instruction: A reply to Haskell and to Stake // *Education Policy Analysis Archives*, 5.
Yen W. M. (1998). Investigating the consequential aspects of validity: Who is responsible and what should they do? // *Educational Measurement: Issues and Practice*, 17(2), p. 5–6.*ces*. Boston: Houghton Mifflin (second edition).

Įteikta 2005 m. gegužės mėn.
Delivered 2005 May

EDITH J. CISNEROS-COHERNOUR
Yucatan autonominio
universiteto docentė
Mokslinių interesų kryptys:
studijų programų įvertinimas, fakulteto akademinio personalo
įvertinimas ir profesinis tobulėjimas, organizacijos plėtra

EDITH J. CISNEROS-COHERNOUR
Associated professor of Education
at the Autonomous University of Yucatán.
Research interests:
educational program evaluation, faculty evaluation
and development, organizational development.

Yucatan autonominis universitetas
Edukologijos fakultetas

Universidad Autónoma de Yucatán
Facultad de Educación

Calle 50 No. 56, Entre 29 y 33
Colonia Francisco de Montejo Mérida
Yucatán, 97200 México
cchacon@tunku.uady.mx