Measurement Scales of Reactions to the Assessment of Graduate Programs: Evidences of Factorial Validity

Kelly Rocha de Queiroz[1]
Amalia Raquel Pérez-Nebra[2]
Fabiana Queiroga[1]

[1]Centro Universitário de Brasília, Brasília, Distrito Federal, Brasil
[2]Universidade de Brasília Brasília, Distrito Federal, Brasil

Abstract

The propose was to seek validity evidences of scales based on the model of reactions of higher education professors about the evaluation of graduate programs conducted by the Brazilian Federal Agency for Support and Evaluation of Graduate Education (Capes). The scales of satisfaction, justice perception, utility perception, and accuracy perception were applied on 814 higher education professors, being 50.36% males, with a mean age of 47.66 years ($SD = 9.34$). Exploratory analysis indicated reliability of the four scales (alphas ranged from .69 to .97 and omegas are from .70). These and other psychometric indicators of the scales indicate that the measures are reliable, and the reaction model was confirmed by the strong correlation between the scales.
Keywords: factor analysis; course evaluation; graduate programs; teaching work

Instrumentos de Mensuração de Reações à Avaliação da Pós-Graduação: Evidências de Validade Fatorial

Resumo

O objetivo deste estudo foi buscar evidências de validade de escalas elaboradas a partir do modelo de reações dos docentes de ensino superior acerca da avaliação da pós-graduação conduzida pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes). Foram aplicadas as escalas de satisfação, percepção de justiça, percepção de utilidade e percepção de precisão a 814 docentes de ensino superior, sendo 50,36% do sexo masculino, com média de idade de 47,66 anos ($DP = 9,34$). As análises exploratórias indicaram confiabilidade para as quatro escalas (alfas entre 0,69 e 0,97 e ômegas a partir de 0,70). Estes e outros indicadores psicométricos das escalas apontam que as medidas são consideradas confiáveis e o modelo de reação se confirma apresentando alta correlação entre as escalas.
Palavras-chave: análise fatorial, avaliação de curso, pós-graduação, trabalho docente

Instrumentos de Medida de Reacciones a la Evaluación del Postgrado: Evidencias de Validación Factorial

Resumen

El objetivo de este estudio fue buscar evidencias de validad de escalas elaboradas a partir del modelo de reacciones de los profesores acerca de la evaluación del postgrado efectuada por la Coordinación de Perfeccionamiento de Personal de Nivel Superior (Capes). Las escalas de satisfacción, percepción de justicia, percepción de utilidad y percepción de precisión fueron aplicadas a 814 docentes de enseñanza superior, siendo el 50,36% del sexo masculino, con promedio de edad de 47,66 años ($DT = 9,34$). Los análisis exploratorios señalaron fiabilidad de las cuatro escalas (alfas variaron entre 0,69 y 0,97 y omegas a partir de 0,70). Estos y demás indicadores psicométricos de las escalas apuntan que las medidas se consideran fiables y el modelo de reacción se confirma presentando una alta correlación entre las escalas.
Palabras clave: análisis factorial; evaluación de curso; posgrado; trabajo docente

Introduction

Since 1976, the Brazilian Federal Agency for Support and Evaluation of Graduate Education (Capes) has been conducting the evaluation of Brazilian graduate studies. This is an evaluation process that involves the participation of the scientific community and seeks to achieve a standard of academic excellence for national *stricto sensu* masters and doctorate courses. This model, besides certifying the quality of the National System of Graduate Programs (SNPG) courses, aims to identify regional asymmetries and strategic areas aiming at directing actions for the development of graduate programs (Brasil, 2018a).

According to the National Plan of Graduate Education 2011-2020 (Brasil, 2010), which aims to define guidelines, strategies, and goals for the development of Brazilian research and graduate studies, the current assessment of graduate programs basically consists of three axes: a) it is made by peers from different areas of knowledge and recognized for their intellectual reputation; b) it has a meritocratic nature

and; c) it associates recognition with encouragement in defining policies and establishing criteria for program funding (Brasil, 2010).

Every four years, the assessment comprises the nearly 4,400 graduate programs in operation, including masters and doctoral courses, academic and professional, from 49 subject areas. In the evaluation conducted in 2017, for the 2013-2016 quadrennium, the results were expressed in grades 1 to 7, and in order to be recommended to stay in SNPG courses they should reach grade 3 (Brasil, 2017). Regarding the evaluation that will be carried out in 2021, for the 2017-2020 quadrennium, the courses that receive grade 3 will be maintained if they are composed only by a master's degree, but those programs with a master's and doctorate degree or those with a doctoral level will be deactivated. (Brasil, 2018b).

Although successful and undoubtedly legitimate in ensuring the good quality of SNPG over the years, assessment of graduate programs does not occur without criticism by higher education professors. Voguel (2015), when interpreting criticism from the academic community through content analysis whose corpus consisted of Capes documents and scientific publications, found that the published comments on the evaluation in 62 articles from 1997 to 2013 were significantly positive (63%), although there were some negatives (21%) and others of suggestive character (16%). Negative criticisms were especially related to criteria such as the use of bibliometric indicators, the strata used for the classification of journals, the definitions of internationalization and endogeny, the attention to teaching, the formation of human resources and the geographical distribution.

In addition to external criticism from the scientific community, there is also internal criticism. In February 2018, the Capes Higher Technical and Scientific Council of Higher Education (CTC-ES) pointed out some considerations in Capes internal document with the purpose of subsidizing the improvement of the assessment of graduate programs, based on the dispositions of 40 coordinators of areas that participated in the evaluation process for the 2013-2016 quadrennium. In general, the recommendations referred to the valorization of other forms of intellectual production, the need to accompany graduates, the reduction of regional asymmetries, the clear disclosure of the excellence parameters for each area, the focus on the merits of the programs, among others (Brasil, 2018c). Given the above, there is evidence that leads to consider

that in the eyes of the academic community, responsible for integrating the academic world with scientific and technological research, the assessment of graduate programs has some criteria that may be considered unsatisfactory, unjust, inaccurate or even useless.

At the same time as it is observed that the evaluation mentioned here refers to *programs* (as Capes is evaluating the graduate programs). It is also noted that there are certain peculiarities when compared to the classical evaluation of a program (Cunha, 2018; Souza, Abbad, & Gondim, 2017). The first point is that the purpose of this paper is not to analyze the effectiveness of programs. This is the objective related to Capes itself. The second point is that, unlike program evaluations, the unit of interest is not within what is classically defined as a *program[1]*, but rather within a program evaluation system. Although we do not intend to judge the effectiveness of the evaluation system, this study focuses on the analysis of part of its effects. It is in this sense that, inspired by models that advocate what is called *vertical compatibility*, professors are now considered the target of the analysis.

The vertical compatibility feature presented in the model by Rico, Hinsz, Burke and Salas (2017) deals with the relationship between different performance levels and how they may be conflicting or collaborative (e.g. performance by the professor, the program and by the higher education institution). In the case of graduate studies, collective performance (program) and individual performance (professor) tend to be collaborative, i.e. researcher performance affects program performance. In this case, the lowest order goals (professor or program) are aligned with the highest order goals (program or SNPG, respectively). But the model presented by the authors is limited to performance and teams. It does not suggest the inclusion of psychological variables, such as reactions, but it discusses possible incongruities between different goals (e.g. goals and values of the program and SNPG's goals and values) and how this misalignment may impact on different performance levels (probably due to reactions that people have with the assessment performed). Thus, the elaboration of reaction scales is an important initial step to understand the effects of Capes evaluation.

According to Fachada (2012) and Keeping and Levy (2000), the opinion (i.e. reactions) about organizational practices, including performance evaluations,

---

[1]   Set of activities organized to be conducted within the specific schedule and budget available for the implementation of policies or conditions that allow the achievement of desirable goals (Cunha, 2018).

has been related to the acceptance, feasibility and success of these evaluative models, although they are neglected in scientific research on the subject. In Brazil, the training and development literature has a long tradition of measurement-related research culminating in a complete book on the subject (Abbad, Mourão, Meneses, Zerbini, Borges-Andrade, & Vilas-Boas, 2012), including, among others, the training reaction (Abbad, Zerbini, Borges-Ferreira, 2012; Borges-Ferreira & Abbad, 2009). While a reaction to training may seem different from a performance appraisal reaction, it contains a common element at its core: the assumption that satisfaction is an important element in achieving positive effects. The object of this satisfaction varies according to what is being evaluated, but in psychological terms the phenomenon is equivalent.

There are several empirical reports of measures of employee reactions to performance reviews by organizations. Different surveys converge with satisfaction as a common element and others varying according to context (Dusterhoff, Cunningham, & MacGregor, 2014; Harrington & Lee, 2015; Keeping & Levy, 2000; Kim, 2016; Kim & Holzer, 2016; Lira, 2014; Lira, Silva, & Viseu, 2016). The most complete model found was the one presented by Keeping and Levy (2000). However, other researches presented similar objects to the present one, such as the study by Fachada (2012) which analyzed the satisfaction of a public performance evaluation system and that of Lira et al., (2016) on justice and accuracy in the context of graduate studies.

Keeping and Levy (2000), when proposing a reaction measure to the performance evaluation, use satisfaction as variables and included other elements: the perception of justice, the perception of utility, and the perception of accuracy, since they were considered relevant constructs in their systematic literature review (Cawley, Keeping, & Levy, 1998). Perceptions of fairness, utility, and accuracy in different management practices are addressed at other stages, but for performance evaluation they find special value and occur after the evaluation.

The Brazilian assessment of graduate studies aims to analyze the results of *stricto sensu* graduate programs, therefore, with macro level of analysis. This means that it requires an instrument that is built from the perception of this evaluation at the same level as well. Traditionally, psychology uses as a multilevel methodology the aggregation of observed individual scores (Bennink, Croon &, Vermunt, 2015), but it is known that this technique can lead to several biases. In

this sense, there is a clear gap in the construction of instruments whose variables are already constructed at another level of analysis, in relation to repercussions or with objects at another level.

Therefore, considering that Capes' assessment of graduate research differs in that it is the assessment of a program that refers to the consequences of the behavior of a group (collegiate) and the individual, a parallel was made with the model formulated by Keeping and Levy (2000) – which addresses the reactions of the evaluated researchers with the individual performance assessment – since it deals with the perception phenomenon about an evaluation system. The Keeping and Levy (2000) model proposes that the reactions be measured from the satisfaction, perception of justice, utility, and precision of the evaluation system. *Satisfaction* with performance assessments is the basis for important decisions in organizations, as that this attitudinal variable corresponds to the recognition of these assessment processes (Judge & Kammeyer-Mueller, 2012), as well as their future perspectives (Fachada, 2012). Corroborating this position, Dusterhoff et al., (2014) in their research showed that those employees who are most satisfied with the performance assessments tend to agree with its outcome, and find this process useful for developing their performance.

The *perception of justice*, within the scope of Psychology, focuses on the personal experiences of individuals in situations which are considered by them to be (un) fair, i.e., by comparing what is subjectively perceived as (un) fair (Greenberg, 2011). Regarding its dimensions, the perception of justice is categorized into four strands: distributive justice, procedural justice, interpersonal justice, and informational justice, depending on the level of research analysis (Colquitt, 2001; Colquitt, Wesson, Porter, Conlon, & Ng, 2001). However, the perception of justice of higher education professors regarding the assessment of graduate programs in the present study was analyzed from a two-dimensional perspective, that is, distributive and procedural, as presented by Keeping and Levy (2000), since the interactional perspective of justice (interpersonal and informational) in graduate evaluation would be better understood by another sample group, such as Capes' area coordinators.

Regarding the *perception of utility*, it is a component that in performance assessments captures the extent to which the employee has learned important information about the assessment, such as how he can do a better job, how to develop his skills, and whether the evaluated meets the expectations of the evaluator (Greller,

1978, quoted in Payne, Horner, Boswell, Schoeder, & Stine-Cheyne, 2009). In this sense, the perception of utility comprises adherence to the function and objectives of the evaluation. It is noteworthy that in the present research the perceptions about the usefulness of the evaluative model for the purpose of improving the SNPG and the graduate programs were measured.

Finally, it is worth highlighting the *perception of accuracy* component. According to Lira (2014), some argue that the individual perception of the evaluated influences the understanding of the accuracy of the evaluation process to which they were submitted. Indeed, the results found by this author indicate that there is a positive relationship between employee satisfaction and the perception of accuracy regarding performance appraisal. Considering the various criticisms of the academic community regarding the assessment of graduate programs (e.g. Brasil, 2018c; Voguel, 2015), we analyzed the perception of accuracy, as a psychological phenomenon and not in relation to the accuracy indicators, which higher education professors have about the assessment of graduate studies.

Thus, considering the importance of both Capes' evaluation of graduate programs and the collegiate's reaction to it, the overall objective of this study was to produce evidence of validity of the scales of satisfaction, perception of justice, perception of utility, and of perception of precision regarding the assessment of graduate programs. These measures aim to foster research on the impact of reactions to program evaluation, based on the premise of relationship between levels that, in this case, individual performance (professor) is an important part of collective performance analysis (program).

Method

Participants

The criterion of participation in the research was that the respondent was linked to a *stricto sensu* academic graduate program in operation, that is, which had a grade equal to or higher than 3 in the evaluation of the graduate program. Professors linked to professional stricto sensu graduate programs were excluded from the analysis. In total, 814 higher education professors participated in this study. Most of the sample consisted of male university teachers (n = 410; 50.36%) who have an institutional relationship of full professor (n = 650; 79.85%), with a mean age of 47.66 years (SD = 9.34) and who worked in public higher education institutions (n = 549; 67.44%). It is noteworthy that,

although there are responses from professors from 152 higher education institutions, the highest incidence was in a university in the state of São Paulo. It is noteworthy that 19.65% of the participants did not fully answer the instrument, leaving some questions of the questionnaire blank. In these cases, only the respondents to each question were considered.

Procedures of design and adaption of scales

The research instrument used to analyze the professor's reactions about the assessment of the graduate program was composed by scales of satisfaction, justice perception, utility perception, and accuracy perception. It is noteworthy that the elaboration of the instrument was based on studies whose scales could be adapted considering the specificity of the evaluation of the graduate program and the participants.

Scale of Satisfaction with Assessment of Graduate Studies. The satisfaction scale regarding the assessment of graduate programs was formulated from the instrument used in the survey conducted by Fachada (2012). The original scale was single factor (α = 0.96). Seven items out of a total of twenty were selected because they were compatible with the study theme due to the specificity of the assessment of the graduate program and the selection criteria were: items that referred to general satisfaction, policy satisfaction and characteristics of the evaluation system, and satisfaction with the evaluation results. Items regarding satisfaction with procedures were not included, as it was an aspect addressed in the perception of justice scale, and satisfaction with the communication of the program, because the original items referred to a level of analysis. In addition, minor adjustments were made in the wording of the items in order to adapt the text to the research subject. As an example, the original item was "How satisfied are you with the existence of performance assessment in your institution?" And in the version of the survey instrument, it was changed to "How satisfied are you with the existence of assessment of graduate programs by Capes?". There was also adaptation to Brazilian Portuguese, as the scale was originally written in Portuguese from Portugal. Furthermore, based on the suggestion of some higher education professors after semantic validation and judges, two items were also included, in addition to those provided in the survey by Fachada (2012): "How satisfied are you with the level of collegiate internal demand from your graduate program?" and "How are you satisfied with the peer review system performed in the assessment of graduate

research?". The scale was composed of nine items and, for the measurement, they all presented amplitude of six points (from 0 to 5), anchored in the extremes in satisfaction: very satisfied and very dissatisfied.

Perception of Justice Scale. With a total of eleven items, the justice perception scale for the assessment of graduate programs was adapted from the constructs of Colquitt (2001) and Lira et al., (2016). Regarding the items that refer to the perception of distributive justice, all four items of the Colquitt (2001) scale (single factor α = 0.92) were used. In addition to translating the items into Portuguese, changes were made in the wording of the items to better adapt to the study theme. As an example, the item "Does your (outcome) reflect what you have contributed to the organization?" became "The grade obtained reflects the contribution of my graduate program to the National System of Graduate Programs (SNPG)." Regarding the items that refer to the perception of procedural justice, we used six items, out of eight, from the scale of Lira et al., (2016) (one-factor scale α = 0.97), as they were the ones that best fit the study theme due to the specificity of the graduate evaluation. Excluded items were "I was able to express my opinions and feelings during these procedures" and "I was able to influence the decisions that resulted from these procedures, namely regarding my final ratings". An item was also included as a result of the recommendation of an evaluating professor after the process of semantic validation and of judges, namely: "The procedures used in the assessment of graduate programs are transparent". Responses were measured by a six-point amplitude scale (0 to 5) with agreement anchor.

Utility Perception Scale. The utility perception scale of the graduate assessment was prepared based on the objectives and criteria of the Quadrennial Evaluation, conducted by Capes in July and August 2017, established in Ordinance n. 59/2017 (Brasil, 2017) by Capes. In total, the scale had nine items with amplitude of six points (from 0 to 5) and agreement anchor.

Accuracy Perception Scale. Regarding the accuracy perception scale of the assessment of graduate programs, the questionnaire was adapted from the scale of Lira et al., (2016) (original α = 0.90). Four of the five items of the original instrument were selected. Therefore, only the item "performance assessments precisely and accurately reflect my true performance" was not used in this research because it is an individual assessment that differs from Capes' program assessment. It is noteworthy that the scale was adapted to Brazilian Portuguese, since the scale was formulated in Portuguese

from Portugal. There was also a change in the wording of the items to better suit the study theme after the validation of judges. As an example, the original item was "Performance appraisals were rigorous and accurate" and was adapted to "Capes assessment of graduate programs is accurate". Importantly, two of the four items on the scale had the inverted wording. Responses were indicated on a six-point amplitude scale (from 0 to 5) with agreement anchor.

Semantic and Judges Validation

In order to perform the semantic validation of the instrument (Pasquali, 2010), the questionnaire was applied in person to four university professors linked to *stricto sensu* graduate programs from three higher education institutions in the Brazilian Federal District (DF) and one in the state of Bahia. In addition, a professor was asked to respond by e-mail to the survey questionnaire with her considerations, as she was teaching at a higher education institution in the state of São Paulo. It is noteworthy that in general there were few semantic suggestions by participants regarding the understanding the items. But, the items were evaluated as relevant and in the expected factors. Moreover, it is noteworthy that, at first, the intention was to perform only a semantic validation, however, considering that the professionals had knowledge and interest in the study theme, some teachers suggested inclusions: three items in the questionnaire, two in the satisfaction scale and one on the perception of justice scale.

Data Collection Procedures

Built from the Survey Monkey electronic platform, the self-applying instrument was mailed to 3,635 academic stricto sensu graduate program coordinators whose email addresses were taken from the Sucupira Platform. In addition, the coordinators were asked to send to the other faculty members of their graduate programs the invitation to access the data collection system in order to provide answers from a larger number of professors who were not graduate coordinators.

Data Analysis Procedures

The data analysis procedures were performed from some steps of the Statistical Package for the Social Sciences software. Data were examined for normality and distribution of discrepant values and considered relatively normal. It should be noted that there was no extreme multivariate case. In addition, the data were subjected to principal component analysis (PC) to verify factorability

and then to principal axis factorization (PAF) with promax oblique rotation, when more than one factor was expected. The reliability of the instruments was measured by calculating Cronbach's alpha (α) and omega (ω) coefficients, and, for the validation of criteria, correlation analyzes were performed between the scales.

Ethical Procedures

This research was submitted and approved by the Research Ethics Committee from UniCEUB (CAAE 77661317.1.0000.0023). It is noteworthy that, prior to the first scale of the instrument, an Informed Consent Form (ICF) was included in which the respondent was required to indicate agreement to participate in the research. University professors were informed of the research objectives, and were warned that their answers would be anonymous and would be analyzed

collectively. The answers were not passed on to the professional' graduate coordinators (or any individual). No analyzes were performed that could recognize the participants.

Results

The scales used in the research were subjected to exploratory and reliability factor analyzes separately.

Scale of Satisfaction with Assessment of Graduate Programs. The scree plot plotted a single factor scale. The results from the Kaiser-Meyer-Olkin test (KMO = 0.86) and Bartlett's sphericity test (approximate Chi-square = 3302.58; gl 36; p <0.000) demonstrated suitability for factor analysis. Thus, a principal factor factorization (PAF) was performed with a single factor solution.

Table 1.

Psychometric Characteristics of the Satisfaction Scale About Assessment of Graduate Programs

Items How satisfied are you with... *Em que medida você está satisfeito com…*	**Factorial Load** (α = 0.87; ω = 0.87)	**h²**
1. ... the existence of graduate assessment. *1. ... a existência da avaliação da pós-graduação.*	0.64	0.42
2. ... the criteria adopted by Capes for the graduate assessment. *2. ... os critérios adotados na avaliação da pós-graduação efetuada pela Capes.*	0.77	0.59
3. ... the periodicity of Capes' graduate assessment. *3. ... a periodicidade da avaliação da pós-graduação efetuada pela Capes.*	0.53	0.29
4. ... the inexistence of self-assessment of graduate programs by Capes. *4. ... a inexistência de autoavaliação da pós-graduação efetuada pela Capes.*	0.49	0.24
5. ... the way Capes disseminates information about the assessment of graduate programs and the results of this process. *5. ... a forma com a Capes divulga as informações acerca da avaliação da pós-graduação e os resultados desse processo.*	0.68	0.46
6. ... the score received considering the quality of work performed by your graduate program. *6. ... a nota recebida considerando a qualidade do trabalho desenvolvido pelo seu programa de pós-graduação.*	0.78	0.62
7. ... the score received considering the amount of work performed by your graduate program. *7. ... a nota recebida considerando a quantidade de trabalho desempenhado pelo programa de pós-graduação.*	0.78	0.61
8. ... the collegiate internal level of demand from your graduate program. *8. ... o nível de cobrança interna do colegiado do seu programa de pós-graduação.*	0.42	0.18
9. ... with the peer review system for the assessment of graduate research. *9. ... com o sistema de avaliação por pares efetuada na avaliação da pós-graduação.*	0.76	0.58

Factor loadings above 0.40 and the relatively high commonalities (h^2) indicate that the factor explains the high percentage of variance between the items. It was found that the explained total variance of the satisfaction scale regarding the graduate assessment composed of nine items was of 44.27%. Finally, Cronbach's alpha and omega coefficients of 0.87 indicated the reliability of the scores generated by the scale.

Perception of Justice Scale. The scree plot indicated that the perception of justice scale regarding assessment of graduate studies can be considered bifactorial, as suggested by the scientific literature (e.g. Colquitt, 2001; Colquitt et al., 2001; Greenberg, 2011). Satisfactory factorability was found by checking the KMO measurement indices (0.94) and Bartlett's sphericity test (approximate Chi-square = 8582.97; gl 55; p <0.000). Thus, these factors were examined by factor analysis of the main axes with promax rotation.

The total variance explained on the two-factor scale is 73.77%. As expected, the items that make up the first factor describe a dimension of Procedural Justice ($\alpha = 0.92$; $\omega = 0.91$), while those that make up the second factor describe a dimension of Distributive Justice ($\alpha = 0, 97$; $\omega = 0.97$).

Utility Perception Scale. The scree plot indicated the existence of two factors on the utility perception scale of the assessment of graduate programs. The KMO measurement index (0.93) and Bartlett's sphericity test (approximate Chi-square = 7327.12; gl 36; p <0.000) indicated the appropriateness of performing the factor analysis. It is noteworthy that the factors were examined through a factor analysis of the main axes with promax rotation.

It was found that the bifactorial structure was quite adequate for the construct, explaining 80.62% of the total variance of the factorial matrix. The items that make up the first factor concern the utility of the assessment of the graduate program oriented to the improvement of the SNPG ($\alpha = 0.96$; $\omega = 0.95$), while the items that make up the second factor refer to the utility of the assessment of graduate studies oriented to graduate programs and their professors ($\alpha = 0.93$; $\omega = 0.91$).

Accuracy Perception Scale. Considering that the scree plot graph does not have the "elbow" - a point that would demonstrate a linear descent trend of the eigenvalues -, and considering that the accuracy perception scale contains only four items, we decided to treat it as a single factor construct. The KMO indicated a mediocre factorability with a value index of 0.57 (Hutcheson

& Sofroniou, 1999, cited in Field, 2009), however no impact on factor loadings was observed (values above 0.30). In addition, Bartlett's sphericity test indicated the following result: Approximate Chi-square = 869.04; gl 6; p <0.000.

It is noteworthy that the items "the grades obtained are based more on favoritism than on the actual performance of the graduate programs" and "there is a tendency for evaluators to give the same rating to all evaluated programs regardless of their actual performance" have the order of the sentences inverted. The total explained variance of the post-graduation rating perception scale is 40.43% and Cronbach's alpha was 0.69 and omega of 0.71, the lowest indices of all scales, but still acceptable.

In order to produce other validity indicators, the constructs were correlated. It was found that all scales correlate with each other, with indices between 0.49 and 0.80.

In summary, all factors have positive and strong (or at least moderate) correlations with each other, since Pearson's correlation coefficients are greater than 0.49 (p <0.05). The correlation between satisfaction and the perception of procedural justice about assessment of graduate programs is noteworthy (r = 0.80; p <0.01), indicating that the perception of evaluation procedures strongly affects the satisfaction with the evaluation process. Another noteworthy relationship refers to perceptions of precision and procedural fairness (r = 0.74; p <0.01), which points out that teachers' perception of precision regarding this evaluation model strongly impacts their perceptions regarding procedural fairness of the assessment.

Discussion

The objective of this study was to search for evidence of the validity of reaction scales of higher education professors regarding the evaluation of graduate studies conducted by Capes and it is understood that the objective has been achieved. To estimate the reactions of the faculty members of the graduate programs about the quadrennial evaluation performed by Capes, the Keeping and Levy (2000) model was used, which comprises the reactions as being determined by the satisfaction with the evaluation system and the perceptions of justice, utility and accuracy. It was found that the four scales can be considered adequate, given that the factor loadings of their items are high, as well as internally consistent: omegas above 0.70 in all scales

Table 2.

Psychometric Characteristics of the Assessment of Graduate Programs Justice Scale

Items	I (α = 0.92; ω = 0.91)	II (α = 0.97; ω = 0.97)	h^2
1. The score obtained reflects the effort that my graduate program has put into work. *1. A nota obtida reflete o esforço que o meu programa de pós-graduação colocou no trabalho.*		0.97	0.91
2. The score obtained is appropriate in view of the accomplishments of my graduate program. *2. A nota obtida é apropriada tendo em vista o realizado pelo meu programa de pós-graduação.*		0,99	0,95
3. The score obtained reflects the contribution of my graduate program to the National System of Graduate Programs (SNPG). *3. A nota obtida reflete a contribuição do meu programa de pós-graduação ao Sistema Nacional de Pós-Graduação (SNPG).*		0,92	0,88
4. The score obtained by my Graduate Program is fair. *4. A nota obtida pelo meu programa de pós-graduação é justa.*		0,91	0,90
5. The procedures used in the assessment of graduate programs are fair, appropriate and impartial. *5. Os procedimentos utilizados na avaliação da pós-graduação são justos, apropriados e imparciais.*	0,67		0,73
6. The procedures were performed within the legally established deadlines. *6. Os procedimentos foram realizados nos prazos legalmente estabelecidos.*	0.52		0.24
7. The procedures used in the assessment of graduate programs were applied consistently. *7. Os procedimentos utilizados na avaliação da pós-graduação foram aplicados de forma consistente.*	0.76		0.73
8. The procedures used for the assessment of graduate programs are based on precise and accurate information. *8. Os procedimentos utilizados na avaliação da pós-graduação se baseiam em informações precisas e rigorosas.*	0.82		0.71
9. These procedures are free from preconceived ideas. *9. Esses procedimentos se encontram em livres de ideias pré-concebidas.*	0.83		0.66
10. The procedures used in graduate assessment are based on ethical and moral standards. *10. Os procedimentos utilizados na avaliação da pós-graduação se apoiam em padrões éticos e morais.*	0.87		0.67
11. The procedures used in the assessment of graduate programs are transparent. *11. Os procedimentos utilizados na avaliação da pós-graduação são transparentes.*	0.85		0.73

Note. I = factorial loads of procedural justice perception items; II – factorial loads of the items of perception of distributive justice.

Table 3.

Psychometric Characteristics of the Graduate Assessment Utility Perception Scale

Items	I (α = 0.96; ω = 0.95)	II (α = 0.93; ω = 0.91)	h^2
1. I find the assessment of graduate programs useful in ensuring the quality of Brazilian graduate research. *1. Considero a avaliação da pós-graduação útil para garantir a qualidade da pós-graduação brasileira.*	0.93		0.80
2. I consider the assessment of graduate programs useful to portray the situation of Brazilian research in the period evaluated. *2. Considero a avaliação da pós-graduação útil para retratar a situação da pós-graduação brasileira no período avaliado.*	0.88		0.78
3. I find the assessment of graduate programs useful for the development of my program. *3. Considero a avaliação da pós-graduação útil para o desenvolvimento do meu programa.*	0.90		0.87
4. I find the assessment of graduate programs useful for the development of the area of knowledge in which my program is inserted. *4. Considero a avaliação da pós-graduação útil para o desenvolvimento da área de conhecimento a qual o meu programa está inserido.*	0.84		0.86
5. I consider assessment of graduate programs useful for the development of the National System of Graduate Programs (SNPG). *5. Considero a avaliação da pós-graduação útil para o desenvolvimento do Sistema Nacional de Pós-Graduação (SNPG).*	0.85		0.83
6. Capes' assessment of graduate programs is useful for me to better understand what is expected of my graduate program regarding its scientific output. *6. A avaliação da pós-graduação efetuada pela Capes é útil para que eu entenda melhor o que se espera do meu programa de pós-graduação acerca de sua produção científica.*		0.66	0.71
7. Capes' assessment of graduate programs is useful for me to better understand what is expected about the training of masters and / or doctors in my graduate program. *7. A avaliação da pós-graduação efetuada pela Capes é útil para que eu entenda melhor o que se espera acerca da formação de mestres e/ou doutores no meu programa de pós-graduação.*		0.91	0.85
8. Capes' assessment of graduate programs is useful for me to better understand what is expected of the social inclusion of my graduate program. *8. A avaliação da pós-graduação efetuada pela Capes é útil para que eu entenda melhor o que se espera da inserção social do meu programa de pós-graduação.*		0.88	0.78
9. Capes' assessment of graduate programs is useful for me to better understand what is expected of my training and updating as a professor. *9. A avaliação da pós-graduação efetuada pela Capes é útil para que eu entenda melhor o que se espera da minha formação e atualização como docente.*		0.89	0.78

Note. I = Factorial loads of the perceived utility of items of the assessment of graduate programs focusing on improving the SNPG. II - Factorial loads of items of perceived perception of assessment of graduate programs focusing on graduate programs (PPG).

Table 4.
Psychometric Characteristics of the Graduate Assessment Precision Perception Scale

Items	Factorial Load ($\alpha = 0.69$; $\omega = 0.71$)	h^2
1. The assessment of Graduate programs, conducted by Capes, is accurate. *1. A avaliação da pós-graduação, efetuada pela Capes, é precisa.*	0.84	0.70
2. The grades obtained are based more on favoritism than on the actual performance of graduate programs. *2. As notas obtidas se baseiam mais em favoritismos do que no real desempenho dos programas de pós-graduação.*	-0.43	0.18
3. Assessment of Graduate programs accurately describes strengths and weaknesses of programs. *3. A avaliação da pós-graduação descreve com precisão as forças e fraquezas dos programas de pós-graduação.*	0.80	0.63
4. There is a tendency for evaluators to rate everyone equally regardless of the actual performance of graduate programs. *4. Existe uma tendência para os avaliadores darem a mesma nota a todos os avaliados independente do real desempenho dos programas de pós-graduação.*	-0.32	0.10

Table 5.
Descriptive Statistics and Bivariate Correlations between Factors

	Mean	SD	1	2	3	4	5
1 Satisfaction *1 Satisfação*	3.15	0.97					
2 Distributive Justice *2 Justiça Distributiva*	3.17	1.53	0.75*				
3 Procedural Justice *3 Justiça Procedimental*	3.06	1.16	0.80*	0.68*			
4 Utility – SNPG *4 Utilidade – SNPG*	3.83	1.18	0.67*	0.49*	0.60*		
5 Utility – PPG *5 Utilidade – PPG*	3.32	1.29	0.65*	0.49*	0.63*	0.72*	
Accuracy *Precisão*	2.87	1.00	0.67*	0.57*	0.74*	0.52*	0.55*

Note. * Correlation is significant at level 0.01.

and similar to the indices found in the scales that served as inspiration for its construction.

The original intention was to construct scales at the supra-individual level, but some phenomena were considered very complex from a point of view of a group (e.g. "my group is pleased with"). Thus, the different scales ended up built on relatively different levels. Some at the individual level (Satisfaction - "are you satisfied with"; Utility - "I consider graduate assessment useful for Brazilian graduate research" and "I consider evaluation to be useful for my program"), group (distributive justice "reflects my program") and

contextual (procedural justice "evaluation procedures" and precision - "evaluation is accurate"). Although the levels of analysis are different, they are strongly related and similar to the original Keeping and Levy (2000) model. This result is consistent with the reactions, from the individual assessment, even if the object of scale assessment is at the supra-individual level, it remains relatively individual.

Satisfaction about graduate assessment is comprised of items that address the criteria, quality and quantity of work, and peer review. Thus, the scale included some of Capes' assessment axes. This scale added the most items in a single factor, suggesting that the construct is varied. It is noteworthy that the scale is about satisfaction with the program evaluation, which differs from the performance evaluation of the researcher/professor.[2] A performance assessment is generally about an organization's objectives, goals, and values, but Capes' assessment may not be consistent with the program's specific objectives, goals, and values, and this distance can create conflict (Rico et al., 2017), besides being mediated by other variables. Moreover, although satisfaction with the assessment is made up of the aforementioned dimensions, it is curious to note that the lack of self-assessment and the degree of internal charge of the graduate program have been positively aggregated (albeit with a lower factor load), since the evaluation system is not composed of these dimensions. In other words, the demands "come in the package" and not the direct listening of the professor in the process.

The perception of procedural justice adds items such as procedures with little bias, transparency, understandable information, that is, with rules established and clear to all. It is noteworthy that the item related to the procedures being fair, appropriate and impartial had a relatively lower load than the others, as well as the item related to the established deadlines. Impartiality in the criteria deserves to be discussed. The criteria established today by Capes are formed by senior researchers who are adapted to the system, so the maintenance of this evaluative model seems rational and natural, since these researchers have status for having been successful with these rules. So why change them? Evidently, this is a provocation that aims to draw attention to the data and also to a possible antagonism that may occur that will be open for future studies. Distributive justice

appears quite coherently and concisely as it has been presented in the scientific literature (e.g. Colquitt, 2001; Colquitt et al., 2001; Greenberg, 2011). This data suggests that the professors' perception regarding the evaluation score, i.e., with the result, is understood in a coherent and similar way among the respondents.

The utility scale having split into two constructs was unexpected. Indeed, the utility orientation for SNPG and graduate programs is clear and distinct. That is, the effect here was from the object analysis level. The first factor describes the utility of the evaluation of Brazilian graduate studies in a broad sense, adding most of the explanation (71.39%). The second factor has a lower explanation (9.23%) and approaches, in conceptual terms, a clarity of roles, because it adds items that address what is expected of the role played by a teacher in graduate school. It is noteworthy that both factors have excellent psychometric quality (Field, 2009).

The precision scale regarding graduate evaluation is composed of lack of favoritism, accurate program description, and differentiation between programs. It is noteworthy that this scale is composed of reverse items. Although already pointed out as a practice that degrades the accuracy of the scale (Suárez-Alvarez, Pedrosa, Lozano, Garcia-Cueto, Cuesta, & Muñiz, 2018), the presentation of items based on critical incidents has helped to write more clearly about the phenomenon. Another point to note about accuracy is that this term is often delicate for researchers, and different colleges - Humanities, Life Sciences, and Exact, Technological, and Multidisciplinary Sciences - may have different definitions and criteria regarding what precision means and how it is operationalized.

Similarly to Keeping and Levy (2000), Dusterhoff et al., (2014), Lira (2014) and Lira et al., (2016), it was found in this research the relationship between satisfaction, fairness, utility and accuracy. It is noteworthy that this relationship has implications from the measure point of view, but also from the practical point of view. In terms of measurement, it validates the reaction proposal about graduate evaluation and validation of scale criteria.

From a practical point of view, it suggests that the teacher's reactions to Capes' graduate assessment come from the various constructs used and that both the assessment system and the area coordinators should consider them when designing and establishing program evaluation policies. In general, the means of the constructs were above the midpoint of the scale (ranging from 0 to 5), except for accuracy (M = 2.87).

---

[2] We thank the reviewer for the suggestion and comment on this point.

The utility means were the highest, so that professors understand that the evaluation is useful, but needs to improve its accuracy, i.e., the criteria used for evaluation. This result is in line with what was found in the different criticisms of graduate studies (Brasil, 2018c;Voguel, 2015).

Although the research has achieved its goal, it is worth pointing out some limitations. The first of these concerns access to participants. It seems incoherent, but researching researchers is a challenge: adherence is low, several coordinators did not pass the research to their collegiate, but also several feedbacks emerged about the importance and impact that this work could have. Another difficulty from the methodological point of view is the adaptation of the scale, which, although careful writing, content, judges occurred, some variables were not addressed at various levels of analysis (e.g. "how much I realize that my collegiate was satisfied with " rather than " I'm satisfied with "). Although this may be a limitation, since the result of correlation with the scales with different levels of analysis were similar to the study that was conducted entirely at the individual level, it suggests that this type of care may ultimately be important but not of fundamental importance for studies of this type and that the aggregate method is a valid output as usual in the area and that correction should occur from other techniques (Bennink, Croon, & Vermunt, 2015). The precision scale was the one that achieved the lowest reliability score and was the most complex to adapt and elaborate items. It is also the major focus of criticism by professors, which may also have influenced their reliability. This scale provides room for review, although it is acceptable.

Finally, it is noteworthy that the proposal to elaborate reaction scales is an innovation within the national context. Thus, the unresolved problems that arise from the results found in this research are: what are the possible consequences of these reactions to the graduate evaluation in terms of the performance of graduate programs? That is, how do micro level variables impact the macro level? Van Veldhoven (2012) describes that research, in general, is concerned with how macro-level variables impact on micro-level variables, but the reverse is still poorly known (micro-macro) and professors' reactions to the assessment of programs may have an influence on the overall performance of the program. Finally, other methodological approaches may deepen issues such as understanding what the evaluation system is, accuracy and justice.

References

Abbad, G. da S., Mourão, L. Meneses, P. P. M., Zerbini, T. Borges-Andrade, J. E., & Vilas-Boas, R. (2012). *Medidas de avaliação em treinamento, desenvolvimento e educação: Fundamentos para gestão de pessoas.* Porto Alegre: Artmed.

Abbad, G. da S., Zerbini, T., & Borges-Ferreira, M. F. (2012). Medidas de reação a cursos presenciais. In G. da S. Abbad, L. Mourão, P. P. M. Meneses, T. Zerbini, J. E. Borges-Andrade, & R. Vilas-Boas (Eds.), *Medidas de avaliação em treinamento, desenvolvimento e educação: Fundamentos para gestão de pessoas* (pp. 78–90). Porto Alegre: Artmed.

Bennink, M., Croon, M. A., & Vermunt, J. K. (2015). Stepwise latent class models for explaining group-level outcomes using discrete individual-level predictors. *Multivariate Behavioral Research*, *50*(6), 662–675. doi: 10.1080/00273171.2015.1074879

Brasil. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (2010). *Plano Nacional de Pós-Graduação – PNPG 2011 - 2020*. Recuperado de http://www.capes.gov.br/images/stories/download/Livros-PNPG-Volume-I-Mont.pdf

Brasil. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (2017). *Portaria Capes nº 59, de 22 março de 2017*. Dispõe sobre o regulamento da Avaliação Quadrienal. Recuperado de http://www.capes.gov.br/images/stories/download/legislacao/27032017-PORTARIA-N-59-DE-22-DE-MARCO-DE-2017.pdf

Brasil. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (2018a). *Sobre a Avaliação*. Recuperado de http://www.capes.gov.br/avaliacao/sobre-a-avaliacao

Brasil. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (2018b). *Portaria Capes nº 182, de 14 de Agosto de 2018*. (2018). Dispõe sobre os processos avaliativos das propostas de cursos novos e dos programas de pós-graduação *stricto sensu* em funcionamento. Recuperado de https://www.capes.gov.br/images/stories/download/legislacao/16082018-PORTARIA-N-182-DE-14-DE-AGOSTO-DE-2018.pdf

Brasil. Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (2018c). *Avaliação da Pós-Graduação.*

Considerações do CTC-ES. Recuperado de https://sei.capes.gov.br

Borges-Ferreira, M. F., & Abbad, G. (2010). Avaliação de aprendizagem em disciplinas de curso técnico a distância. *Estudos de Psicologia*, *14*(2), 141–149. doi: 10.1590/s1413-294x2009000200007

Cawley, B. D., Keeping, L. M., & Levy, P. E. (1998). Participations in the performance appraisal process and employee reactions: A meta-analytic review of filed investigations. *Journal of Applied Psychology*, *83*(4), 615-633. doi: 10.1037/0021-9010.83.4.615

Cunha, C. G. S. (2018). Avaliação de políticas públicas e governamentais: tendências recentes e experiências no Brasil. *Revista Estudos de Planejamento*, *12*, 27-57. Recuperado de https://revistas.dee.spgg.rs.gov.br/index.php/estudos-planejamento/article/view/4298/4056

Dusterhoff, C., Cunningham, B., & MacGregor, J. N. (2014). The effects of performance rating, leader-member exchange, perceived utility, and organizational justice on performance appraisal satisfaction: Applying a moral judgment perspective. *Journal of Business Ethics*, *119*(2), 265-273. doi: 10.1007/s10551-013-1634-1

Colquitt, J. A. (2001) On the dimensionality of organizational justice: A construct validation of a measure. *Journal of Applied Psychology*, *86*(3), 386-400. doi: 10.1037//0021-9010.86.3.386

Colquitt, J. A., Wesson, M. J., Porter, C. O. L. H., Conlon, D. E., & Ng, K. Y. (2001). Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. *Journal of Applied Psychology*, *86*(3), 425-445. doi: 10.1037//0021-9010.86.3.425

Fachada, D. F. C. (2012). *Avaliação de desempenho – Satisfação dos funcionários da administração pública* (Dissertação de mestrado). Faculdade de Economia da Universidade do Porto, Porto. Recuperado de https://repositorio-aberto.up.pt/handle/10216/80853.

Field, A. (2009). *Descobrindo a estatística usando o SPSS*. Porto Alegre: Artmed.

Greenberg, J. (2011). Organizational justice: The dynamics of fairness in the workplace. In S. Zeleck (Ed.), *Handbook of industrial and organizational psychology*, *3*, pp. 271-327. Washington: American Psychological Association.

Harrington, J. R. & Lee, J. H. (2015). What drives perceived fairness of performance appraisal? Exploring the effects of psychological contract fulfillment on employees perceived fairness of performance appraisal in U.S. federal agencies. *Public Personnel Management*, *44*(2), 214-238. doi: 10.1177/0091026014564071

Judge, T. A., & Kammeyer-Mueller, J, D. (2012). Job Attitudes. *Annual Review of Psychology*, *63*, 341-367. doi: 10.1146/annurev-psych-120710-100511

Keeping, L. M., & Levy, P. E. (2000). Performance appraisal reactions: Measurement, modeling, and method bias. *Journal of Applied Psychology*, *85*(5), 708-723. doi: 10.1037/0021-9010.85.5.708

Kim, J. (2016). Impact of performance appraisal justice on the effectiveness of pay-for-performance systems after civil service reform. *Public Personnel Management*, *45*(2), 148-170. doi: 10.1177/0091026016644625

Kim, T., & Holzer, M. (2016). Public employees and performance appraisal. *Review of Public Personnel Administration*, *36*(1), 31-56. doi: 10.1177/0734371X14549673

Lira, M. (2014). Satisfaction with a performance appraisal system in the Portuguese public sector: The importance of perceptions of justice and accuracy. *Review of Applied Management Studies*, *12*, 30-37. doi: 10.1016/j.tekhne.2015.01.005

Lira, M., Silva, V. P. G, & Viseu, C. (2016). Performance appraisal as a motivational tool in the portuguese public administration. *Portuguese Journal of Finance, Management and Accounting*, *2*(3), p. 91-118. Recuperado de http://u3isjournal.isvouga.pt/index.php/PJFMA/article/view/142/77

Pasquali, L. (2010). *Instrumentação psicológica: Fundamentos e práticas*, Porto Alegre: Artmed.

Payne, S. C., Horner, M. T., Boswell, W. R., Schroeder, A. N., & Stine-Cheyne, K. J. (2009). Comparison of online and traditional performance appraisal systems. *Journal of Managerial Psychology*, *24*(6), 526-544. doi: 10.1108/02683940910974116

Rico, R., Hinsz, V. B., Burke, S., & Salas, E. (2017). A multilevel model of multiteam motivation and performance. *Organizational Psychology Review*, *7*(3), 197–226. doi: 10.1177/2041386616665456

Souza, D. B. L.; Abbad, G., & Gondim, S. M. G. (2017). Modelos lógicos na avaliação de um mestrado profissional: um exemplo de aplicação. *Revista Brasileira de Pós-Graduação, 14*, 1-19. doi: 10.221713/2358-2332.2016.v14.14291

Suárez-Alvarez, J., Pedrosa, I., Lozano, L. M., García-Cueto, E., Cuesta, M., & Muñiz, J. (2018). Using reversed items in Likert scales: A questionable practice. *Psicothema, 30*(2), 149–158. doi: 10.7334/psicothema2018.33

Van Veldhoven, M. J. P. M. (2012). *About tubs and tents: Work behavior as the foundation of strategic human resource management.* Tilburg: Prismaprint.

Voguel, M. J. M. (2015). *Avaliação da Pós-Graduação Brasileira: Análise dos quesitos utilizados pela CAPES e das críticas da comunidade acadêmica* (Tese de doutorado). Escola de Comunicação e Artes da Universidade de São Paulo, São Paulo. Recuperado de http://www.teses.usp.br/teses/disponiveis/27/27151/tde-29062015-150747/pt-br.php

About the authors:

Kelly Rocha de Queiroz
Holds a master's degree in Psychology from the University Center of Brasilia (Centro Universitário da Brasília). She is a science and technology analyst and is currently the Coordinator of Induction and Innovation programs at the Brazilian Coordination of Higher Education Personnel (CAPES).
ORCID: https://orcid.org/0000-0002-1182-422X
E-mail: kelly.queiroz@capes.gov.br

Amalia Raquel Pérez-Nebra
Visiting Professor at the Universidade de Brasília (UnB). The author coordinates several research and intervention projects in Social Psychology applied in institutional and organizational contexts. Her research themes are focused on sustainable labor welfare, culture, media influence, and over-indebtedness. She has also conducted research projects at the Universitat de València and the Universidad Autônoma de Madrid.
Publons: www.researcherid.com/rid/D-6030-2018
ORCID: https://orcid.org/0000-0001-8386-1233
E-mail: amaliaraquel.perez@campusviu.es

Fabiana Queiroga
PhD in Social, Organizational and Work Psychology from the Universidade de Brasília. She is linked to the masters program of the University Center of Brasilia (Centro Universitário de Brasília), where she teaches subjects related to the field of Organizational Behavior. She has experience in building Psychology Assessment Instruments and Measures, with an emphasis on large-scale educational assessment, learning assessment, and performance.
ORCID: https://orcid.org/0000-0002-3811-8202
E-mail: queiroga.unice@gmail.com

Contact:

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
SBN Quadra 02 Lote L Lote 06 9º andar, Asa Norte
Brasília-DF, Brasil
Telefones: (61)981790906/(61)2022-6341