Kirkpatrick Four-level Model Evaluation: An Evaluation Scale on the Preservice Teacher's Internship Program

Muhammad Bukhori Dalimunthe[1] * (iD)

[1] Department of Economics Education, Faculty of Economics, Universitas Negeri Medan, Indonesia

A R T I C L E   I N F O

A B S T R A K

Program magang yang diselenggarakan tergantung pada penyedia program (sekolah), sehingga banyak siswa memiliki perbedaan yang besar dalam pengalaman magang. Di sisi lain, kampus memiliki keterbatasan dalam menjalankan program, salah satunya dalam mengevaluasi pencapaian program magang. Tujuan penelitian ini yaitu membangun validitas pada skala evaluasi program magang untuk guru prajabatan menggunakan model evaluasi empat tingkat Kirkpatrick. Jenis penelitian ini yaitu kuantitatif dengan pendekatan kuantitatif konfirmatori. Sampel sebanyak 212 guru prajabatan. Metode yang digunakan dalam mengumpulkan data yaitu observasi, tes dan kuesioner. Instrument yang digunakan untuk lembar tes dan kuesioner. Analisis data menggunakan pendekatan confirmatory factor analysis (CFA) dengan aplikasi SmartPLS v3.0. Hasil analisis yaitu model evaluasi empat tingkat memenuhi validitas konstruk dan membuktikan validasi yang akurat pada skala evaluasi program magang. Secara khusus, skala evaluasi program magang sangat efektif dalam melacak pencapaian program di setiap tingkatan. Implikasinya bagi para pemangku kepentingan akademik di kampus dapat menggunakan skala evaluasi ini sebagai alat ukur untuk bereksperimen dengan keberhasilan program magang. Selain itu, mereka dapat meninjau capaian keberhasilan guru prajabatan dalam melaksanakan program magang berdasarkan setiap tingkatan pada model Kirkpatrick.

A B S T R A C T

Internship programs are organized depending on the program provider (school), so many students have a big difference in the internship experience. On the other hand, the campus has limitations in running the program, one of which is evaluating the achievement of the internship program. The purpose of this study is to build validity on the evaluation scale of the apprenticeship program for pre-service teachers using Kirkpatrick's four-level evaluation model. This type of research is quantitative with a confirmatory quantitative approach. The sample is 212 pre-service teachers. The methods used in collecting data are observation, tests, and questionnaires. Instruments used for test sheets and questionnaires. Data analysis used a confirmatory factor analysis (CFA) approach with the SmartPLS v3.0 application. The analysis results are that the four-level evaluation model meets construct validity and proves accurate validation on the internship program evaluation scale. In particular, the internship program evaluation scale effectively tracks program achievements at each level. The implication is that academic stakeholders on campus can use this evaluation scale as a measuring tool to experiment with the success of the internship program. In addition, they can review the achievements of pre-service teachers in implementing apprenticeship programs based on each level of the Kirkpatrick model.

1. INTRODUCTION

Since 2017, the government of Indonesia has issued regulations related to the Bachelor of Education Program as stated in the regulation of the Minister of Research, Technology and Higher Education number 55 of 2017 concerning Teacher Education Standards. The regulation regulates the development of teacher education curricula, one of which is the Field School Program or internships in schools in stages. The internship program is carried out in two stages, namely Internship I, which aims to build the foundation of the teaching profession through various activities carried out in schools. Furthermore, Internship II aims to finalize the academic competence of education and fields of study through activities carried out in schools. The internship activities are described in detail in the Internship Program Guide (Baharuddin, 2021; Marufi et al., 2018; Sunardi & Sudjimat, 2016). The internship program is implicitly translated and implemented by campuses in Indonesia, where its implementation is in line with and leads to an increase in key performance indicators (Makarim, 2021; Ramadhani & Rahayu, 2020). So that all campuses manage their resources to make the national program a success. Each campus manages the evaluation of the internship program independently. So there have been many studies that

Corresponding author.
E-mail addresses: daliori86@unimed.ac.id (Muhammad Bukhori Dalimunthe)

have found an evaluation of the effectiveness of preservice teachers internships (Dewanto, 2018; Ismail et al., 2018). The evaluation of internships in previous studies was massively carried out concerning the overall achievement of an organized program. Unfortunately, organized internship programs are dependent on program providers (schools), so many students have a vast disparity in internship experience (Nghia, 2018; Nghia & My Duyen, 2018). On the other hand, it is undeniable that the campus has limitations in running the program, one of which is evaluating the internship program's achievement (Tindowen et al., 2019; Tovey, 2001). Thus, there is a gap in the analysis of program evaluation information that is reported to be limited to the program achievement area but has not yet targeted student learning achievement. This research offers a construction measurement of the internship based on Kirkpatrick's evaluation model.

Kirkpatrick's model is a program evaluation form that uses four levels developed by Donald Kirkpatrick (Effendi et al., 2022; Kirkpartrick, 1959). The trainees involved are directed to capacity building through the transfer of knowledge, skills, or attitudes through structured learning and follow-up activities to solve identified problems or performance gaps (Badu, 2013; Putman & Polly, 2021). In the context of the internship program, preservice teachers are given direct field practice in schools for a certain period to achieve predetermined learning outcomes (Bush & Grotjohann, 2020). Then a series of items on the evaluation scale of the internship program were given to the preservice teachers to review the achievement of the program (Kazaz & Alagözlü, 2020; Nghia & Duyen, 2019). Kirkpatrick's four-level evaluation model consists of reaction, learning, behavior, and results. The reaction and learning levels are seen as internal because they focus on the processes that occur in the internship program. In comparison, the level of behavior and results is seen as external because it focuses on the changes that occur after completing the internship program (Dewi & Kartowagiran, 2018; D. L. Kirkpatrick & Kirkpatrick, 2007; Praslova, 2010). Several relevant studies have examined the validity of the measurement of internship instruments for preservice teachers (Gracia et al., 2021; Nghia & Duyen, 2019; Praslova, 2010). However, it has not provided reliable information in the internship program for preservice teachers in Indonesia. The evaluation instrument for the internship program is not yet available credibly and comprehensively to evaluate internship programs that focus on internal and external aspects of apprentices, based on Kirkpatrick's evaluation model. The urgency of this research is to present an internship program evaluation instrument from Kirkpatrick's four-level evaluation model based on the context of internship for preservice teachers at Indonesian campuses. The aim paper is to construct validity of the internship program evaluation scale using the Kirkpatrick model, which can be widely used for preservice teachers. The novelty offered focuses on two things. First, the measurement scale is aimed at the construction of the evaluation of internship programs for preservice teachers, where research over the past decade has majorly discussed the evaluation of an academic program such as teaching practicum (Kazaz & Alagözlü, 2020), measuring learning outcomes (Nghia & Duyen, 2019), teacher identity for the preservice teacher (Aykac et al., 2017), the relevance of four factors (i.e., teacher education program, experience from field practice, psychological factors, and implementation of technology integration) to the internship program (Liu, 2016), and the determinants of internship effectiveness (Narayanan et al., 2010). Finally, the internship program evaluation uses Kirkpatrick's four-level evaluation model that focuses on the reaction, learning, behavior, and results. So that the measurement scale developed produces an adaptive and comprehensive instrument that reviews each stage of the internship program. Therefore, this study aims to test the construct validity of the evaluation of the pre-service teacher pre-service program using Kirkpatrick's four-level evaluation model.

2. METHODS

This paper uses a confirmatory quantitative approach in comparative causality to explain the relationship between variables involving reaction, learning, behavior, and results from variables as a construction of the internship program evaluation (Creswell, 2014). Construction validity uses confirmatory factor analysis (CFA) in which the variable structure is based on credible theories and research results to develop postulates or assumptions about the relationship between the observed variables and the underlying factors from credible theories and research results (Clark & Watson, 1995; Jöreskog et al., 2016). Participants are the preservice teachers who have completed an internship program. A sample of 212 the preservice teachers consisting of study programs from economics education (n=93), accounting education (n=50), business education (n=38), and administration education (n=31) came from the Faculty of Economics, Universitas Negeri Medan-Indonesia. The sample is determined by total sampling, in which the entire population is the sample in this research. The measurement scale of the internship program evaluation instrument was adopted from various relevant and significant sources used in education (Aykac et al., 2017; Dewi & Kartowagiran, 2018; Kazaz & Alagözlü, 2020; D. L.

Kirkpatrick & Kirkpatrick, 2007; Liu, 2016; Narayanan et al., 2010; Nghia & Duyen, 2019; Praslova, 2010).
In addition, the questionnaire items have been adjusted to the internship program carried out on campus
by considering the objectives and evaluation of the internship program (see Appendix A). The
measurement scale uses a differential semantic scale (points 1-6) in the form of a bipolar continuum
(Widhiarso, 2017). The differential semantic scale available in the questionnaire consists of three
dimensions, namely the evaluation dimension (poor-good), the potential dimension (weak-strong), and
the activity dimension (passive-active). Each dimension is adapted to its use by considering the context of
the statement on each item in the questionnaire. The grid of evaluation instruments for the internship
program is presented in Table 1.

Table 1. Kirkpatrick's Four-Level Evaluation Instrument For The Internship Program

Reaction

1	How would you rate your interest in the internship program?
2	How would you rate the internship program as helpful in describing student characteristics?
3	How would you rate the internship program to help describe the rules and regulations of the partner school?
4	How do you rate the preparation of tutors (teacher) in guiding during the implementation of the internship program?
5	How would you rate the convenience and assistance of partner schools supporting the internship program?
6	How would you rate an internship program as an educational experience to help you do your job or assignment better?

Learning

1	Due to the internship program, you become better able to communicate with other people.
2	Due to the internship program, you become more capable of seeking more feedback on strengths and weaknesses to improve significantly on the teacher's identity.
3	Due to the internship program, you will feel more empowered personally.
4	Due to the internship program, you can respond to aggressive behavior better.
5	Due to the internship program, you can respond to passive behavior better.
6	Due to the internship program, you become better able to set goals.
7	Due to the internship program, you will feel more valued as a preservice teachers.

Behavior

1	To what extent has your confidence in using your knowledge or skills increased due to the internship program?
2	To what extent does the content on the internship program accurately reflect what conditions exist in the school?
3	To what extent do you have access to the necessary resources (e.g., equipment and information) at the partner school to apply your knowledge or skills when implementing an internship program?
4	To what extent have you received assistance through mentoring and feedback in applying your knowledge or skills to the course after completing the internship program?
5	As a result of the learning outcomes of the internship program, how distant have you performed on the course objectives?

Results

1	After completing the internship program, you can describe the general characteristics of the students.
2	After completing the internship program, you can describe the school's organizational structure and working procedures.
3	After completing the internship program, you can describe the school's rules and regulations.
4	After completing the internship program, you can describe the ceremonial-formal activities at school (e.g., ceremonies and meetings).
5	After completing the internship program, you can describe routine activities in the form of intra-curricular (e.g., laboratory practice, remedial), co-curricular (e.g., museum visits, industry visits), and extra-curricular activities (e.g., football/basketball clubs, Pencak silat, arts, scout).

Evaluation of the internship program using Kirkpatrick's four-level evaluation model is more
responsive to reviewing the development of abilities and challenges faced by students during the
internship program. Briefly explain the definition of each level of the internship program evaluation

model. First, the reaction is the preservice teachers' perception regarding the benefits (affective reactions) and how much utility they feel about the internship program. Second, learning measures the internship learning outcomes that they feel directly. Third, behavior is a behavior change criterion on the knowledge and skills used by students to do assignments based on the internship experience they have obtained. Finally, results are an increase in understanding and stability of students' personalities related to the teaching profession and the culture and tasks played by the teacher. Items at the reaction-learning-behavior-results level consist of six, seven, five, and five items, respectively, to test the construct validity of the internship evaluation measurement scale. The stages of the reaction, learning, behavior, results, which are the evaluation model constructs, comprehensively describe the developmental achievements at each level. Likewise, challenges and problems that occur can be immediately evaluated partially from each level. The measurement scale of the internship program evaluation in this paper provides this information. So that measurements made by stakeholders on campus can provide assessments and recommendations for internship programs effectively and comprehensively revealing the success or failure of preservice teachers in implementing the internship.

The students carry out an internship program at the school, a mandatory program for preservice teachers. The preservice teachers intensively and proactively participate in the internship program with various challenges. The internship program is one semester. The authors observe the progress of the preservice teachers' internships through program progress reports that are consulted with the supervisor every month. Each supervisor guides eight the preservice teachers intensively in responding to all the challenges they face at the school where they are interns. One week after the internship program ended, the authors gave a questionnaire to all participants to fill out all the electronic questionnaire items (google form). Data collection was carried out through a one-time procedure for filling out the instrument to a group of teacher students who had completed the internship program. All participants (N=212) completed the questionnaire. The tabulation of the data obtained is still in ordinal form, then the ordinal data is transformed into intervals using the successive interval (MSI) method. This is because the CFA test in partial least squares requires data in the form of intervals or ratios to continue the transformed interval data at the analysis stage (Hair Jr et al., 2016). Data analysis used a confirmatory factor analysis (CFA) approach with the SmartPLS v3.0 application. All data collected were analyzed to test the internship program evaluation questionnaire's construct validity; no data was lost. Techniques to test the construct validity using internal consistency, convergent, and discriminant validity. Internal consistency testing using Cronbach's alpha coefficient criteria > 0.70 (Hair Jr et al., 2016; Viladrich et al., 2017). Cronbach's alpha provides a reliability estimate based on the intercorrelation of the observed variable indicators (Streiner, 2003). Furthermore, the convergent validity test uses the criteria of average variance extracted (AVE) > 0.50, and the outer loading value > 0.70 (Hair Jr et al., 2016). Convergent validity measures the extent to which a measure (indicator) is positively correlated with other alternative measures of the same construct. The high value of outer loadings on a construct shows that the indicators have a good similarity described by the construct. Finally, the discriminant validity test uses the confidence interval bias of the heterotrait-monotrait (HTMT) does not include 1, which can be calculated by doing the bootstrapping (Hair Jr et al., 2016; Henseler et al., 2015). Discriminant validity measures the extent to which a construct is unique and captures different phenomena, and is not represented by other constructs in a model.

3. RESULT AND DISCUSSION

Results

The demographics of the participants (see Table 2) involved in this study were dominated by a female (84.91%) than male (15.09%). The preservice teachers from the economics education (43.87%) are the department with the most internships, followed by accounting education (23.58%), business education (17.93%), and administration education (14.62%). The origin of the dominant preservice teachers is newcomer (64.62%) who come from outside the Medan (35.38%). Futhermore, tutors' expertise is still in the same direction (86.32%) with preservice teachers. This is an essential spotlight for program managers at the faculty and university levels so that the guidance process can successfully support student learning outcomes. This information can be followed up through more in-depth research by scholars in the future.

Table 2. Participant Demographics

Demographics	Value (%)
Gender	
Male	32 (15.09)

Demographics	Value (%)
Female	180 (84.91)
Department	
Economic education	93 (43.87)
Eccounting education	50 (23.58)
Business education	38 (17.93)
Administration education	31 (14.62)
Origin	
Resident (Medan)	75 (35.38)
Newcomer	137 (64.62)
Tutors (teacher) expertise with the preservice teacher interns	
In line	183 (86.32)
No	29 (13.68)

Based on the tabulation and testing of data analysis, statistical values were obtained that proved the validity of the evaluation scale construct of the internship program using the Kirkpatrick model. The testing approach uses CFA with various construct validity techniques, namely internal consistency, convergent validity, and discriminant validity. The test results are presented comprehensively and discussed in detail based on the criteria for the proof of the test. The statistical value of the construct validity test is briefly shown in Table 3.

Table 3. Summary of Construct Validity Value on the Internship Program Evaluation Scale

Constructs	Internal Consistency	Convergent Validity		Discriminat Validity	
	Cronbach's	AVE	Loadings	HTMT confidence interval bias does not include 1	
	> 0.70	> 0.50	> 0.70	2.5%	97.5%
Reaction	0.899	0.666			
Item X11			0.859		
Item X12			0.852		
Item X13			0.814		
Item X14			0.780		
Item X15			0.758		
Item X16			0.829		
Learning	0.957	0.797			
Item X21			0.888		
Item X22			0.888		
Item X23			0.905		
Item X24			0.902		
Item X25			0.908		
Item X26			0.889		
Item X27			0.867		
Behavior	0.912	0.740			
Item X31			0.854		
Item X32			0.858		
Item X33			0.813		
Item X34			0.867		
Item X35			0.908		
Results	0.922	0.763			
Item X41			0.824		
Item X42			0.900		
Item X43			0.898		
Item X44			0.899		
Item X45			0.844		
Learning – Behavior				0.740	0.889
Reaction – Behavior				0.690	0.866
Reaction – Learning				0.746	0.857
Results – Behavior				0.794	0.907

Constructs	Internal Consistency	Convergent Validity		Discriminat Validity	
	Cronbach's	AVE	Loadings	HTMT confidence interval bias does not include 1	
	> 0.70	> 0.50	> 0.70	2.5%	97.5%
Results – Learning				0.752	0.874
Results – Reaction				0.713	0.838

The value on the four-level internship evaluation scale measures will show high internal consistency (α > 0.70). Cronbach's alpha value of the constructs of reaction (0.899), learning (0.957), behavior (0.912), and results (0.922) is more significant than 0.70, so the test results have proven excellent and representative reliability based on the intercorrelation of the measured construct indicators. The test results provide information on the consistency between the items in the questionnaire, indicating that the evaluation scale of the internship program with the Kirkpatrick model has a reliable measurement function. The value on Kirkpatrick's four-level evaluation model shows strong convergent validity (measured using AVE). In addition, the indicators on each construct have good similarities (measured using outer loadings). For instance, the AVE value of the constructs of reaction (0.666), learning (0.797), behavior (0.740), and results (0.763) are more significant than 0.50, proving that each indicator has a positive correlation on the same construct. Subsequent testing was carried out by paying attention to the outer loadings scores of all indicators in each construct, it was found that the item scores (indicators) of each construct (reaction-learning-behavior-results) on the internship evaluation scale had a score greater than 0.07. These pieces of evidence give us confidence that the evaluation scale developed in this paper can review the four-level internship program, which has a significant correlation of the items at each level.
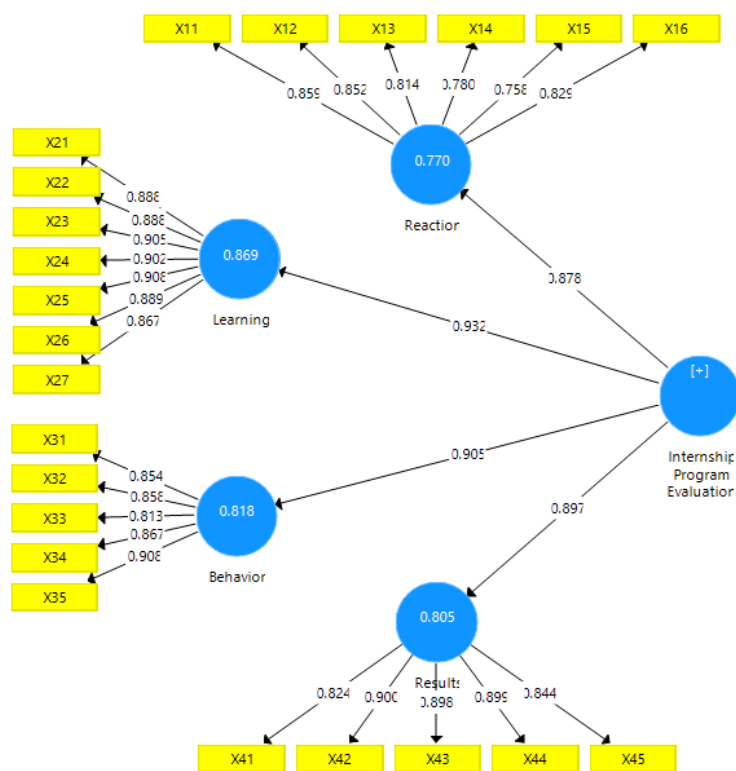

Figure 1. Reflective Measurement Model – Internship Program Evaluation Scale

The discriminant validity test uses the HTMT correlation score confidence interval bias (2.5%; 97.5%), which is not more than a score of 1 (Henseler et al., 2015). The correlation score between the constructs of the internship program evaluation scale is away from a score of 1 which means the correlation ratio between indicators in each construct has a strong discriminant. So that the four constructs used on the internship program evaluation scale have unique discriminants and differ from one construct to another. Each level in Kirkpatrick's evaluation model can capture different phenomena and

has a different role from one level to another. The results of this test provide evidence that the reaction level is successful in describing the participants' perceptions regarding the benefits and utility of the internship program. Furthermore, the learning level on the evaluation scale of the internship program has succeeded in revealing the size of the internship learning outcomes that they feel have a direct impact on their profession as teachers. Likewise, the behavior level on the evaluation scale of the internship program has succeeded in revealing the criteria for changing behavior in learning activities that impact completing tasks based on internship experiences at school. Finally, the level results on the evaluation scale of the internship program have succeeded in describing an increased understanding of the teaching profession. Based on Figure 1, information is obtained regarding the reflective measurement model that depicts the results of the CFA estimation, which is displayed graphically. The measurement model provides a strengthening of confidence in testing the construct validity of the internship program evaluation scale using the Kirkpatrick model. The direct effect values of each construct of reaction (0.878), learning (0.932), behavior (0.905), and evaluation (0.897) have a strong influence on the evaluation of the internship program. This finding is in line with various researches on developing internship program evaluation instruments in which Kirkpatrick's four-level evaluation model has effectiveness and is relevant to student internships (Ahanchian et al., 2017; Dewi & Kartowagiran, 2018; Jamil et al., 2013). The learning level is the most powerful construct for evaluating the internship program compared to other program evaluation levels. The preservice teachers enjoy the learning process in the school internship program.

Discussion

The findings in this article complement previous research that focused on developing a measurement scale for internship programs evaluation (Aykac et al., 2017; Gracia et al., 2021; Narayanan et al., 2010; Nghia & Duyen, 2019; Praslova, 2010). Several previous investigators highlighted the strengthening of teacher identity for preservice teachers via internships (Aykac et al., 2017; Gracia et al., 2021). The measurement scale of the Kirkpatrick internship evaluation model in this study also reveals the identity of teachers in internal aspects (namely level of reaction and learning) in the context of the Field School Program in Indonesia (Direktorat Pembelajaran, 2017). Furthermore, this study examines the external aspects (namely level of behavior and results) comprehensively successfully as an evaluation construct of the Kirkpatrick model program, which was not found in the research accomplished by Aykac et al., and Gracia et al., (Aykac et al., 2017; Gracia et al., 2021). Research findings from other investigators investigate the construction of a scale for measuring the effectiveness and impact of internships on preservice teachers learning outcomes (Narayanan et al., 2010; Nghia & Duyen, 2019; Praslova, 2010). Their investigation findings have succeeded in developing a measurement scale related to internship evaluation using the Kirkpatrick model. Nevertheless, the findings in this study give a different color from the findings they obtained, specifically on preservice teacher's activities. The measurement scale of the internship evaluation in this study highlights the operational activities of preservice teachers at each level. These activities refer to the internship objectives in Internship Guidelines in the Indonesian context. Furthermore, a comprehensive measurement scale can be used and reflects student activities during internships so that program evaluations can be assessed in a relevant and credible manner. Exploration of findings based on Figure 1 provides information that the measurement model is reflective, explaining that indicators related to latent variables (reaction, learning, behavior, results) have the same variance (Hair et al., 2017; Henseler et al., 2016). The latent variable theoretically constructs the four-level evaluation model developed by Kirkpatrick so that the examination results successfully prove the four-level evaluation model (Kirkpartrick, 1959). Relevant research findings from other scholars are in line with the findings in this paper, which explains that four-levels have significance as a teacher apprenticeship evaluation construct (Amir & Ismail, 2014; Mahmoodi et al., 2019; Piryani et al., 2018). Furthermore, this research complements the context of internships at the Field School Program in Indonesia. This research has succeeded in presenting novelty when viewed from various relevant researches, in which the measurement scale of teacher internship evaluation uses Kirkpatrick's theoretical basis with preservice teacher's internship programs in the context of Indonesian campuses. Kirkpatrick's four-level evaluation model was used to develop a program evaluation measurement scale that considered the levels of reaction, learning, behavior, and results. Kirkpatrick's four-level model, which is a construct of evaluating an internship program, has been proven to measure the evaluation of an internship program based on construct validity testing. Miscellaneous relevant studies confirm that the scale development process involves systematic and complex procedures that demand methodological rigor (Kazaz & Alagözlü, 2020; Nghia & Duyen, 2019). A sequence of construct validity and reliability tests, including internal consistency, convergent validity, and discriminant validity, were used to prove that the evaluation scale of internships teachers could be used credibly and comprehensively.

Using the preservice teacher's perceptions of the services, mentoring, and reporting of internship programs that they feel from supervisors (lecturer) and tutors (teacher), it is revealed that the learning level has a higher outer loading value than other levels, so that construct is the most meaningful when they have the internship programs. It provides information internally that the learning process in the internship program has succeeded in providing insight into the learning paradigm applied by preservice teachers to their activities while studying in class (Dalimunthe et al., 2019; Putman & Polly, 2021). The same applies to the teaching profession (Gravett et al., 2019), classroom management (Adams et al., 2020), academic culture, and school organization (Tindowen et al., 2019), which they will play. Furthermore, when the evaluation of the internship program is viewed as externally focused on the changes that occur, the behavior level plays an essential role over other levels (D. L. Kirkpatrick & Kirkpatrick, 2007; Praslova, 2010). At the behavior level, the preservice teachers felt a change in the actions in the learning process that they felt after completing the internship. Significant changes occurred in interns such as communicating and behavior with supervisors and tutors, improving academic performance, and increasing their ability to interact with students in class related to learning to become a teacher (Goldhaber, 2018; M. Kirkpatrick et al., 2019; Mahmoodi et al., 2019). Based on the results of the construct validity test of the internship program evaluation scale, it was found that the questionnaire had strong validity and reliability to evaluate the performance of the internship program at each level of the Kirkpatrick evaluation model. This finding has been proven using the CFA approach through a series of tests of internal consistency, convergent validity, and discriminant validity (Hair et al., 2017; Henseler et al., 2015; Viladrich et al., 2017). The three techniques have different purposes and benefits that complement each other to develop a measurement scale for evaluating internship programs. All construct validity tests provide confidence and feasibility of a measurement scale that can effectively and comprehensively evaluate the preservice teacher's internship programs. This research has limitations on internship, namely preservice teachers in the reaction-learning-behavior-results activity frames. Meanwhile, the study has not explored the perceptions of supervisors (lecturers) and tutors (teachers) regarding preservice teachers' performance during internships. Finally, the limitations of this research are in the context of the Field School Program in Indonesia, so that the internship instrument in this research cannot be used in other internship programs in higher education, such as internships in companies, institutions, and laboratories; or modification is required if used in higher education institutions in other countries. However, in general, the measurement scale has been successfully tested for construct validity. Therefore, the campus can use it as an instrument to measure the success of preservice teachers in carrying out internships based on the level of reaction, learning, behavior, results.

4. CONCLUSION

Implicative information explains that internally viewed reaction rates and externally viewed behavior levels recreate important roles in Kirkpatrick's model games. Universities can use an evaluation scale to measure prayer's success, which pays attention to the internal-external aspects of prayer in learning outcomes so that the research findings have implications for campuses in Indonesia that carry out teacher internships. The scale of measurements and findings are very helpful for the development of psychometric science and educational scholars who will apply the novelty to future research.

5. REFERENCES

Adams, T., Koster, B., & Brok, P. Den. (2020). Student teachers' classroom management during the school internship. *European Journal of Teacher Education*, 1–19. https://doi.org/10.1080/02619768.2020.1860011.

Ahanchian, M., Sharafi, S., Vafaee, M., & Hajiabadi, F. (2017). Evaluate the Effectiveness of internship program in nursing student using Kirkpatrick's model TT. *gums-rme*, *9*(1), 17–19. https://doi.org/10.18869/acadpub.rme.9.1.17.

Amir, T., & Ismail, O. (2014). An evaluation of the internship program at the college of education, Sultan Qaboos University, Muscat. *International Journal of Humanities, Social Sciences and Education*, *1*(10), 5–10. https://doi.org/10.13140/RG.2.2.10633.47200.

Aykac, N., Yildirim, K., Altinkurt, Y., & Marsh, M. M. (2017). Understanding the underlying factors affecting the perception of pre-service teachers' teacher identity: A new instrument to support teacher education. *Üniversitepark Bülten*, *6*(1), 67. https://doi.org/10.22521/unibulletin.2017.61.6.

Badu, S. Q. (2013). Implementasi Evaluasi Model Kirkpatrick Pada Perkuliahan Masalah Nilai Awal Dan Syarat Batas. *Jurnal Penelitian dan Evaluasi Pendidikan*, *16*, 102–129. https://doi.org/10.21831/pep.v16i0.1108.

Baharuddin, M. R. (2021). Adaptasi Kurikulum Merdeka Belajar Kampus Merdeka (Fokus: Model MBKM Program Studi). *Jurnal Studi Guru dan Pembelajaran*, *4*(1). https://doi.org/10.30605/jsgp.4.1.2021.591.

Bush, A., & Grotjohann, N. (2020). The impact of a long-term internship on pre-service teacher collaboration. *Journal of Innovation in Psychology, Education and Didactics*, *24*(1).

Clark, L. A., & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment*, *7*(3), 309–319. https://doi.org/10.1037/1040-3590.7.3.309.

Creswell, J. W. (2014). *Research design: Quantitative, qualitative and mixed methods* (4th ed). Thousand Oaks, CA: Sage.

Dalimunthe, M. B., Nugrahadi, E. W., Fitrawaty, & Dalimunthe, R. Z. (2019). *Evaluation of Teacher Professional Training Program with Contextual, Input, Process, and Product (CIPP) Models*. RedWhite Press PP - Jakarta. https://doi.org/10.32698/hum0213.

Dewanto, A. C. (2018). Efektivitas PPL untuk meningkatkan efikasi diri mengajar pada mahasiswa pendidikan matematika. *Delta: Jurnal Ilmiah Pendidikan Matematika*, *6*(1), 1–9. https://doi.org/10.31941/delta.v6i1.662.

Dewi, L. R., & Kartowagiran, B. (2018). An evaluation of internship program by using Kirkpatrick evaluation model. *REID (Research and Evaluation in Education)*, *4*(2), 155–163. https://doi.org/10.21831/reid.v4i2.22495.

Direktorat Pembelajaran. (2017). Panduan Program Pengenalan Lapangan Persekolahan. In *Direktorat Jenderal Pembelajaran dan Kemahasiswaan*. Kementerian Riset, Teknologi, dan Pendidikan Tinggi.

Effendi, Zainuddin, Z., & Ahmad, M. S. (2022). Implementasi evaluasi model kirkpatrick terhadap kualitas sistem pembelajaran. *Jurnal Educatio*, *8*(1). https://doi.org/10.29210/1202221160.

Goldhaber, D. (2018). Evidence-Based Teacher Preparation: Policy Context and What We Know. *Journal of Teacher Education*, *70*(2), 90–101. https://doi.org/10.1177/0022487118800712.

Gracia, E. P., Rodríguez, R. S., Pedrajas, A. P., & Carpio, A. J. (2021). Teachers' professional identity: validation of an assessment instrument for preservice teachers. *Heliyon*, *7*(9), e08049. https://doi.org/10.1016/j.heliyon.2021.e08049.

Gravett, S., Petersen, N., & Ramsaroop, S. (2019). A University and School Working in Partnership to Develop Professional Practice Knowledge for Teaching . In *Frontiers in Education* (Vol 3, bl 118).

Hair, J., Hollingsworth, C. L., Randolph, A. B., & Chong, A. Y. L. (2017). An updated and expanded assessment of PLS-SEM in information systems research. *Industrial Management & Data Systems*, *117*(3), 442–458. https://doi.org/10.1108/IMDS-04-2016-0130.

Hair Jr, J. F., Hult, G. T. M., Ringle, C., & Sarstedt, M. (2016). *A primer on partial least squares structural equation modeling (PLS-SEM)*. Sage publications.

Henseler, J., Hubona, G., & Ray, P. A. (2016). Using PLS path modeling in new technology research: updated guidelines. *Industrial Management & Data Systems*, *116*(1), 2–20. https://doi.org/10.1108/IMDS-09-2015-0382.

Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science*, *43*(1), 115–135. https://doi.org/10.1007/s11747-014-0403-8.

Ismail, I., Hasan, H., & Musdalifah, M. (2018). Pengembangan Kompetensi Mahasiswa Melalui Efektivitas Program Magang Kependidikan. *Edumaspul: Jurnal Pendidikan*, *2*(1), 124–132. https://doi.org/10.33487/edumaspul.v2i1.48.

Jamil, N. A., Shariff, S. M., & Abu, Z. (2013). Students' Practicum Performance of Industrial Internship Program. *Procedia - Social and Behavioral Sciences*, *90*, 513–521. https://doi.org/10.1016/j.sbspro.2013.07.121.

Jöreskog, K. G., Olsson, U. H., & Wallentin, F. Y. (2016). *Multivariate analysis with LISREL*. Springer. https://doi.org/10.1007/978-3-319-33153-9.

Kazaz, I., & Alagözlü, N. (2020). Evaluation of teaching practicum for pre-service English language teachers: A scale development study. *Dil ve Dilbilimi Çalışmaları Dergisi*, *16*(3), 1562–1593. https://doi.org/10.17263/jlls.803902.

Kirkpartrick, D. (1959). Techniques for evaluation training programs. *Journal of the American Society of Training Directors*, *13*, 21–26.

Kirkpatrick, D. L., & Kirkpatrick, J. D. (2007). *Implementing the Four Levels: A Practical Guide for Effective Evaluation of Training Programs*. Berrett-Koehler Publishers, Inc.

Kirkpatrick, M., Akers, J., & Rivera, G. (2019). Use of Behavioral Skills Training with Teachers: A Systematic Review. *Journal of Behavioral Education*, *28*(3), 344–361. https://doi.org/10.1007/s10864-019-09322-z.

Liu, S.-H. (2016). Teacher education programs, field-based practicums, and psychological factors of the implementation of technology by pre-service teachers. *Australasian Journal of Educational Technology*, *32*(3 SE-Articles). https://doi.org/10.14742/ajet.2139.

Mahmoodi, M., Rashtchi, M., & Abbasian, G. R. (2019). Evaluation of in-service teacher training program in iran: Focus on the Kirkpatrick model. *Education and Self Development*, *14*(4), 20–38. https://doi.org/10.26907/esd14.4.03.

Makarim, N. A. (2021). *Indikator Kinerja Utama Perguruan Tinggi Negeri dan Lembaga Layanan Pendidikan Tinggi*. Kementerian Pendidikan dan Kebudayaan.

Marufi, M., Ilyas, M., Anas, A., & Islamiah, R. Y. (2018). Program Pengembangan Kewirausahaaan Fakultas Keguruan dan Ilmu Pendidikan. *Caradde: Jurnal Pengabdian Kepada Masyarakat*, *1*(1). https://doi.org/10.31960/caradde.v1i1.13.

Narayanan, V. K., Olk, P. M., & Fukami, C. V. (2010). Determinants of internship effectiveness: An exploratory model. *Academy of Management Learning & Education*, *9*(1), 61–80. https://doi.org/10.5465/amle.9.1.zqr61.

Nghia, T. L. H. (2018). External stakeholders' roles and factors influencing their participation in developing generic skills for students in Vietnamese universities. *Journal of Education and Work*, *31*(1), 72–86. https://doi.org/10.1080/13639080.2017.1386774.

Nghia, T. L. H., & Duyen, N. T. M. (2019). Developing and validating a scale for evaluating internship-related learning outcomes. *Higher Education*, *77*(1), 1–18. https://doi.org/10.1007/s10734-018-0251-4.

Nghia, T. L. H., & My Duyen, N. T. (2018). Internship-related learning outcomes and their influential factors. *Education + Training*, *60*(1), 69–81. https://doi.org/10.1108/ET-02-2017-0030.

Piryani, R. M., Dhungana, G. P., Piryani, S., & Sharma Neupane, M. (2018). Evaluation of teachers training workshop at Kirkpatrick level 1 using retro-pre questionnaire. *Advances in Medical Education and Practice*, *9*, 453–457. https://doi.org/10.2147/AMEP.S154166.

Praslova, L. (2010). Adaptation of Kirkpatrick's four level model of training criteria to assessment of learning outcomes and program evaluation in Higher Education. *Educational Assessment, Evaluation and Accountability*, *22*(3), 215–225. https://doi.org/10.1007/s11092-010-9098-7.

Putman, S. M., & Polly, D. (2021). Examining the Development and Implementation of an Embedded, Multi-Semester Internship: Preliminary Perceptions of Teacher Education Candidates, Clinical Educators, and University Faculty. *Peabody Journal of Education*, *96*(1), 99–111. https://doi.org/10.1080/0161956X.2020.1864250.

Ramadhani, M. A., & Rahayu, A. (2020). Competency Improvement through Internship: An Evaluation of Corporate Social Responsibility Program in Vocational School. *International Journal of Evaluation and Research in Education*, *9*(3), 625–634. https://doi.org/10.11591/ijere.v9i3.20571.

Streiner, D. L. (2003). Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency. *Journal of Personality Assessment*, *80*(1), 99–103. https://doi.org/10.1207/S15327752JPA8001_18.

Sunardi, S., & Sudjimat, D. A. (2016). Magang Industri Untuk Meningkatkan Relevansi Kompetensi Profesional Guru Produktif SMK. *Teknologi dan Kejuruan*, *39*(2). https://doi.org/10.17977/tk.v39i2.7791.

Tindowen, D. J., Bangi, J., & Parallag Jr, C. (2019). Pre-Service Teachers' Evaluation on their Student Internship Program. *International Journal of Learning, Teaching and Educational Research*, *18*(10), 279–291. https://doi.org/10.26803/ijlter.18.10.18.

Tovey, J. (2001). Building Connections between Industry and University: Implementing an Internship Program at a Regional University. *Technical Communication Quarterly*, *10*(2), 225–239. https://doi.org/10.1207/s15427625tcq1002_7.

Viladrich, C., Angulo-Brunet, A., & Doval, E. (2017). A journey around alpha and omega to estimate internal consistency reliability. *Annals of Psychology*, *33*(3), 755–782. https://doi.org/https://doi.org/10.6018/analesps.33.3.268401.

Widhiarso, W. (2017). Semantik Diferensial. *Fakultas Psikologi UGM. Yogyakarta*, *10*.