Archives of
Rehabilitation

Winter 2024. Vol 24. Num 4

Research Paper

Designing and Assessing the Validity and Reliability of the Bayley-III Test Examiner Clinical Performance Scale

Farin Soleimani[1], *Nahideh Hasani Khiabani[2], Leila Yazdi [3],Hamidreza Lornejad[4], Naria Aboulghasemi[4], Ghazal Shariatpanahi[5]

1. Pediatric Neurorehabilitation Research Center, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran.
2. Children Development Section, Zahra Mardani Azar Children's Hospital, Tabriz University of Medical Sciences, Tabriz, Iran.
3. Homayoun Evolution Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran.
4. Child Health Office, Ministry of Health and Medical Education, Tehran, Iran.
5. Department of Pediatric Infectious Disease, Bahrami Hospital, Tehran University of Medical Sciences, Tehran, Iran.

Use your device to scan and read the article online

Citation Soleimani F, Hasani Khiabani N, Yazdi L, Lornejad H, Aboulghasemi N, Shariatpanahi Gh. [Designing and Assessing the Validity and Reliability of the Bayley-III Test Examiner Clinical Performance Scale (Persian)]. Archives of Rehabilitation. 2024; 24(4):566-585. https://doi.org/10.32598/RJ.24.4.3676.1

https://doi.org/10.32598/RJ.24.4.3676.1

ABSTRACT

Objective In recent years, the infant development and assessment programs have received attention by the health system of Iran. In this regard, the training of examiners for performing the Bayley tests has become one of the needs of the health system. This study aims to design and evaluate the validity and reliability of the Bayley-III examiner clinical performance scale.

Materials & Methods In this study, participants were Bayley-III test examiners from medical universities in Iran, who were selected using a purposive sampling method. First, in focused group discussion sessions with 10 Bayley-III test experts, the items of instrument were formulated. Then, to evaluate the face validity, the questionnaire was sent to a panel of expert consisting of 8 Bayley-III experts and examiners and they were asked to rate the items from 1 to 5 in terms of importance. The items with an impact score of less than 1.5 were deleted, and other items were revised, if needed. To evaluate content validity, the questionnaire was sent to 10 Bayley-III test experts, and then the content validity ratio (CVR) and content validity index (CVI) were measured. The items with CVR <0.6 and CVI <0.7 were deleted. The reliability of the instrument was assessed by calculating the inter-rater reliability, test re-test reliability, and internal consistency. Quantitative data analysis was done in SPSS software, version 22.

Results The instrument with 66 items, including 32 items for the cognitive domain, 22 items for the language domain (13 for receptive and 9 for expressive components), 12 items for the motor domain (9 for fine and 3 for gross motor skills), and 14 items for the general section, was assessed in stability and reliability. The intraclass correlation coefficient (ICC) for the test re-test reliability was obtained 0.83 for the specific section and 0.86 for the general section. The ICC for the inter-rater reliability was 0.80 (95% CI: 58-82). The highest correlation value was related to the cognitive scale (0.88), and the lowest value was for the gross motor skill subscale (0.76). The kappa coefficient for the inter-rater reliability of the general section was above 0.7. The kappa coefficient for the specific section ranged from 1 to 0.35. The Cronbach's α for the internal consistency of the general section was 0.76.

Conclusion The designed questionnaire is a valid and reliable tool to evaluate the clinical performance of Bayley-III test examiners in Iran.

Keywords Bayley-III, Validity, Reliability, Infant development, Clinical performance

Received: 08 Dec 2022
Accepted: 16 Jul 2023
Available Online: 01 Jan 2024

Corresponding Author:
Nahideh Hasani Khiabani
Address: Children Development Section, Zahra Mardani Azar Children's Hospital, Tabriz University of Medical Sciences, Tabriz, Iran.
Tel: +98 (914) 3006285
E-Mail: nahid_md2000@yahoo.com

English Version

Introduction

The value of early diagnosis of developmental disorders and providing intervention services for infants and toddlers has gained increasing significance. Timely and periodic evaluation of development offers the possibility of early diagnosis and treatment and prevents the loss of developmental potential of the child [7, 8]. The reason for increasing efforts in diagnosis at an early age is the cost-effectiveness and higher impact of intervention programs at this stage [9]. The early interventions improve children's developmental prognosis [10] with short- and long-term positive effects [11, 12]. Currently, most specialists monitor a child's development based on their clinical judgment, but several studies have shown that clinical judgment alone is ineffective in diagnosing developmental delays [9, 13, 14]. The evidence indicates that, on average, only 30% of children with developmental or behavioral problems are diagnosed by a doctor without a screening tool [11, 15, 16]. For this purpose, many developmental screening tests have been designed and provided to children's caregivers. Diagnostic tests of developmental disorders should be used to definitively diagnose the cases referred to as developmental delay in the screening tests.

Compared to the several screening tests, such as the ages and stages questionnaires (ASQ-3) and the Denver developmental screener test (Denver-II), diagnostic tests for developmental disorders are limited and, in most cases, only diagnose developmental disorders in one area. These tests include Battelle developmental inventory screener (BDI-2), MacArthur-Bates short-forms (SFI and SFII) tests, and World Health Organization (WHO) motor milestones.

The "Bayley developmental scales" diagnostic test is one of the few globally valid diagnostic tests that, in addition to being comprehensive in all developmental areas, has high psychometric indicators. This test is an individual evaluation tool that assesses the developmental performance of children aged 1 to 42 months in 3 developmental domains: Cognitive, language, and motor. The evaluation of these three areas is done objectively by the examiner.

Therefore, caregivers in Iran's health system have chosen and implemented Bayley's test to diagnose and evaluate developmental disorders in children. To do this screening, first, the instructors train examiners to perform the test in a theoretical and practical training course. After completing the training course, the testers should practically implement the test in healthcare centers to acquire the necessary skills. They should send three videos of the test done on a healthy child to be evaluated by the instructors and get a certificate of skill in performing the test. Because there is no tool to assess the competence in conducting the test inside and outside the country, there have always been challenges in accepting or rejecting the clinical qualification of these testers by different trainers.

Therefore, the current research was conducted at the request of the Ministry of Health and Medical Education to design a standard instrument with acceptable validity and reliability to evaluate the clinical performance of Bayley test examiners.

Materials and Methods

This study was carried out in two parts: Building the tool and determining its validity and reliability in six steps. These steps are as follows: Conducting a review of the literature, explaining the concepts of clinical practice according to the results of the review of the literature and the opinions of experts, compiling the initial draft and the final items of the tool, adjusting the scoring structure and its interpretation, determining the face validity of the items by calculating the impact score, determining the quantitative content validity of the items by calculating the relative content validity ratio (CVR) and the content validity index (CVI).

The target population was the examiners of the Bayley test, and the statistical population was the trainers or evaluators. The inclusion criteria included national instructors and examiners with 5 years of experience in conducting the test. The exclusion criterion was an unwillingness to participate in the study. A purposive sampling was used to select the participants. The sample comprised 10 people for calculating face and content validity and 8 for internal consistency, inter-rater reliability, and test re-test.

After a focused group discussion with a team of experts, collecting their opinions, and reviewing the literature, we concluded that this concept includes two parts: Specific clinical performance (the performance of the examiner in implementing the test items according to the instructions in three cognitive, linguistic, and motor scales) and general clinical performance (examiner's performance in complying with the general requirements of the test).

The initial draft had 111 items, compiled in the specific section (94 items) and the general section (17 items) using the Delphi method. Specific items were scored on a 5-point Likert scale: 1) Does not comply with the guidelines at all, 2) Does not comply with the guidelines in many cases, 3) Does not comply with the guidelines in some cases, 4) Is acceptable but needs to be improved, and 5) It is favorable and following the instructions. To determine quantitative face validity, the initial draft was sent to the group of experts, and they were asked to give each item a score of 1 to 5 according to its importance. In this evaluation, a score of 1 indicates the lowest, and a score of 5 indicates the highest level of importance. In the next step, items with an impact score of less than 1.5 were removed, and the structure of other items was reviewed and revised. CVR and CVI were used to determine the content validity of the tool. The final items of the tool were sent to a group of 10 trainers and examiners, and they were asked to evaluate each item in terms of three aspects, including "necessary", "helpful but not necessary", and "not necessary", to determine the CVR of each item. According to the minimum acceptable CVR based on the number of experts, items with a CVR score of less than 0.6 were excluded. To calculate the CVI of each item, experts were asked to determine the degree of clarity, simplicity, and relevance of each item with a 4-part spectrum. If the CVI of an item was less than 0.7, the item was removed; if it was between 0.7 and 0.79, it was revised; and if it was greater than 0.79, it was considered acceptable.

Determining the reliability of the tool was implemented in three stages. Six complete videos of performing the Bayley test of six age groups (3-4 months, 5-6 months, 12 months, 20 months, 31-32 months, and 42 months) were prepared and sent to 8 trainers.

In the first step, to determine the reliability of the tests, the evaluators were asked to evaluate the films twice, with a 2-week interval (without referring to the results of the first time), and complete the checklist (test re-test). The scores of the evaluators were summarized and subjected to statistical analysis by calculating the intra-cluster correlation coefficient (ICC).

In the second step, ICC was used to determine the instrument's reliability between evaluators with a confidence interval (CI) of 95%. According to studies, ICC values less than 0.5 indicate poor reliability, values between 0.5 and 0.75 indicate moderate reliability, values between 0.75 and 0.90 indicate good reliability and values above 0.9 indicate an excellent estimate of reliability [17].

In the third step, the Cronbach's α coefficient was used to check the internal consistency of the items. This coefficient ranges between 0 and 1, and the values closer to 1 indicate that the instrument under study has a higher internal consistency. The acceptable Cronbach's α coefficient is often considered more than 0.70.

Finally, the Cohen kappa coefficient was used to determine the inter-rater agreement between each observer and the reference observer. The size of the kappa coefficient in statistical analysis ranges from -1 to +1. The closer this number is to 1, the more proportional the agreement is, and a kappa value above 0.6 is acceptable [18].

Results

Quantitative face validity

Based on the analysis of Likert scores, the items with an impact score of less than 4 (11 items from the cognitive part, 3 items from receptive communication, 4 items from expressive communication, 4 items from gross movements, and 2 items from the general section) were removed. The structure of other items was reviewed and revised (Table 1).

Content validity

According to the minimum acceptable CVR based on the number of experts (10 people), none of the items had a CVR less than 0.6, and no item was deleted. Based on the analysis of CVI scores, the items whose content validity index is less than 0.7 should be removed, and the items between 0.7 and 0.79 should be revised. In this study, all items had a content validity index higher than 0.79, except for item 29 of the cognitive scale, whose content validity index was equal to 0.7 and was revised.

The final tool was compiled with 66 items in the specific section (including 32 cognitive items, 13 receptive communication items, 9 expressive communication items, 9 fine movements items, and 3 gross movements items) and 14 items in the general section.

Reliability

The reliability of the tool was measured in two parts:

1. Time consistency (test re-test) with the ICC method was calculated for 6 age groups and two sections, including cognitive linguistic (receptive and expressive communication), movement (fine and gross movements),

Table 1. The number of deleted items of the initial draft of the tool based on the impact score and its importance

Area	Number of Deleted Items
Cognitive	11
Receptive communication	3
Expressive communication	4
Fine movements	-
Gross movements	4
General	2

Archives of Rehabilitation

and general sections (Table 2). The ICC rate in the specific and general sectors was 0.83 and 0.86, respectively.

2. Inter-rater reliability by ICC method with 95% CI (Table 3) for movement items was 0.76 (95% CI, 54%-88%), linguistic items 0.81 (95% CI, 60%, 87%), cognitive items 0.88 (95% CI, 71%, 91%), the general section 0.81 (95% CI, 61%, 86%), and 0.80 for the whole instrument (95% CI, 58%, 82%).

3. The internal consistency of the items was calculated by the Cronbach's α obtained at 0.76 for all the items in the general section of the tool (Table 4).

4. The degree of agreement between 7 observers and the reference observer was calculated with the Cohen kappa coefficient. The highest rate was in movement items, and the general section in 3 cases, the cognitive scale in 1 case (equal to 1), and the lowest rate was in observer number 2 (Table 5). Generally, the Cohen kappa coefficient in different scales indicates the appropriate agreement between each observer and the reference observer.

Discussion

Studies show that a standard tool is essential for evaluating the skills of training recipients, and evaluation with

Table 2. Test re-test reliability by scales and age groups

Intra Cluster Correlation Coefficient (ICC)	Age Group 1 (3-4 m)	Age Group 2 (5-6 m)	Age Group 3 (12 m)	Age Group 4 (20 m)	Age Group 5 (31-32 m)	Age Group 6 (42 m)	Total
Cognitive scale	0.81	0.86	0.84	0.86	0.88	0.88	0.85
Linguistic scale (expressive and receptive communication)	0.88	0.88	0.91	0.86	0.87	0.89	0.85
Receptive	0.85	0.87	0.88	0.86	0.81	0.87	0.85
Expressive	0.81	0.86	0.84	0.86	0.86	0.87	0.84
Motor scale (fine movements-gross movements)	0.88	0.87	0.88	0.86	0.86	0.89	0.85
Fine movements	0.89	0.88	0.90	0.88	0.88	0.90	0.86
Gross movements	0.84	0.83	0.84	0.81	0.83	0.87	0.84
Special skills	0.85	0.80	0.84	0.86	0.86	0.82	0.83
General skills	0.88	0.87	0.84	0.90	0.91	0.88	0.86

Archives of Rehabilitation

Table 3. Inter-rater reliability (intra-cluster correlation coefficient)

Component	ICC	95% CI
Gross movements domain	0.76	61-84
Fine movements domain	0.87	66-96
Motor scale	0.74	54-88
Expressive communication domain	0.86	62-87
Receptive communication domain	0.87	60-89
Linguistic scale	0.81	60-87
Cognitive scale	0.88	71-91
General	0.81	61-86
Total	0.80	58-82

ICC: Intra-cluster correlation coefficient; CI: Confidence interval.

Table 4. Internal consistency of items by six age groups

Cronbach's α Internal Consistency	Age Group 1 (3-4 m)	Age Group 2 (5-6 m)	Age Group 3 (12 m)	Age Group 4 (20 m)	Age Group 5 (31-32 m)	Age Group 6 (42 m)	Total
General component	0.78	072	0.78	0.82	0.88	0.78	0.76

a standard tool minimizes the possibility of error and the bias of the evaluator's opinions. Finally, it improves the quality of the evaluation. Therefore, this study was conducted for the first time in Iran to design a valid and reliable tool to evaluate the clinical performance of the Bayley test examiner at the request of the Ministry of Health and Medical Education in Iran.

The review of domestic and foreign studies showed no tool specifically designed to evaluate the clinical performance of the Bayley tester. For this reason, similar tools have been investigated.

In a research conducted by Shayan et al. [3] to design and compile the logbook of 22 healthcare worker learners in Iran, a questionnaire with 17 modules and 205 skills was developed based on the approved curriculum of healthcare workers. The opinions of experts confirmed the face and content validity of the questionnaire. To determine the reliability of the questionnaire using the Cronbach's α, only the internal consistency of

Table 5. Inter-rater agreement (Cohen's kappa coefficient) of the instrument

Cohen's Kappa Coefficient	Observer 1	Observer 2	Observer 3	Observer 4	Observer 5	Observer 6	Observer 7
Cognitive items	0.65	0.65	0.73	0.83	0.83	0.83	1
Linguistic items	0.71	0.51	0.55	0.75	0.45	0.72	0.83
Motor items	0.75	0.35	1	1	0.75	1	1
General component	0.75	1	1	1	0.65	0.75	0.83

the items was evaluated, and the test re-test and inter-rater reliability were not measured [3]. Our study developed a tool consisting of specific (66 items) and general (14 items) sections according to the skills the examiner should have in evaluating children aged 1 to 42 months in the Bayley test. Similar to the structure of the Bayley test, the specific part was taken from the content of the test to measure cognitive, linguistic, and motor scales, and the general part included the items that the examiner must follow while performing the test. The steps of conducting the above study are carefully designed, but unlike our study, validity, reliability, and agreement between evaluators have not been calculated.

Another study was conducted in three stages by Mansourian et al. (2015), including the preparation of a checklist for evaluating the practical skills of dental students and determining its validity and reliability in the first stage, the application of the checklist in the second stage, and examination of the results in the third stage. First, according to the goals in the approved educational curriculum, a draft of the checklist, including clinical skills, was prepared. To determine its validity, the checklist was approved by 5 professors and lecturers. The Lawshe method was used to check the content validity of the instrument. The test re-test method was used to check the reliability of the checklist. A video of performing the desired clinical skill was prepared and provided to 5 professors. Two weeks later, the same professors reviewed and graded the film. The correlation coefficient of the test and re-test scores was equal to 0.98. [4]. In this study, as in our study, the content validity and reliability of the checklist have been determined using exact statistical methods. In our study, the ICC at two times (test and re-test) was 0.83 in the specific sector and 0.86 in the general sector, and the ICC in the evaluators' observations (95% CI, 82%, 58%) was 0.80.

In another study conducted by Mokarami H et al. (2019) to design a tool to evaluate the internship course in the field of occupational health engineering, during 4 sessions, three areas necessary for a graduate of this field were defined. The initial questionnaire was designed with 44 items. In the next step, the importance of each item was checked by 10 experts. CVR and CVI were used to check quantitative content validity. The average CVI was equal to 0.85, the average CVR was equal to 0.76, and the impact scores of all the items of the tool were higher than 1.5. Internal consistency and test re-test methods were utilized to determine the reliability of the checklist. Finally, the final questionnaire included three fields and a total of 40 items. The Cronbach's α coefficient of the final questionnaire was 0.835 [5]. It is worth mentioning that this study seeks to evaluate an internship course from an educational level. Therefore, it considers three areas: Education and learning, behavioral and management goals. However, in our study, the competence of the testers in conducting an observational test was considered. The specific part of our tool also corresponds to the educational and learning scope of the above study, and the general part also corresponds to the behavioral goals in the above research.

Another study was conducted by Yaghini et al. [6] to design a tool for medical students' clinical competence to determine the minimum training and the number of clinical encounters during the pediatric internship. First, the educational minimums of the pediatric department of general medicine were extracted from the logbook. Then, these educational minimums were designed as a checklist using experts' opinions. In the second step, the face and content validity of the checklist was confirmed by the faculty members of the pediatrics department. To confirm the reliability, this checklist was used to determine the number of clinical encounters, and the reliability of this checklist was calculated with the Cronbach's α coefficient at 0.82 [6]. It is noteworthy that in this study, the method of determining face and content validity and reliability is not clear and detailed.

Another study was conducted by Sahebalzamani et al. [19] (2012) to investigate the validity and reliability of using the direct observation of practical skills to evaluate the clinical skills of nursing students. A list of practical nursing procedures was provided to 45 professionals to determine content validity. Then, the evaluation checklist was developed. To evaluate inter-observer reliability, 20 students were assessed by two testers. The Cronbach's α coefficient was used to determine reliability by the total internal consistency method. The reliability of the test was measured by Cronbach's α coefficient of 0.94. The lowest and highest correlation coefficient values in the reliability between evaluators were 0.42 and 0.84, respectively, which were significant in all cases (P=0.001) [19]. It is worth mentioning that in this study, 20 students were evaluated by two testers to assess the inter-observer reliability. It seems that the determination of reliability with the similarity of the opinion of only two examiners should be investigated.

Another study was conducted by Jabbari et al.[20] entitled "designing and determining the validity and reliability of the tool for evaluating the clinical competence of 24 occupational therapists". In this research, the initial questionnaire was prepared with 128 items. After reviews and revisions by the expert group, the checklist

was reduced to 66 items. The Lawshe method was used to check face and content validity, and CVR and CVI were used to determine content validity. The questionnaire was reduced from 66 items to 54 items. The correlation coefficient of two tests in two weeks was used to confirm the tool's reliability. The Cronbach's α coefficient was also calculated to determine the internal correlation of the questionnaire. The cut-off point of this questionnaire was calculated as 162 [20]. The above study used a suitable method in designing and determining validity and reliability.

Conclusion

One of the limitations of our study was the small number of qualified trainers as evaluators in the country. Also, according to the structure of the Bayley test, it was impossible to measure the internal consistency in the specific section.

Considering that communication skills are an important part of health services and the ability to establish proper communication forms the basis of clinical practice, it is suggested that a suitable tool be designed to evaluate the communication skills of the Bayley tester.

So far, the evaluation of the clinical performance of the Bayley test examiner has been done only by watching the videos sent by the examiner and without a standard tool. Due to the appraiser's opinions, the results are not acceptable. Therefore, for the first time, this study designed a comprehensive and valid tool to measure the clinical performance of the Bayley test examiner in Iran.

According to the qualitative stages of tool design, the tool designed to evaluate the clinical performance of the Bayley test examiner has the necessary validity and reliability.

Ethical Considerations

Compliance with ethical guidelines

Before starting the study, the ethical code number was obtained from the Ethics Committee of the University of Social Welfare and Rehabilitation Sciences (Code: IR.USWR.REC.1400.273). Written consent was obtained in simple and fluent language to film the children and send them to the expert group. The children's parents were fully justified about the reason for filming and how to use these videos. The parents were assured that this consent was completely optional and they would not be deprived of the center's services if they did not consent to filming.

Funding

This study was conducted with the financial support of the Children's Health Department of the Ministry of Health and Medical Education.

Authors' contributions

Conceptualization, methodology, and writing the original draft: All authors; Validation, analysis, research, review, and sources: Farin Soleimani, Nahida Hasani Khiabani, and Leila Yazdi; Editing: Farin Soleimani and Nahida Hasani Khiabani; Project management: Farin Soleimani.

Conflict of interest

All authors declared no conflict of interest.

Acknowledgments

The authors would like to express their sincere gratitude to the Pediatric Neurorehabilitation Research Center, research vice president of Rehabilitation and Social Health Science University, Tabriz University of Medical Sciences vice president of health, Shahid Beheshti University of Medical Sciences vice president of health and mentors and examiners of the Bayley test of the countries medical sciences universities who participated as expert groups in the design of instrument items and evaluation of videos and also children's parents are appreciated for this project.

مقاله پژوهشی

طراحی و تعیین روایی و پایایی ابزار ارزیابی عملکرد بالینی آزمونگر بیلی-۳

فرین سلیمانی[۱]، *ناهیده حسنی خیابانی[۲]، لیلا یزدی[۳]، حمیدرضا لرنژاد[۴]، ناریا ابوالقاسمی[۴]، غزال شریعت‌پناهی[۵]

۱. مرکز تحقیقات توانبخشی اعصاب اطفال، دانشگاه علوم توانبخشی و سلامت اجتماعی، تهران، ایران.
۲. بخش تکامل کودکان، بیمارستان کودکان زهرا مردانی‌آذر، دانشگاه علوم‌پزشکی تبریز، تبریز، ایران.
۳. مرکز تکامل همایون، دانشگاه علوم‌پزشکی شهید بهشتی، تهران، ایران.
۴. اداره سلامت کودکان، وزارت بهداشت، درمان و آموزش پزشکی، تهران، ایران.
۵. گروه بیماری‌های عفونی کودکان، بیمارستان بهرامی، دانشگاه علوم‌پزشکی تهران، تهران، ایران.

Use your device to scan and read the article online

Citation Soleimani F, Hasani Khiabani N, Yazdi L, Lornejad H, Aboulghasemi N, Shariatpanahi Gh. [Designing and Assessing the Validity and Reliability of the Bayley-III Test Examiner Clinical Performance Scale (Persian)]. Archives of Rehabilitation. 2024; 24(4):566-585. https://doi.org/10.32598/RJ.24.4.3676.1

https://doi.org/10.32598/RJ.24.4.3676.1

چکیده

هدف در چند سال اخیر برنامه ارزیابی و ارتقای تکامل کودکان، موردتوجه و تأکید نظام سلامت واقع شده است. در این راستا آموزش آزمونگران تست بیلی (Bayley) و ارزیابی ایشان، به یکی از نیازهای نظام سلامت تبدیل شده است. این پژوهش با هدف طراحی یک ابزار جامع و معتبر و تعیین روایی و پایایی آن برای ارزیابی عملکرد بالینی آزمونگر بیلی انجام شد.

روش بررسی جامعه پژوهش آزمونگران تست بیلی در دانشگاه‌های علوم‌پزشکی کشور بودند. در این مطالعه نمونه‌گیری مبتنی بر هدف انجام شد. ابتدا مفاهیم ارزیابی بالینی از طریق مرور متون و جلسات بحث گروهی متمرکز با ۱۰ نفر از خبرگان تبیین شد و گویه‌های ابزار طراحی شد. جهت بررسی روایی ظاهری از نمره تأثیر استفاده شد. پرسش‌نامه به گروه خبرگان متشکل از ۸ نفر از مربیان (ارزیابان) و آزمونگران باتجربه (ارزیابی‌شدگان) ارسال شد و از آن‌ها خواسته شد تا به گویه‌ها از نظر اهمیت از ۱ تا ۵ نمره دهند.گویه‌هایی که نمره تأثیر کمتر از ۱/۵ داشتند حذف شدند و ساختار سایر گویه مورد اصلاح و بازبینی ظاهری قرار گرفت. برای بررسی روایی محتوا از دو ضریب نسبی روایی محتوا و شاخص روایی محتوا استفاده شد و پرسش‌نامه به یک گروه ۱۰نفری از خبرگان ارسال شد. با توجه به تعداد گروه خبرگان، گویه‌هایی که میزان ضریب نسبی روایی محتوای کمتر از ۰/۶ و شاخص روایی محتوای کمتر از ۰/۷ داشتند حذف شدند. پایایی ابزار از طریق تکرارپذیری دفعات آزمون (آزمون ـ بازآزمون)، ضریب توافق در مشاهدات ارزیابان، مقایسه ارزیابان با ارزیاب مرجع و هم‌خوانی درونی تعیین شد. آنالیز داده‌های کمی با نرم‌افزار SPSS نسخه ۲۲ انجام گرفت.

یافته‌ها ابزار در بخش اختصاصی با ۶۶ گویه شامل ۳۲ گویه شناختی، ۲۲ گویه زبانی (۱۳ گویه درکی، ۹ گویه بیانی)، ۱۲ گویه حرکتی (۹ گویه حرکات ظریف، ۳ گویه حرکات درشت) و در بخش عمومی با ۱۴ گویه وارد مرحله سنجش ثبات زمانی و پایایی شد. میزان پایایی آزمون ـ بازآزمون با ضریب همبستگی درون‌رده‌ای در بخش اختصاصی ۰/۸۳ و در بخش عمومی ۰/۸۶ و ضریب همبستگی درون‌رده‌ای در مشاهدات ارزیابان ۰/۸۰ (فاصله اطمینان ۹۵ درصد: ۵۸-۸۲) بود. بیشترین ضریب همبستگی مربوط به مقیاس شناختی (۰/۸۸) و کمترین در حرکات درشت (۰/۷۶) بود. ضریب کاپا در بین مشاهده‌گران در بخش عمومی بالای ۰/۷ بود. بالاترین ضریب کاپا در بین مشاهده‌گران در حیطه اختصاصی برابر ۱ و کمترین ۰/۳۵ بود. میزان هم‌خوانی درونی گویه‌ها (آلفای کرونباخ) در بخش عمومی ۰/۷۶ بود.

نتیجه‌گیری ابزار طراحی‌شده در ارزیابی مهارت‌های بالینی آزمونگر بیلی دارای روایی و پایایی است.

کلیدواژه‌ها آزمون بیلی، روایی و پایایی، تکامل کودک، عملکرد بالینی

تاریخ دریافت: ۱۷ آذر ۱۴۰۱
تاریخ پذیرش: ۲۵ تیر ۱۴۰۲
تاریخ انتشار: ۱۱ دی ۱۴۰۲

* **نویسنده مسئول:**
ناهیده حسنی خیابانی
نشانی: تبریز، دانشگاه علوم‌پزشکی تبریز، بیمارستان کودکان زهرا مردانی‌آذر، بخش تکامل کودکان.
تلفن: ۱۵۹۱۳۵۸ (۹۱۴) ۹۸+
رایانامه: nahid_md2000@yahoo.com

مقدمه

صلاحیت از موضوعات چالش‌برانگیز در بخش مراقبت‌های سلامت است که در حیطه‌های مختلف درمانی ازجمله آموزش و بالین و مدیریت از اهمیت بسیاری برخوردار است [۱]. صلاحیت بالینی به‌کارگیری مدیرانه مهارت‌های تکنیکی و ارتباطی، دانش، استدلال بالینی، عواطف و ارزش‌ها در محیط بالینی است. ارزیابی صلاحیت بالینی، فرایندی پیچیده و ترکیبی از مراحل مختلف ارزیابی است [۲].

تا کنون ارزیابی صلاحیت بالینی در مطالعات مختلف موردبررسی قرار گرفته است. مثلاً در زمینه ارزیابی آموزش بهورزان در کشور، شایان و همکاران، پژوهشی با عنوان «طراحی و تدوین کارنامه عملکرد[1] فراگیران بهورزی کشوری» انجام دادند. پس از مطالعه متون، پرسش‌نامه اولیه تدوین و از نمونه‌ها نظرسنجی شد. روایی ظاهری و محتوایی پرسش‌نامه با استفاده از نظرات صاحب‌نظران و مربیان بهورزی تعیین شد. قابلیت اعتماد یا پایایی پرسش‌نامه با استفاده از آزمون آلفای کرونباخ تعیین شد. پرسش‌نامه‌ها با ۱۷ پودمان و ۲۰۵ مهارت تدوین شدند. پس از نظرسنجی سطح شایستگی تک‌تک مهارت‌ها به دست آمد. نتیجه این پژوهش ابزار ارزشمندی را در جهت ارزیابی توانمندی فراگیران بهورزی معرفی می‌کند [۳].

در پژوهشی، «طراحی چک‌لیست ارزیابی مهارت عملی دانشجویان دندانپزشکی» در ۳ مرحله اجرا شد: مرحله اول، تهیه چک‌لیست و تعیین روایی و پایایی آن، مرحله دوم تکمیل چک‌لیست برای یک گروه از دانشجویان توسط ۴ نفر از اساتید و گروه دیگر به روش رتبه‌بندی جهانی[2] و مرحله سوم بررسی پیامدها. نتیجه این مطالعه عینی‌تر بودن و توجه بیشتر به اجزای موردارزیابی در روش چک‌لیست و رضایت بیشتر دانشجویان را نشان داد [۴].

در پژوهش «تدوین و اعتبارسنجی ابزار ارزش‌یابی دوره کارآموزی در عرصه رشته مهندسی بهداشت حرفه‌ای» که در دانشکده بهداشت دانشگاه علوم‌پزشکی شاهرود انجام شد، ابزاری برای ارزش‌یابی دانشجویان بهداشت حرفه‌ای طراحی شد. نتایج حاکی از آن بود که ابزار ارزش‌یابی تدوین‌شده، دارای ویژگی‌های روان‌سنجی مناسبی است و به‌عنوان یک مقیاس استاندارد برای ارزش‌یابی کارآموزان دوره کارآموزی بهداشت حرفه‌ای قابل استفاده است [۵].

پژوهش «طراحی ابزاری برای ارزیابی صلاحیت بالینی دانشجویان پزشکی در بخش اطفال» نشان داد با توجه به حجم زیاد دانش در رشته پزشکی و ضرورت یادگیری مهارت یا دانش، لازم است این اصول در ارزیابی صلاحیت بالینی و طراحی ابزار ارزش‌یابی جهت بهبود عملکرد افراد نسبت به وظایف شغلی آینده و نیازهای جامعه در نظر گرفته شود [۶].

امروزه اهمیت تشخیص زودرس اختلالات تکاملی و ارائه خدمات مداخله‌ای در شیرخواران و نوپایان ارزش روزافزونی پیدا کرده است. ارزیابی به‌موقع و دوره‌ای تکامل، امکان تشخیص و درمان زودرس را فراهم کرده و مانع از دست رفتن پتانسیل‌های تکاملی کودک می‌شود [۷، ۸]. دلیل تلاش فزاینده در جهت تشخیص در سنین پایین‌تر آن است که برنامه‌های مداخله‌گرانه مقرون‌به‌صرفه هستند و چنانچه در سنین پایین‌تر شروع شوند تأثیر بیشتری دارند [۹]، پیش‌آگهی تکاملی کودکان را بهبود می‌بخشند [۱۰] و دارای اثرات مثبت کوتاه‌مدت و بلندمدت هستند [۱۱، ۱۲]. در حال حاضر اغلب متخصصین براساس قضاوت بالینی خود این پایش را انجام می‌دهند. ولی مطالعات متعددی نشان داده‌اند قضاوت بالینی به‌تنهایی در تشخیص تأخیر تکامل کارایی لازم را ندارد [۹، ۱۳، ۱۴]. شواهد حاکی از آن است که به‌طور متوسط تنها ۳۰ درصد از کودکان دارای مشکلات تکاملی یا رفتاری توسط پزشک بدون استفاده از ابزار غربالگر تشخیص داده می‌شوند [۱۱، ۱۵، ۱۶]. به همین منظور آزمون‌های غربالگر تکاملی متعددی، طراحی و در اختیار مراقبین سلامت کودکان قرار گرفته است. به‌منظور تشخیص قطعی مواردی که در آزمون‌های غربالگری به‌عنوان تأخیر تکامل ارجاع شده‌اند، می‌بایست از آزمون‌های تشخیصی اختلالات تکاملی استفاده کرد. در مقایسه با تعدد آزمون‌های غربالگری مثل پرسش‌نامه سنین و مراحل[3] و غربالگری تکاملی دنور[4]، آزمون‌های تشخیصی اختلالات تکاملی محدود بوده و در اغلب موارد تنها به تشخیص اختلال تکاملی در یک حیطه اکتفا می‌کنند مثل فهرست تکاملی باتل[5]، آزمون مک‌آرتور[6] و تست مراحل تکامل حرکتی[7].

آزمون تشخیصی «مقیاس‌های تکاملی بیلی[8]» از معدود آزمون‌های تشخیصی معتبر جهانی است که علاوه بر جامعیت در تمامی حیطه‌های تکاملی از شاخص‌های روان‌سنجی[9] بالایی نیز برخوردار است. این آزمون، یک ابزار ارزیابی فردی است که عملکرد تکاملی کودکان ۱-۴۲ماهه را در ۳ حیطه تکاملی شناختی، زبانی و حرکتی ارزیابی می‌کند. ارزیابی این ۳ حیطه به‌شکل عینی توسط آزمونگر انجام می‌پذیرد.

بنابراین آزمون بیلی جهت تشخیص اختلالات تکاملی در نظام سلامت کشور انتخاب شد و هم‌اکنون در حال اجراست. آموزش آزمونگران جهت انجام تست توسط مربیان در یک دوره آموزش نظری و عملی انجام می‌گیرد. پس از اتمام دوره آموزشی،

1. Logbook
2. Global rating
3. Ages and Stages Questionnaires (ASQ-3)
4. Denver Developmental Screening Test (Denver-II)
5. Battelle Developmental Inventory screener (BDI-2)
6. MacArthur-Bates Short-Forms (SFI and SFII) tests
7. WHO Motor Milestones
8. Bayley developmental scales
9. Psychometric properties including validity and reliability

آزمونگران به‌منظور کسب مهارت‌های لازم، اجرای آزمون را در مراکز بهداشتی به‌صورت عملی تمرین کرده و پس از کسب آمادگی، ۳ فیلم از اجرای تست برای کودک سالم جهت ارزیابی توسط مربیان و کسب گواهی جهت مهارت و توانمندی انجام تست، ارسال می‌کنند. با توجه به اینکه ابزار ارزیابی صلاحیت در انجام تست در داخل و در خارج از کشور وجود ندارد، همیشه چالش‌هایی در قبولی یا رد صلاحیت بالینی این آزمونگران توسط مربیان مختلف مطرح بوده است.

بنابراین پژوهش حاضر به درخواست وزارت بهداشت با هدف طراحی ابزاری استاندارد با روایی و پایایی قابل‌قبول به‌منظور ارزیابی مهارت و توانمندی عملکرد بالینی آزمونگران تست بیلی-۳ انجام شد.

روش‌ها

این مطالعه در ۲ بخش ساخت ابزار و سپس تعیین روایی و پایایی آن، در ۶ گام طراحی شد. جامعه هدف، آزمونگران آزمون بیلی وزارت بهداشت در کشور بودند. جامعه آماری، مربیان یا ارزیابان (منتخب وزارت در سطح کشورکه دارای توانایی بالا در انجام آزمون هستند) و آزمونگران باتجربه یا ارزیابی‌شدگان در دانشگاه‌های علوم‌پزشکی مختلف ایران در سال‌های ۱۴۰۰ تا ۱۴۰۱ بودند. معیارهای ورود به مطالعه عبارت بودند از: مربیان کشوری، آزمونگرانی با سابقه ۵ سال انجام آزمون که قبلاً این دوره آموزشی را گذرانده و جهت انجام تست موردتأیید بودند. معیارهای خروج از مطالعه، عدم تمایل به شرکت در مطالعه و جلسات یا پاسخ‌دهی تعیین شد.

نمونه‌گیری به‌صورت مبتنی بر هدف بود. حجم نمونه در محاسبه روایی صوری و محتوایی ۱۰ نفر، همبستگی درونی ۴۰ نفر و پایایی نمره‌دهی بین ارزیابان و دفعات آزمون ۸ نفر بود.

گردآوری داده‌ها با ابزار «ارزیابی عملکرد بالینی آزمونگر بیلی» انجام شد که براساس متغیرهای مدنظر در فاز اول مطالعه طراحی شد و سپس روایی ظاهری و محتوایی کمی، همبستگی درونی گویه‌ها و پایایی نمره‌دهی بین ارزیابان و دفعات آزمون سنجیده شد.

جهت طراحی ابزار ۷ گام انجام شد: ۱) مرور متون، ۲) تبیین مفاهیم عملکرد بالینی با توجه به نتایج حاصل از مرور متون و نظرات خبرگان، ۳) تدوین پیش‌نویس اولیه و گویه‌های نهایی ابزار، ۴) تنظیم ساختار نمره‌دهی و تفسیر آن، ۵) تعیین روایی ظاهری گویه‌ها با محاسبه نمره تأثیر، ۶) تعیین روایی محتوایی کمی گویه‌ها با محاسبه ضریب نسبی روایی محتوا[10] و شاخص روایی محتوا[11].

10. Content validity ratio (CVR)
11. Content validity index (CVI)

مرور متون، با کلیدواژه‌های: ارزیابی عملکرد بالینی[12]، ارزیابی مهارت[13]، ارزیابی آموزشی[14]، ارزیابی عملکرد[15]، توانمندی بالینی[16]، توانمندی حرفه‌ای[17]، ارزیابی خبرگی[18]، کارپوشه[19]، عملکرد بالینی[20]، آزمون مشاهده مستقیم مهارت‌های عملی[21]، کتابچه گزارش روزانه عملکرد یا کارنامه عملکرد[22]، ارزیابی توسط همتایان[23]، مهارت‌های عملی، مشاهده، آزمون بیلی-۳ و براساس معادل‌های مش در پایگاه‌های اینترنتی پاب‌مد، اوویدمدلاین، گوگل‌اسکولار، مگ‌ایران، مدلاین و اس‌آی‌دی به زبان‌های فارسی و انگلیسی بدون محدودیت زمانی انجام گرفت. هدف از انجام این مرحله، شناسایی مؤلفه‌های ارزیابی آزمونگران در گروه‌های سنی مختلف به‌منظور بخش‌ها و استخراج گویه‌های ارزیابی بود.

پس از بحث متمرکز گروهی با گروه خبرگان ۱۰نفری (شامل مربیان و آزمونگران باتجربه و متخصصان در زمینه تکامل کودک) و جمع‌آوری نظرات آن‌ها و مرور متون، در تبیین عملکرد بالینی آزمونگر به این نتیجه رسیدیم که این مفهوم شامل ۲ بخش است:

عملکرد بالینی اختصاصی: عبارت است از عملکرد آزمونگر در اجرای گویه‌های آزمون مطابق با دستورالعمل در ۳ مقیاس شناختی، زبانی (درکی و بیانی)، و حرکتی (حرکات ظریف و حرکات درشت).

عملکرد بالینی عمومی: عبارت است از عملکرد آزمونگر در رعایت موارد کلی انجام آزمون شامل رعایت سکوت هنگام انجام آزمون، حذف عوامل حواس‌پرتی کودک، محاسبه دقیق سن تقویمی کودک و سن اصلاح‌شده در صورت نارسی، رعایت قانون بازگشت و رعایت قانون توقف آزمون.

پیش‌نویس اولیه دارای ۱۱۱ گویه بود که در بخش اختصاصی مشتمل بر شناختی (۴۵ گویه)، ارتباط درکی (۱۸ گویه)، ارتباط بیانی (۱۴ گویه)، حرکات ظریف (۹ گویه) و حرکات درشت (۸ گویه) و در بخش عمومی (۱۷ گویه) توسط گروه تحقیق و خبرگان به روش دلفی تدوین شد.

پس از بحث گروهی با گروه خبرگان ۸نفری (شامل مربیان و آزمونگران باتجربه آزمون)، با توافق همه اعضا، نمره‌دهی گویه‌های بخش اختصاصی به شکل لیکرت پنج‌تایی (با تفسیر نمره ۱= هیچ‌گونه مطابقتی با دستورالعمل ندارد، نمره ۲= در بسیاری از

12. Clinical performance assessment
13. Training assessment
14. Educational assessment
15. Performance assessment
16. Clinical competency
17. Professionalism
18. Assessment of expertise
19. Portfolios
20. Clinical function
21. Direct observation of procedural skills (DOPS)
22. logbook
23. Peer assessment

موارد با دستورالعمل مطابقت ندارد، نمره ۳= در بعضی موارد با دستورالعمل مطابقت ندارد، نمره ۴= قابل‌قبول است ولی نیاز به ارتقا دارد، نمره ۵= مطلوب و مطابق با دستورالعمل است) تدوین شد. درمورد گویه‌های بخش عمومی طیف لیکرت پنج‌تایی (با تفسیر ضعیف، متوسط، خوب، بسیارخوب و عالی) موردتوافق قرار گرفت.

جهت تعیین روایی ظاهری کمی، پیش‌نویس اولیه به گروه خبرگان ۸نفری (شامل مربیان و آزمونگران باتجربه آزمون) ارسال و از آن‌ها درخواست شد که با توجه به میزان اهمیت، به هر گویه نمره ۱ تا ۵ بدهند. در این ارزیابی، نمره ۱ نشان‌دهنده کمترین و نمره ۵ نشان‌دهنده بیشترین میزان اهمیت بود. برای تعیین نمره اهمیت[24] و نمره تأثیر[25] هر گویه از **فرمول‌های شماره ۱** و **۲** استفاده شد:

۱. نمره اهمیت= تعداد افراد ارزیابی‌کننده ÷ مجموع نمرات گویه

۲. نمره تأثیر= نمره اهمیت × درصد فراوانی کسانی که امتیاز ۴ و ۵ به گویه داده‌اند.

در مرحله بعد گویه‌هایی که نمره تأثیر کمتر از ۱/۵ داشتند حذف شدند و ساختار سایر گویه‌ها مورد اصلاح و بازبینی قرار گرفت.

به‌منظور تعیین روایی محتوایی ابزار، از دو ضریب CVR و CVI استفاده شد.

گویه‌های نهایی ابزار به یک گروه ۱۰نفری از مربیان و آزمونگران باتجربه ارسال و از آن‌ها خواسته شد جهت تعیین CVR، هر یک از گویه‌ها را (ضروری است، مفید است ولی ضروری نیست، ضرورتی ندارد) ارزیابی کنند.

برای محاسبه CVR از **فرمول شماره ۳** استفاده شد:

$$3.\ CVR=\frac{ne-N/2}{N/2}$$

با توجه به حداقل CVR قابل‌قبول براساس تعداد خبرگان، گویه‌هایی که نمره CVR کمتر از ۰/۶ داشتند حذف شدند.

برای محاسبه CVI هر یک از گویه‌ها، از خبرگان خواسته شد میزان واضح بودن، سادگی و مرتبط بودن هر گویه را با طیف چهارقسمتی زیر تعیین کنند:

- واضح بودن (کاملاً واضح است، واضح اما نیاز به بازبینی دارد، نیاز به بازبینی جدی دارد، مبهم است)
- سادگی (کاملاً ساده است، ساده اما نیاز به بازبینی دارد، نیاز به بازبینی جدی دارد، مبهم است)
- مرتبط بودن (کاملاً مرتبط است، مرتبط اما نیاز به بازبینی دارد، نیاز به بازبینی جدی دارد، مرتبط نیست)

برای محاسبه CVI، تعدادی که گزینه اول و دوم را انتخاب کرده بودند بر تعداد کل تقسیم، و اگر حاصل آن از ۰/۷ کوچک‌تر بود گویه حذف و اگر بین ۰/۷ تا ۰/۷۹ بود بازبینی، و اگر از ۰/۷۹ بزرگ‌تر بود قابل‌قبول در نظر گرفته شد. با در نظر گرفتن نمرات CVR و CVI هر یک از گویه‌ها و همچنین براساس نظرات پیشنهادی خبرگان مبنی بر خلاصه کردن متن گویه‌ها، ابزار نهایی تدوین شد.

تعیین پایایی ابزار، در ۳ مرحله انجام شد. ابتدا طبق نظر تیم تحقیق، ۶ فیلم کامل از اجرای آزمون بیلی که توسط آزمونگران با سطح توانایی‌های متفاوت (به‌دلیل تعیین تفاوت نمره‌دهی ابزار در شناسایی آزمونگران با مهارت‌های متفاوت) از ۶ گروه سنی (۳-۴ماهه، ۵-۶ماهه، ۱۲ماهه، ۲۰ماهه، ۳۱-۳۲ماهه و ۴۲ماهه) تهیه شده بود به ۸ نفر از مربیان آزمون بیلی جهت ارزیابی ارسال شد. شایان ذکر است که آزمون بیلی در ۱۷ گروه سنی از ۱۵روزگی تا ۴۲ماهگی قابل ارزیابی است و با افزایش سن، گویه‌ها پیچیده‌تر می‌شوند. بنابراین در هر گروه سنی تعدادی از گویه‌ها توسط آزمونگر مورد مشاهده قرار می‌گیرد. ۶ گروه سنی با توجه به پوشش تمام گویه‌ها در ابزار نهایی انتخاب شد تا تمام گویه‌ها موردارزیابی قرار گیرند.

در مرحله اول، جهت تعیین پایایی در دفعات آزمون، از ارزیابان خواسته شد فیلم‌ها را دوبار، با فاصله زمانی ۲ هفته (بدون رجوع به نتایج بار اول) ارزیابی کرده و چک‌لیست را تکمیل کنند (آزمون ـ بازآزمون). نمرات ارزیابان در بار اول و دوم به‌طور جداگانه جمع‌بندی شد و با روش محاسبه ضریب همبستگی داخل‌خوشه‌ای[26] مورد تجزیه‌وتحلیل آماری قرار گرفت.

در مرحله دوم جهت تعیین پایایی ابزار بین ارزیابان[27] از ICC با فاصله اطمینان ۹۵ درصد استفاده شد. طبق مطالعات، مقادیر ICC کمتر از ۰/۵ نشانه پایایی ضعیف، مقادیر ۰/۵-۰/۷۵ نشانه پایایی متوسط و مقادیر ۰/۷۵-۰/۹۰ نشانه پایایی خوب و بالای ۰/۹ نشانه پایایی عالی است [۱۷].

در مرحله سوم، جهت هم‌خوانی درونی گویه‌ها از روش آلفای کرونباخ استفاده شد. هم‌خوانی درونی شاخصی است که همگنی موجود بین تغییرات نمرات خرده‌مقیاس‌ها بین افراد یک نمونه را در یک مقطع زمانی نشان می‌دهد. ضریب آلفای کرونباخ رایج‌ترین شاخصی است که در این گونه مطالعات برای ارزیابی هم‌خوانی درونی استفاده می‌شود. این ضریب بین صفر تا ۱ متغیر بوده و هرچه به ۱ نزدیک‌تر باشد نشان‌دهنده این است که ابزار موردمطالعه از هم‌خوانی درونی بالاتری برخوردار است.

24. Importance score
25. Impact score
26. Intra cluster correlation coefficient (ICC)
27. Intra observer

غالباً ضریب آلفای کرونباخ مناسب بیش از ۰/۷۰ در نظر گرفته می‌شود.

درنهایت، جهت تعیین میزان توافق بین هر مشاهده‌گر با مشاهده‌گر مرجع (که توسط تیم تحقیق مشخص شد) از ضریب کاپای کوهن[۲۸] استفاده شد. اندازه ضریب کاپا در تحلیل‌های آماری ۱- تا ۱+ است. هرچه این عدد به ۱ نزدیک‌تر باشد بیانگر وجود توافق متناسب است و اندازه کاپای بالای ۰/۶ موردقبول است [۱۸].

یافته‌ها

روایی ظاهری کمی

براساس نتایج حاصل از تحلیل نمرات لیکرت، گویه‌هایی که نمره تأثیر کمتر از ۴ داشتند (بخش شناختی ۱۱ گویه، ارتباط درکی ۳ گویه، ارتباط بیانی ۴ گویه، حرکات درشت ۴ گویه و عمومی ۲ گویه) حذف شدند و ساختار سایر گویه‌ها مورد اصلاح و بازبینی قرار گرفت **(جدول شماره ۱)**.

روایی محتوایی

با توجه به حداقل CVR قابل‌قبول براساس تعداد خبرگان (۱۰ نفر)، هیچ‌کدام از گویه‌ها CVR کمتر از ۰/۶ نداشتند و هیچ گویه‌ای حذف نشد.

براساس تجزیه‌وتحلیل نمرات CVI گویه‌هایی که شاخص روایی محتوای آن‌ها کمتر از ۰/۷ هستند باید حذف شوند و گویه‌هایی که بین ۰/۷ تا ۰/۷۹ هستند باید مورد بازبینی قرار گیرند. در این مطالعه همه گویه‌ها شاخص روایی محتوای بالاتر از ۰/۷۹ داشتند به‌غیر از گویه ۲۹ مقیاس شناختی که شاخص روایی محتوایی آن معادل ۰/۷ بود و مورد بازبینی قرار گرفت. میزان CVR و CVI هر یک از بخش‌های اختصاصی و عمومی ابزار مطابق **جدول شماره ۲** است.

ابزار نهایی با ۶۶ گویه در بخش اختصاصی (شامل ۳۲ گویه شناختی، ۱۳ گویه ارتباط درکی، ۹ گویه ارتباط بیانی، ۹ گویه حرکات ظریف و ۳ گویه حرکات درشت) و ۱۴ گویه در بخش عمومی تدوین شد.

پایایی

سنجش پایایی ابزار در ۲ بخش انجام گرفت:

ثبات زمانی (آزمون ـ بازآزمون) با روش ICC به تفکیک ۶ گروه سنی و در ۲ بخش اختصاصی شامل گویه‌های شناختی، زبانی (ارتباط درکی و بیانی)، حرکتی (حرکات ظریف و درشت) و بخش عمومی محاسبه شد **(جدول شماره ۳)**. میزان ICC در بخش اختصاصی و در بخش عمومی به‌ترتیب۰/۸۳ و ۰/۸۶ بود.

پایایی بین ارزیابان نیز با روش ICC با فاصله اطمینان ۹۵ درصد محاسبه شد **(جدول شماره ۴)** و در گویه‌های حرکتی ۰/۷۶ (فاصله اطمینان ۹۵ درصد معادل ۵۴-۸۸)، گویه‌های زبانی ۰/۸۱ (فاصله اطمینان ۹۵ درصد معادل ۶۰-۸۷)، گویه‌های شناختی ۰/۸۸ (فاصله اطمینان ۹۵ درصد معادل ۷۱-۹۱) و در بخش عمومی ۰/۸۱ (فاصله اطمینان ۹۵ درصد معادل ۶۱-۸۶) و در کل ابزار ۰/۸۰ (فاصله اطمینان ۹۵ درصد معادل ۵۸-۸۲) بود.

هم‌خوانی درونی گویه‌ها به تفکیک ۶ گروه سنی با استفاده از آلفای کرونباخ محاسبه شد و در کل گویه‌های بخش عمومی ابزار معادل ۰/۷۶ بود **(جدول شماره ۵)**.

میزان توافق بین ۷ مشاهده‌گر با مشاهده‌گر مرجع با ضریب کاپای کوهن محاسبه شد. بالاترین میزان، در گویه‌های حرکتی و بخش عمومی هر کدام در ۳ مورد و مقیاس شناختی در ۱ مورد (معادل ۱) و کمترین میزان در مشاهده‌گر شماره ۲ بود **(جدول شماره ۶)**. به‌صورت کلی ضریب کاپای کوهن در مقیاس‌های مختلف، بیانگر توافق مناسب بین هر کدام از مشاهده‌گران با مشاهده‌گر مرجع است.

28. Cohen's kappa coefficient

جدول ۱. تعداد گویه‌های حذف‌شده پیش‌نویس ابزار براساس نمره تأثیر و اهمیت

حیطه	تعداد گویه‌های حذف‌شده
شناختی	۱۱
ارتباط درکی	۳
ارتباط بیانی	۴
حرکات ظریف	–
حرکات درشت	۴
عمومی	۲

توانبخشی

جدول ۲. میزان ضریب نسبی روایی محتوا و شاخص روایی محتوای گویه‌های ابزار

حیطه شناختی (شماره گویه)	CVR*	واضح بودن	سادگی	مرتبط بودن
		CVI**	CVI	CVI
۱	۰/۸	۱	۱	۰/۹
۲	۰/۸	۱	۱	۰/۹
۳	۱	۱	۱	۰/۸
۴	۰/۸	۱	۱	۰/۹
۵	۰/۸	۱	۱	۰/۹
۶	۱	۱	۱	۰/۸
۷	۰/۶	۱	۱	۱
۸	۱	۱	۱	۱
۹	۱	۱	۱	۱
۱۰	۰/۸	۱	۱	۰/۹
۱۱	۱	۱	۱	۱
۱۲	۰/۸	۰/۹	۱	۰/۹
۱۳	۱	۱	۱	۰/۹
۱۴	۱	۱	۱	۰/۹
۱۵	۰/۸	۱	۱	۱
۱۶	۰/۶	۰/۹	۰/۹	۰/۸
۱۷	۰/۸	۱	۱	۱
۱۸	۱	۱	۱	۱
۱۹	۱	۱	۱	۰/۹
۲۰	۰/۸	۱	۱	۱
۲۱	۰/۸	۱	۱	۱
۲۲	۰/۸	۱	۱	۱
۲۳	۰/۶	۱	۱	۱
۲۴	۰/۸	۱	۱	۰/۹
۲۵	۰/۸	۱	۱	۰/۹
۲۶	۰/۶	۱	۱	۱
۲۷	۰/۸	۱	۱	۰/۹
۲۸	۰/۸	۱	۱	۱
۲۹	۰/۶	۰/۷	۰/۷	۰/۷
۳۰	۰/۸	۱	۱	۰/۹

حیطه شناختی (شماره گویه)	CVR*	واضح بودن	سادگی	مرتبط بودن
		CVI**	CVI	CVI
۳۱	۰/۸	۰/۹	۰/۹	۰/۹
۳۲	۱	۱	۱	۱
حیطه ارتباط درکی (شماره گویه)	**CVR**	**واضح بودن**	**سادگی**	**مرتبط بودن**
		CVI**	CVI	CVI
۱	۰/۸	۱	۱	۰/۹
۲	۰/۸	۱	۱	۰/۹
۳	۰/۸	۱	۱	۰/۹
۴	۱	۱	۱	۰/۹
۵	۰/۸	۱	۱	۰/۹
۶	۱	۰/۹	۰/۹	۰/۸
۷	۰/۸	۰/۹	۰/۹	۰/۹
۸	۰/۶	۰/۸	۰/۹	۰/۸
۹	۰/۶	۱	۱	۱
۱۰	۰/۶	۱	۱	۰/۸
۱۱	۰/۶	۱	۱	۰/۸
۱۲	۰/۸	۰/۹	۱	۰/۹
۱۳	۰/۶	۱	۱	۰/۹
۱۴	۰/۸	۱	۱	۰/۹
حیطه ارتباط بیانی (شماره گویه)	**CVR**	**واضح بودن**	**سادگی**	**مرتبط بودن**
		CVI**	CVI	CVI
۱	۱	۱	۱	۱
۲	۱	۱	۱	۰/۹
۳	۰/۸	۱	۱	۱
۴	۱	۱	۱	۱
۵	۱	۱	۱	۱
۶	۱	۱	۱	۰/۹
۷	۱	۱	۱	۱
۸	۰/۸	۱	۱	۱
۹	۱	۱	۱	۰/۸
۱۰	۱	۱	۱	۰/۸

حیطه حرکات ظریف (شماره گویه)	CVR	واضح بودن	سادگی	مرتبط بودن
		CVI**	CVI	CVI
۱	۰/۸	۱	۱	۱
۲	۱	۱	۱	۱
۳	۰/۸	۱	۰/۹	۱
۴	۰/۸	۱	۰/۹	۱
۵	۰/۸	۱	۱	۱
۶	۰/۸	۱	۱	۱
۷	۰/۶	۱	۱	۰/۹
۸	۰/۸	۱	۱	۱
۹	۰/۶	۱	۱	۱

حیطه حرکات درشت (شماره گویه)	CVR	واضح بودن	سادگی	مرتبط بودن
		CVI**	CVI	CVI
۱	۰/۸	۱	۱	۰/۹
۲	۱	۱	۱	۱
۳	۱	۱	۱	۱

مهارت‌های عمومی (شماره گویه)	CVR	واضح بودن	سادگی	مرتبط بودن
		CVI**	CVI	CVI
۱	۱	۱	۱	۱
۲	۰/۶	۱	۱	۱
۳	۰/۶	۱	۱	۱
۴	۰/۶	۱	۱	۱
۵	۱	۱	۱	۱
۶	۱	۱	۱	۱
۷	۱	۱	۱	۱
۸	۱	۱	۱	۱
۹	۰/۶	۱	۱	۱
۱۰	۰/۸	۱	۱	۱
۱۱	۱	۱	۱	۱
۱۲	۱	۱	۱	۱
۱۳	۱	۱	۱	۱
۱۴	۱	۱	۱	۱

توانبخشی

جدول ۳. پایایی دفعات آزمون به تفکیک مقیاس‌ها و گروه‌های سنی

ضریب همبستگی داخل‌خوشه‌ای*	گروه سنی ۱ (۴-۳ ماه)	گروه سنی۲ (۶-۵ ماه)	گروه سنی ۳ (۱۲ ماه)	گروه سنی۴ (۲۰ ماه)	گروه سنی ۵ (۳۲-۳۱ ماه)	گروه سنی ۶ (۴۲ ماه)	کل
مقیاس شناختی	۰/۸۱	۰/۸۶	۰/۸۴	۰/۸۶	۰/۸۸	۰/۸۸	۰/۸۵
مقیاس زبانی (ارتباط درکی و بیانی)	۰/۸۸	۰/۸۸	۰/۹۱	۰/۸۶	۰/۸۷	۰/۸۹	۰/۸۵
درکی	۰/۸۵	۰/۸۷	۰/۸۸	۰/۸۶	۰/۸۱	۰/۸۷	۰/۸۵
بیانی	۰/۸۱	۰/۸۶	۰/۸۴	۰/۸۶	۰/۸۶	۰/۸۷	۰/۸۴
مقیاس حرکتی (حرکات ظریف ـ حرکات درشت)	۰/۸۸	۰/۸۷	۰/۸۸	۰/۸۶	۰/۸۶	۰/۸۹	۰/۸۵
حرکات ظریف	۰/۸۹	۰/۸۸	۰/۹۰	۰/۸۸	۰/۸۸	۰/۹۰	۰/۸۶
حرکات درشت	۰/۸۴	۰/۸۳	۰/۸۴	۰/۸۱	۰/۸۳	۰/۸۷	۰/۸۴
مهارت‌های اختصاصی	۰/۸۵	۰/۸۰	۰/۸۴	۰/۸۶	۰/۸۶	۰/۸۲	۰/۸۳
مهارت‌های عمومی	۰/۸۸	۰/۸۷	۰/۸۴	۰/۹۰	۰/۹۱	۰/۸۸	۰/۸۶

توانبخشی

جدول ۴. پایایی بین ارزیابان (ضریب همبستگی داخل‌خوشه‌ای)

بخش	ضریب همبستگی داخل خوشه‌ای *	فاصله اطمینان ۹۵ درصد**
حیطه حرکات درشت	۰/۷۶	۸۴-۶۱
حیطه حرکات ظریف	۰/۸۷	۹۶-۶۶
مقیاس حرکتی	۰/۷۴	۸۸-۵۴
حیطه ارتباط بیانی	۰/۸۶	۸۷-۶۲
حیطه ارتباط درکی	۰/۸۷	۸۹-۶۰
مقیاس زبانی	۰/۸۱	۸۷-۶۰
مقیاس شناختی	۰/۸۸	۹۱-۷۱
بخش عمومی	۰/۸۱	۸۶-۶۱
کل ابزار	۰/۸۰	۸۲-۵۸

توانبخشی

بحث

مطالعات نشان می‌دهند وجود یک ابزار استاندارد برای ارزیابی مهارت گیرندگان آموزش ضروری است و ارزیابی با یک ابزار استاندارد، احتمال خطا و اعمال سلیقه شخصی ارزیاب را به حداقل رسانده و کیفیت ارزیابی را ارتقا می‌دهد. بنابراین، این مطالعه برای اولین‌بار در ایران با هدف طراحی یک ابزار روا و پایا جهت ارزیابی عملکرد بالینی آزمونگر تست بیلی به درخواست وزارت بهداشت انجام شد.

گویه‌های این ابزار با اجماع نظر محققین و گروه خبرگان شامل مربیان (ارزیابان) و آزمونگران باتجربه (ارزیابی‌شدگان)، تدوین شد. روایی ظاهری کمی گویه‌ها، با استفاده از نمره تأثیر و روایی محتوایی با ضریب نسبی روایی محتوی و شاخص روایی محتوی بالا تعیین شد. پایایی ابزار با ضریب همبستگی داخل‌خوشه‌ای ابزار

جدول ۵. هم‌خوانی درونی گویه‌های ابزار به تفکیک ۶ گروه سنی

هم‌خوانی درونی گویه‌ها (آلفای کرونباخ)	گروه سنی ۱ (۴-۳ ماه)	گروه سنی۲ (۶-۵ ماه)	گروه سنی ۳ (۱۲ ماه)	گروه سنی۴ (۲۰ ماه)	گروه سنی ۵ (۳۲-۳۱ ماه)	گروه سنی ۶ (۴۲ ماه)	کل
بخش عمومی	۰/۷۸	۰/۷۲	۰/۷۸	۰/۸۲	۰/۸۸	۰/۷۸	۰/۷۶

توانبخشی

جدول ۶. میزان توافق (ضریب کاپا کوهن) ابزار

میزان توافق (ضریب کاپا)	مشاهده‌گر					
	۱	۲	۳	۴	۵	۶	۷
گویه‌های شناختی	۰/۶۵	۰/۶۵	۰/۷۳	۰/۸۳	۰/۸۳	۰/۸۳	۱
گویه‌های زبانی	۰/۷۱	۰/۵۱	۰/۵۵	۰/۷۵	۰/۴۵	۰/۷۲	۰/۸۳
گویه‌های حرکتی	۰/۷۵	۰/۳۵	۱	۱	۰/۷۵	۱	۱
بخش عمومی	۰/۷۵	۱	۱	۱	۰/۶۵	۰/۷۵	۰/۸۳

توانبخشی

در دفعات آزمون و بین ارزیابان اندازه‌گیری شد که نشان‌دهنده پایایی خوب و قابل اطمینان ابزار بود. هم‌خوانی درونی گویه‌ها بالا بود و ضریب کاپای کوهن مناسب بیانگر توافق مناسب و قابل‌قبول بین ارزیابان بود.

بررسی مطالعات داخل و خارج از کشور نشان داد ابزاری که به‌طور خاص برای ارزیابی عملکرد بالینی آزمونگر بیلی طراحی شده باشد وجود ندارد. به همین دلیل ابزارهای مشابه موردبررسی قرار گرفت.

در پژوهش شایان و همکاران، با هدف طراحی و تدوین کارنامه عملکرد فراگیران بهورزی کشور، پس از مرور متون و تشکیل گروه کانونی، پرسش‌نامه‌ای با ۱۷پودمان و ۲۰۵ مهارت براساس کوریکولوم مصوب آموزش بهورزی تدوین شد. روایی ظاهری و محتوایی پرسش‌نامه با استفاده از نظرات صاحب‌نظران آموزش بهورزی وزارت بهداشت و مسئولین آموزش بهورزی و چند تن از مدیران و مربیان بهورزی تعیین شد. این مطالعه شامل ۷ مرحله بود: ۱) مطالعه گسترده و میدانی، ۲) تعیین مؤلفه‌های اساسی و مهارت‌های موردنیاز فراگیران در قالب ۱۷ پودمان و ۲۰۵ مهارت، ۳) مطالعه و تدوین پرسش‌نامه در ۵ سطح، ۴) تعیین روایی محتوایی و ظاهری پرسش‌نامه با یک جلسه کانونی، ۵) توزیع پرسش‌نامه‌ها با پست الکترونیک، ۶) جمع‌بندی اطلاعات دریافتی، و ۷) تجمیع اطلاعات دریافتی با آمار توصیفی. با نهایی شدن اطلاعات، درنهایت کارنامه عملکرد پیشنهادی در ۲ بخش پودمان‌ها و مهارت‌های پروسیجرال تدوین شد. در این پژوهش به‌منظور تعیین پایایی پرسش‌نامه با روش آلفای کرونباخ، تنها هم‌خوانی درونی گویه‌ها ارزیابی شده است و پایایی دفعات آزمون و پایایی بین ارزیابان سنجیده نشده است [۳]. در مطالعه ما ابزاری مشتمل بر ۲ بخش اختصاصی (۶۶ گویه) و عمومی (۱۴ گویه) با توجه به مهارت‌هایی که آزمونگر باید در ارزیابی کودکان ۱-۴۲ماهه در تست بیلی داشته باشند تبیین شد. همانند ساختار تست بیلی، بخش اختصاصی برگرفته از محتوای تست جهت سنجش ۳ مقیاس شناختی، زبانی و حرکتی بود و بخش عمومی شامل گویه‌هایی بود که آزمونگر در حین انجام کار باید در انجام تست رعایت کند. مراحل انجام مطالعه فوق دقیق طراحی شده، ولی برخلاف مطالعه ما بررسی روایی و پایایی و توافق بین ارزیابان محاسبه نشده است.

مطالعه منصوریان و همکاران در ۳ مرحله انجام شد. مرحله اول: تهیه چک‌لیست ارزیابی مهارت‌های عملی دانشجویان دندانپزشکی و تعیین روایی و پایایی آن، مرحله دوم: کاربرد چک‌لیست، مرحله سوم: بررسی پیامدها (الف: تعیین میزان رضایت‌مندی ارزیابان و ارزیابی‌شوندگان، ب: آنالیز آماری). جهت تهیه چک‌لیست، ابتدا با توجه به اهداف آموزشی مندرج در کوریکولوم آموزشی مصوب، پیش‌نویسی از چک‌لیست مشتمل بر مهارت‌های بالینی تهیه شد. این چک‌لیست جهت تعیین روایی ظاهری به تأیید ۵ نفر از اساتید و مدرسین رسید. برای تعیین روایی محتوایی از روش لاوشه استفاده شد. بدین‌ترتیب که نسخه‌ای از چک‌لیست در اختیار ۵ تن از اساتید متخصص در رشته دهان و فک و صورت قرار گرفت و از آن‌ها خواسته شد تا هر سؤال را به ۳ شکل «ضروری»، «مفید» و «غیرضروری» مورد داوری قرار دهند. نظر داوران با CVR و CVI تعیین شد. برای بررسی پایایی چک‌لیست، از روش آزمون ـ بازآزمون استفاده شد. فیلمی از انجام مهارت بالینی موردنظر تهیه شد و در اختیار ۵ نفر از اساتید قرار گرفت و اساتید با مشاهده فیلم اقدام به نمره‌دهی کردند. ۲ هفته بعد، فیلم مزبور توسط همان اساتید بررسی و نمره‌دهی شد. میزان ضریب همبستگی نمرات قبل و بعد برابر ۰/۹۸ بود [۴]. در این مطالعه نیز مانند مطالعه ما روایی محتوایی و پایایی چک‌لیست با استفاده از روش‌های دقیق آماری تعیین شده است. ICC قبل و بعد در مطالعه ما در بخش اختصاصی ۰/۸۳ و در بخش عمومی ۰/۸۶ و ضریب همبستگی درون‌رده‌ای در مشاهدات ارزیابان ۰/۸۰ (فاصله اطمینان ۹۵ درصد: ۵۸-۸۲) بود.

در مطالعه مکرمی و همکاران که به‌منظور طراحی ابزاری جهت ارزش‌یابی دوره کارآموزی در عرصه مهندسی بهداشت حرفه‌ای انجام شد، در ابتدا ۴ جلسه با متخصصان آموزش پزشکی، آموزشی و پژوهشی با محوریت وظایف حرفه‌ای دانش‌آموختگان تشکیل شد و درنهایت ۳ حیطه تعریف شد که ضروری است تا یک فارغ‌التحصیل این رشته برمبنای آن تربیت شود. حیطه

اول: اهداف آموزشی و یادگیری، حیطه دوم: مهارت‌های مدیریتی و فردی (اهداف رفتاری)، و حیطه سوم: توسعه صلاحیت‌های حرفه‌ای شغلی و کارآفرینی (اهداف مدیریتی). سپس معیارهای ارزش‌یابی جامع و متناسب با این اهداف سه‌گانه تهیه شد. این معیارها به‌عنوان یک چارچوب مفهومی برای طراحی گویه‌ها مورداستفاده قرار گرفتند. پرسش‌نامه اولیه با ۴۴ گویه طراحی شد. از ۴۰ متخصص تقاضا شد تا درمورد رعایت دستور زبان، جمله‌بندی و قرارگیری عبارات در جای مناسب نظر خود را اعلام کنند. در مرحله بعد میزان اهمیت هر گویه از نظر ۱۰ متخصص با استفاده از نمرات تأثیر آیتم بررسی شد. برای بررسی روایی محتوایی کمی، از CVR و CVI استفاده شد. میانگین CVI برابر با ۰/۸۵، میانگین CVR برابر با ۰/۷۶ و نمرات تأثیر همه گویه‌های ابزار بالاتر از ۱/۵ بود. برای تعیین پایایی چک‌لیست، همسانی درونی و بازآزمایی برآورد شد. درنهایت پرسش‌نامه نهایی شامل ۳ حیطه و درمجموع ۴۰ گویه تدوین شد. میزان ضریب آلفای کرونباخ کل پرسش‌نامه نهایی ۰/۸۳۵ به دست آمد [۵]. شایان ذکر است که این مطالعه به‌دنبال ارزش‌یابی یک دوره کارورزی از یک مقطع آموزشی است و بنابراین ۳ حیطه آموزشی و یادگیری، اهداف رفتاری و اهداف مدیریتی را در نظر گرفته است. ولی در مطالعه ما صلاحیت آزمونگران در انجام یک تست مشاهده‌ای مدنظر بود. البته بخش اختصاصی ابزار ما نیز منطبق بر حیطه آموزشی و یادگیری مطالعه فوق و بخش عمومی نیز منطبق بر اهداف رفتاری در مطالعه فوق است.

مطالعه یقینی و همکاران با هدف طراحی ابزاری برای بررسی صلاحیت بالینی دانشجویان پزشکی در بخش اطفال با تعیین حداقل‌های آموزشی[۲۹] و تعداد مواجهات بالینی دوره کارآموزی اطفال انجام شد. بدین‌منظور ابتدا حداقل‌های آموزشی گروه اطفال رشته پزشکی عمومی از لاگ‌بوک استخراج شد و سپس این حداقل‌های آموزشی با نظرخواهی از اعضای شورای آموزشی گروه اطفال و معاونت آموزشی به‌صورت چک‌لیستی طراحی شد. در مرحله دوم این چک‌لیست در اختیار ۲۷ نفر از اعضای هیئت علمی ۳ دانشگاه قرار گرفت و از طریق نظرسنجی میزان مواجهات بالینی موردنیاز تعیین شد. چک‌لیست طراحی‌شده به‌صورت ۱۰ گروه گردش‌های آموزش بالینی شامل: نوزادان، عفونی، گوارش، کلیه، تنفس، خون، اورژانس، عمومی، غدد و درمانگاه و نیز ۶۳ حداقل آموزش بالینی تعیین شد. روایی ظاهری و محتوایی چک‌لیست توسط اعضای هیئت علمی گروه اطفال تأیید شد. برای تأیید پایایی، این چک‌لیست در تعیین میزان مواجهات بالینی استفاده شد و پایایی این چک‌لیست با ضریب آلفای کرونباخ ۰/۸۲ تأیید شد [۶]. آنچه در این مطالعه جای بحث و بررسی دارد این است که روش تعیین روایی ظاهری و محتوایی و همچنین روش‌های تعیین پایایی با جزئیات مشخص نیست.

مطالعه صاحب‌الزمانی و همکاران با هدف بررسی روایی و پایایی استفاده از روش مشاهده مستقیم مهارت‌های عملی[۳۰] برای ارزیابی مهارت‌های بالینی دانشجویان پرستاری انجام شد. جهت تعیین روایی محتوایی، فهرستی از پروسیجرهای DOPS رشته پرستاری در اختیار ۴۵ نفر افراد حرفه‌ای شامل اعضای هیئت علمی گروه داخلی جراحی و سرپرستاران باتجربه قرار گرفت و از آنان خواسته شد پروسیجرهای اساسی در پرستاری را رتبه‌بندی کنند. سپس چک‌لیست ارزیابی براساس این رتبه‌بندی تدوین شد. برای همگن شدن قضاوت آزمونگرها، دستورالعمل‌های نمره‌دهی و راهنمای استفاده از چک‌لیست در اختیار آنان قرار گرفت. دانشجویان نیز براساس راهنمای نوشتاری که شامل اهداف پژوهش، نحوه ارزیابی با روش DOPS، نوع پروسیجرها، اسامی آزمونگرها و چک‌لیست ارزیابی بود در یک جلسه توجیهی، توجیه شدند. برای هر دانشجو ۸ آزمون در طی ۶ ماه انجام شد. برای ارزیابی پایایی بین مشاهده‌گران ۲۰ نفر دانشجو توسط ۲ آزمونگر ارزیابی شدند. جهت تعیین پایایی به روش همسانی درونی کل از ضریب آلفای کرونباخ استفاده شد.

همبستگی نمرات DOPS با میانگین نمرات نظری و بالینی دانشجویان به‌ترتیب ۰/۱۱۷ (P=۰/۴۲۹) و ۰/۳۷۶ (P=۰/۰۰۸) به دست آمد. پایایی آزمون توسط ضریب آلفای کرونباخ ۰/۹۴ محاسبه شد. کمترین و بیشترین مقدار ضریب همبستگی در پایایی بین ارزیابان به‌ترتیب ۰/۴۲ و ۰/۸۴ بود که در تمام موارد معنی‌دار (P=۰/۰۰۱) بودند [۱۹]. شایان ذکر است که در این مطالعه برای ارزیابی پایایی بین مشاهده‌گران ۲۰ نفر دانشجو توسط ۲ آزمونگر ارزیابی شدند که به نظر می‌رسد تعیین پایایی با همسانی نظر تنها ۲ آزمونگر و اینکه آیا این ۲ آزمونگر نسبت به یکدیگر بی‌اطلاع بودند یا نه، جای بررسی دارد.

مطالعه جباری و همکاران با هدف طراحی و تعیین روایی و پایایی ابزار ارزیابی صلاحیت بالینی کاردرمانگران انجام شد. در این پژوهش، ابتدا در دوره ۲ماهه، براساس مرور متون و چک‌لیست‌ها و ابزارهای موجود در منابع، صلاحیت بالینی تعریف و حیطه‌های آن مشخص شد. سپس براساس مطالعات انجام‌شده و نقش و وظایف کاردرمانی، گویه‌های هر حیطه تدوین و پرسش‌نامه اولیه با ۱۲۸ گویه تنظیم شد. هیئت خبرگان این پرسش‌نامه را دوبار در ۲ جلسه ۲ساعته بررسی کردند و پس از اصلاحات و بازبینی به ۶۶ گویه کاهش یافت. سپس به‌منظور بررسی روایی ظاهری و محتوایی به روش لاوشه در اختیار ۱۵ نفر از خبرگان قرار گرفت. در این روش CVR تک‌تک گویه‌ها محاسبه شد و گویه‌هایی که CVR کمتر از ۰/۴۹ داشتند حذف شد و پرسش‌نامه از ۶۶ گویه به ۵۴ گویه کاهش یافت. برای تأیید پایایی

29. Core curriculum

30. Direct observation of procedural skills (DOPS)

ابزار روی ۳۰ کاردرمانگر، ضریب همبستگی دو آزمون در فاصله ۲ هفته محاسبه شد و آزمون همبستگی پیرسون (r=۰/۹۹۵) پایایی ابزار را تأیید کرد. برای تعیین همبستگی درونی پرسش‌نامه نیز ضریب آلفای کرونباخ محاسبه شد. نقطه برش این پرسش‌نامه ۱۶۲ محاسبه شد [۲۰]. مطالعه فوق از روش مناسبی در طراحی و تعیین روایی و پایایی استفاده کرده است.

نتیجه‌گیری

تا کنون ارزیابی عملکرد بالینی آزمونگر تست بیلی فقط با مشاهده فیلم‌های ارسالی توسط آزمونگر و بدون وجود یک ابزار استاندارد انجام گرفته و به‌سبب اعمال سلیقه ارزیاب، نتایج هم‌خوانی قابل‌قبولی نداشته‌اند. بنابراین، این مطالعه برای اولین‌بار به طراحی یک ابزار جامع و معتبر برای سنجش عملکرد بالینی آزمونگر تست بیلی در ایران پرداخت که می‌تواند چالش‌های وزارت بهداشت در ارزیابی آزمونگران بعد از اتمام دوره آموزشی نظری و عملی را مرتفع نماید. همچنین ارزیابی دوره‌ای مربیان و آزمونگران را جهت تأیید صلاحیت بالینی در طول زمان ممکن می‌سازد.

با توجه به مراحل کیفی طراحی ابزار، استفاده از مرور متون گسترده و نظرات خبرگان در طراحی گویه‌ها و با توجه به روایی و پایایی بالایی که در روان‌سنجی مشخص شد، ابزار طراحی‌شده جهت ارزیابی عملکرد بالینی آزمونگر تست بیلی از اعتبار و پایایی لازم برخوردار است.

از محدودیت‌های مطالعه ما، تعداد کم مربیان واجد شرایط به‌عنوان ارزیاب در سطح کشور بود و ما توانستیم از ۸ ارزیاب جهت پایایی بین ارزیابان، هم‌خوانی درونی گویه‌ها و میزان توافق بین مشاهده‌گران در ۶ گروه سنی استفاده کنیم. با توجه به اینکه هر گروه سنی به بررسی گویه‌هایی خاص که مربوط به همان گروه سنی است می‌پردازد، امکان سنجش هم‌خوانی درونی در بخش اختصاصی امکان‌پذیر نبود.

پیشنهاد می‌شود با توجه به اینکه مهارت‌های ارتباطی به‌عنوان جزئی مهم از خدمات سلامت شناخته شده و توانایی برقراری ارتباط مناسب، پایه و اساس عملکرد بالینی را تشکیل می‌دهد ابزار مناسب جهت ارزیابی مهارت ارتباطی آزمونگر بیلی طراحی شود.

ملاحظات اخلاقی

پیروی از اصول اخلاق پژوهش

قبل از شروع مطالعه، کد اخلاقی به شماره IR.USWR.REC.1400.273 از کمیته اخلاق دانشگاه علوم توانبخشی و سلامت اجتماعی اخذ شد. به‌منظور فیلم‌برداری از کودکان و ارسال آن‌ها به گروه خبرگان، یک رضایت‌نامه کتبی به زبان کاملاً ساده و روان اخذ شد. والدین کودکان در زمینه علت فیلم‌برداری و نحوه استفاده از این فیلم‌ها کاملاً توجیه شدند. والدین اطمینان یافتند که این رضایت کاملاً اختیاری است و در صورت عدم رضایت آن‌ها برای فیلم‌برداری، از خدمات مرکز محروم نخواهند شد.

حامی مالی

این مطالعه با حمایت مالی اداره سلامت کودکان وزارت محترم بهداشت و درمان انجام شد.

مشارکت نویسندگان

مفهوم‌سازی و روش‌شناسی: همه نویسندگان؛ اعتبارسنجی، تحلیل، تحقیق و بررسی و منابع: دکتر فرین سلیمانی، ناهیده حسنی خیابانی، لیلا یزدی؛ نگارش: همه نویسندگان؛ ویراستاری: دکتر فرین سلیمانی، ناهیده حسنی خیابانی؛ مدیریت پروژه: دکتر فرین سلیمانی.

تعارض منافع

این مطالعه هیچ‌گونه تعارض منافع ندارد.

تشکر و قدردانی

از مرکز تحقیقات توانبخشی اعصاب اطفال و معاونت پژوهشی دانشگاه علوم توانبخشی و سلامت اجتماعی و مربیان و آزمونگران آزمون بیلی دانشگاه‌های علوم‌پزشکی کشور که به‌عنوان گروه خبرگان در طراحی گویه‌های ابزار و ارزیابی فیلم‌ها مشارکت داشتند و همچنین والدین کودکان تقدیر و قدردانی می‌شود.

References

[1] Parsa Yekta Z, Ahmadi F, Tabari R. [Factors defined by nurses as influential upon the development of clinical competence (Persian)]. Journal OF Guilan University of Medcical Sciences. 2005; 14(54):9-23. [Link]

[2] Carr SJ. Assessing clinical competency in medical senior house officers: How and why should we do it? Postgraduate Medical Journal. 2004; 80(940):63-6. [DOI:10.1136/pmj.2003.011718] [PMID]

[3] Shayan S, Rafieyan M, Kazemi M. [Design and develop Logbook student's nationwide health providers training courses of the entire country (Persian)]. Teb va Tazkie. 2018; 27(2):124-32. [Link]

[4] Mansourian A, Shirazian S, Jalili M, Vatanpour M, Arabi LPM. [Checklist development for assessing the dental students' clinical skills in oral and maxillofacial medicine course and comparison with global rating (Persian)]. Journal of Dental Medicine. 2016; 2016; 29(3):169-76. [Link]

[5] Mokarami H, Javid AB, Zaroug Hossaini R, Barkhordari A, Gharibi V, Jahangiri M, et al. [Developing and validating tool for assessing the field internship course in the field of occupational health engineering (Persian)]. Iran Occupational Health. 2019; 16(3):58-70. [Link]

[6] Yaghini O, Parnia A, Monajemi A, Daryazadeh S. [Designing a tool to assess medical students' clinical competency in pediatrics (Persian)]. Research in Medical Education. 2018; 10(1):39-47. [DOI:10.29252/rme.10.1.39]

[7] Briggs-Gowan MJ, Carter AS, Irwin JR, Wachtel K, Cicchetti DV. The brief infant-toddler social and emotional assessment: Screening for social-emotional problems and delays in competence. Journal of Pediatric Psychology. 2004; 29(2):143-55. [DOI:10.1093/jpepsy/jsh017] [PMID]

[8] Halfon N, Regalado M, Sareen H, Inkelas M, Reuland CH, Glascoe FP, et al. Assessing development in the pediatric office. Pediatrics. 2004; 113(6 Suppl):1926-33. [DOI:10.1542/peds.113.S5.1926] [PMID]

[9] Rydz D, Srour M, Oskoui M, Marget N, Shiller M, Birnbaum R, et al. Screening for developmental delay in the setting of a community pediatric clinic: A prospective assessment of parent-report questionnaires. Pediatrics. 2006; 118(4):e1178-86. [DOI:10.1542/peds.2006-0466] [PMID]

[10] Mayson TA, Harris SR, Bachman CL. Gross motor development of Asian and European children on four motor assessments: A literature review. Pediatric Physical Therapy. 2007; 19(2):148-53.[DOI:10.1097/PEP.0b013e31804a57c1] [PMID]

[11] Soleimani F, Dadkhah A. Validity and reliability of Infant Neurological International Battery for detection of gross motor developmental delay in Iran. Child. 2007; 33(3):262-5. [DOI:10.1111/j.1365-2214.2006.00704.x] [PMID]

[12] Levine DA. Guiding parents through behavioral issues affecting their child's health: The primary care provider's role. Ethnicity & Disease. 2006; 16(2 Suppl 3):S3-21-8. [PMID]

[13] Ristovska L, Jachova Z, Trajkovski V. Early detection of developmental disorders in primary health care. Paediatria Croatica. 2014; 58:8-14. [DOI:10.13112/PC.2014.2]

[14] Marks K, Hix-Small H, Clark K, Newman J. Lowering developmental screening thresholds and raising quality improvement for preterm children. Pediatrics. 2009; 123(6):1516-23. [DOI:10.1542/peds.2008-2051] [PMID]

[15] King TM, Glascoe FP. Developmental surveillance of infants and young children in pediatric primary care. Current Opinion in Pediatrics. 2003; 15(6):624-9. [DOI:10.1097/00008480-200312000-00014] [PMID]

[16] Sand N, Silverstein M, Glascoe FP, Gupta VB, Tonniges TP, O'Connor KG. Pediatricians' reported practices regarding developmental screening: Do guidelines work? Do they help? Pediatrics. 2005; 116(1):174-9. [DOI:10.1542/peds.2004-1809] [PMID]

[17] Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. Journal of Chiropractic Medicine. 2016; 15(2):155-63. [DOI:10.1016/j.jcm.2016.02.012] [PMID]

[18] McHugh ML. Interrater reliability: The kappa statistic. Biochemia Medica. 2012; 22(3):276-82. [DOI:10.11613/BM.2012.031] [PMID]

[19] Sahebalzamani M, Farahani H. [Validity and reliability of direct observation of procedural skills in evaluating the clinical skills of nursing students of Zahedan nursing and midwifery school (Persian)]. Zahedan Journal of Research in Medical Sciences. 2012; 14(2):e93588. [Link]

[20] Jabbari A, Hosseini MA, Fatoureh-Chi S, Hosseini A, Farzi M. [Designing a valid & reliable tool for assessing the occupational therapist's clinical competency (Persian)]. Archives of Rehabilitation. 2014; 14(4):44-9. [Link]