

Employing Kirkpatrick's framework to evaluate nurse training: an integrative review*

Fernanda Maria de Miranda^{1,2}

 <https://orcid.org/0000-0003-2198-2827>

Bruna Vasconcelos dos Santos¹

 <https://orcid.org/0000-0002-3967-9327>

Vicki Leigh Kristman³

 <https://orcid.org/0000-0001-5508-0552>

Vivian Aline Mininel⁴

 <https://orcid.org/0000-0001-9985-5575>

Highlights: **(1)** Kirkpatrick's framework is effective for evaluating various nursing training. **(2)** The framework upholds the choice of measuring instruments for each level. **(3)** The four-levels or a combination of the first three were the most commonly used to evaluate training. **(4)** The four-levels or a combination of the first three were the most common to evaluate training. **(5)** Evaluation of results in organizational practices is the most challenging level.

Objective: to evaluate the evidence on the use of Donald Kirkpatrick's framework in nursing training evaluation. **Method:** integrative literature review in the Latin American and Caribbean Health Sciences Literature, Medical Literature Analysis and Retrieval System and Web of Science databases. Studies that answered the review question "Which is the evidence in using Donald Kirkpatrick's framework to evaluate training in the nursing workplace?" published in Portuguese, English, or Spanish were included. **Results:** out of 108 studies retrieved, thirteen were included. The majority evaluated the four levels proposed in the model (reaction, learning, behavior, and results) or, at least, a combination of the first three ones. Different instruments were used to evaluate nursing training, mainly in quantitative approaches for reaction and learning levels and qualitative for behavior and results levels. This approach highlights the flexibility of the model and the importance of choosing a reliable set of instruments, which is crucial to qualify the analysis at each level. **Conclusion:** Kirkpatrick's model has been used worldwide to evaluate training in the nursing field and has been shown to be suitable for it, as long as there is an appropriate selection of instruments at each level.

Descriptors: Continuing Education; Nurses; Evaluation of the Efficacy-Effectiveness of Interventions; Review; Methods; Inservice Training.

* Paper extracted from doctoral dissertation "Training Brazilian nurses to promote mental health at work", presented to Universidade Federal de São Carlos, São Carlos, SP, Brazil. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - Finance Code 001, Brazil.

¹ Universidade Federal de São Carlos, São Carlos, SP, Brazil.

² Scholarship holder at the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brazil.

³ Lakehead University, Department of Health Sciences, Thunder Bay, ON, Canada.

⁴ Universidade Federal de São Carlos, Departamento de Enfermagem, São Carlos, SP, Brazil.

How to cite this article

Miranda FM, Santos BV, Kristman VL, Mininel VA. Employing Kirkpatrick's framework to evaluate nurse training: an integrative review. Rev. Latino-Am. Enfermagem. 2025;33:e4431 [cited   ]. Available from:  <https://doi.org/10.1590/1518-8345.7250.4431>

year month day

Introduction

Training healthcare personnel has been incorporated as a strategy for talent retention in human resources management. Training and development practices at the workplace are described as dynamic and continuous processes aimed at promoting social advances, increasing resolution and fostering better health outcomes⁽¹⁾.

Training can be an opportunity for professional development and increased productivity, creativity, and innovation. These are the main aspects of high-quality services and teams effectiveness⁽²⁾. Lack of training can negatively influence perception of stress, increase turnover and stress, and reduce work performance⁽³⁾.

Training strategies are key to improve health systems in low-income countries⁽⁴⁾. The development of workers through training is a critical indicator of effective management⁽⁵⁾. In Brazil, training is also political, since knowledge interacts as an essential part of the power struggle and meaningful learning could lead to deep-rooted changes in healthcare⁽⁶⁾.

Evaluating training programs can help managers to understand the effectiveness and sustainability of the learning process to produce changes in the short or long term. However, selecting a suitable framework for evaluation is not an easy task and many interventions struggle with low or very low levels of evidence for targeted outcomes⁽⁴⁾.

In 1959, Donald Kirkpatrick published a framework to support managers to evaluate the results of training and development practices among workers and organizational systems⁽⁷⁻⁸⁾, based on summative assessments. The model is composed of four levels of evaluation: Reaction, Learning, Behavior and Results⁽⁷⁾. Reaction (level I) encompasses the participant's perceptions of their learning experience, program structure, content, teaching methods, and instructional aspects, such as materials and the quality of instruction. Learning (level II) is related to changes in attitudes, perceptions, knowledge and skills. Behavior (level III) evaluates the knowledge transfer to the workplace through behavioral changes. Finally, the Results level (level IV) evaluates the applicability of intervention results to change the organizational practice and improve the outcomes⁽⁷⁻⁸⁾.

Managers and researchers should carefully consider adopting an evaluation system based on this framework, as the implementation of isolated levels is not recommended and higher level of evaluation do not lead to more valuable results if they are isolated⁽⁷⁾.

The Kirkpatrick model has been widely accepted in the scientific community to support development and

training in several areas, including health⁽⁹⁻¹⁴⁾. Several improvements were described by the Kirkpatrick Group since the first publication, arriving at a new model called the New World Kirkpatrick model in 2016⁽¹⁵⁾. Despite the improvements, there are criticisms regarding the applicability of the Kirkpatrick model⁽¹⁶⁻¹⁸⁾, the major concern being the difficulty in implementing all four levels⁽¹⁶⁾. These criticisms prompted this review to ensure the method's applicability in nursing. The aim of this study is to evaluate the evidence on the use of Donald Kirkpatrick's framework in the evaluation of nursing training.

Methods

Type of study

We used an integrative literature review, based on the Whittemore and Knafl framework, carried out in five stages: problem identification; literature search/screening; data evaluation; data analysis; presentation of the review⁽¹⁹⁾. We choose this methodological structure to ensure a systematic approach to evidence synthesis. The intended outcome is to obtain a comprehensive understanding of the use of Kirkpatrick's framework in nursing and the identification of gaps for future studies⁽¹⁹⁾.

This integrative review was registered on the Open Science Framework (OSF) platform on May 20th, 2022, and the protocol can be accessed at osf.io/uprv7, with the following DOI 10.17605/OSF.IO/JQ8U9⁽²⁰⁾.

Locus

This integrative review was conducted in the municipality of Thunder Bay, located in the province of Ontario (ON), Canada.

Period

The study period was September 2022 to November 2023.

Population

We used the review question: "Which is the evidence in the use of the Donald Kirkpatrick framework to evaluate nursing workplace training?" was used. The formulation of the research question was developed from the PCC strategy namely: Population (P): nursing; Concept (C):

Donald Kirkpatrick's four-level framework; Context (C): evaluation of workplace training.

Selection criteria

Studies were included if they addressed the research question and if they were published in English, Spanish, or Portuguese between 2006 to 2023, with full content available. The time frame was based on the year of publication of the third edition of the book "Evaluating training programs" where some interpretations of the traditional framework were updated⁽⁷⁾. Studies that did not meet the inclusion criteria or did not clearly describe the instruments used were excluded. We also removed duplicates, literature reviews, and gray literature.

Sample definition

The search was performed in November 2022 from Thunder Bay, ON, Canada by FMM, end updated in August 2023. The Latin American and Caribbean Health Sciences Literature (LILACS), Medline (via PubMed), and Web of Science (WoS) bibliographic research databases were used.

Search strategies (strings) were based on the Health Sciences Descriptors (DeCS) and the Medical Subject Heading (MeSH). After analysis of exploratory searches, we decided to include the keyword "Kirkpatrick" to reduce the scope of the search. The final strategy used was: ("Education, continuing" OR "Continuous Learning" OR "Learning, Continuous" OR "Lifelong Learning" OR "Learning, Lifelong" OR "Life-Long Learning" OR "Learning, Life-Long" OR "Learnings, Life-Long" OR "Life Long Learning" OR "Life-Long Learnings" OR "Continuing Education") AND (("nurses" OR "Nurse" OR "Personnel, Nursing" OR "Nursing Personnel" OR "Registered Nurses" OR "Nurse, Registered" OR "Nurses, Registered" OR "Registered Nurse") OR ("nursing, team")) AND ("Kirkpatrick") for MEDLINE and WoS. For LILACS, it was decided to use only the keyword "Kirkpatrick" considering the theme specificity and the quantity of studies retrieved in the original string.

After searching the databases, the files generated in the BIBTXT or RIS format were exported to the Rayyan[®] review manager⁽²¹⁾ which supported the initial screening of abstracts and titles using a semi-automation process. The selection of articles was carried out in two blind stages by two different researchers (FMM and BVS); the first screening was completed

by reading the titles and abstracts of the retrieved documents; the second screening involved a complete review. Conflicts were discussed until a consensus was reached. A third researcher (VAM) was available to decide any persistent conflict.

Data collection

The data was extracted using an analytical matrix built in Microsoft Excel[®] containing information of the characteristics of the studies (authors, year of publication, country of publication, objective, type of study, sample and results) and information on the use of the Kirkpatrick model (levels used, elements, timing, evaluation criteria and measures).

Data treatment and analysis

We analyzed the data through reduction and comparison to identify patterns, themes, or relationships⁽¹⁹⁾. Subsequently, a descriptive analysis based on the theoretical-conceptual framework adopted was carried out, considering the precepts of Donald Kirkpatrick's framework⁽⁷⁾.

This review followed the guidelines of the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) for reporting systematic reviews⁽²²⁾. PRISMA was designed for systematic reviews of studies evaluating the effects of health interventions; however, the items on the checklist are applicable to reports of systematic reviews in other areas or evaluating other types of interventions⁽²²⁾. Thus, all PRISMA section were used to guide the review report, except for topics not applicable to this review's approach, including analyses of the risk of bias, determination of measures of effect, and analyses of possible causes of heterogeneity. The effectiveness of the Kirkpatrick framework through the training effectiveness results shown in each article included in the sample selected for this study was analyzed. To do this, both the training design and the use of Kirkpatrick levels were considered, the configuration of the instrument used, the timing of the evaluation and other aspects.

Results

One hundred and eight articles were retrieved and 13 were included in the final sample (Figure 1).

Figure 2 shows the 13 articles characterized by country, year of publication, objective and methods.

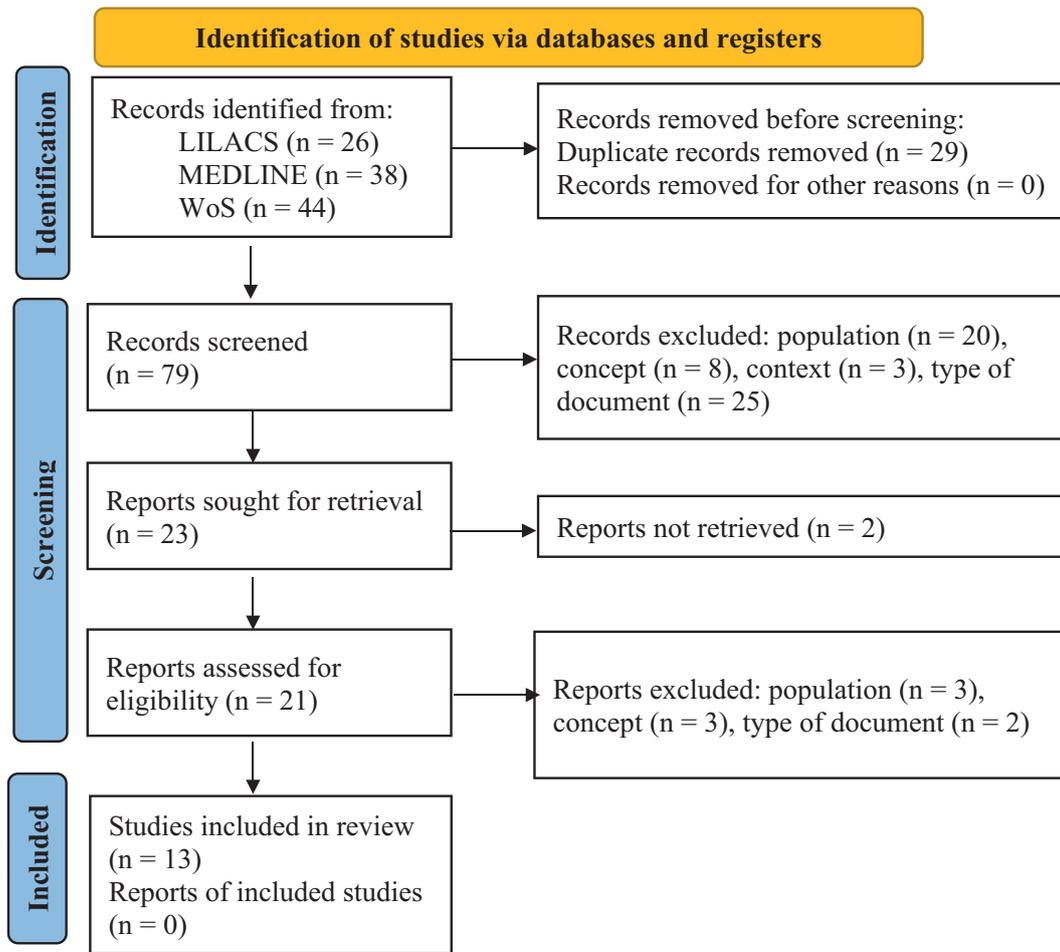


Figure 1 - PRISMA⁽²²⁾ workflow for sample selection. Thunder Bay, ON, Canada, 2023

Country, year	Objective	Methods
Canada, 2022 ⁽²³⁾	To report on a mixed-methods evaluation of a person-centered competency-based educational program on a Caregiver-Centered Care for the healthcare workforce	Mixed-methods evaluation (n = 161 healthcare workers, including 60 nurses)
Cambodia, 2022 ⁽²⁴⁾	To evaluate the behavior and outcomes, which are considered the medium-term results of the two training programs, and to identify influencing factors, emphasizing the institutional development of educational and clinical facilities in Cambodia.	Qualitative study (n = 37 nurses).
Brazil, 2021 ⁽²⁵⁾	To evaluate a specialization course focused on health quality and patient safety.	Cross-sectional mixed methods (n = 46 young professionals, those 7 nurses).
Brazil, 2021 ⁽²⁶⁾	To analyze how nurses who care for bedridden psychiatric patients in a general hospital perceive mental health training using active methodologies.	Qualitative study, with descriptive and exploratory approach (n = 5 nurses).
China, 2021 ⁽¹⁰⁾	To design a 100-hour training program for nursing innovation teams and evaluate the effect of this training program using Kirkpatrick's model.	Quasi experimental with pre- and post- tests (n = 61 nurses).
Germany, 2020 ⁽²⁷⁾	To implement and evaluate CIPE* intervention in three pilot courses.	Quasi-experimental single group, pre-posttest study (n = 39 healthcare workers).
Brazil, 2020 ⁽²⁸⁾	To describe the development process and present the results of a pilot study on the use of low-cost handmade simulators to teach and learn Obstetrics.	Cross-sectional pilot study (n = 10 resident doctors, 12 undergraduate medical students, and 9 nurses).
Brazil, 2018 ⁽²⁹⁾	To evaluate the effectiveness of the UNA-SUS/UERJ ¹ Specialization Course in Elderly Health.	Mixed-methods evaluation (n = 444 participants)
Iran, 2018 ⁽³⁰⁾	To evaluate the effectiveness of a continuing education program for the prevention of occupational exposure to needlestick injury, blood, and body fluids in a nursing team based on the Kirkpatrick model.	Two-group before-after quasi-experimental design (n = 120 nurses): 60 nurses in the IG ² and 60 in the CG ³ .

(continues on the next page...)

(continuation...)

Country, year	Objective	Methods
USA, 2017 ⁽³¹⁾	To investigate a VJC effect on satisfaction, knowledge, and practice.	Pilot study (n = 16 nurses)
Singapore, 2016 ⁽³²⁾	To evaluate the impact of WB [¶] simulation on nurses' recognition of and response to patient deterioration in clinical settings.	Before-and-after study (n = 99 nurses, classified as enrolled nurses and registered nurses).
Finland, 2016 ⁽³³⁾	To use Kirkpatrick's four-level model to evaluate an e-learning continuing education course from the perspectives of nurse managers in psychiatric hospital organizations.	Qualitative descriptive evaluation study (n = 28 managers nurses)
Australia, 2011 ⁽³⁴⁾	To compare two approaches to training clinicians in prescribing exercises to prevent falls.	Randomized trial with concealed allocation and blind outcome assessment (n = 134 workers, those 21 nurses. They were allocated in 2 groups: WB [¶] and FF ^{**}).

*CIPE = Continuing interprofessional education; [¶]UNA-SUS/UERJ = *Universidade Aberta do Sistema Único de Saúde/Universidade Estadual do Rio de Janeiro*; ^{||}IG = Intervention group; [§]CG = Control group; ^{||}VJC = Virtual journal club's; [¶]WB = Web-based; ^{**}FF = Face-to-face

Figure 2 - Sample characterization (n = 13). Thunder Bay, ON, Canada, 2023

The Kirkpatrick model has been useful in training, continuing education and updating courses, as well as in specialization programs with individual-centered and team-based interventions. Figure 3 shows the use of the four-levels Kirkpatrick model to evaluate training results:

Training	Outcomes by level
Specialization course ⁽²⁹⁾	I: more positive than negative statements. Most prolific: "manifestation about the course" (38.49%) and "content" (32.07%). Less prolific: "personal issues" (38.10%) as health problems, family issues and difficulty organizing time. II: 71,85% of retention. III: 93,7% of approval in the course.
Specialization course ⁽²⁵⁾	I: majority of "totally agree" in 7 out of 8 satisfaction statements. II: 69% with grade A/excellent. III: 90% agreed that the learnings were incorporated into practice. IV: 100% created a Patient Safety Committee or equivalent.
Knowledge upgrade program ⁽²⁴⁾	III: improved teaching and learning activities, built capacity, improved clinical practices, strengthened clinical management, and more involved in professional activities. IV: enhanced institutional capacity, improved learning and behavior and strengthened clinical performance at the hospital.
Low-cost handmade simulation ⁽²⁸⁾	I: 100% of satisfaction. II: perceiving significant gains in theoretical knowledge; III/IV: perception of ability to solve clinical problems and decreased anxiety to deal with similar situations.
WB* simulation ⁽³²⁾	I: scores 3.78 (max. 5) on motivation. II: significant increase in knowledge: RN [†] pre 18.80 and posttest 22.47; EN [‡] - pre 16.57 and posttest 19.57. III: positive attitudes (RN [†] 3.82; EN [‡] 4.06, max. 5), with no significant difference between groups. IV: the number of deteriorating patients triggered increased significantly (8.96% to 14.58%).
Team-based program ⁽¹⁰⁾	I: trainees scored 66.28 pre- and 81.97 posttest in climate (range 21-105). II: 26.54 and 32.31 in self-efficiency (range 8-40). III: 25.07 and 32.90 in behavior (range 10-50). IV: 133.50 and 168.02 in ability (max. 205).
CIPE ^{§(27)}	I: high level of satisfaction - mean score of 38.9 (range 30-44). II: positive learning effect (7.2% higher in posttest).
VJC ⁽³¹⁾	I: statistically significant satisfaction. II: improvement in knowledge, however, only answers to 2 out of 5 questions demonstrated statistical significance. III: 50% stated that the VJC did have an impact on their practice, and 50% noted an increase in confidence in caring for patients or in awareness.
Active methods ⁽²⁶⁾	I: good satisfaction. II: increase skills, such as therapeutic approach and active listening. III: more humanized and integrated psychiatric patient care.
PCC ^{¶(23)}	I: means 4.5 to 4.8 in satisfaction (max. 5). II: increased from 38.90 to 46.60 in knowledge and confidence (max. 50). III: three themes emerge: (1) usable skills, (2) reinforced practice, and (3) requires leadership and a culture change.
FF** seminar and WB* training ⁽³⁴⁾	Both groups had similar results. I: satisfaction with the content and its relevance: mean scores 25.73 (WB*) and 26.11 (FF**) (range 0-35); satisfaction with course facilitation and support: 11.61 (WB) and 12.08 (FF**) (range 0-15). II: median score for the exercise assignment: 78.6 for both groups (range 0-100). III: changes in clinical behavior: 21.75 (WB*) and 21.88 (FF**) (range 0-30).
CEC ^{††(30)}	I: 90% of IG ^{‡‡} were satisfied. II: significant increase in IG ^{‡‡} awareness (IG ^{‡‡} : scores 8.32 in pre- and 13.98 in posttest; CG ^{§§} : scores 8.45 pre- and 8.86 posttest; max. 15). III: increase of good performance response in IG ^{‡‡} (IG ^{‡‡} : from 50% to 75%; CG ^{§§} : decreased from 45% to 30%). IV: significant decrease of exposure to needlestick injuries in IG ^{‡‡} (from 40% to 15%).

(continues on the next page...)

(continuation...)

Training	Outcomes by level
E-learning CEC ^{††(33)}	I: varied feelings, being positive (enthusiasm, motivation, and fascination) more frequent. Less usual: neutral or negative (anger, fear, doubt, resistance, irritation, underestimation, feelings of being coerced, distraction, and laziness). II: evolving in learning (legislation, ethical issues, and cope mechanisms). III: attitudes toward patient care and treating became more patient centered and collaborative and reduction in the use of coercive methods. IV: increase of cooperation, anticipation of aggression, overall difference in attitudes, acknowledgment of patients, awareness of one's work, and new knowledge.

*WB = Web-based; *RN = Registered nurse; *EN = Enrolled nurse; *CIPE = Continuing interprofessional education; *VJC = Virtual journal club's; *PCC = Person-centered and competency based program **FF = Face-to-face; *CEC = Continuing education course; *IG = Intervention group; *CG = Control group

Figure 3 - Using the four-levels Kirkpatrick model to evaluate training results (n = 13). Thunder Bay, ON, Canada, 2023

Kirkpatrick's framework was effective in evaluating all the training identified, considering the measurable and prolific outcomes of the samples by level in different types of training. Consistency in compliance with the guidelines was observed, demonstrating their versatility and flexibility.

Brazil was the only country with more than one publication^(25-26,28-29). Quasi-experimental approaches^(10,27-28,30-32) followed by qualitative research^(24,26,33), mixed-

methods research^(23,25,29) and one experimental study⁽³⁴⁾ were found.

Most of the studies used all four levels of Kirkpatrick's Framework^(10,25,28,30,32-33) or a combination of the first three levels^(23,26,29,31,34). One article evaluated the intervention with a combination of levels I and II⁽²⁷⁾ and another with a combination of levels III and IV⁽²⁴⁾.

Figure 4 shows a timeline for the implementation of these levels.

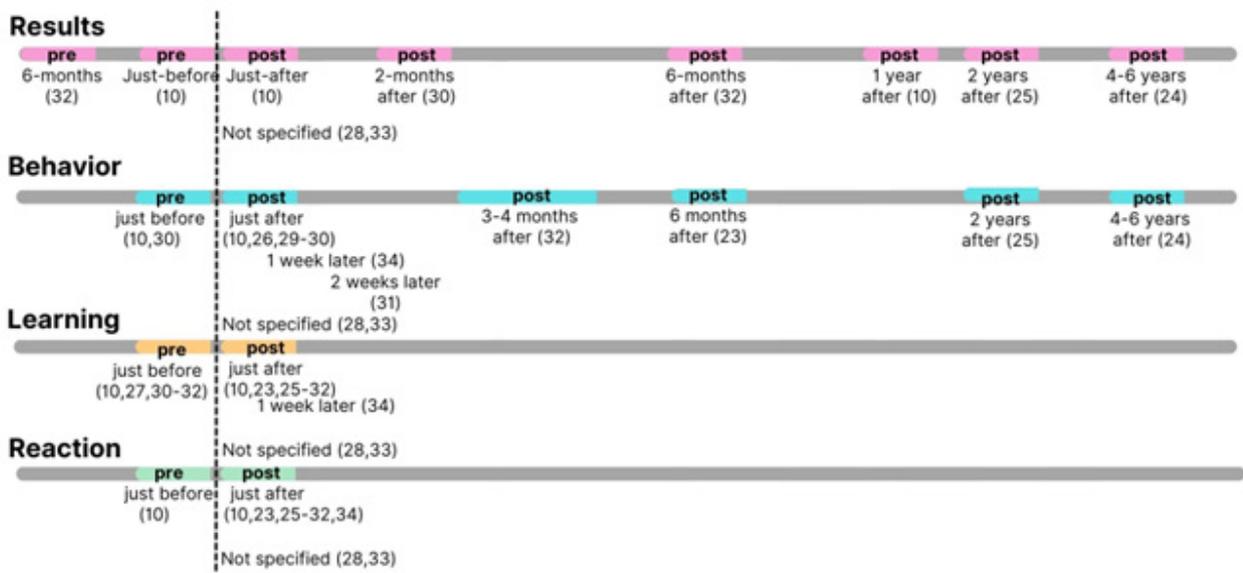


Figure 4 - Multiple data points were collected before and after the intervention using the four levels of Kirkpatrick's model (n = 13). Thunder Bay, ON, Canada, 2023

Two studies^(28,33) did not identify the timeline for data collection. They only described that there was cross-sectional data collection considering the availability of the intervention for a period of time such as "since 2014"⁽²⁸⁾ or "between 2008 and 2011"⁽³³⁾. None of the studies had explicit follow-up designs, although some of them collected data long after the intervention^(10,24-25,32). In one of them, the authors cited a previous data collection after the intervention, but did not present or compare the data, nor did they characterize a follow-up study⁽²⁴⁾.

Different instruments have been used to evaluate interventions, based on the specificity of the investigation; however, a main approach or a standard instrument for

evaluating any level of the Kirkpatrick Framework was not found. Considering the relevance of measurement tools for the four levels of Kirkpatrick's model and the lack of consensus on the best way to do it (and this is the richness of the model – it can be customized according to with study), it was decided to synthesize the instruments that were used in each selected study to help future nurses in this topic.

Quantitative instruments were the most common. Table 1 shows the quantitative instruments used to evaluate the training programs, their measures (items and range), and whether the study described some source of reliability using Cronbach's α .

Table 1 - Quantitative instruments applied to evaluate the four-level Kirkpatrick's framework (n = 13). Thunder Bay, ON, Canada, 2023

Level: Instrument Description	Item	Range	α^*
I: Five-point Likert scale for learner's satisfaction ⁽²³⁾	5	1-5	n/s [†]
I: Five-point Likert scale's questionnaire: "material and syllabus", "assignments", "tutoring", and "learning support structure" ⁽²⁵⁾	4	1-5	n/s [†]
I: Reaction assessing tool including exogen aspects, (e.g., internet access) and endogenous aspects (e.g., material) for each module and the entire course ⁽²⁹⁾	9	1-4	n/s [†]
I: NOIC [‡] with 5-point Likert scale ⁽¹⁰⁾	21	1-5	0.94
I: ABC-SAT [§] , domains: a = affective, b = behavioral, c = cognitive ⁽²⁷⁾	a = 3	0-12	0.60
	b = 2	0-8	0.66
	c = 6	0-24	0.83
I: Five-point Likert scale statements related to satisfaction, potentiality for optimal learning and bad feelings' prevention ⁽²⁸⁾	3	1-5	n/s [†]
I: Five-point Likert's scale questionnaire, domains: c = content, t = teachers, f = facilities ⁽³⁰⁾	c = 5		
	t = 4	1-5	0.87
	f = 3		
I: Five-point Likert scale's survey related to usability, format, discussion opportunities, participation and time adequacy ⁽³¹⁾	6	1-5	n/s [†]
I: IMMS ⁽³²⁾	4	1-5	0.79
I: Self-reported satisfaction questionnaire, domains c = "content and its relevance" and s = "support and facilitation" ⁽³⁴⁾	c = 35	0-35	0.92
	s = 15	0-15	0.77
II: 10-questions education knowledge and confidence test ⁽²³⁾	10	5-50	0.92
II: Four-Processual tests (one for each learning module) ⁽²⁵⁾	n/s [†]	0-10	n/s [†]
II: The innovation self-efficacy questionnaire ⁽¹⁰⁾	8	1-5	0.86
II: Self-developed written knowledge test regarding key topics of the intervention, each with 4 possible answers ⁽²⁷⁾	9	0-18	n/s [†]
II: Five-point Likert scale's statements: "(...) was used to increase my theoretical knowledge" and "(...) was used to increase my ability to solve clinical problems" ⁽²⁸⁾	2	1-5	n/s [†]
II: Awareness questionnaire ⁽³⁰⁾	15	0-15	0.87
II: 5-question closed-end test (2 had more than one correct answer) ⁽³¹⁾	5	0-5	n/s [†]
II: Metacognition Questionnaire ⁽³²⁾	30	n/s [†]	n/s [†]
II: 1-hour knowledge test ⁽³⁴⁾	30	0-30	n/s [†]
II: Assignment submission requiring a description of an exercise program tailored to a hypothetical client scenario ⁽³⁴⁾	1	0-100	n/s [†]
III: Self-reported questionnaire about behavior regarding patient safety protocols ⁽²⁵⁾	9	1-5	n/s [†]
III: NIBS [¶] ⁽¹⁰⁾	10	1-5	0.90
III: Two closed-ended statements about decreased of anxiety/stress and ability of solve a clinical problem/situation ⁽²⁸⁾	2	1-5	n/s [†]
III: The performance questionnaire ⁽³⁰⁾	15	0-15	0.78
III: Self-reported Knowledge transfer at workplace questionnaire ⁽³²⁾	14	1-5	0.94
III: Self-reported satisfaction questionnaire, domain b = "change in clinical behavior" ⁽³⁴⁾	b = 30	0-30	0.84
IV: Questionnaire of agreement about the implantation of five goals for patient safety workplace after the course ⁽²⁵⁾	6	1-5	n/s [†]
IV: Scale of Clinical Nursing Staff Innovation Ability ⁽¹⁰⁾	41	0-205	0.82
IV: The questionnaire for exposure to sharp objects, blood and body fluids ⁽³⁰⁾	53	n/s [†]	0.89

* α = Cronbach's α ; [†]n/s = Not specified; [‡]NOIC = The Nurse Organizational Innovation Climate Scale; [§]ABC-SAT = The 'Affective-Behavioral-Cognitive-Satisfaction questionnaire'; ^{||}IMMS = The Instructional Materials Motivation Survey; [¶]NIBS = Nurse Innovation Behaviour Scale

A great variability of instruments regarding quantitative approaches was found. Level I was mostly evaluated by a five-point Likert scale for items such as satisfaction^(23,25,27-28,30,32,34), immersion^(28,31), climate⁽¹⁰⁾, relevance and confidence⁽³²⁾. Both knowledge and awareness approaches were used on level II. The most common was closed-end learning tests^(25,27,29-31,34), followed by self-rated questionnaires^(10,23,32). Level III and IV quantitative measures included trainees' perception of knowledge transfer to the workplace through closed-end questionnaires^(10,25,28,30,32-34) or scales⁽¹⁰⁾.

Although less frequent than quantitative, the use of a qualitative approach was found in behavior level^(23-24,26,29,31,33) followed by reaction^(26,29,33), results^(24,28,33) and learning^(26,33). Seven^(23-24,26,28-29,31,33) studies used qualitative instruments to evaluate training programs, however only two^(24,26) of them did so without a quantitative piece. Semi-structured interviews^(23-24,26), questionnaires with open questions^(28,33), descriptive analysis of the key words in students' projects⁽²⁹⁾, qualitative research⁽³¹⁾, qualitative analysis based on the participants' narratives⁽²⁹⁾, and adherence after the first month of the intervention⁽²⁹⁾ were identified. The most prevalent type of qualitative analysis was thematic analysis^(24,26,28-29). Most of the studies did not describe reliability or validity of the qualitative instruments, with the exception of Koto-Shimada⁽²⁴⁾ who carried out a pilot test. Other work-related measures have been also reported, i.e., clinical records and patient information⁽³²⁾, self-reported exposure to needlestick injury questionnaire⁽³⁰⁾.

Choosing a measurement tool is crucial to evaluating the effectiveness of any intervention, especially when it comes to training. This step is embodied into Kirkpatrick's model and should be carefully selected by nurses in order to qualify the analysis of each level. The more precise the tool (i.e., more in line with the object of study), the greater the chances of a reliable evaluation.

Discussion

This review has shown that the Kirkpatrick Framework has been widely used in the evaluation of nurses' training and also in other areas of health⁽¹¹⁾. Our sample depicted in this study indicated consistency in compliance with the guidelines and the main training outcomes indicate that this model is useful to evaluate effectiveness, with positive results at each level in different types of training interventions.

The evaluation process converged with the orientation of the framework which focuses on participant outcomes⁽⁷⁾. However, one study approached the participants' managers and a complementary view emerged, as exemplified in

the quote: "the nurses did not experience the change in the atmosphere and attitudes in the same way as the nurse managers had observed"⁽³³⁾. The inclusion of the leadership perspective can be positive, especially to obtain feedback on knowledge transfer to the workplace (level III) or changes in practice or organizational outcomes (level IV) considering their supervisory role.

The most common type of baseline information was on Learning level^(10,23,27,30,32). The Behavior and Results levels were the most divergent levels regarding timing. There are no guidelines in Kirkpatrick's Framework on the best time to evaluate each level^(7,15) and we cannot support a more common or assertive time for data collection, considering that none of the studies converge on the best time. However, it is important to point out that training programs expect short-, medium- and long-term outcomes. Therefore, selecting a specific time (e.g., before, during, soon after, longer after) or period (e.g., follow-ups) is essential for a comprehensive evaluation that considers that feeling of reaction will fade over time and the most important outcomes will appear over time⁽⁸⁾.

The Kirkpatrick model is a flexible and adjustable framework that embeds different tools for evaluating training outcomes (Reaction, Learning, Behavior and Results), which reinforces its power, since the realities, training and contexts of each organization are completely different. However, precisely because it does not indicate a specific methodological standard, authors must be careful to ensure the validity and reliability of the information collected⁽¹⁷⁾.

Both the number of studies and the diversity of instruments decreased as the level of evaluation increased, it has been observed. The literature review indicates that the Results' level is the most difficult to achieve. There is acknowledged that level IV "could identify the added value to society of a given educational program, as it makes it possible to evaluate not only of the application of intervention projects in the practice of health professionals, but also the results of their use in a local context"⁽²⁹⁾. However, half of the samples in this review did not report measuring this level. The most common reason reported for the exclusion of level IV was the need for medium- and long-term follow-up periods⁽²⁶⁾ and the complexity of relating the outcomes with the training program, distinguishing them from many other factors that could affect the results⁽³¹⁾. Results evaluation (level IV) is the most challenging because it must consider the transformation or impacts of training on organizational practice, which requires other types of instruments, organizational information and the point of view of other stakeholders, as well as more time. In this sense, similar

data was found regarding the perceived relevance of levels III and IV and the difficulty to implement these levels⁽¹⁷⁾.

There is some criticism about the rigidity of this framework, which can lead to essential aspects of the evaluation, such as the formative vision being missed^(16,18). Only one study reported data collection in each module and at the end of the course at the reactions level⁽²⁹⁾. No other experience was identified at the follow-up. This evidence shows that, despite the diversity in the perspective of outcomes evaluation tools, especially with the use of qualitative approaches to capture the trainees' perspective, all the experiences reported were in a summative or transversal view.

Trainers are not using only the highest levels, which is a concern of the criticism of Kirkpatrick's model⁽¹⁶⁻¹⁸⁾ it has been observed. Six articles managed to reach Level four and another five used a combination of the first three levels; accounting for 85% of the sample. This indicates that nursing trainers are attempting to use all four levels of the model to collect information on outcomes, despite the challenge regarding the Results level. This reaffirms the researchers' commitment to the framework's recommendation to use a complete model for greater accuracy⁽⁷⁻⁸⁾.

The relevance of this review is the fact that it synthesizes the recognition of the operationalization of the framework. A large bibliometric analysis found the Kirkpatrick model to be a trending topic of interest after the 2000s, including 20.7% of its sample as some kind of literature review publications. However, most studies aimed to criticize the model or identify its benefits, and little review evidence was found on the types and characterization of evaluation tools and their application in a specific field, such as nursing⁽¹¹⁾.

This study has methodological limitations. The multiple research designs of the publications included made it impossible to produce evidence through meta-analysis or meta-synthesis techniques. Considering the nature of the integrative review, the aim is not to critically appraise the quality of the studies and therefore, the quality of the study was not used an inclusion criterion.

We encourage further research to identify the best constructs for evaluating each level, to expand the implementation of level IV and to foster scientific conversation about the methodological orientation for composing the chain of evidence between the levels, as proposed by the New World Kirkpatrick model⁽¹⁵⁾. Moreover, our literature review identified limited longitudinal studies with consecutive follow-ups for data collection, which is highly recommended for achieving high-level evidence regarding causal pathways.

Conclusion

Kirkpatrick's four-level Framework is a suitable choice in terms of adapting evaluation instruments to the training design. The selection of a reliable set of instruments is crucial to qualify the analysis of each level, with Likert scales being the most common choice for trainers. This model is prolific for evaluating training in the nursing field but requires a careful choice of instruments for each level. Many qualitative and quantitative measurement tools have been identified that can be useful for practitioners and academics in further evaluation and research.

References

1. Nahm ES, Archibald M, Mills ME, Costa L, Warren J, Nair P, et al. Continuum of nursing education and practice: Time to close the chasm between academia and practice. *J Prof Nurs.* 2023;46:134-40. <https://doi.org/10.1016/j.profnurs.2023.02.012>
2. Buljac-Samardzic M, Doekhie KD, van Wijngaarden JDH. Interventions to improve team effectiveness within health care: a systematic review of the past decade. *Hum Resour Health.* 2020;18(2). <https://doi.org/10.1186/s12960-019-0411-3>
3. Ertan ŞS, Şeşen H. Positive organizational scholarship in healthcare: The impact of employee training on performance, turnover, and stress. *JMO.* 2022;28(6):1301-20. <https://doi.org/10.1017/jmo.2019.61>
4. Pantoja T, Opiyo N, Ciapponi A, Dudley L, Gagnon MP, Herrera CA, et al. Implementation strategies for health systems in low-income countries: an overview of systematic reviews. *Cochrane Database Syst Rev.* 2017;9(9):CD011086. <https://doi.org/10.1002/14651858.CD011086.pub2>
5. Rivaldo Y, Nabella SD. Employee Performance: Education, Training, Experience and Work Discipline. *Qual Access Success.* 2023;24(193):182-8. <https://doi.org/10.47750/QAS/24.193.20>
6. Ogata MN, Silva JAM, Peduzzi M, Costa MV, Fortuna CM, Feliciano AB. Interfaces between permanent education and interprofessional education in health. *Rev Esc Enferm USP.* 2021;55:e03733. <https://doi.org/10.1590/S1980-220X2020018903733>
7. Kirkpatrick DL, Kirkpatrick JD. Evaluating training programs. 3rd ed. San Francisco, CA: Berrett-Koehler Publisher; 2006.
8. Yardley S, Dornan T. Kirkpatrick's levels and education 'evidence'. *Med Educ.* 2012;46(1):97-106. <https://doi.org/10.1111/j.1365-2923.2011.04076.x>
9. Heydari MR, Taghva F, Amini M, Delavari S. Using Kirkpatrick's model to measure the effect of a new

- teaching and learning methods workshop for health care staff. *BMC Res Notes*. 2019;12(1):388. <https://doi.org/10.1186/s13104-019-4421-y>
10. Huang X, Wang R, Chen J, Gao C, Wang B, Dong Y, et al. Kirkpatrick's evaluation of the effect of a nursing innovation team training for clinical nurses. *J Nurs Manag*. 2022;30(7):2165-75. <https://doi.org/10.1111/jonm.13504>
 11. Alsalamah A, Callinan C. The Kirkpatrick model for training evaluation: bibliometric analysis after 60 years (1959–2020). *Ind Commer Train*. 2022;54(1):36-63. <https://doi.org/10.1108/ICT-12-2020-0115>
 12. Choy CL, Liaw SY, Goh EL, See KC, Chua WL. Impact of sepsis education for healthcare professionals and students on learning and patient outcomes: a systematic review. *J Hosp Infect*. 2022;122:84-95. <https://doi.org/10.1016/j.jhin.2022.01.004>
 13. Khurshid Z, De Brún A, McAuliffe E. Protocol for an integrated evaluation framework to study training, curricular and contextual factors impacting the success of a measurement for improvement training programme for healthcare staff in Ireland. *BMJ Open*. 2022;12(2):e047639. <https://doi.org/10.1136/bmjopen-2020-047639>
 14. Malhotra S, Negi G, Sharma SK, Kaur R. A prospective interventional study to assess the impact of a 'structured compact training' on knowledge and skills of safe blood transfusion practices among nurses working in a tertiary care institute. *Transfus Med*. 2022;32(1):32-7. <https://doi.org/10.1111/TME.12835>
 15. Kirkpatrick JD, Kirkpatrick WK. *Kirkpatrick's Four Levels of Training Evaluation*. 1st ed. Alexandria, VA: Association for Talent Development; 2016.
 16. Reio TG Jr, Rocco TS, Smith DH, Chang E. A Critique of Kirkpatrick's Evaluation Model. *New Horiz Adult Educ*. 2017;29(2):35-53. <https://doi.org/10.1002/nha3.20178>
 17. Rosa A. Kirkpatrick's four-level model for evaluating training programs: the weak links in a "chain of evidence". *LLL*. 2018;14(31):17-37. <https://doi.org/10.19241/lll.v14i31.113>
 18. Cahapay M. Kirkpatrick Model: Its Limitations as Used in Higher Education Evaluation. *Int J Assess Tool Educ*. 2021;8(1):135-44. <https://doi.org/10.21449/ijate.856143>
 19. Whitemore R, Knafel K. The integrative review: updated methodology. *J Adv Nurs*. 2005;52(5):546-53. <https://doi.org/10.1111/j.1365-2648.2005.03621.x>
 20. Miranda FM, Mininel VA. Use of Kirkpatrick's framework to evaluate nursing team's training: integrative literature review protocol. *Open Science Framework*. 2023. <https://doi.org/10.17605/OSF.IO/JQ8U9>
 21. Ouzzani M, Hammady H, Fedorowicz Z, Elmagarmid A. Rayyan—a web and mobile app for systematic reviews. *Syst Rev*. 2016;5:210. <https://doi.org/10.1186/s13643-016-0384-4>
 22. Page MJ, McKenzie JE, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71. <https://doi.org/10.1136/bmj.n71>
 23. Parmar JK, L'Heureux T, Anderson S, Duggleby W, Pollard C, Poole L, et al. Optimizing the integration of family caregivers in the delivery of person-centered care: evaluation of an educational program for the healthcare workforce. *BMC Health Serv Res*. 2022;22(1):364. <https://doi.org/10.1186/s12913-022-07689-w>
 24. Koto-Shimada K, Fujita N, Matsuoka S, Jimba M, Touch S, Zwi AB. Medium-term outcomes of a program to upgrade the nursing faculty in Cambodia: A qualitative study. *Nurse Educ Today*. 2022;116:105438. <https://doi.org/10.1016/j.nedt.2022.105438>
 25. Santos RA, Mendes WV Júnior, Martins M. Quality in healthcare and patient safety: evaluation of the results of a distance learning program. *Cien Saude Colet*. 2021;26(10):4553-68. <https://doi.org/10.1590/1413-812320212610.11082021>
 26. Machado MGO, Sampaio CL. Training on common mental disorders at a hospital ward: the use of active methodologies in care construction. *SMAD, Rev Eletrônica Saúde Mental Álcool Drog*. 2021;17(1):75-83. <https://doi.org/10.11606/issn.1806-6976.smad.2021.168134>
 27. Jobst S, Windeisen M, Wuensch A, Meng M, Kugler C. Supporting migrants and refugees with posttraumatic stress disorder: development, pilot implementation, and pilot evaluation of a continuing interprofessional education for healthcare providers. *BMC Med Educ*. 2020;20(1):311. <https://doi.org/10.1186/s12909-020-02220-3>
 28. Knobel R, Menezes MO, Santos DS, Takemoto MLS. Planning, construction and use of handmade simulators to enhance the teaching and learning in Obstetrics. *Rev. Latino-Am. Enfermagem*. 2020;28:e3302. <https://doi.org/10.1590/1518-8345.3684.3302>
 29. Castro JA Filho, Motta LB. Evaluation in distance learning: a case report of the UNASUS/UERJ postgraduate course in elderly health. *Rev Bras Geriatr Gerontol*. 2018;21(5):513-22. <https://doi.org/10.1590/1981-22562018021.180021>
 30. Bijani M, Rostami K, Momennasab M, Yektatalab S. Evaluating the Effectiveness of a Continuing Education Program for Prevention of Occupational Exposure to Needle Stick Injuries in Nursing Staff Based on Kirkpatrick's Model. *J Natl Med Assoc*. 2018;110(5):459-63. <https://doi.org/10.1016/j.jnma.2017.11.002>
 31. LaMar RA. Using a Virtual Journal Club for Sharing Evidence-Based Practice Recommendations in Critical Care Registered Nurses. *Teach Learn Nurs*. 2017;12(1):53-8. <https://doi.org/10.1016/j.teln.2016.08.006>

32. Liaw SY, Wong LF, Lim EYP, Ang SBL, Mujumdar S, Ho JTY, et al. Effectiveness of a Web-Based Simulation in Improving Nurses' Workplace Practice With Deteriorating Ward Patients: A Pre- and Postintervention Study. *J Med Internet Res.* 2016;18(2):e37. <https://doi.org/10.2196/jmir.5294>
33. Lahti ME, Kontio RM, Välimäki M. Impact of an e-Learning Course on Clinical Practice in Psychiatric Hospitals: Nurse Managers' Views. *Perspect Psychiatr Care.* 2016;52(1):40-8. <https://doi.org/10.1111/ppc.12100>
34. Maloney S, Haas R, Keating JL, Molloy E, Jolly B, Sims J, et al. Effectiveness of Web-Based Versus Face-To-Face Delivery of Education in Prescription of Falls-Prevention Exercise to Health Professionals: Randomized Trial. *J Med Internet Res.* 2011;13(4):e116. <https://doi.org/10.2196/jmir.1680>

Authors' contribution

Study concept and design: Fernanda Maria de Miranda, Vicki Leigh Kristman, Vivian Aline Mininel. **Obtaining data:** Fernanda Maria de Miranda, Bruna Vasconcelos dos Santos. **Data analysis and interpretation:** Fernanda Maria de Miranda, Bruna Vasconcelos dos Santos, Vicki Leigh Kristman, Vivian Aline Mininel. **Obtaining financing:** Fernanda Maria de Miranda, Vivian Aline Mininel. **Drafting the manuscript:** Fernanda Maria de Miranda, Bruna Vasconcelos dos Santos. **Critical review of the manuscript as to its relevant intellectual content:** Fernanda Maria de Miranda, Bruna Vasconcelos dos Santos, Vicki Leigh Kristman, Vivian Aline Mininel. **Others (Supervision):** Vicki Leigh Kristman, Vivian Aline Mininel.

All authors approved the final version of the text.

Conflict of interest: the authors have declared that there is no conflict of interest.

Received: Dec 20th 2023
Accepted: Aug 7th 2024

Associate Editor:
Maria Lúcia Zanetti

Corresponding author:
Vivian Aline Mininel
E-mail: vivian.aline@ufscar.br
 <https://orcid.org/0000-0001-9985-5575>

Copyright © 2025 Revista Latino-Americana de Enfermagem
This is an Open Access article distributed under the terms of the Creative Commons (CC BY).

This license lets others distribute, remix, tweak, and build upon your work, even commercially, as long as they credit you for the original creation. This is the most accommodating of licenses offered. Recommended for maximum dissemination and use of licensed materials.