Paths and Reflections on the Construction of an Evaluation Index System for Core Competencies in Physical Education

Fangyong Liang*

Heilongjiang University, Harbin, 150080, China
**Corresponding author: 13936277776@139.com*

Abstract: *To construct an evaluation index system for core competencies in physical education needs to understand the logical relation between epistemology and methodology. Based upon the integrative and generative features of competency and its three-tier structure from conceptual clarity; exploring the applicable boundaries of major theoretical resources within physical education contexts; unveiling the epistemological blind spots of two research paths (quantification/interpretation); we proposed screening criteria through cross-validation of three kinds of evidence, a combined weighting strategy, and a hierarchical differentiation scheme with complementary dual anchoring according to the progressive logic "indicator selection—weight assignment—standard anchoring". Reflective analysis of the evaluation system reveals the validity dilemma stems from the deep incompatibility between measurement logic and the essence of competency. Through analyzing the value tensions among functions, it proposes a context-sensitive framework switching mechanism, delineating the applicable boundaries of our model and outlining the direction for its future evolution. This study suggests that the crux of the evaluation system does not lie in streamlining indicators or getting accurate weights but in matching the evaluation logic with the embodied/situated nature of competency.*

Keywords: *core competencies in physical education; evaluation index system; validity; anchoring strategy; weight assignment*

Introduction

The assessment of core competencies in physical education has received great attention recently. The division of them into three dimensions according to curriculum standards are motor ability, healthy behavior and sports morality. However, the answer about "what competency contains" by dimensional division cannot solve the question how it can be identified and measured. The embodied, situated and generative features of physical education competency will face adaptability problems when transplanting general evaluation theories. After reviewing existing researches, most focused on technical issues such as indicator screening and weight calculation but there was no systematic investigation regarding epistemology of evaluation's premise, disciplinary appropriateness of theoretical resources and functional realization of the evaluation index system in real scenario; therefore the construction of the evaluation system is mostly reduced to a technical problem. Our study tries to fill up this gap. Based on our clarification of the connotation of core competencies in physical education we systematically demonstrated the theoretical path for constructing the evaluation index system and did a reflective examination of it from three dimensions (validity, function, and boundaries).

1. Conceptual Clarification and Logical Starting Point of Core Competency Evaluation

1.1 Clarification of the Connotation and Analysis of the Structure of Core Competencies in Physical Education

The evolution of the concept of core competence definitions in physical education passed from vagueness to structural definition; curriculum standards sum it up in three dimensions – motor ability, healthy behavior, and sports morality. Three dimensions are not juxtapositional (the foundation of motor ability, motivation provided by sports morality) but just listing them does not clarify its internal

structure. According to meta-theoretical aspect of competency research, we define core competencies in physical education as a comprehensive quality system accumulated by those who participate in physical education. Core characteristics include: integrativeness and generativity – the first one means an organic integration rather than mere addition of various elements while the second implies that competency is a dynamic result which is gradually built up continuously in bodily practice[1].

Structure point of view There is a three tier progressively architectural framework for the core competencies in physical education: "Skill level – Cognitive level – Attitudinal level", Skill level means motor representation and bodily control ability, Cognitve level means tactical understanding, internalization of rules and application of health knowledge, Attitudinal level implies motivation for sports participation, identification with sports values and character cultivation. The three are mutually constrained hierarchically—skill level provide the bodily basis for deepening of the cognitive level which in its turn optimize the performance of the skill level while attitudinal level regulate the direction and intensity of the former two. It avoid the narrowness of the core competencies as a quantified list of skills and it could be used as a concept premise when we construct the evaluation index system later on.

1.2 Theoretical Resources and Applicable Boundaries for the Construction of the Evaluation Index System

In building up an evaluation index system, we need a guiding theory, but they have great difference in their logical methods and applicability: The competency-based education model emphasizes the observable competencies as unit of evaluation. Its goal is clear and it is highly operable; some dimensions such as tactical awareness, role perception, embodiment of sports morality cannot be completely decomposed into individual behavior indicators, relying too much on the competency checklist obscures the holistic nature of competency. Dynamic assessment theory can capture the potential development level of learners through intervention-feedback-reassessment. It has explanatory power about non-linear skill leap in physical education learning, but there exists a tension between operational cost and standardization if evaluated at scale. Item response theory provides psychometric support for developing evaluation instruments, however, the assumption of unidimensionality or local independence contradicts the intrinsic association among dimensions of physical education competency.

After reviewing all these theoretical resources, we found that most of them are based on general educational contexts or cognitively-oriented disciplines: thus dividing indicators by the logic can't fully understand the embodiment, situation and generation characteristics unique in physical education as a discipline. The SOLO taxonomy divides various degrees of structuring of responses at different levels of thinking, it offers a cognitive development reference when setting up the level of evaluation standards but with low explanatory power about such implicit characteristics (like embodied performance) and affective attitudes. Borrowing theory should have its own boundaries so that there will not be any direct transplant resulting in the disconnection between evaluation content and the essence of the discipline. Take matching the embodied cognitive attributes of physical education learning as the basic criterion when choosing theories, make necessary adjustment if needed[2].

1.3 Knowledge Map and Paradigmatic Limitations of Existing Evaluation Research

Systematic literature review gives us an initial start for our research. Based on bibliometric analysis we can see there are two general development paths when researchers study the evaluation of core competencies in physical education: one is quantitative evaluation which focuses on selecting and weighting observable behavioral indicators using mainly the Delphi method, the analytic hierarchy process and the entropy method; the obvious advantages of quantitative evaluation include its transparency and reproducibility, however, the tendency to fragment the selection of indicators easily breaks down the organically integrated structure of competency into isolated items while relying too much on expert's consensus in assigning weights leads to the ossification of the evaluation system regarding the experience of a particular group without context adaptability. The other is interpretive evaluation which aims at generating mechanisms of meanings and individualized developmental trajectory during the learning process through means like classroom observation, learning portfolios, etc. capturing the dynamic of competency, however due to the ambiguity of evaluation standards and subjective interpretation of results makes it hard to get stable and consistent professional judgment.

Each route has their structural epistemological blind spots (quantitative route assumes that

competency can be fully externalized and quantified; interpretive route supposes that professional intuition alone is enough for holistic grasp), but it's important to note that most studies consider constructing an evaluation index system as a technical task without any critical examination on the value presuppositions and epistemological premises embedded in evaluation. A trend goes from macro-level conceptual analysis into concrete tool development in researches in this field, however, there are little reflective results at the methodological level. Most of them don't have coherent theory logic among evaluation objectives, evaluation content, and evaluation methods – they just end up with "construction" of the evaluation index system and no in-depth inquiry about the premises of construction—paradigmatic limitation becomes the logical starting point for further path generation and reflective examination.

2. Path Generation and Logical Progression in the Construction of the Evaluation Index System

2.1 Multi-Source Bases and Screening Criteria for Indicator Selection

Cornerstone in building up an evaluation system is the indicator selection that decides the credibility of further research. Sources of indicators must follow the cross-validation logic of a threefold basis: extracting consensus-based constituent elements of competency from the knowledge system of physical education (theoretical basis); point out the expected stage-specific outcomes of physical education (objective basis), ensure consistency with indicators and learning specifications at different learning stages; include experiential judgment of frontline teaching subjects on competency performance (subject basis) compensate for possible contextual distortion due to pure theoretical deduction. Candidate indicators extracted from the overlapping region of the threefold basis not only possess the normative constraint of disciplinary logic but also the ecological information from educational settings, whose content validity far surpasses those of indicator sets originated from one source alone[3].

After determining a candidate indicator set, then there is need to screen for superior elimination and streamlining of indicators: observability refers that the competency performance mentioned by each indicator could be directly or indirectly captured in evaluations; discriminability indicates meaning variation on scores between learners at different levels without ceiling effect or floor effect; sensitivity means good response to instruction and development (those which do not show any significant change even after long-term learning) which should be eliminated; mutual exclusivity refers clear boundary among indicators avoiding overlapping concepts leading to repetitive inclusion of evaluation information and distortion of weight assignment. The coordinated use of above criteria would effectively reduce the redundancy of the evaluation index system with broad content coverage but achieving its economic operation.

2.2 Mechanism Selection and Calculation Logic of Weight Assignment

The weighting for multi-indicators are not only allocating weights but also making value judgments among indicators. The subjectivity of weighting relies on the experiential judgment and professional consensus from a group of experts; even though very valid when dealing with those areas whose connotation of indicators cannot be quantified completely, but can be affected by various systemic factors (composition of experts' background, difference in judgment scale and group negotiation process); while the objective weighting allocates weights based on the discrete information of the evaluation data itself: It enables reproducible testing and cross-context comparison, however, there might be loss of cross-group stability of weight coefficients due to fluctuation of sample characteristics. Evaluating core competencies in physical education, if we rely exclusively on the subjective method, then the result will lead to over-fitting the cognitive preference of one expert group or else if we rely solely on the objective method, it leads us towards direction which is statistically significant yet lack explanatory power in the pedagogical sense.

Solution is the combination of weighting: Multiple rounds of Delphi survey are conducted for getting the judgment matrix of experts on the importance of each indicator by the analytic hierarchy process to get the subjective weight vector; Meanwhile, calculate the variability information entropy of each indicator from the evaluation data matrix, the objective weight vector will be obtained via the entropy method. If two kinds of weights meet the requirement of statistical consistency, then we can calculate the combined weight coefficient through multiplicative synthesis or linear weighting such as the final result has the dual constraint of professional consensus and data evidence at the same time.

Integrates experiential information (value judgment) in the level of value and factual information (empirical level), avoid systematic bias caused by only one information source, improves the anti-interference capacity and transfer adaptability of weight interpretation in multiple evaluation context[4].

2.3 Hierarchical Differentiation and Anchoring Strategies of Evaluation Standards

Evaluation standards connect abstract indicators with concrete judgment; its quality determines degree of inter-rater consistency and interpretation validity. The hierarchical distinction shall be based upon the non-linear features of competency development rather than an equidistant progression mechanically. Generally speaking, the growth curve of the learning process for motor skills grows quickly firstly then slowly later whereas the internalization of sports attitudes has alternating plateaus and leap periods. Uniformly dividing grades violates the inner rhythm of competency development thereby weakening the discriminative role of middle grades causing imbalance in grade density at both ends. Hierarchical division conforms to the empirical developmental trajectory of each dimension: denser where the rate of change is fast and relaxes thresholds when changing gently so that the increment of development corresponded to each level approximately equal.

Both behavior and outcome anchoring should exist in the anchoring strategy. Former is meaning the typical behavioral performance is used as the reference standard for grading. We preset behaviors based on each level which allow evaluator to judge by matching degree with these examples; it's intuitive to establish a reference framework to enhance inter-rater consistency. Latter means the verifiable learning outcomes are used as the reference for judgment. An evaluated person does some motor tasks or solving problems, who gets graded scores depending upon the quality of his/her task completion. Both anchoring modes complement each other (former captures process performance whereas latter verifies outcome-based competency) so we could use evaluation standards that are process-sensitive but outcome-verifiable so we could avoid evaluation blind spots due to only using one anchoring mode.

3. Reflective Examination and Paradigmatic Calibration in the Construction of the Evaluation Index System

3.1 The Epistemological Dilemma of Evaluation Validity and the Possibility of Its Resolution

Validity is the foundation of a quality of an index system; however, about the evaluation of core competencies in physical education, validity goes no further than technical reliability test. The quantitative evaluation paradigm assumes complete externalization and quantification of competency hiding validity - correspondence accuracy between measurement instruments and the traits measured-embeddedness and holistic generativity of core competencies. Standardization of evaluation disintegrated the organically integrated structure of competency into isolated behavior units leading to structural representational gap between "competency" represented by evaluation results with competency in real settings (measurement might be accurate at the technical dimension, but the object being measured has already deviated from the original form of competency). This dilemma comes from the deep incompatibility between measurement logic and the essence of competency which can't be solved even if we increase our sample size or optimize our statistical models.

The introduction of the notion of "ecological validity" will solve this dilemma --the demand for ecological validity requires that we satisfy both the requirements of internal validity as well as ecological validity at the same time---the former ensure measurement precision and consistency; latter guarantee structural similarity and meaningful coherence between the evaluation context with real contexts of physical education participation. In order to place our evaluation tasks into authentic sports scenarios (evaluating tactics decision-making ability in game contexts, evaluating sports morality performance in team collaboration, or evaluating healthy behavior habits in self-directed exercise) so that the evaluation results can be transferred well to infer daily competency states. The epistemological turn is away from the single orientation of "precise measurement", to the composite orientation of "effective representation"--seeking a new balance between measurement precision and ecological authenticity.

3.2 Value Tensions and Adjustment Orientations of Evaluation Functions

Deep conflicts of values between different functions of evaluation index systems: Differentiation

orientation focusing discrimination function (ranks comparison with norm-reference condition), choose highly variable behavior variables for designing indicators; focus on hierarchically rank learners according to their ability level. Contrary to Developmental orientation focusing intra-individual difference and growth trajectory (self-comparison with criterion reference condition), need indicators sensitive to longitudinal tracking and oriented towards education improvement; focus on degree of progress rather than horizontal ranking. Two conflicting orientations are essentially divergent from "for selection" logic vs. "for growth", accommodate both functions in one index system usually canceling out the principle of indicator design resulting in an ambiguous result interpretation.

Basic idea: Adjustment is context sensitive switch of evaluation framework. Different purposes leads to change of reference systems/scoring rules (norm-referenced scoring – emphasis on relative ranking) vs. criterion-referenced scoring (focuses on magnitude/direction of individuals' own changes), without changing any content composition/weight structure of the index system, but adapt the way of presenting evaluation information and interpretative framework to unleash its differentiated functional value under different contexts by using one system; save repetitive job of setting up a new evaluation system for every function with coherent interpretation across contexts. Assumes that index system has sufficient richness in dimensionality/hierarchy refinement[5]

3.3 Theoretical Boundaries and Directions of Evolution of the Evaluation Model

Each evaluation model reduces the complexity of competency – it is a vital academic self-awareness ensuring the rational use of the model by defining applicable boundaries. Applicable educational stages: Focuses on the basic education stage that needs to examine the applicability for preschool or higher education stages according to the age appropriateness of descriptor details and covering the range of development in grade setting. Applicable event type: Motor skills indicators match better with the features of open-skill events (ball games and track and field), while descriptors and anchoring examples of closed-skill events (gymnastics and martial arts) need recalculation. Applicable evaluation purpose: Good adaptability in small-scale formative evaluation but the feasibility and validity performance in large-scale monitoring still needs empirically tested. Defining these boundary conditions prevents an inappropriate extension of the model when the theoretical premise does not hold.

The evolution of the evaluation model has to confront issues both internally (internal optimization) and externally (external responsiveness). Internally: if there is change in the consensus on disciplinary connotations or new evidence from learning science emerges – expansion/organization of indicator dimensions, regular calibration processes for the weight system and empirical revisions of threshold grades of divisions will be the main topics of self-renewal of this model. Externally regarding technology: The maturity of emerging technologies like artificial intelligence-assisted evaluation and wearable devices offers new possibilities for capturing and automatic analysis of the process of motor performance that force structural adjustment of the model from the perspective of information acquisition methods. A vital model academically cannot be closed and self-sufficient – it needs space for interface flexibility and revision approaches towards a more fitting representation of the essence of competency via continuous self-examination and interdisciplinary dialogue.

Conclusion

We investigated the construction process of an evaluation index system for core competencies in physical education through levels of conceptual clarity, path generation, and reflective examination. Identified the integrative and generative characteristics of competency and its three-tier structure (the theoretical selection should adopt embodied cognitive attributes), showed the progressive logic of three links—the indicator selection, weight assignment, and standard anchoring—a coherent chain "of what to select—how important—how to judge", revealed the epistemology roots of the validity dilemma, put forward a way out with ecological validity, analyzed the functional tensions, offered a context-sensitive switching strategy, clarified the bounds and direction of evolution of the model. We show that there are some basic issues need to be addressed at the epistemological level when constructing the evaluation system such as whether competency can be measured and if so, how?. In our next work we will clarify the operationalized testing methods for ecological validity, verify the applicable thresholds of the framework switching mechanism under different contexts, pay attention to the impact of technology embedding like artificial intelligence and wearable devices on the model's structure.

References

[1] Zhang Lei. "Design of Teaching Evaluation for Students' Core Competencies in Physical Education Based on Three Dimensions." Teaching and Management, no. 20 (2024): 58–61.
[2] Liu Haohui, et al. "The Reasons for Proposing Grand Evaluation, Analysis of Its Elements, and Promotion Strategies Under the Guidance of Core Competencies in Physical Education and Health." Journal of Capital University of Physical Education and Sports 35, no. 4 (2023): 381–390.
[3] Qiu Xiao and Chen Xianghe. "Current Status, Mechanism, and Practice of Physical Education Learning Assessment Based on Core Competencies." Youth Sports, no. 7 (2023): 81–83.
[4] Zou Liang and Liu Xianfa. "Construction and Application of an Evaluation Index System for Middle School Students' Physical Literacy: Based on Big Data from the Comprehensive Evaluation of Education Quality in City A." Educational Measurement and Evaluation, no. 6 (2023): 33–47.
[5] Ye Changfa and Yu Minling. "Discussion on the Construction of a High School Physical Education Learning Quality Evaluation System from the Perspective of Core Competencies." Sports Vision, no. 18 (2021): 95–96.