frontiers | Frontiers in Education

TYPE Conceptual Analysis
PUBLISHED 25 April 2025
DOI 10.3389/feduc.2025.1556993

OPEN ACCESS

EDITED BY
Surette Van Staden,
University of Innsbruck, Austria

REVIEWED BY
Jose Manuel Salum Tome,
Temuco Catholic University, Chile
Magdalena Kohout-Diaz,
Université de Bordeaux, France

*CORRESPONDENCE
Matias Urrutia-Jorde
murrutia@uoregon.edu

RECEIVED 07 January 2025
ACCEPTED 08 April 2025
PUBLISHED 25 April 2025

CITATION
Urrutia-Jorde M (2025) Prospects for development of culturally inclusive models for education: diagnostic classification models.
Front. Educ. 10:1556993.
doi: 10.3389/feduc.2025.1556993

Prospects for development of culturally inclusive models for education: diagnostic classification models

Matias Urrutia-Jorde*

Department of Special Education, University of Oregon, Eugene, OR, United States

Diagnostic classification models (DCM) have been a recent topic of conversation in the development of educational materials. Specifically, there has been significant criticism of their validity and use within the educational system. For this reason, I identify avenues for the development of DCMs by taking a global perspective. Current literature on adapting pedagogy to culture is presented to consider multidimensional models, like DCMs, as useful tools for global development. This publication aims to present DCMs in a more accessible format, as well as illustrate that multidimensional models are useful for large scale data aggregation. I challenge criticisms that DCMs has failed to establish reliability and validity by looking at the foundation of the model. Although not all issues are resolved, DCMs are presented as practical tools for future research, and as prospects for educational development in low- and middle-income countries.

KEYWORDS

diagnostic classification models, cultural pedagogy, development, international assessment, education

Introduction

There is a well established relationship between socioeconomic status (SES) and academic achievement. In the United States, children with higher SES incur advantages that can be identified as early as preschool. The United States has higher SES inequality than other countries, but it has been suggested that this factor alone cannot account for the poor performance on international assessments (Schmidt et al., 2015). Some studies analyzing academic performance of children on international assessments have focused instead on structural characteristics like class size, standardization and teacher quality Other studies focus on the implementation of intervention without sufficient evidence on effectiveness (Schmidt et al., 2015; U. S. Department of Education, 2013). Increasing attention has been drawn to the curriculum as representing students' opportunity to learn (OTL).

AUS study by Abedi and Herman (2010), was able to demonstrate that SES and language proficiency jointly affected OTL. Low SES was associated with poor performance and low OTL, and the disadvantage was compounded based on the learner's English language proficiency. Schmidt et al. (2015) then posed the question of how educational achievement levels differ across countries, and the role curriculum plays in explaining these

differences. Using data from the Program for International Student Assessment (PISA), and the Trends in International Mathematics and Science Study (TIMSS), they were able to analyze and compare trends internationally and within the US. The research group was able to demonstrate that the US are not alone. Although there is higher variability of OTL in the US, globally, students from disadvantaged backgrounds receive less rigorous curriculum content (Schmidt and Burroughs, 2015), and that higher alignment between intended and implemented curriculum was correlated to student achievement (Schmidt et al., 2015).

Global development has become a pressing issue; research focused on adapting pedagogy to cultural context has identified problems addressing the suitability of western educational materials for use in other countries (Jukes et al., 2021). In addition, the lack of infrastructure is a barrier to educational reform, leaving developing countries with major interruptions in education (Lawn et al., 2008; Duff et al., 2020). Lack of infrastructure has made implementation of special education and individualized instruction a major challenge in low- and middle-income countries, lacking any common acknowledgment of disability in regular instruction.

The use of hierarchical linear modeling (HLM), or the nested structure of data aggregation on an international scale, might demonstrate that diagnostic classification models (DCMs) have promise for addressing specific problems in the development of education programs in low- and middle-income countries (LMIC). Specifically, DCMs promise to provide assessments that can inform instruction and aggregate data at a classroom, school, and state level, as well as promising early screening for disability, and individualized learner profiles to inform remedial instruction (Bradshaw, 2016; U. S. Department of Education, 2023).

However, recent criticism of DCM has focused on the failure to establish appropriate reliability and validity for the measure (Sessoms and Henson, 2018). DCMs are nested large-scale multidimensional models that are capable of aggregating data used for international assessment. My goal in writing is to present DCM as a potential avenue for global development and as a tool for culturally inclusive methods of psychometric and academic assessment for countries with a large number of diverse cultural groups. The latter goal can be done by investigating successful development of multidimensional and inclusive models, like those developed in Australia, as well as providing an example of what development looks like in the present moment. In this way, DCMs can be presented in a more accessible format. However, before DCM can be considered for the international stage, its validity and reliability need to be established and so the publication will look at the foundations of DCMs.

DCMs: what they are and where they came from

DCMs are distinguished by their ability to make appropriate categorical inferences for mastery and non-mastery criterion. DCMs differ from traditional criterion based measures (CBM) in that they aim to describe learning as a toolbox that identifies specific aspects that a learner needs to improve on. This means that unlike previous unidimensional models, DCMs are multidimensional models considered appropriate for categorical tasks. DCMs could be promising for refining cognitive theories or as educational tools for describing diverse student reasoning (Bradshaw, 2016).

In this sense DCMs can be thought of as an ideal case of criterion referenced measurement (CRM) as described in the original publication by Popham and Husek (1969). An ideal test that not only ties response to criterion, but also presents scores that represent an individual's response pattern and mastery. DCMs are unambiguous, are tied to criterion, and are homogenous, meaning that based on score, administrators should know within error limits what a person can and cannot do. Yet, if diagnostic classification models are really as promising as has been proposed, we should begin by addressing important aspects that have not been met in their development: validity and reliability.

It has also been previously noted that DCM requires large sample sizes and that a prevalent problem has been the failure to report results for reliability analysis (Sessoms and Henson, 2018). Reliability (consistency) is important and should be considered as a statistic to describe a measure's truthfulness. However, further issues with reliability cannot be addressed until there can be more adequate data on specific problems and greater reporting. Instead, focus will shift toward DCM's validity foundations in CRMs.

Messick (1994) identified issues of accessibility and the need for rigorous scientific analysis of assessment. These are considered in DCM's technical foundations which presents DCM as a measure that should be attaining validity, and should indicate the issue is one of theoretical understanding and not mathematical rigor. For example, Messick described the importance of rigorous definition and appropriateness of the criterion in relationship to the specific construct. Although the appropriateness of criterion is addressed below, it remains useful to think of these criteria as built upon previous efforts developing CRM.

CRM began with Popham & Husek from the original publication and many of the goals for the development remain the same. CRM promised to change the model from above the norm (overachiever) and below the norm (underachiever) to one that describes if a student has met the criterion for mastery (Popham and Husek, 1969). DCMs developed as a refinement of previous efforts rather than an entirely new development. For example, the original publication describes an ideal case of CRM as identifying specific portions of material that an individual learner has mastered (Popham and Husek, 1969). DCM can be conceived as an *ideal* criterion referenced test.

Just as they could be considered an advancement in CRM, DCMs also represented an advancement toward a multidimensional model. Previous models had focused on a conception of learning where learners are conceived of as, behind, at mastery, or ahead (Bradshaw, 2016). This conception limits learners and instructors, sometimes leading to redundancy in supplemental instruction. Multidimensional models deviate from previous iterations by identifying mastery in additional aspects that serve as indicators for performance on material. CRM, like DCM, was originally conceived of as a more comprehensive method for measuring learner progress, and without development, CRM would've never achieved appropriate validity or reliability.

Over the years, CRM has undergone extensive development and refinement. Probably the most significant example is curriculum based measures (CBM), which has in turn undergone extensive development for over 30 years. CBMs faced similar challenges in development that were described in a review by

Tindal (2005) for early reading and math assessment, as well as early screening for disability (Tindal, 2005). Originally, work focused on the identification of specific criterion (reading fluency, comprehension, etc.) and development of appropriate tools for analyzing CBM which became an important piece of response to intervention (RTI). Tindal (2005) systematically applied Messick's framework to CBM and described the ways in which CBM was developed to provide a thoughtful and thorough analysis of validity. In this way both technical and practical aspects are considered, and Tindal develops a strong argument in favor of CBM.

These examples are analogous to the promises and challenges proposed by DCMs. New tools need to be developed to study the effectiveness and appropriateness of DCMs, which can then be conceived as the next stage of CRMs. Further development is needed to address "*sub-criterion*" relevant to identify learning zones in DCM assessment. Finally, development of DCMs mark another step in the framework described by moving from a typical criterion referenced test to an ideal criterion referenced test (Popham and Husek, 1969).

Ensuring the solution can be sustained

To address validity, technical foundations of DCM need to present the theory behind their development. Validity can be thought of as the measure's ability to describe the learner's mastery. Once validity is established, the measure can then be used to identify differences in growth and mastery between individuals in specific aspects of literacy and numeracy. Previous unidimensional models like CBM provide clues to definition and identification of important aspects of validity.

The parallels drawn between CBM and DCM are intentional, primarily because the challenges in developing DCMs are universally present in identifying appropriate *sub-criterion*, that require robust support for their validity and use within a framework. The usefulness of DCMs can be considered in relation to the existing literature, because these sub-criteria need to be identified *a priori* (Bradshaw, 2016). And, as Messick stated, it is critical to have sufficient discriminant validity among measures. Likewise, DCM needs to address aspects of construct validity with sufficient rigor that support relevant domains and theoretical understanding.

The single most important factor to be understood in the development of DCMs is that they are mathematically sophisticated models that require development with assistance from a statistician (Sessoms and Henson, 2018). DCMs promise to classify students based on skill mastery and would enable targeted feedback for remediation. Recent review by Sessoms and Henson (2018) supported criticism of low validity and reliability, however, the authors proposed some of significant changes that need to take place, explicitly calling for researchers to report reliability and validity where most publications are not.

Further development of validity remains beyond the scope of this publication, as identification of factors would require careful review of literature to address what publications are available on CRMs to possibly aid in identification of necessary components for DCM. So far, the issues remain totally unaddressed, as a serious attempt at describing validity would require expertise in *criterion* and assistance from a *statistician*. In lieu of these experts DCM's validity would be better thought of as addressed by the technical underpinnings of the framework.

Technical foundations

In this last section, I address two main technical aspects of DCMs that should be considered. The first is the testing and falsification of a hypothesis. The second is a description of what is contained in a learner profile. The DCM framework contains within it a system for researchers to test their hypothesis. To do this, researchers need to identify what sub-criterion are required for each item on the test and develop a matrix. The matrix serves as the working hypothesis and is falsifiable through a log-linear analysis which would report whether the item was found to correlate with the criterion. These items are able to produce the learner profile once the best items and the relevant criterion have been selected from a pilot study. As stated above, expertise is required for development because analysis of each item grows exponentially based on the number of sub-criterion required. This sets a theoretical limit for each item to a number of sub-criterion for practical purposes, a more in depth view will be addressed below.

Testing hypothesis within the framework

Multidimensional models require more complex algorithms, longer tests, and larger samples. DCMs also have promise for high dimensionality under feasible testing conditions. In modeling DCM, traits are typically known as attributes "mastery" or "non-mastery" denoted as present or not present (1 or 0) these attributes are represented in a Q-matrix in which attributes are represented in each column and items are represented in rows. Therefore, a researcher uses the Q-matrix to represent a detailed hypothesis about how attributes relate and how they interact to yield a task response. For example, an item would be hypothesized to contain within it one or several attributes defined by a 1 or a 0 (Bradshaw, 2016).

The Q-matrix is used in confirmatory factor analysis so that each item can be denoted as possessing or not possessing a specific trait. DCMs are confirmatory latent class models that have two requirements: latent class definition and class specific response behavior (Leighton and Gierl, 2007). Other latent class models include Cognitive Diagnostic Models, a category which includes DCMs; this means DCMs assume responses are conditionally independent. DCMs are appropriate when the task traits are *categorical inferences*: classifications such as non-mastery, partial mastery, or complete mastery. A sample distribution might include for example a math problem that includes attributes of *multiplicative comparisons*, or *referent units* which, depending on the relationship hypothesized beforehand between item and criterion, correlate with item response.

The results can be expressed by a simple bar graph known as an Item Characteristic Bar Chart (ICBC) that denotes a threshold for mastery and can inform the probability that given the response of a student, that student has mastery of a trait. Additional information for each individual item would also include the difficulty of the item

and a value to denote how well the item discriminates between two groups of examinees: groups like masters or non-masters. Results can be useful for aggregating mastery of specific attributes at a classroom, school and state level (Bradshaw, 2016). If an item with two attributes were analyzed, it would then undergo a process analogous to an analysis of variance (ANOVA) called a log-linear cognitive diagnosis model (LCDM). In the same way an ANOVA for two traits would produce a 2×2 model, with four main effects and an interaction, LCDM produces the same data using methods appropriate for DCM. Item response probabilities dependent on the individual's personal attribute profile. These analyses are done for every item and the information grows exponentially for every factor and item possess, creating a practical limit to the number of relationships that researchers can test between items.

DCMs can be thought of as testing a hypothesis relating variables to item response, and like all hypotheses these need to be falsifiable. Therefore, once a researcher has developed a Q-matrix which includes all attributes and items, they can test, and express results using an ICBC. This would include a significance level or *p*-value and a value for item discrimination, which again is used to determine the item's ability to discriminate between categories (master or non-master). Then these items can undergo a more statistically appropriate measure to determine response variance using the LCDM which, as stated above, is analogous to an ANOVA (Bradshaw, 2016).

Interpreting DCMs

The parallels addressed earlier between CBMs and DCMs draw attention to those familiar well documented challenges. Originally, the idea of using CRMs was a step that required development (Popham and Husek, 1969). DCMs mark an additional step toward an ideal criterion referenced test that can identify various aspects of learner development and changes the current unidimensional model of learning in which a student is either "ahead" or "behind" to one which identifies which specific skills a student needs to develop. This idea, proposed in the 70s, was as far from being a reality then as it is now. Similar challenges were present in describing an appropriate level of rigor when developing the measure's validity all of which were overcome.

Review of available literature on DCMs by Sessoms and Henson (2018) found that DCMs have insufficient reliability and validity. However, as described above in the technical foundations, DCM are mathematically rigorous. Development of DCMs from a mathematical perspective is unnecessary, because the models are already mathematically sophisticated. This would indicate that results producing low levels of reliability could be due to the insufficient development criterion from a theoretical standpoint.

Latent Class Models like DCMs require two specifications: *latent class definitions and class specific response behavior*. *Latent class definitions* are the set of attributes to be measured by a test and are determined *a priori*. This is important because these are the traits that then classify examinees. This means that a test that measures "A" number of attributes produces 2^A possible learner profiles (Bradshaw, 2016). Then, each individual would have a learner profile in which they have a unique pattern of mastery over each attribute. This unique pattern of mastery would be denoted by a pattern of either 1 or 0. For example, [1001] would indicate a learner who has mastery over the first and fourth traits, but not the second or third.

The model's Class Specific Response behavior denotes the relationship between items and attributes that are specified prior to analysis in the hypothesis. Each item is designed to elicit a specific subset of attributes, which should align with the Q-matrix. Entries in the matrix are denoted as 1 if they are hypothesized to systematically influence response to an item i or otherwise denoted as 0. The Q-matrix is specified to a statistical model and represents the researcher's hypothesis and is used to refute a hypothesis by examining the model fit (Bradshaw, 2016).

In this way the model Q-matrix can be applied to a statistical framework that analyzes an individual's scores on item response. In this case, is a structural parameterization is expressed with summation of all possible learner profiles (2^A), the total proportion of which is equal to one. This first part of the statistical equation is simply to denote the *base rate* of mastery in a population. This is then completed by the second half of the statistical model which is a function of the product. This equation uses both sigma and pi to denote these. As stated above, the sigma notation is used to denote mastery base rates, while the pi notation, a function of product, is the probability an examinee will provide a correct response given their specific learner profile (For a more detailed explanation, refer to Bradshaw, 2016).

Simplified view of DCMs

DCMs are mathematically complicated models, simplified only with the intention of aiding some foundational understanding for the implications of their development. DCMs are currently in a state of underdevelopment primarily because of the failure to report both *reliability and validity*. The development of DCM's validity is one that requires expertise in the literature that concerns criterion because these attributes are determined *a priori*. These need to be aided in development by a statistician because each item needs to be designed within the framework and the relationship between items and attributes needs to be specified prior to analysis.

The use of DCMs involves methods that are statistically sophisticated. The framework for development would be tested using a Q-matrix that operates as a working hypothesis, requiring the above collaboration to develop all items and identify all traits prior to any assessment. In summary, DCMs use an equation made of components sigma and pi which is specific to the general framework for which the Q-matrix is adapted. The equation includes a working base rate of mastery in a population (sigma), and an expression of the probability of a learner's response given their specific profile (pi), producing multimodal distributions. The probability of correct response is defined differently for each DCM as given the specifications of the q-matrix.

Responses are then analyzed using the LCDM allowing researchers to test and falsify the hypothetical relationship of correct response and a learners specific profile. In this way, DCMs can provide analyses for multiple traits in a single item and give us results similar to a 2×2 ANOVA, with four main effects and an interaction. Additionally, each item produces a bar graph or ICBC that denotes item discrimination, an important piece in

development to determine which items are unnecessary or poor predictors when conducting pilot studies. The ICBC also has an associated *p*-value with it that should aid in determining statistical validity of each individual item.

International stages for education

In addressing the application of DCM for development, some of the advantages of DCM are the ability to provide aggregated performance data at various levels. Both Trends in International Mathematics and Science Study (TIMSS) and Progress in International Reading Literacy Study (PIRLS) provide hierarchical data (student, classroom, school, country), which suits the nested structure required for DCMs. International collaboration of these institutions have been making huge progress in the development of their methodologies, recently transitioning to digital formats, as well as adopting new methodologies that asses not only learner performance but compare the results to questionnaire responses about their school and home environments (International Association for the Evaluation of Educational Achievement [IEA], 2023; International Association for the Evaluation of Educational Achievement [IEA], 2026). These organizations also focus on collaboration to ensure that these assessments remain adaptable to the local contexts.

However, it has been noted that the number of measures that relate to OTL has declined since 1995 in which the most extensive collection of measures was aggregated by the TIMSS (Schmidt et al., 2015). For this reason, the emphasis on factors that address the role of curriculum coverage, teacher quality, and school environment fail to account for the curriculum in analysis of the OTL. Results, then represent other results that are well established, that parental involvement, access to resources, and school quality are related to SES. This means that SES is shaping both content coverage and student learning through different channels (Schmidt et al., 2015; International Association for the Evaluation of Educational Achievement [IEA], 2023: International Association for the Evaluation of Educational Achievement [IEA], 2026).

Schmidt and McKnight (2011) outline the dangers of neglecting these other facets, warning against attitudes that conceive of math achievement as a product of cognitive ability due to SES and not OTL, that means equal curriculum would not solve these discrepancies in performance. They were able to analyze data on OTL and recognized that although school and SES do matter, the effects did not occlude the additional impact of curriculum. By comparing US school systems to Japanese school systems, they were able to demonstrate various aspects of content coverage do in fact converge to produce higher achievement. The authors then go on to make a strong case for the fact that these factors are further exacerbated by SES and data accrued at district level which sets guidelines for curriculum.

The Programme for International Student Assessment (PISA) is an international framework for, math, reading, and science competencies that aims to inform educational policy and practice. Recent comparison between international and US standards (PISA & NAEP) found that trends in significance were comparable internationally to those in the US suggesting that failures to produce statistical validity might include an increased sample size and refinement of evaluation criteria (Mazzeo and Von Davier, 2009). Other research has focused on the strengths of the PISA framework in integrating Collaborative Problem Solving () and the advancements and collaborations in making these programs culturally sensitive (He et al., 2017). However, the considerations of large scale multidimensional models leave large organizations like PISA, TIMSS and PIRLS looking for avenues to refine their methods.

Given PISA's focus on assessing education across diverse cultural contexts, diagnostic classification models (DCMs) present an innovative method to enhance the assessment of student skills. By providing detailed, culturally sensitive profiles of student mastery, DCMs could address the limitations of one-size-fits-all models, offering a more nuanced approach to international educational evaluation. Recently, a review by Ravand and Baghaei (2019) echoed this sentiment by analyzing the use of DCM as something that has rarely been done in the original intention. That is, DCM has rarely been used to develop educational assessments right from the start.

This neglect of DCM for their intended use is likely due to the consideration of costs, and the current practices of adapting educational materials to other cultural contexts. Reasons for the inadequacy of foreign materials is discussed in the following section. The need to consider OTL focuses primarily on curriculum, multidimensional models like DCM promise to aid development not only by refining methods for comparative international study but that data collected could be used to produce individualized learning profiles that can directly identify specific areas in which a learner is struggling. The additional advantage of DCM is that this data can then further be aggregated at various levels for use by international organizations like PISA that focus on educational policy and practice.

Adapting to a global stage

Recently, Jukes et al. (2021) identified factors failing in educational reform, arguing these have largely failed due to the lack of consideration of cultural context. It has also previously been suggested that contemporary issues of development should be considered in a historical context that acknowledges previous efforts and consequences of global development policy (Sakata et al., 2021; Lawn et al., 2008). For example, educational reform in Sub Saharan Africa, should be considered within the context of the neoliberal policy that has affected the development of LMICs (Jukes et al., 2021) These would include the problems in global development considered a product of neoliberal policy that has been responsible for the misallocation of funding that has largely produced negative results in other sectors, diverting funds from education (Lawn et al., 2008). For this reason, it has been important to identify how specific cultural behaviors and beliefs interact with teaching practices to identify compatible teaching practices, and to consider how external intervention will affect a country's autonomy.

Jukes et al. (2021) also noted previous work in development and adaptation of pedagogy focused on teaching practices in primarily WEIRD countries and cautioned against the use of these in LMICs

(Jukes et al., 2021). Their work was based on previous iterations of cultural frameworks that have been proposed by Greenfield (2016); in which countries change predictably as they develop, conceiving of culture as a spectrum. Specifically becoming more individualistic, gender egalitarian societies, becoming more child-centered and valuing multiple perspectives (Greenfield, 2016). Competing claims, however, focus on the persistence of agricultural values in a community after industrialization, as proposed by Alberto Alesina in his work on the emergence of gender roles through plow use (Heine, 2010), which has generally remained an unaddressed possibility in the adaptation of pedagogy to other cultural contexts.

The framework proposed by Greenfield (2016) is based on analysis of trends in the last 100 years of urbanization in China and America and is a largely *post hoc* conjecture, which alone is not enough. The need for identification of specific cultural factors is what motivated a seminal study conducted by Robin Alexander (2001), to identify true universals in education. Jukes et al. (2021) had developed their review by addressing the replication and expansion of work done by Alexander and supported the work done by the Research Triangle Institute (RTI). He found similar factors like those proposed in Greenfield and Alexander. For example, children in Tanzania are not encouraged to speak in front of adults, fear embarrassment, value togetherness and cooperation, and concede to age-graded authority. These features are consistent with less industrialized collectivist values (Jukes et al., 2021). Based on this comparison it might hold true that, as cultures become more industrialized, they change in predictable ways (Greenfield, 2016).

Further research has focused on differences in autonomy and relatedness (Keller, 2016), which emerge before children begin school and serve as traits that are adaptive to their cultural environment. However, significant effort has not resulted in identifying factors that lead to the ontogeny of these differences. Yet, multidimensional models for confirmatory factor analysis like DCM have potential. Using models like these, researchers can expect to better understand relevant factors through psychometric testing as well as aid in developing educational materials.

Overly narrow use of methodology has resulted in weaknesses in research on cultural factors used to adapt pedagogy. With reliance on observational data and *post hoc* analysis of previous trends, predictions cannot be made on the way culture develops. As a result, indicators of different aspects of development require a multidimensional measure and further confirmatory factor analysis. The next step toward understanding development would be to use multidimensional models to support different developmental and socialization pathways in children. Keller (2016) also reminds us that culture is always changing, and more research is needed to identify critical factors. Ideally, these findings found consistencies that point to development on a spectrum like those previously proposed (Keller, 2016; Greenfield, 2016).

Finally, Jukes et al. (2021) concluded that the use of materials developed using WEIRD samples are likely inadequate for use in LMICs. Additionally, the current approach could be described as color blind or operating under the assumption that race and ethnicity do not directly impact behavior disregarding people's race. In contrast, a multicultural approach acknowledges differences and appreciates aspects of various cultures (Heine, 2010). For this reason, the implementation of multidimensional models and factor analysis are potentially necessary for future research to ensure the falsifiability of these claims.

Major challenges of education in the United States include the use of materials before sufficient evidence has been provided for their effectiveness (U. S. Department of Education, 2013). Therefore, it is likely that WEIRD materials are also unsuited for LMICs. In contrast, previous attempts at developing materials using multidimensional models have been promising. Although these have been criticized as having insufficient reliability and validity (Sessoms and Henson, 2018), the study in Australia recently succeeded at developing a multidimensional model that solved both issues of reliability and validity.

Hypothetical example

Gartland et al. (2022) outlined the development and validation of a multidimensional, culturally and socially inclusive questionnaire using a community based approach that included a diverse population in Australia. Specifically, this measure included aboriginal Australians and at-risk populations in its development. A major strength of the study was the use of confirmatory factor analysis (CFA) which is considered a gold standard for psychometric testing (an advantage of DCMs), and its success at documenting statistical significance and reliability (Garland et al., 2022). The authors identified three stages in development: generation of items and conceptual subscales, pilot tests, and a refined validation study. This framework, when applied to DCM, has been addressed previously as problems in identifying items and appropriate conceptual subscales for factor analysis (Bradshaw, 2016).

The inclusion of diverse populations in development in turn allowed identification of indicators of factors representative of a diverse population capable of being aggregated at various levels. As a result, Gartland et al. (2022) produced items with exceptional validity and reliability ($\alpha = 0.7–0.9$). The sample sizes used ($n = 489$ and $n = 1{,}114$) increased between pilot and validation studies and used exceptionally large samples when compared to the audit of educational research in the United States which found comparably low sample sizes (US Department of Education). Under these conditions the study was able to narrow the original 169 items and 19 subscales down to 43 items and 11 scales deemed to be most relevant and impactful (Gartland et al., 2022).

This example presents a feasible and realistic understanding of the size and development needed for DCMs to serve as a psychometric and criterion referenced test, or to use DCM as a hypothetical framework for global development. In this work, DCMs show promise to address many issues in the development of education in developing countries. Specifically, they promise early disability screening, testing, and the aggregation of testing data on a classroom, school, and regional scale (Sessoms and Henson, 2018; Bradshaw, 2016). The proposed numbers for development should indicate why development of DCM is not ready for classroom implementation, due to the large number or participants needed for development. However, the example above speaks to the importance of DCM as one of development for feasible multidimensional testing as the next step in criterion referenced measures. As well as aiding the development of educational

measures, which have previously been linked to improved life expectancy and growth in the health sector in LMICs (Popham and Husek, 1969; Lawn et al., 2008).

DCM appears to be a potential avenue for cost effective development and adaptation of educational materials. The development of multidimensional models can help identify culturally relevant educational factors of classroom behavior and develop educational testing items as indicators of various aspects of learner mastery. DCM is the next step in this development of educational measures and testing, given that developed measures using diverse samples in developing countries would hypothetically remain valid as populations become more urban and countries industrialize. The interpretation of the continued validity of the measure arrived at by the inclusion of diverse samples, means that even if evoked culture persists as the populations urbanize, the developed measure would include items already considered indicators of relevant factors. Based on the inclusion of populations already representative of these changes, and the proposition by Greendfield (2016) that populations change predictably when analyzing these factors, it is possible that once developed, DCMs could be used in a country for an extended period of time as well as remaining useful by aggregating data at various levels.

Discussion

After having reviewed the technical foundations, it should be clear why the comparison was drawn to the original development of CRMs to DCMs. As each developed as the aim of the other it is simpler to think of the work that needs to be done as analogous to previous efforts. DCMs promise an ideal assessment that was originally thought to be unattainable. DCMs are multidimensional models that require complex algorithms, longer tests, and larger samples. The development of validity, through a framework like Messick's, is also of importance because the original publication extended the conversation of validity to a more rigorous one that takes into account ethical considerations. The required expertise and statistical sophistication as well as identification of appropriate criterion measures, is far beyond the scope of this paper.

However, this publication primarily presents the promising aspects of development within DCMs: They are particularly suited to development on a large scale that addresses the needs of educational development on a global stage. They have promise for assessment that have high dimensionality under feasible testing conditions, to identify disability, or to identify skills in which a student may need remedial instruction.

This is suited to learning environments of LMIC that, at present, have little to no infrastructure to identify and diagnose students with specific disabilities. Additionally, they remain a valuable tool for the development of international assessment techniques both in the US and internationally, as they have the necessary data aggregation capabilities required for use in assessments like PISA. DCM, however, has failed to be implemented in the way that it was intended, it has not been used in development but instead has been primarily applied to retrofitting of other measures. For this reason DCM would benefit from the opportunity to develop at a global scale.

It was additionally proposed that the failure of DCM to be implemented in an international setting has been due to the current practice of adapting materials to other cultural contexts. This was argued to be insufficient but likely motivated by practical consideration of the cost of development. Because DCM is effective for use both at an individual level, as it is able to inform remedial instruction and provide early disability screening, and at a large scale aggregating data at various levels, it seems DCM is a cost effective way that LMIC can begin development of educational materials.

Conclusion

Since DCMs remain a recent development within the literature, this publication aims to identify a specific framework for development and application of DCMs in their most practical aspects. DCMs could, for example, be used in large scale assessment that could support early screening for disability, as well as identifying areas for which students have yet to achieve mastery. DCMs promise to address many necessary dimensions including assessment and early screening, informing both instruction and intervention. Additionally, DCMs have promise for development in low-income countries where implementation could address critical issues in development of education programs in countries with little to no infrastructure for early screening for disability.

The promise for development in education also has been demonstrated to have a positive impact on other domains in the health sector. For example, assessment of progress in primary health care in the last thirty years since the declaration of Alma-Ata, has found that commitment to health goals has not led to direct improvement in the health sector (Lawn et al., 2008). Although these findings have been in large part due to governmental and global organization; major problems with the Millenium Development Goals (MDGs) have presented themselves during accelerated scale ups in health. DCMs serve as an efficient intervention that can be implemented "top-down" and be used to improve health outcomes and should be able to remain viable in use during scale up (Lawn et al., 2008; Gartland et al., 2022).

Lawn et al. (2008) describes intersectoral cooperation that has lacked but supports that improvements in agricultural and educational sectors have an often greater impact in improving health outcomes and life expectancy. Because other sectors have contributed to public-health gains, a primary goal of this publication was to emphasize the potential usefulness of DCMs in addressing issues with large scale development of infrastructure, while also framing education as part of an iterative process of community involvement. This may be a necessary paradigm shift (Rifkin, 1996), as well as potential solution to problems with budgeting and data aggregation when addressing a "top-down" vs. "bottom-up" approach to educational development.

Author's note

In this publication both DCMs and education must be addressed as an issue of development; specifically they must be looked at through a lens that supports their importance for potential funding and to meet goals of development set out by Alma-ata, The World Health organization, as well as the United

Nations Development Goals, and Millennium Development Goals. Social determinants, bad governance, climate change, financial crisis, etc., are causes of ill health. DCMs promise to make progress in lessening inequity, improving health, and providing empowerment. Still, other efforts are needed to address social issues such as rural employment, food security, health care, social identification, and education.

Author contributions

MU-J: Conceptualization, Writing – original draft, Writing – review and editing.

Funding

The author declares that no financial support was received for the research and/or publication of this article.

Conflict of interest

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Generative AI statement

The author declares that no Generative AI was used in the creation of this manuscript.

Publisher's note

References

Abedi, J., and Herman, J. (2010). Assessing English language learners' opportunity to learn mathematics: Issues and limitations. *Teach. Coll. Rec.* 112, 723–746. doi: 10.1177/016146811011200301

Alexander, R. (2001). Border crossings: Towards a comparative pedagogy comparative education, Nov., 2001, Vol. 37, No. 4, Special Number (24): Comparative education for the twenty-first century: An international response. *Comp. Educ.* 37, 507–523.

Bradshaw, L. (2016). "Diagnostic classification models," in *The Wiley Handbook of Cognition and Assessment*, eds A. A. Rupp and J. P. Leighton (Hooken, NJ: Wiley), doi: 10.1002/9781118956588.ch13

Duff, H., Faerron Guzmán, C., Almada, A., Golden, C., and Myers, S. (2020). "Typhoid and torrents: The link between downstream health and upstream actions," in *Planetary health case studies: An anthology of solutions*, eds S. Myers and H. Frumkin (Island Press). doi: 10.5822/phanth9678_6

Gartland, D., Riggs, E., Giallo, R., Glover, K., Stowe, M., Mongta, S., et al. (2022). Development and validation of a multidimensional, culturally and socially inclusive Child Resilience Questionnaire (parent/caregiver report) to measure factors that support resilience: A community-based participatory research and psychometric testing study in Australia. *BMJ Open* 12:e061129. doi: 10.1136/bmjopen-2022-061129

Greenfield, P. M. (2016). Social change, cultural evolution, and human development. *Curr. Opin. Psychol.* 8, 84–92. doi: 10.1016/j.copsyc.2015.10.012

He, Q., von Davier, M., Greiff, S., Steinhauer, E. W., and Borysewicz, P. B. (2017). "Collaborative problem solving measures in the Programme for International Student Assessment (PISA)," in *Innovative Assessment of Collaboration*, eds A. A. von Davier, M. Zhu, and P. C. Kyllonen (Dordrecht: Springer), 95–111. doi: 10.1007/978-3-319-33261-1_7

Heine, S. J. (2010). "Cultural psychology," in *Handbook of Social Psychology*, 5th Edn, eds S. T. Fiske, D. T. Gilbert, and G. Lindzey (Hoboen, NJ: John Wiley & Sons, Inc), 1423–1464. doi: 10.1002/9780470561119.socpsy002037

International Association for the Evaluation of Educational Achievement [IEA] (2023). *TIMSS 2023 Assessment Framework*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

International Association for the Evaluation of Educational Achievement [IEA] (2026). *PIRLS 2026 Assessment Framework*. Chestnut Hill, MA: TIMSS & PIRLS International Study Center.

Jukes, M. C. H., Sitabkhan, Y., and Tibenda, J. J. (2021). *Adapting Pedagogy to Cultural Context*. Research Triangle Park, NC: RTI Press, doi: 10.3768/rtipress.2021.op.0070.2109

Keller, H. (2016). Psychological autonomy and hierarchical relatedness as organizers of developmental pathways. *Phil. Trans. R. Soc. B* 371:20150070. doi: 10.1098/rstb.2015.0070

Lawn, J. E., Rohde, J., Rifkin, S., Were, M., Paul, V. K., and Chopra, M. (2008). Alma-Ata 30 years on: revolutionary, relevant, and time to revitalise. *Lancet* 372, 917–927. doi: 10.1016/S0140-6736(08)61402-6

Leighton, J., and Gierl, M. (eds.) (2007). *Cognitive diagnostic assessment for education: Theory and applications*. Cambridge University Press.

Mazzeo,, J, and Von Davier, M. (2009). Review of the Programme for International Student Assessment (PISA) test design: Recommendations for fostering stability in assessment results. *Educ. Working Papers* 28, 23–24.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessment. *Educ. Research.* 23, 13–23. doi: 10.3102/0013189X023002013

Popham, W. J., and Husek, T. R. (1969). Implications of criterion-referenced measurement. *J. Educ. Meas.* 6, 1–9. doi: 10.1111/j.1745-3984.1969.tb00654.x

Ravand, H., and Baghaei, P. (2019). Diagnostic classification models: Recent developments, practical issues, and prospects. *Int. J. Test.* 20, 24–56. doi: 10.1080/15305058.2019.1588278

Rifkin, S. B. (1996). Paradigms lost: Toward a new understanding of community participation in health programmes. *Acta Trop.* 61, 79–92. doi: 10.1016/0001-706x(95)00105-n

Sakata, N., Oketch, M., and Candappa, M. (2021). Pedagogy and history: Ujamaa and learner-centered pedagogy in Tanzania. *Comp. Educ. Rev.* 65, 56–75. doi: 10.1086/712052

Schmidt, W., and McKnight, C. (2011). *Chapter 9 Content Coverage Matters in Inequality for all: The challenge of unequal opportunity in American schools*. Teachers College Press.

Schmidt, W. H., and Burroughs, N. A. (2015). Puzzling out PISA: What can international comparisons tell us about american education? *Am. Educ.* 39, 24–31.

Schmidt, W. H., Burroughs, N. A., Zoido, P., and Houang, R. T. (2015). The role of schooling in perpetuating educational inequality: An international perspective. *Educ. Research.* 44, 371–386. doi: 10.3102/0013189X15603982

Sessoms, J., and Henson, R. A. (2018). Applications of diagnostic classification models: A literature review and critical commentary. *Measurement* 16, 1–17. doi: 10.1080/15366367.2018.1435104

Tindal, G. (2005). Curriculum-based measurement: A brief history of nearly everything from the 1970s to the present. *Assess. Effect. Intervent.* 30, 6–18. doi: 10.1155/2013/958530

U. S. Department of Education (2013). *Institute of Education Sciences, What Works Clearinghouse (2013, November). Beginning Reading Intervention Report: Reading Mastery*. Washington, DC: U. S. Department of Education

U. S. Department of Education (2023). *Institute of Education Sciences, ExcelinEd Policy tool Kit Comprehensive K-3 Early Literacy Policy*. Washington, DC: U. S. Department of Education