Communications for Statistical Applications and Methods
2024, Vol. 31, No. 5, 467–485

https://doi.org/10.29220/CSAM.2024.31.5.467
Print ISSN 2287-7843 / Online ISSN 2383-4757

The alignment between contextual and model generalization: An application with PISA 2015

Wan Ren[1,a], Wendy Chan[a]

[a]Graduate School of Education, University of Pennsylvania, USA

Abstract

Policymakers and educational researchers have grown increasingly interested in the extent to which study results generalize across different groups of students. Current generalization research in education has largely focused on the compositional similarity among students based on a set of observable characteristics. However, generalization is defined differently across various disciplines. While the concept of compositional similarity is prominent in causal research, generalization among the statistical learning community refers to the extent to which a model produces accurate predictions across samples and populations. The purpose of this study is to assess the extent to which concepts related to contextual generalization (based on compositional similarity) are associated with the ideas related to model generalization (based on accuracy of prediction). We use observational data from the Programme for International Student Assessment (PISA) 2015 wave as a case study to examine the conditions under which contextual and model generalization are aligned. We assess the correlations between statistical measures that quantify compositional similarity and prediction accuracy and discuss the implications for generalization research.

Keywords: internal validity, external validity, generalizability, model predictions, covariates, PISA

1. Introduction

With the proliferation of experimental studies in education, policymakers and practitioners have grown increasingly interested in understanding the extent to which the results of a study apply or generalize to a population or target group of students (Tipton and Olsen, 2018). When the goal is to generalize the results from a sample to a larger population of students, the so-called "narrow to broad" perspective (Shadish *et al.*, 2002), probability or random sampling is the most powerful tool to facilitate the generalizability of study results (O'Muircheartaigh and Hedges, 2014). Additionally, if treatment is randomly assigned within the study, random sampling and random assignment of treatment collectively strengthen the internal and external validity of the study results (Shadish *et al.*, 2002). Alternatively, when the goal is to generalize across samples, or across units of a "similar level" (Shadish *et al.*, 2002), random assignment or selection is the strongest tool to facilitate this type of generalization. However, in both cases, random sampling and selection are rare in educational studies (Olsen *et al.*, 2013) and as a result, much of the current generalization research has focused on methods to improve generalizations, primarily using propensity score methods (Stuart *et al.*, 2011; Tipton, 2013a). For studies that focus on generalizations from a sample to a population, the propensity score estimates the conditional probability of selecting into the sample as a function of a set of observable covariates.

[1]Corresponding author: Graduate School of Education, University of Pennsylvania, 3700 Walnut St, Philadelphia, PA 19104. Email: renwanmichelle@gmail.com

Published 30 September 2024 / journal homepage: http://csam.or.kr

Importantly, the covariates used in propensity score methods are assumed to moderate both the treatment effect and the sample selection process (Kern *et al.*, 2016). Similarly, when generalizing across samples, the covariates used in the estimation of propensity scores should affect both the treatment effect and the selection process. If certain assumptions are met, propensity score methods produce bias-reduced estimates of treatment impacts that can be used to generalize across student populations (Tipton and Olsen, 2018).

Propensity score approaches have made an important contribution in improving the contextual generalizability of experimental results. Specifically, in narrow to broad generalizations, propensity scores are primarily used to match or reweight students in a sample with those in the population with the understanding that when the two groups are compositionally similar (based on the observable treatment effect moderators), treatment effect estimates will be generalizable. A similar assessment is made when generalizing across samples. However, generalization based on contextual (compositional) similarities is not the only type of generalization of interest in studies. The notion of generalization has different meanings in various research fields. For example, in the statistical learning context, generalizability refers to the extent to which a statistical model can be used to make out-of-sample predictions (Linden and Yarnold, 2016a, 2016b; Cai *et al.*, 2020). In this context, generalizability is associated with the predictive accuracy of a model across various samples of students. As another example, generalization in studies with international large-scale datasets refers to the extent to which the coefficients of predictive models are all positive or all negative across countries (Marsh *et al.*, 2015; Marsh, 2016).

Given the differences between how generalizability is defined, an important question is whether different notions of generalization in the contextual and model framework are associated with each other. The purpose of this study is to explore the connection between measures of contextual generalization (based on compositional similarity) and measures of model generalization, the latter of which refers to statistical models and predictive accuracy across various samples. This exploration is motivated by two factors. One, contextual generalization and propensity score methods are often used in studies where causal inference and the estimation of the causal impact of an intervention is the goal. In contrast, model generalization is related to prediction and the extent to which a fitted model can be applied across different samples of students or units to produce accurate predictions. Because causal inference and prediction are generally not discussed in the same context, a goal of this study is to assess whether a relationship exists between the two frameworks within the context of generalizability. Two, while both contextual and model generalization are important strands of generalization research, the concepts related to the former are often not present in studies that focus on the latter; namely, studies that center around the generalizability of models generally do not include discussions of contextual generalization. As a result, this study aims to bridge the two frameworks of generalization by providing the first empirical evidence of the association (or absence of one) between measures of contextual and model generalization. If a relationship exists between the two frameworks, this has important implications for generalization research as it would allow researchers to identify the conditions under which the results of a study apply across various groups of students.

The article is organized as follows. First, we provide a brief review of the concepts related to contextual and model generalization. In the same section, we also discuss the assumptions and existing measures used to assess each type of generalization (contextual and model) and highlight some important differences between the goals of each framework. Then, we explore the relationship between contextual and model generalization using a case study based on a subset of data from the Programme for International Student Assessment (PISA) 2015 study. PISA is a widely used, publicly available educational data set that assesses various aspects of 15-year-olds' educational experiences across a sample of countries. Because the empirical example is based on observational data from PISA, our study differs from prior generalization research in an important way. We deviate from prior generalization studies by using an observational data set, in place of an experimental study, as the empirical example. This was done because PISA provides a rich source of covariate information that can be used to address a variety of research questions and because PISA is publicly available, the results of our study may motivate future work on the connection between contextual and model generalization. However, while our empirical example is based on an international large-scale dataset, the relationship between contextual and model generalization has broader implications for educational research. In the final section, we conclude with a discussion of these implications in relation to our findings based on the PISA data.

2. Review of contextual generalizability

In this section, we provide a brief review of contextual generalization. Although the sample to population (narrow to broad) framework is the common one of interest in education studies, we center our review on the sample to sample (units of similar levels) perspective to be consistent with the PISA empirical example. However, extensions of the assumptions and approaches can be made to the narrow to broad perspective. Our review is based on the causal inference model used to estimate treatment effects in non-experimental studies (Rosenbaum and Rubin, 1983). In non-experimental or observational studies, the goal is to estimate an average treatment effect in the presence of self-selection (non-random selection) into the treatment and comparison groups. In our context, we refer to membership in a specific country (in PISA) as the equivalent of a "treatment" and assess the contextual generalizability across individuals in the treatment and comparison countries (groups).

Given this setup, consider a population P in which a sample S of n units are selected. Throughout, the units in our study refer to students, but they may also represent schools, communities, or aggregates of individuals. For each unit i, we define the binary treatment variable T_i, which is equal to one if a student is a member of a given country and zero otherwise. We also observe a p-dimensional column vector of observed pretreatment covariates $\mathbf{X}_i$, whose support is denoted by $\mathcal{X}$. Let τ denote the parameter of interest in the generalization study, where τ can refer to an average treatment effect (Rubin, 1974) based on an observable outcome Y or it can refer to a model parameter. Examples of outcomes include assessment scores and graduation rates. In the sample to sample framework, when units are randomly assigned to samples, and certain assumptions hold (Tipton, 2013a; Imai and Ratkovic, 2014), several methods lead to unbiased and generalizable parameter estimates $\hat{\tau}$.

2.1. Propensity score methods

The challenge in many generalization studies in education is that the selection into samples is not random. This creates a selection bias in parameter estimates since units that select into a sample may be compositionally different from units that select into different samples. In this case, to make valid generalizations, model-based methods are needed (Olsen *et al.*, 2013; O'Muircheartaigh and Hedges, 2014). Propensity score methods are a common approach to improve generalizations from nonrandom samples (Stuart *et al.*, 2011). Propensity score methods were originally developed to address the bias in treatment effect estimates from observational studies when the units were not randomly assigned to treatments (Rosenbaum and Rubin, 1983). Propensity scores are the conditional probabilities of selection into a sample, conditional on the values of a set of observable covariates. Once estimated, propensity scores can be used to match or reweight a sample to be compositionally similar (on the observable covariates) to individuals in the comparison sample (or group) (Stuart, 2010). In practice, to facilitate the most bias reduction, the covariates used for propensity scores in generalization studies are variables that moderate both the outcome of interest and selection.

Under the given framework, let $\pi(\mathbf{X})$ denote the true propensity score given by:

$$\pi(\mathbf{X}) = \Pr(T = 1 \mid \mathbf{X}). \tag{2.1}$$

Because propensity scores are a function of the covariates $\mathbf{X}$, they serve as unidimensional summaries of the covariate set. Importantly, propensity scores are balancing scores where matching by propensity scores is equivalent to matching by all the covariates $\mathbf{X}$ used in the propensity score model (Rosenbaum and Rubin, 1983). A common method to estimate the propensity scores is with logistic regression based on $\mathbf{X} = (X_1, X_2, \ldots, X_p)$:

$$\operatorname{logit}(\hat{\pi}(\mathbf{X})) = \log\left(\hat{\pi}\left(\mathbf{X}/1 - \hat{\pi}(\mathbf{X})\right)\right) = \alpha_0 + \alpha_1 X_1 + \cdots + \alpha_p X_p. \tag{2.2}$$

2.2. Assumptions for propensity scores

The validity of propensity score-based estimators depends on several assumptions. First, the Stable Unit Treatment Value Assumption (SUTVA) (Rubin, 1978, 1980, 1990; Tipton, 2013a) for the sample is required. Under this assumption, the outcomes of each student do not depend on membership in a specific sample and there is no interference among students across samples. Second, the treatment assignment must be strongly ignorable (Rosenbaum and Rubin, 1983). Under this assumption, the parameter τ is conditionally independent of the treatment indicator T, given the propensity scores. This assumption implies that the covariates $\mathbf{X}$ explain all the variation in parameter estimates. Additionally, strong ignorability of treatment requires that the true propensity scores be bounded away from zero and one.

3. Generalizability and propensity scores

3.1. Contextual generalization

If the assumptions for propensity score methods are met, they can be used to produce bias-reduced estimates of the parameter of interest. In addition to their use in estimation, matching, and reweighting, propensity scores have also been used to assess generalization among students of different populations (Tipton, 2014; Chan, 2017; Tipton and Olsen, 2018). In this framework, contextual generalization is assessed by examining the extent of compositional similarity among groups of students or schools based on the propensity score distributions. Note that because propensity scores are univariate summaries of multiple covariates, similarity on the propensity scores is equivalent to similarity in the distributions of covariates that are used to estimate the propensity scores (Rosenbaum and Rubin, 1983). Formally, compositional similarity and contextual generalization is equivalent to the condition:

$$\pi(\mathbf{X}) \perp T. \tag{3.1}$$

Equation (3.1) states that, under contextual generalization, the true propensity scores are independent of the treatment indicator. Note that this condition is also equivalent to:

$$(\pi(\mathbf{X}) \mid T = 1) \overset{d}{=} (\pi(\mathbf{X}) \mid T = 0), \tag{3.2}$$

where the distributions of propensity scores are perfectly balanced among the treatment and control samples.

In practice, several statistical measures are available to quantify the extent of contextual generalization or compositional similarity. Among these, the generalizability index β, or B-index, represents the level of "distributional similarity between the propensity scores" of two samples (Tipton, 2014). The B-index has two important properties. First, it is bounded between zero and one where values close to one represent stronger compositional similarity. Note that the B-index is equal to one under condition (3.1) where the propensity scores are independent of T. Second, the B-index does not require any distributional assumptions, so its validity does not depend on the specific assumptions related to the sample data. These two properties are important as they make the B-index a statistic that is easy to interpret and flexible to use. Tipton (2014) suggested the following cutoffs to assess the extent of contextual generalizability: (i) a B-index between 0.9 to 1 suggests that the samples are very similar on the given set of covariates, (ii) a B-index between 0.8 to 0.9 indicates that while there is some similarity, some reweighting is needed to derive unbiased parameter estimates from one sample, (iii) a B-index between 0.5 to 0.8 implies that even with reweighting, there would still be bias and/or large inflation of standard errors, (iv) a B-index smaller than 0.5 suggests that the two samples are too different to warrant comparison.

Other measures used to quantify contextual generalization include standardized mean differences for individual covariates and for the propensity score logits (Stuart *et al.*, 2011) as well as measures of overlap in the propensity score distributions (Chan, 2021). In this study, we focus on the B-index alone as a measure of contextual generalization. We do so because the B-index offers both an interpretable and flexible measure to quantify the compositional similarity between two samples of units. However, because the alternative measures are based on the same quantities (namely, the propensity scores and the covariates used in the propensity score models), the assessments of contextual generalizability should be similar among all the measures.

3.2. Model generalization

Contextual generalization focuses on compositional similarity among samples based on the distributions of covariates. Model generalization focuses on the extent to which a predictive model fitted among units in one sample yields accurate predictions of an outcome for another sample. As a result, model generalization is assessed using measures of predictive accuracy between two samples; namely, the better the predictive accuracy, the more generalizable the model across various samples of students or schools. Formally, model generalization requires that:

$$\begin{pmatrix}\mathbf{X}\\ Y\end{pmatrix} \perp T. \tag{3.3}$$

In practice, to assess model generalization, k-fold cross-validation (Grandvalet and Bengio, 2004) is commonly used. This method repeatedly designates one sample as the training sample and the holdout $1/k$ data as a validation sample. k-fold cross-validation, also regarded as rotation estimation, splits the dataset into k mutually exclusive folds or subsets with approximately equal sample sizes per subset. The model is then trained iteratively by the training set (usually k–1 folds of the data) and then validated by the holdout fold. The cross-validation accuracy, in classification settings, can be computed based on the total amount of correct predictions divided by the sample size of the dataset. To measure the predictive accuracy, the Root-Mean-Square Error (RMSE) is used. The RMSE represents the square root of the squared differences between the predicted outcomes generated by the model and the observed values of the outcome values. The RMSE is given by the square root of the following:

$$\mathrm{E}(Y - E(Y \mid \mathbf{X}))^2, \tag{3.4}$$

where the expression is the expected value of the squared differences between the observed outcomes Y and the estimated outcomes $\mathrm{E}(Y|\mathbf{X})$ based on the data. If a predictive model is generalizable from the training dataset to the validation dataset, then the cross-validation RMSE will be relatively small.

3.3. Research questions

Given the differences between contextual and model generalizability, the current study addresses the following research questions.

1. Are measures of contextual and model generalizability associated with each other? That is, if two samples are contextually generalizable, does this imply that a model fit to one sample will generalize to units in the other sample (and have a small RMSE)?
2. Under what conditions, if any, does contextual generalization imply model generalization?

Collectively, the two research questions inform practitioners and policymakers of the conditions under which two frameworks for generalization are aligned with each other. If the compositional similarity between two groups is associated with accuracy in predictions, this can have implications for the choice of analytic approach to use when generalizing parameter estimates across populations.

4. Case study: PISA 2015

To explore the relationship between contextual and model generalization, we use data from the Programme for International Student Assessment (PISA) 2015 study. PISA is a widely used publicly available educational data set that assesses various aspects of 15-year-olds' education across a sample of countries. To date, seven waves of PISA data are publicly available, and the 2015 wave PISA data focuses on science as the learning outcome, which we also use in the current study. The original sample size of the PISA 2015 data is $N = 519,334$ students, sampled from 72 countries and economies.

Given the large number of variables and observations among all participating countries, we reduced the original size of the data by restricting the focus of our study to 30 countries and to students (in each country) who were categorized as low socioeconomic status (SES). Our decision to focus on a subset of the data was motivated by two factors. One, for our statistical analyses, it was essential to reduce the dimension of the observational data as including all variables significantly impacted the computational speed in the analyses. Second, we focused on low-SES students (rather than high-SES students) because researchers of education equity are strongly interested in understanding the factors that affect academic achievement among these student groups. Note that we also conducted the same analyses using the subpopulation of high SES students. The results were similar and can be provided upon request. Students were identified as low-SES using the PISA Economic, Social and Cultural Status (ESCS) index in the specified country. Using the subset of low-SES students among 30 countries, our final analytic sample comprised $N_S = 48,903$ students.

Table 1: Covariate description (PISA 2015 data)

Variable	Variable abbreviation	Variable description
Student-level Covariates [a]		
ESCS	ESCS	Index of economic, social and cultural status (a composite score built by the indicators parental education (PARED), highest parental occupation (HISEI), and home possessions (HOMEPOS) including books in the home via principal component analysis (PCA))

Mother_Edu	ST005Q01TA	What is the highest level of schooling completed by your mother?
		(1 = ISCED level 3A; 2 = ISCED level 3B, 3C;
		3 = ISCED level 2; 4 = ISCED level 1;
		5 = She did not complete ISCED level 1)

Father_Edu	ST007Q01TA	What is the highest level of schooling completed by your father?
		(1 = ISCED level 3A; 2 = ISCED level 3B, 3C;
		3 = ISCED level 2; 4 = ISCED level 1;
		5 = He did not complete ISCED level 1)

Vocational_Edu	ISCEDO	Programme orientation (ISCEDO) indicates whether the programme's curricular content was general, pre-vocational or vocational (1 = General; 2 = Pre-Vocational; 3 = Vocational; 4 = Modular) For the following item: (1 = No, never; 2 = Yes, once; 3 = Yes, twice or more)
Retention_PrimaryEdu	ST127Q01TA	Have you ever repeated a grade? At ISCED 1: primary education
Retention_LowSecEdu	ST127Q02TA	Have you ever repeated a grade? At ISCED 2: lower secondary education For the following items: How often do you do these things? (1 = Very often; 2 = Regularly; 3 = Sometimes; 4 = Never or hardly ever)
Freq_HaveSciBooks	ST146Q02TA	Borrow or buy books on broad science topics
Freq_GoSciClub	ST146Q05TA	Attend a science club
Freq_SimNaturalSciLab	ST146Q06NA	Simulate natural phenomena in computer programs virtual labs
Freq_SimTechSciLab	ST146Q07NA	Simulate technical processes in computer programs virtual labs
Freq_VisitEcologyWebPage	ST146Q08NA	Visit web sites of ecology organisations
Freq_FollowNewsBlog	ST146Q09NA	Follow news via blogs and microblogging **The following are numeric variables**
Test_Anxiety	ANXTEST	Personality: Test Anxiety (WLE [b]), derived from ST118 based on IRT scaling.
Enjoy_Teamwork	COOPERATE	Collaboration and teamwork dispositions: Enjoy cooperation (WLE), including answers to items ST082Q02NA, ST082Q03NA, ST082Q08NA, and ST082Q12NA.
Value_Teamwork	CPSVALUE	Collaboration and teamwork dispositions: Value cooperation (WLE), including answers to items ST082Q01NA, ST082Q09NA, ST082Q13NA and ST082Q14NA.
Achy_Motivat	MOTIVAT	Student Attitudes, Preferences and Self-related beliefs: Achieving motivation (WLE), derived from ST119 based on IRT scaling.
Sch_Belonging	BELONG	Subjective well-being: Sense of Belonging to School (WLE), derived from ST034 based on IRT scaling.
Num_SciClsAWeek	ST059Q03TA	Number of class periods required per week in science
Num_ClsAWeek	ST060Q01NA	In a normal, full week at school, how many class periods are you required to attend in total?
Min_MathLearnAWeek	MMINS	Learning time (minutes per week) - Mathematics
Min_LearnAWeek	TMINS	Learning time (minutes per week) - in total
School-level Covariates [a]		**The following are numeric variables**
Sch_Size	SCHSIZE	School Size (Sum)
Prop_ComputerInternet	RATCMP2	Proportion of available computers that are connected to the Internet
Staff_Short	STAFFSHORT	Shortage of educational staff (WLE)
Prop_CertTcher	PROATCE	Index proportion of all teachers fully certified
Prop_SciTertiaryGrad	PROSTMAS	Index proportion of science teachers with ISCED level 5A and a major in science
Num_TcherSch	TOTAT	Total number of all teachers at school
Num_SciTcherSch	TOTST	Total number of science teachers at school

a The covariates were selected using GBM and RF models.

b WLE refers to weighted likelihood estimates (Warm, 1989).

Table 2: Subsamples of Countries

Grouping criterion	Subsamples	Description	Countries
Geographic region	1	Central & Eastern Europe	Montenegro, Bulgaria, Turkey, Croatia, Czech Republic, Estonia, Lithuania, Latvia, Poland, Russian Federation
	2	Western Europe	Switzerland, Spain, Ireland, Iceland, Luxembourg, Portugal, Greece, Finland
	3	Asia	United Arab Emirates (UAE), B-S-J-G (China), Korea, Chinese Taipei, Hong Kong, Macao
	4	Americas	Costa Rica, Mexico, Colombia, Peru, Uruguay, United States
GDP per capita	1	GDP per capita $< 10,000$	Colombia, Peru, Montenegro, Bulgaria, B-S-J-G (China)*,Russian Federation, Mexico
	2	GDP per capita $< 15,000$	Turkey, Costa Rica, Croatia, Poland, Latvia, Lithuania
	3	GDP per capita $< 20,000$	Uruguay, Estonia, Czech Republic, Greece, Portugal
	4	GDP per capita $< 50,000$	Spain, Korea, United Arab Emirates, Hong Kong, Finland
	5	GDP per capita $> 50,000$	Iceland, United States, Ireland, Macao, Switzerland, Luxembourg
k-means clustering	1	Low ESCS but high science achievement on average	B-S-J-G (China), Macao, Hong Kong, Chinese Taipei, Costa Rica, Luxembourg, Portugal, Spain, Uruguay, Colombia
	2	Low ESCS and low science achievement on average	Bulgaria, Mexico, Montenegro, Peru, Russian Federation, Turkey, United Arab Emirates
	3	High ESCS and high science achievement on average	Croatia, Czech Republic, Estonia, Finland, Greece, Iceland, Ireland, Korea, Latvia, Lithuania, Poland, Switzerland, United States

ESCS refers to the PISA economic, social, and cultural status index. Because of data limitations, the GDP per capita for China was used to reflect the GDP per capita in the four provinces (B-S-J-G: Beijing, Shanghai, Jiangsu, Guangdong) in China, but the GDP per capita for the four richer provinces in China could be higher than the GDP per capita at the national level.

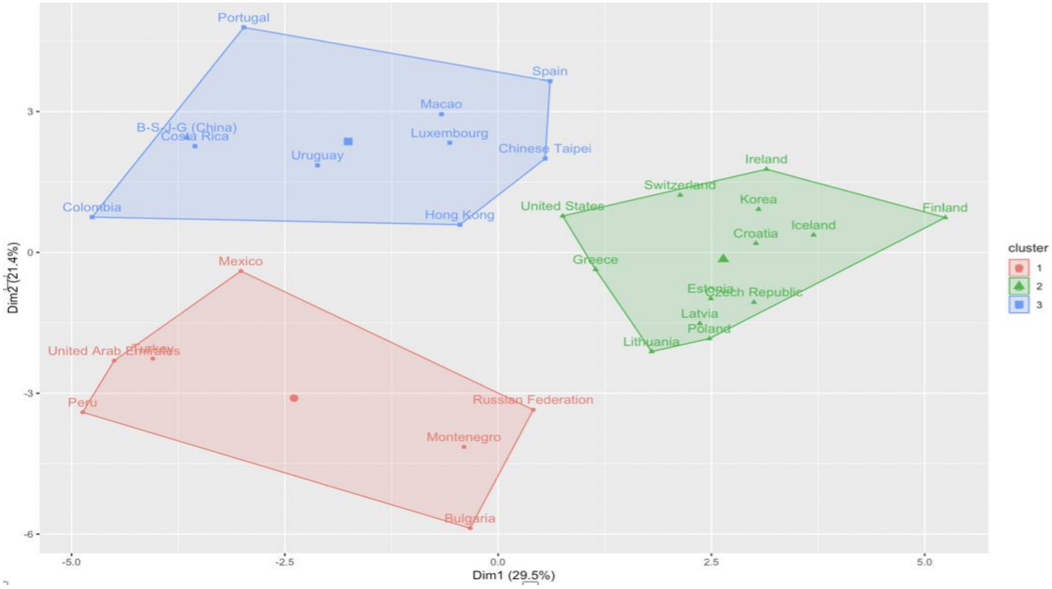
5. Data preprocessing

Preliminary analyses revealed that most of the PISA covariates had missing data. However, among the variables that were used in our statistical analyses, most (90%) had rates of missingness of less than 10%. Because the identified variables were crucial for the generalization analyses, we performed missing data imputation using Multiple Imputation by Chained Equations (MICE) (Van Buuren and Groothuis-Oudshoom, 2010). MICE was chosen for its flexibility in dealing with different types of variables and it was used for covariates with missing data for each country and economy in the analytic sample. Note that alternative approaches to MICE for imputation can also be used, but these methods may depend on different assumptions (Murray, 2018).

5.1. Variable selection

The PISA 2015 study included a rich set of covariates for each participating country and economy. One of the key assumptions required to facilitate generalizations (of any form) is that the covariates be moderators of the outcome. As a preliminary analysis, we identified the variables that were significant predictors (as a result, potential moderators) of the outcome of science achievement. We used random forest and gradient boosting methods (Freund and Schapire, 1996; Breiman, 2001) to identify the covariates from the original PISA set that included hundreds of variables at the student, teacher, and school level. Random forest (RF) is an ensemble machine learning approach that uses classification and regression trees (CART) (Breiman, 2001) to identify a set of relevant predictors. Similarly, gradient boosting methods (GBM) (Friedman, 2001) use CARTs to build prediction models to rank the covariates based on the accuracy of predictions. Both methods were used as they can accommodate high dimensional data and facilitate model interpretability (Guelman, 2012; Natekin and Knoll, 2013;

Figure 1. Grouping of countries/economies based on *k*-means clustering using 28 covariates



Note: For this plot, all variables are scaled at the country level. Since more than two dimensions (i.e., 28 covariates) were used to generate the three *k*-means clusters, principal component analysis (PCA) was performed to visualize the clusters (note that the data is standardized before PCA). The axes refer to the first two principal components which explain the majority of the variance in the data (x-axis explaining 29.5% of the variance, y-axis explaining 21.4% of the variance).

Figure 1: *Grouping of countries/economies based on k-means clustering using 28 covariates.*

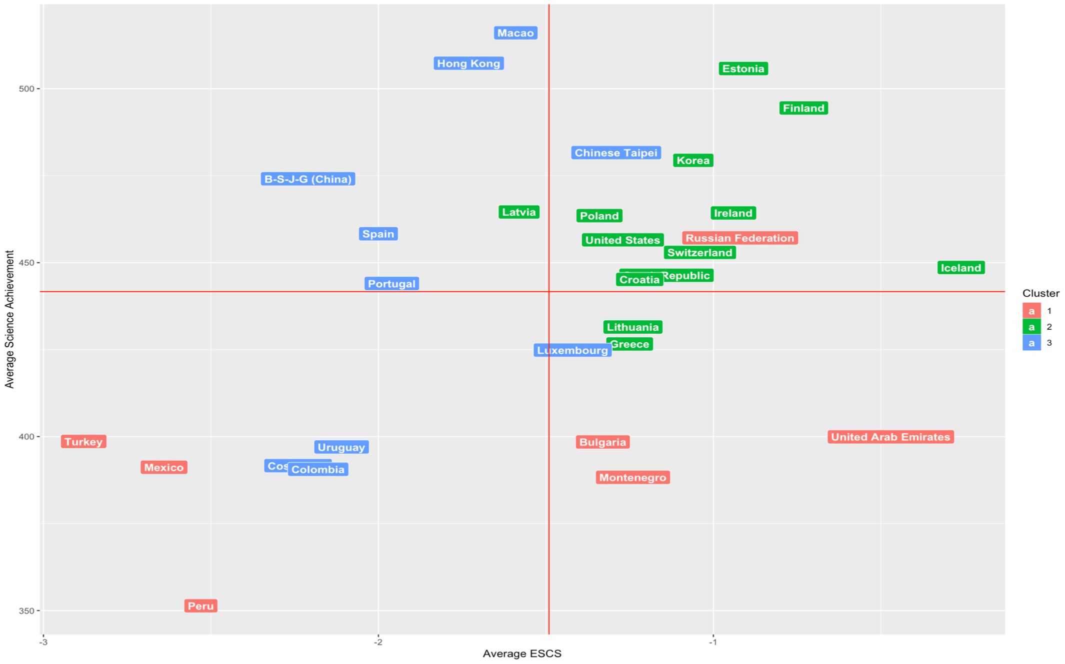
Taieb and Hyndman, 2014; Zhang and Haghani, 2015). Using the combination of RF and GBM, we identified 28 covariates that served as potential outcome moderators. Of these, 21 were at the student level and 7 were at the school level. A description of the variables can be found in Table 1. In the next section, we describe how the covariates were used to construct subsamples of the countries and compute statistics for contextual generalization.

5.2. Subsamples of countries

Since the current study involves multiple countries, we organized our analysis and discussion of contextual and model generalization among three groups based on: (i) geographic region, (ii) GDP per capita and (iii) *k*-means clustering. Because the variability among countries can affect the assessment of generalization and pose a challenge in identifying trends, the purpose of the groups is to compare measures of generalization among smaller (and potentially more homogenous) subsets of countries. Note that our criteria for constructing the groups are a subset of possible approaches. Alternative groups can be based on similarities in education systems and population size.

Table 2 lists the three main groups of the 30 countries and economies based on the three criteria. Within each group, we identified multiple subsamples of countries. For the subsample based on geographic region (first row), we used a coarse grouping method based on the continents, which created four main subsamples of countries. For the second approach based on GDP per capita (second row), we created five subsamples based on ranges that quantified the wealth of the countries. Note that

Figure 2. Description of clusters from *k*-means clustering subgroup



Note: ESCS refers to the Economic, Social and Cultural Status. In this plot, ESCS and science achievement are based on the original scale.

Figure 2: *Description of clusters from k-means clustering subgroup.*

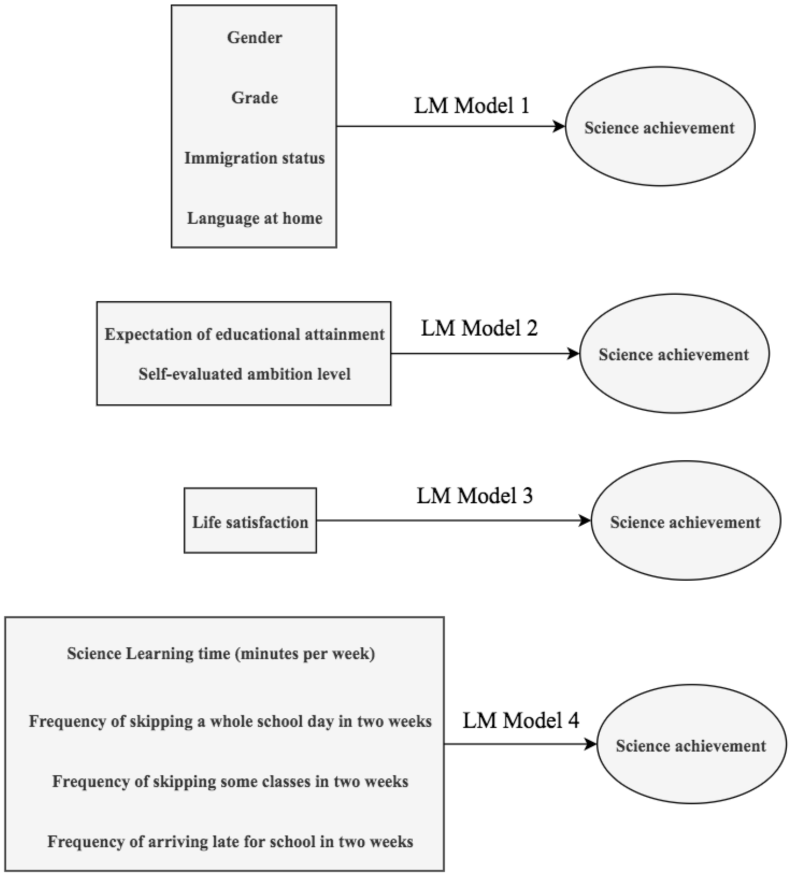
the 2015 GDP per capita used in this approach are in current US dollars from the World Bank, updated as of October 2019 (World Bank, 2019). The third grouping method was based on the statistical method of *k*-means clustering. The *k*-means clustering method is an unsupervised learning technique used to group a set of observations based on a pre-specified distance metric (Everitt and Hothorn, 2011). Given a set of *p* continuous variables, a common metric used in *k*-means is the weighted Euclidean distance given by:

$$d^{e}_{i,i'} = \sqrt{\sum_{h=1}^{p} w_h \left(X_{ih} - X_{i'h}\right)^2}, \tag{5.1}$$

where a $X_{ih}, X_{i'h}$ is the h^{th} covariate for units i and i' and w_h is the weight assigned to the covariate. If $w_h = 1$, the distance metric in (5.1) is the Euclidean distance. If the covariate set includes both categorical and continuous variables, an alternative distance metric is based on the Gower (1971) measure of similarity:

$$d^{g}_{ii'} = \sum_{h=1}^{p} w_{ii'h} d_{ii'h} / \sum_{h=1}^{p} w_{ii'h}, \tag{5.2}$$

Figure 3. Types of linear outcome models used for PISA 2015 data


Gender
Grade
Immigration status
Language at home
LM Model 1
Science achievement
Expectation of educational attainment
Self-evaluated ambition level
LM Model 2
Science achievement
Life satisfaction
LM Model 3
Science achievement
Science Learning time (minutes per week)
Frequency of skipping a whole school day in two weeks
Frequency of skipping some classes in two weeks
Frequency of arriving late for school in two weeks
LM Model 4
Science achievement

Note: LM refers to linear model.

Figure 3: *Types of linear outcome models used for PISA 2015 data.*

where $d_{ii'h}$ is the similarity between units i and i' for covariate X_h. For categorical variables, the similarity $d_i i' h = 1$ if the two units have the same value and 0 otherwise. For continuous variables, the similarity $d_{ii'h}$ is generally based on a standardized difference (Tipton, 2013b).

To perform k-means clustering, we used the covariates in Table 1 that were significantly predictive of the science achievement outcome. The final row of Table 2 and the plot in Figure 1 shows the three resulting clusters (subsamples) of countries under the k-means approach. To assess the types of countries that were placed into each cluster (subsample), Figure 2 shows the average ESCS and science achievement score by cluster. ESCS refers to the PISA economic, social, and cultural status index. Note that Figure 2 plots ESCS and science achievement as the latter was the main outcome of PISA 2015 and we were interested in trends in science achievement scores based on various ESCS values. Figure 2 suggests that most countries in the second cluster had high values of ESCS (high SES) and high science achievement scores while most countries in cluster 3 had low values of ESCS (low SES) but high science achievement scores. There was notably more variability in science and ESCS values among the countries in cluster 1. However, with the exception of the Russian Federation, most countries in cluster 1 had comparatively lower science achievement scores.

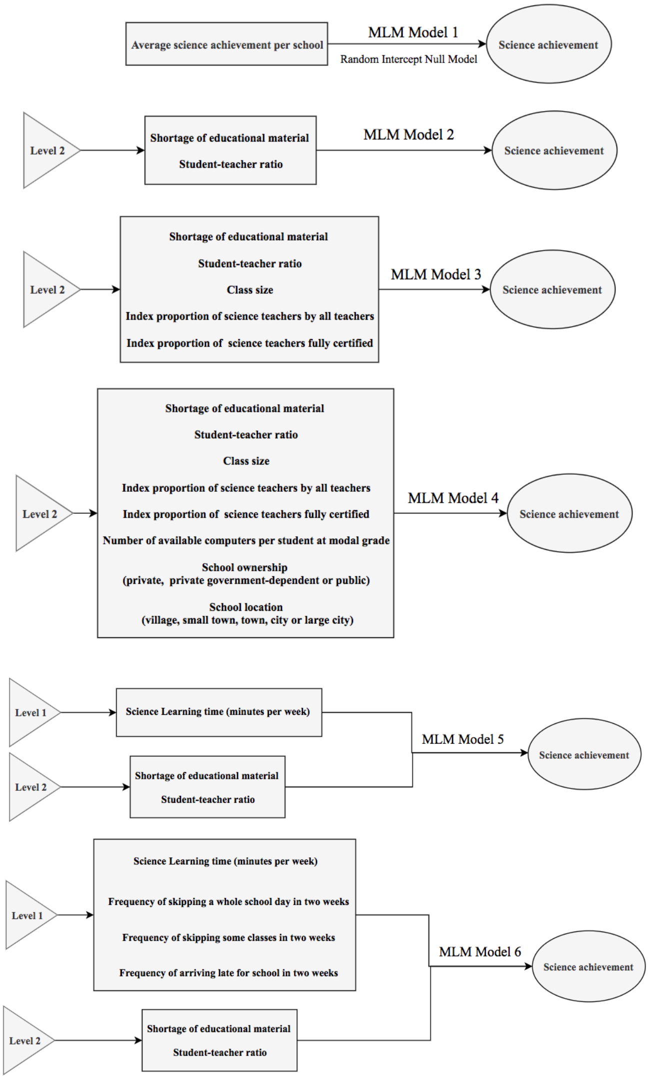
6. Assessing the relationship between contextual and model generalization with PISA

Using the 30 countries and three main groups (by geographic region, GDP, and *k*-means clustering), we addressed the first research question by examining the extent to which contextual generalizability was associated with model generalizability. Contextual generalizability refers to the extent to which countries in the same subsample (for example, Montenegro and Bulgaria in the geographic group) are compositionally similar based on a set of covariates. Thus, we consider the "treatment" to be membership in a country within a subsample and the comparison as membership in the other countries in the same subsample. We used the B-index (Tipton, 2014) to quantify the contextual generalizability and the values of the index are based on the same 28 covariates used in the *k*-means clustering (Table 1). To assess model generalization, we fit a series of 10 models, of which four were linear regression models and six were multilevel models using science achievement as the outcome. The multilevel models were specified based on students (Level 1) nested within schools (Level 2). Although these 10 models are only a subset of the possible outcome models, we included them as examples of the types of regression models that may be used. Additionally, we examined 10 different models to determine whether the measures of model generalization varied and depended on the specific approach used. Model generalization was assessed via *k*-fold cross-validation and the RMSE was used to quantify the predictive accuracy. Within each subsample, the RMSE was computed by training the model on one country and applying it to another country. Figures 3 and 4 provide visuals of the linear and multilevel models, respectively, and the covariates used to predict the given outcome. In both figures, the covariates are given in the left columns and the outcome, science achievement, is given in the far right.

We assessed the relationship between contextual and model generalization in the following way. We computed values of the B-index and RMSE using an individual to subsample of countries approach. Under this method, generalization is assessed between a single country and the remaining countries in the subsample. Thus, to assess contextual generalization, the B-index is computed by comparing the propensity score distributions between students of a single country with students of the remaining countries in the subsample. Similarly, with model generalization, the model is trained to all countries but one in a subsample and it is validated using the holdout country (Grandvalet and Bengio, 2004).

To determine whether there is a relationship between contextual and model generalization, we computed correlations between the B-indices and RMSEs. Because a higher B-index implies stronger contextual generalization while a lower RMSE is associated with stronger model generalization, we expect that a significant negative correlation between the two measures would suggest an alignment between the two types of generalization.

Figure 4. Types of multilevel outcome models used for the PISA 2015 data


Average science achievement per school
MLM Model 1
Random Intercept Null Model
Science achievement
Level 2
Shortage of educational material
Student-teacher ratio
MLM Model 2
Science achievement
Level 2
Shortage of educational material
Student-teacher ratio
Class size
Index proportion of science teachers by all teachers
Index proportion of science teachers fully certified
MLM Model 3
Science achievement
Level 2
Shortage of educational material
Student-teacher ratio
Class size
Index proportion of science teachers by all teachers
Index proportion of science teachers fully certified
Number of available computers per student at modal grade
School ownership (private, private government-dependent or public)
School location (village, small town, town, city or large city)
MLM Model 4
Science achievement
Level 1
Science Learning time (minutes per week)
Level 2
Shortage of educational material
Student-teacher ratio
MLM Model 5
Science achievement
Level 1
Science Learning time (minutes per week)
Frequency of skipping a whole school day in two weeks
Frequency of skipping some classes in two weeks
Frequency of arriving late for school in two weeks
Level 2
Shortage of educational material
Student-teacher ratio
MLM Model 6
Science achievement

Note: MLM refers to multilevel model.

Figure 4: *Types of multilevel outcome models used for the PISA 2015 data.*

7. Results

In this section, we discuss the results from the contextual and model generalization analyses. For the sake of parsimony, we summarize the overall findings from the 90 combinations of B-index and

Table 3: Average correlations between B-index and RMSE for linear regression models

	Geographic region	GDP per capita	$k-$means
All Models	−0.17	−0.16	0.24^{*}
LM Model 1	−0.19	−0.07	0.29
LM Model 2	−0.29	−0.20	0.06
LM Model 3	−0.08	−0.19	0.32
LM Model 4	−0.12	−0.18	0.34

Note: $^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$. The p values for all models have been corrected for multiple hypothesis testing.

RMSE computations (30 countries × 3 main groups). We center our discussion around three main trends: (i) values in the B-index for contextual generalization among the countries in each subsample, (ii) values of the RMSE for model generalization among the subsamples and, (iii) the correlation between values of the B-index and RMSE. In the following section, we begin with the values of the contextual and model generalization statistics for the linear regression case.

7.1. Contextual and model generalization - linear regression

The results suggest that the B-indices that describe contextual generalization varied among the three main groups of countries. When grouped by geographic region, the B-indices were highest (B-index ranging from 0.49 to 0.78) for the countries in Western Europe and in the Central and Eastern European subsample. Note, however, that while the values were considered highest for these two subsamples, the range from 0.49 to 0.78 suggests moderate compositional similarity at best (Tipton, 2014). Thus, when grouped by geographic region, most countries in the subsample were less compositionally similar based on the potential outcome moderators. When grouped by GDP per capita, there was no discernible pattern among the values in each subsample with the exception that countries with high GDP per capita generally had small values of the B-index. This suggests that there was less compositional similarity among countries grouped by GDP. In the k-means clustering group, the B-indices were highest (B-index > 0.50) among members in cluster 2, which included countries that had both high ESCS index values and high science achievement scores. Overall, geographic proximity accounted for some of the compositional similarity or contextual generalization among the sample of 30 countries, but there was not a single method of grouping that was associated with consistently high values in the B-index.

We analyzed the values of the RMSE for model generalization in the same framework. As mentioned, we fit 10 models on the subsamples, of which four were linear regression models. When grouped by geographic region, the patterns of cross-validation RMSEs were similar overall across the four models. Countries in the Americas had the lowest RMSE values across the four models and as a result, these countries had the strongest generalizability in model predictions. In contrast, countries in Asia had the lowest level of model generalization with the highest values in RMSE. When grouped by GDP per capita, the countries with the lowest GDP had the lowest RMSEs and hence, the strongest model generalizability. This is consistent with the trends in the k-means group in which countries that had the lowest ESCS index values and lowest science achievement scores had the highest extent of model generalization.

An important question is how the values of the B-index for contextual generalization compared with the values of the RMSE for model generalization when fitting predictive models based on linear regression. Table 3 provides the overall correlations between the B-indices and RMSE values in each of the three main groups. The first row provides the average correlation across all the models.

Table 4: Average correlations between B-index and RMSE for multilevel linear regression models

	Geographic region	GDP per capita	$k-$means
All Models	-0.23^{*}	-0.26^{**}	0.24^{**}
MLM Model 1	−0.18	−0.26	0.17
MLM Model 2	−0.20	−0.27	0.19
MLM Model 3	−0.31	−0.22	0.39
MLM Model 4	−0.29	−0.30	0.41
MLM Model 5	−0.19	−0.27	0.17
MLM Model 6	−0.21	−0.27	0.16

Note: $***p < 0.001$, $**p < 0.01$, $*p < 0.05$. The p values for all models have been corrected for multiple hypothesis testing.

These were computed by taking the average of each model by country combination across the subsamples within each group. Because we assessed the statistical significance of the correlations across multiple combinations, we applied a Benjamini-Hochberg correction to the p-values of the correlations (Benjamini and Hochberg, 1995). Table 3 illustrates that the correlations between the B-indices and RMSE were negative for some models and groups, while the values were positive in others. A negative correlation between the B-index and RMSE implies some degree of alignment between contextual and model generalization. The results from Table 3 suggest that among the given four models and three main groups, there was inconsistent evidence of alignment so that the presence of strong compositional similarity on covariates did not necessarily imply generalization in model predictions. Additionally, none of the correlations, with the exception of the average correlation under k-means, is statistically significant. Interestingly, the correlations for all models were negative in both the group based on geographic region and GDP per capita, but they were all positive under k-means. Thus, with respect to alignment between the two definitions of generalization, the extent of alignment in contextual and model generalization when comparing the individual to remaining countries in the subsamples is mixed.

7.2. Contextual and model generalization - multilevel linear regression

In this section, we assess whether the association between contextual and model generalization depends on the type of model. In addition to the four linear regression models, we fit six multilevel models in which student data was nested within schools in each participating country. The models are depicted in Figure 4. Like the linear regression case, we assessed contextual and model generalization by referring to individual countries in the subsamples as “treatment” and the remaining countries as the comparison.

We first analyzed the trends in B-index and RMSE values. Because the propensity scores were estimated using the same covariates and countries in each subsample, the B-index values were the same as in the linear regression cases. As a result, we focus on the trends among RMSE values in the subsamples. When grouped by geographic region, the average RMSE varied across the subsamples with some of the smallest RMSE values (implying high model generalization) seen in the Americas and Eastern European countries. This trend was consistent across the six multilevel models. When grouped by GDP per capita, there was notably more variability in RMSE values though the smallest values were largely seen in the low GDP subsample. Finally, when grouped by k-means, RMSE values also varied considerably across clusters, but countries in the high ESCS index and high science achievement scores had the lowest RMSEs overall, implying strongest model generalization.

Using the estimated B-index and RMSE values for the multilevel models, we estimated the correlations between the two measures. Table 4 shows the correlations for the six models and three main groups of countries. The first row provides the average correlation across the multilevel models, which were computed in a similar way as the values in the linear regression models. Interestingly, the correlations for all models were negative but insignificant when countries were grouped by geographic region and GDP per capita. In contrast, they were all positive (but still insignificant) when countries were grouped by *k*-means. Additionally, only the average correlations across all groups were significant. This suggests that the empirical evidence for alignment between contextual and model generalization, when generalizing in the individual country to remaining countries framework (within subsamples), is inconsistent for the multilevel models, which was also the case for the linear regression models.

8. Discussion

The goal of this application to the PISA 2015 data was to assess the extent to which concepts of contextual generalization aligned with concepts associated with model generalization. The analyses above sought to address this question by focusing on three main groups of countries from the data set. Within each group, we created subsamples of countries to base our generalization assessments. The results suggest two main implications. First, the results are somewhat mixed with respect to an alignment between contextual and model generalization. While the correlations between the B-index and RMSE were largely negative among the geographic region and GDP groups, few were significant and the correlations were all positive when countries were grouped by *k*-means. While not explored in the current study, generalization analyses suggest that the association between contextual and model generalization may be stronger with “local” generalizations between pairs of countries, rather than between a single country and a subsample of countries. A second implication is that while the type of model potentially plays a role, the magnitude of the correlations in the two cases (linear and multilevel) suggests that the conditions for alignment may not necessarily depend on it. This was seen in the comparable values of the correlations between the linear and multilevel cases in Tables 3 and 4. However, this should be interpreted with caution as the correlations given in the tables are based on averages and the cross-validation RMSEs depend on the specific training and validation data sets.

9. Conclusion

Generalization research continues to play an important role in informing educational research as policymakers and practitioners have grown increasingly interested in identifying approaches and best practices to support populations of students. Because generalization is defined and assessed differently across various disciplines, the current study sought to assess the extent to which two definitions of generalizability were aligned. This study focused on the relationship between contextual and model generalization that, at a broad level, relates to the relationship between causal inference and prediction. This connection, if present, is important in several ways. One, if compositional similarity was related to model and predictive accuracy, then precise estimates of parameters of interest can be derived for various populations of students or individuals. Two, if a relationship exists between contextual and model generalization, prediction can be used to derive parameter estimates for students or individuals not sampled in a study. This can have implications for ways to handle missing data. Finally, if contextual and model generalization are aligned, this relationship can potentially facilitate causal inference across multiple populations of students, particularly when a model that is used to derive precise causal estimates can be applied in various groups of individuals.

The results of our study suggest that contextual and model generalization are potentially aligned, but the results are inconsistent. Because the current study focused on generalizations from individual to subsamples of countries, the results suggest that any potential alignment between contextual and model generalization is difficult to observe in this perspective. This finding may not necessarily be surprising as model generalization is dependent on the specific samples used in the training set. However, the empirical evidence that an alignment between the two definitions of generalization potentially exists is useful. For researchers that use model-based methods for statistical inference, the results of this study can inform ways of deriving inference across different samples of students, particularly when data on potential outcome moderators is available. Future research should explore whether the empirical evidence of alignment is stronger in other perspectives of generalization, such as sample to population or population to sample (Shadish *et al.*, 2002). If found, the connection between contextual and model generalization may inform both generalization research on program impacts across local communities and the development of local policy based on the results of these generalization studies.

References

Benjamini Y and Hochberg Y (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing, *Journal of the Royal Statistical Society: Series B (Methodological)*, **57**, 289–300.

Breiman L (2001). Random forests, *Machine Learning*, **45**, 5–32.

Cai XL, Xie DJ, Madsen KH, Wang YM, Bogemann SA, Cheung EF, MOller A, and Chan RC (2020). Generalizability of machine learning for classification of schizophrenia based on resting-state functional MRI data, *Human Brain Mapping*, **41**, 172–184.

Chan W (2017). Partially identified treatment effects for generalizability, *Journal of Research on Educational Effectiveness*, **10**, 646–669.

Chan W (2021). The sensitivity of small area estimates under propensity score subclassification for generalization, *Journal of Research on Educational Effectiveness*, **15**, 178–215.

Everitt B and Hothorn T (2011). *An Introduction to Applied Multivariate Analysis with R.*, Springer Science & Business Media, New York.

Freund Y and Schapire RE (1996, July). Experimentswith a new boosting algorithm, *icml*, **96**, 148–156.

Friedman JH (2001). Greedy function approximation: A gradient boosting machine, *Annals of Statistics*, **29**, 1189–1232.

Gower JC (1971). A general coefficient of similarity and some of its properties, *Biometrics*, **27**, 857–871.

Grandvalet Y and Bengio Y (2004). Semi-supervised learning by entropy minimization, *Advances in Neural Information Processing Systems*, **17**, Available from: https://proceedings.neurips.cc/paper/2004/hash/96f2b50b5d3613adf9c27049b2a888c7-Abstract.html

Guelman L (2012). Gradient boosting trees for auto insurance loss cost modeling and prediction, *Expert Systems with Applications*, **39**, 3659–3667.

Imai K and Ratkovic M (2014). Covariate balancing propensity scores, *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **76**, 243–263.

Kern HL, Stuart EA, Hill J, and Green DP (2016). Assessing methods for generalizing experimental impact estimates to target populations, *Journal of Research on Educational Effectiveness*, **9**, 103–127.

Linden A and Yarnold PR (2016a). Using data mining techniques to characterize participation in observational studies, *Journal of Evaluation in Clinical Practice*, **22**, 839–847.

Linden A and Yarnold PR (2016b). Using machine learning to identify structural breaks in single-group interrupted time series designs, *Journal of Evaluation in Clinical Practice*, **22**, 855–859.

Marsh HW, Abduljabbar AS, Morin AJ, Parker P, Abdelfattah F, Nagengast B, and Abu-Hilal MM (2015). The big-fish-little-pond effect: Generalizability of social comparison processes over two age cohorts from Western, Asian, and Middle Eastern Islamic countries, *Journal of Educational Psychology*, **107**, 258–271.

Marsh HW (2016). Cross-cultural generalizability of year in school effects: Negative effects of acceleration and positive effects of retention on academic self-concept, *Journal of Educational Psychology*, **108**, 256–273.

Murray JS (2018). Multiple imputation: A review of practical and theoretical findings, *Statistical Science*, **33**, 142–159.

Natekin A and Kroll A (2013). Gradient boosting machines, a tutorial, *Frontiers in Neurorobotics*, **7**, 21.

O'Muircheartaigh C and Hedges LV (2014). Generalizing from unrepresentative experiments: A stratified propensity score approach, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **63**, 195–210.

Olsen RB, Orr LL, Bell SH, and Stuart EA (2013). External validity in policy evaluations that choose sites purposively, *Journal of Policy Analysis and Management*, **32**, 107–121.

Rosenbaum PR and Rubin DB (1983). The central role of the propensity score in observational studies for causal effects, *Biometrika*, **70**, 41–55.

Rubin DB (1974). Estimating causal effects of treatments in randomized and nonrandomized studies, *Journal of Educational Psychology*, **66**, 688–701.

Rubin DB (1978). Baysian inference for causal effects: The role of randomization, *The Annals of Statistics*, **6**, 34–58.

Rubin DB (1980). Randomization analysis of experimental data: The Fisher randomization test comment, *Journal of the American Statistical Association*, **75**, 591–593.

Rubin DB (1990). Formal mode of statistical inference for causal effects, *Journal of Statistical Planning and Inference*, **25**, 279–292.

Shadish WR, Cook TD, and Campbell DT (2002). *Experimental and Quasi-experimental Designs for Generalized Causal Inference*, Houghton Mifflin, Boston.

Stuart EA (2010). Matching methods for causal inference: A review and a look forward, *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, **25**, 1–21.

Stuart EA, Cole SR, Bradshaw CP, and Leaf PJ (2011). The use of propensity scores to assess the generalizability of results from randomized trials, *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **174**, 369–386.

Taieb SB and Hyndman RJ (2014). A gradient boosting approach to the Kaggle load forecasting competition, *International Journal of Forecasting*, **30**, 382–394.

Tipton E (2013a). Improving generalizations from experiments using propensity score subclassification: Assumptions, properties and contexts, *Journal of Educational and Behavioral Statistics*, **38**, 239–266.

Tipton E (2013b). Stratified sampling using cluster analysis: A sample selection strategy for improved generalizations from experiments, *Evaluation Review*, **37**, 109–139.

Tipton E (2014). How generalizable is your experiment? Comparing a sample and population through a generalizability index, *Journal of Educational and Behavioral Statistics*, **39**, 478–501.

Tipton E and Olsen RB (2018). A review of statistical methods for generalizing from evaluations of educational interventions, *Educational Researcher*, **47**, 516–524.

Van Buuren S and Groothuuis-Oudshoorn K (2010). Mice: Multivariate imputation by chained equations in R, *Journal of Statistical Software*, **45**, 1–68.

World Bank (2019). World Bank national accounts data: GDP per capita (current US$), Available from: https://data.worldbank.org/indicator/NY.GDP.PCAP.CD

Zhang Y and Haghani A (2015). A gradient boosting method to improve travel time prediction, *Transportation Research Part C: Emerging Technologies*, **58**, 308–324.

Received July 23, 2023; Revised January 10, 2024; Accepted April 23, 2024