
Xingyuan Gao, Jiangang Xia, Jianping Shen, and Xin Ma

Abstract: Successful school leadership is highly contextually dependent. However, few studies focused on the comparisons of school leadership across different countries. Even among the existing studies, comparisons tend to be conducted with the assumption that the underlying factorial structure of the construct is the same. In this study, school principal’s decision-making power in 12 decision-making areas from the PISA database are compared between China and the United States. The test of measurement invariance of the construct of principal decision-making power is conducted. Findings and implications are discussed.

Keywords: cross-cultural comparison, measurement invariance, PISA, principal power

INTRODUCTION

School leadership plays a vital role in school effectiveness (Fullan 2010; Leithwood, Harris, and Hopkins 2008). Over the past decades, scholars have developed a large body of knowledge to guide school leadership and principal’s work. Researchers, policymakers, and education practitioners share this knowledge in an unprecedented way at the age of accelerating globalization. Although this existing body of knowledge might greatly improve the research efficiency of leadership studies around the world, scholars have cautioned that production and flow of such knowledge remains one way—mainly from Western countries to non-Western countries (Walker and Dimmock 2002; Walker, Hu, and Qian 2012). A small body of studies has suggested that successful educational leadership is highly contextually dependent (e.g., Hallinger 2003; Moos, Krejsler, and Kofod 2008).

It is rather idealistic to think much existing principalship knowledge developed in a Western context is universal across countries and could be directly applied to any local school settings, given the disparity in the cultural, economic, and political background. Therefore, before adapting, transplanting, and even reproducing school leadership knowledge from Western to non-Western countries, sufficient work on cross-nation comparisons on various leadership elements, including job responsibilities, attitudes, experience, local practice, and metal models should be conducted first. In this
study, we aim to compare school principal’s power in 12 decision-making areas in China and the United States. Based on an international database, we examine the measurement invariance of the construct of principal decision-making power.

LITERATURE REVIEW

Comparison Studies in Leadership

Cross-nation comparison studies are critical to understand how different social contexts of various countries can form people’s perceptions and constructs. Especially in the field of organizational analyses, numerous studies had been developed to discover commonalities and differences in leadership across countries. For example, learning from the Global Leadership and Organizational Behavior Effectiveness (GLOBE) research program, Javidan et al. (2006) reported in-depth comparison of leadership styles across the world. Among others, they found that while Chinese and U.S. leaders both scored high in performance orientation, Chinese leaders were less future oriented and less assertive, but were more collectivist and more rules oriented. Dorfman et al. (1997) empirically compared Western and Asian countries’ view on effective leadership behaviors, they found that Westerners and Asians agreed on supportive, contingent reward, and charismatic, but disagreed on directive, participative, and contingent punishment. Silverthorne (2001) compared effective and ineffective leaders in the United States and Taiwan by measuring five major dimensions of personality, including neuroticism, extraversion, openness, agreeableness, and conscientiousness. Data from the U.S. sample supported the five-factor personality model, and indicated that effective leaders were more emotionally stable, more extraverted, more open to experience, more agreeable and more conscientious; the Taiwanese sample indicated there were no significant differences between effective and ineffective leaders in openness. The author conjectured that the big five model of personality was less useful in eastern cultures as Chinese population had different view on openness in both social and moral senses.

In the field of education, Bush and Jackson (2002) studied the preparation of school leadership from an international perspective. They reported England’s National College for School Leadership’s visits to 15 international leadership centers in nine countries. Across the 15 leadership centers, from Western to eastern, each nation and state used very different models to improve school leadership. Using Hong Kong’s “Leadership Training Programme for Principals” an example, Bush and Jackson suggested that it was vital to recognize that cloning good practice directly from a different political, social and professional context might not provide similar good results.

To sum up, this stream of literature implied that leadership was conceptualized and operationalized in different ways around the world. In the field of education, it is especially important to realize that school leaders from different backgrounds tend to have different ideas about effective leadership and different focuses of work.

Comparison of School Leadership between China and the United States

In the literature, relatively few studies have been developed to compare Chinese and U.S. principals. Chan and Du (2008) used a 30-item Likert-type scale questionnaire to compare
Chinese and U.S. principals in seven leadership areas, including character, professional knowledge, professional skill, administrative style, administrative duties, personnel management, and student affairs management. In the four areas that have significant differences (administrative style, administrative duties, personnel management, and student affairs management), U.S. principals consistently obtained higher scores. Su, Adams, and Mininberg (2003) compared Chinese principals with U.S. principals in their perception of importance of principals’ responsibilities. Among 20 areas surveyed, the top five areas selected by Chinese principals were supervise teaching and learning; provide leadership to all teachers and all staffs, legal and regulatory applications, motivate others; and “effectively distribute workload, and the top five areas selected by U.S. principals were provide leadership to all teachers and all staffs, motivate others; personal sensitivity; good judgment on job assignments; and effectively carry out tasks. While principals from both countries highlighted the importance of provide leadership to all teachers and all staffs and motivate others, principals seemed to value other three areas differently.

Besides direct comparisons mentioned previously, some studies tried to attribute the difference to external factors. Moos et al. (2008) pointed out that the Chinese school system, unlike most school systems in Western countries, was essentially top-down instead of decentralized. Chinese principals, therefore, worked as middle managers whose major tasks were to implement decisions and plans determined at the state-level government. Nevertheless, Moos, Krejsler, and Kofod noted that in Western countries with strong accountability demands (e.g., the United States), their principals shared some similarities with Chinese principals, including setting goals to follow external standards, and establishing rigid rules for internal managerial control. Although it was not an exclusive comparison between the United States and China, an international comparative study conducted by Lee and Hallinger (2012) revealed some interesting findings. Lee and Hallinger analyzed data on 5,927 principals in 34 societies drawn from the Progress in International Reading Literacy Study 2006, they found higher gross domestic product per capita was significantly related to the decrease in principals’ time allocation for instructional leadership.

School Principal’s Decision-making Power

In the practice of schooling, school principalship was reflected by how much power a principal has in various decision-making areas. From a micro perspective, principals’ decision-making power could be constrained by the teachers, school board, central office, and state department of education (Portin, Shen, and Williams 1998). From a macro perspective, decision-making power could be influenced by the country’s historical tradition and educational policy. As Bush (2009) noted, principals’ roles were expanding in schools around world.

In the past two decades, it appeared that two major themes have emerged to support the expansion of principal’s decision-making power—increasing accountability and decentralization. The first theme was more related to principal-teacher power relationship (internally) (Shen and Xia 2012). As principals were increasingly held accountable for educational quality, principals gained more responsibilities, influence, and power within schools (Fullan and Watson 2000; Portin and Shen 1998; Portin et al. 1998; Shen and Associates 2005; Wildy and Louden 2000; Witziers, Bosker, and Krüger 2003). The second theme was more in
relative to principal-district/state power relationship (externally)—in many Asian countries (e.g., China, Indonesia, Japan, Korea, Malaysian, Singapore) that had traditionally centralized, top-down educational system, researchers observed a trend of increasing school autonomy (Bjork 2006; Huang and Wiseman 2011). As state governments attempted to decentralize educational system, more decision-making power in various areas was released and distributed to the local school level.

**Principal’s decision-making power in China**

In China, laws and regulations were established to point out the direction for reform—decentralization (Huang and Wiseman 2011). In 1993, the Program for Education Reform and Development in China (1993 Program) stated that central government encouraged local government taking more responsibility in educational management and finances, and gradually establishing community sponsored schools. In 2001, the Basic Education Curriculum Reform Outline (trial) set an objective of “change from centralization in curriculum control to a joint effort between central government, local authorities, and schools to strengthen the relevance of the curriculum to local situations” (see Cui and Zhu 2014, 3). In 2010, the National Outline for Medium and Long-term Education Reform and Development (2010–2020) stated to build a modern education system that increases school autonomy (2010). In November 2013, the Decision on Some Major Issues Concerning Comprehensively Deepening the Reform highlighted that greater power was given to schools to make their own decisions about school affairs, and improve the internal governance structure of the schools (see CCCPC 2013). Through this gradual process of decentralization, decision-making power in many domains was slowly but continually distributed to local and school levels. Evidently, although Chinese principals still had limited decision-making power in fiscal and human resources, it appeared that they were gaining increasing decision-making power in curriculum development and school management.

**Principal’s decision-making power in America**

U.S. schools had a tradition of decentralization. Local schools were loosely coupled with the federal and state governments, which allowed them to have sufficient autonomy in decision making across various domains (Weick 1976). While the increasing accountability pressure from the No Child Left Behind Act of 2001 (NCLB) seemed to cutback some principal power in school control (Rouse et al. 2007), the more recent Every Student Succeeds Act 2015, which replaced NCLB, was believed to allow greater decision-making power at the local school level (Brown et al. 2016).

**CONCEPTUAL FRAMEWORK**

Our previous work, published in *Chinese Education and Society* in 2017, compared the school autonomy in China and the United States with data from the Programme for International Student Assessment (PISA) 2012 (Xia, Gao, and Shen 2017). In that article, we briefly reviewed the historic tracks of educational reform over the past decades in both
countries. Although it was found that U.S. schools were more autonomous across most indicators than Chinese schools, we argued that the two countries were undergoing educational change with opposite objectives—while the United States was trying to tighten the traditionally loose educational system, China realized the limitation of being too centralized.

The present study seeks to continue our line of work in producing new understanding of school leadership difference and similarity across the two countries. It partially replicated our last study on the principal power/responsibility. As our previous study found significant differences between U.S. and Chinese principals with raw data from PISA 2012, our goal of this study was to further explore whether the concepts/constructs being compared were interpreted in a conceptually similar manner by respondents (principals) from both countries. We attempted to establish measurement invariance by conducting multiple-group confirmatory factor analysis (CFA). Measurement invariance is a statistical technique, which was initially developed by Byrne, Shavelson, and Muthen (1989) to make meaningful cross-group comparisons. The goal is to ensure that the scale scores are measuring the same construct or trait in all groups. Horn and McArdle (1992) suggested that “if there is no evidence indicating presence or absence of measurement invariance—the usual case—or there is evidence that such invariance does not obtain, then the basis for drawing scientific inference is severely lacking: findings of differences between individuals and groups cannot be unambiguously interpreted” (117). For the present study, we conducted a series of tests to assess four levels of measurement invariance of principal decision-making power, including configural invariance, metric/weak invariance, scalar/strong invariance, and residual/strict invariance. As the minimum condition, strong measurement invariance needs to be held in order to make meaningful interpretation of measurement data. In case that full measurement invariance at each level could not be obtained, tests for partial measurement invariance would be conducted. Evidence of partial measurement invariance fulfills the requirement for comparing groups on a latent variable (see Byrne et al. 1989). As long as a majority of items on the latent variable have factor loadings invariant across group, we can make meaningful interpretation of measurement data (e.g., Reise, Widaman, and Pugh 1993).

Conceptually speaking, this study attempted to answer following research questions:

Research Question 1: From a measurement perspective, is the factorial structure of principal decision-making power comparable between American and Chinese principals? In other word, can the items from PISA 2015 measure the same construct/latent traits of power across the two countries?

Research Question 2: If so, who have more decision-making power across the given areas, American principals or Chinese principals?

METHOD

Data Source and Samples

The most recent 2015 PISA school data were used in the present study. We selected both school samples from mainland China and the United States. It was worth noting that only four provinces/cities of mainland China—Beijing, Shanghai, Jiangsu, and Guangdong
(B-S-J-G)—participated PISA in 2015. Therefore, the 2015 PISA data of mainland China were not nationally representative. We suggested that readers interpret and discuss results with caution.

The main purpose of the present study was cross-national comparison. Therefore, we applied senate weight created by the Organization for Economic Cooperation and Development (OECD). According to OECD (2014), senate weights were based on a constant scalar (in this case, 5,000), and allowed contributions from each of the countries to be equal regardless of the size of the population. As displayed in Table 1, 268 schools from China (B-S-J-G) and 174 schools from the United States participated in the PISA 2015 school survey. After applying the weighting factor, both countries had around 5,000 respondents.

Measures

The theoretical model of principal decision-making power was first tested against observed construct with CFA. A group of surveyed items were extracted from PISA 2015 school survey to perform CFA to test the fit of hypothesized one-factor model for both U.S. principals and Chinese principals, as well as measurement invariance across two groups.

In the 2015 PISA, sampled schools responded to question “who has a considerable responsibility” for 12 decision-making domains (coded as SC33 in the questionnaire). The 12 decision-making domains include (1) selecting teachers for hire, (2) firing teachers, (3) establishing teachers’ starting salaries, (4) determining teachers’ salary increases, (5) formulating the school budget, (6) deciding on budget allocations within the school, (7) establishing student disciplinary policies, (8) establishing student assessment policies, (9) approving students for admission to the school, (10) choosing which textbooks are used, (11) determining course content, and (12) deciding which courses are offered. Available options for these items include (a) principal, (b) teachers, (c) school governing board, (d) regional or local education authority, and (e) national education authority. PISA coded the above response into two categories, where 1 = ticked, indicating that the stakeholder has the power to make decision in the area, 2 = not ticked, indicating the stakeholder does not have the power to make decision in the area. Since the focus of the present study is principal decision-making power only, we only selected principal-related options, which resulted in 12 dichotomous variables for data analysis. Table 2 presented an initial comparison of principal power between China and the United States.

Table 2 presented principal’s power in 12 decision-making areas between China and the United States. The descriptive statistics showed that for all 12 areas, China had a relatively lower percentage of principals who had the power for decision-making. The chi-square tests

<table>
<thead>
<tr>
<th>TABLE 1</th>
<th>Sample sizes of sampled and weighted schools from China and the United States</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sampled</td>
</tr>
<tr>
<td>China</td>
<td>268</td>
</tr>
<tr>
<td>United States</td>
<td>174</td>
</tr>
</tbody>
</table>
further showed statistically significant differences between American and Chinese principals in the 12 areas, with small to medium effect sizes.

Statistical Analysis Procedure

The main purpose of the present study was to assess the differences between latent traits/constructs of decision-making power as held by Chinese and American principals. To begin with, single-group CFAs were conducted to determine the fit of the conceptualized one-factor model in two samples from China and the United States, respectively. Chi-square test, comparative fit index (CFI), Tucker-Lewis index (TLI), and the root mean square error of approximation (RMSEA) are used as a group of fit indices to determine model fit because CFI, TLI, and RMSEA have been proven to be less sensitive to large sample than the chi-square value (Fan, Thompson, and Wang 1999; Hu and Bentler 1999; Marsh, Balla, and McDonald 1988). The cutoff values of good model fit were close to 0.05, 0.96, 0.95, 0.06 for the chi-square, CFI, TLI, and RMSEA, respectively (Yu 2002). As the data from both groups showed acceptable good model fit, we proceeded with a multigroup CFA to test measurement invariance for the hypothesized model across two groups.

Multigroup CFA models for testing of measurement invariance were computed with Mplus 7.4 (Muthén and Muthén 2017). For the purposes of this paper measurement invariance was examined in a nested model approach. Four levels of measurement invariance—configural invariance, metric/weak invariance, scalar/strong invariance, and residual/strict invariance—were nested (by means of restriction) and assessed (Vandenberg and Lance 2000). Furthermore, when full measurement invariance could not hold, partial measurement invariance was still established by removing equality constraints based on largest modification index (chi-square >3.42, \( p < .05 \)). We referred to the results from chi-square difference tests for nested model comparisons. Weighted WLSMV was chosen as the estimator for

### TABLE 2
Initial comparison of principal power in 12 decision-making areas

<table>
<thead>
<tr>
<th>Decision-making areas</th>
<th>United States</th>
<th>China</th>
<th>Effect size (Cramer’s V)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M</td>
<td>SD</td>
<td>M</td>
</tr>
<tr>
<td>1. Hiring teachers</td>
<td>0.82</td>
<td>0.381</td>
<td>0.35</td>
</tr>
<tr>
<td>2. Firing teachers</td>
<td>0.72</td>
<td>0.45</td>
<td>0.22</td>
</tr>
<tr>
<td>3. Teacher start salaries</td>
<td>0.22</td>
<td>0.412</td>
<td>0.07</td>
</tr>
<tr>
<td>4. Teacher salary increase</td>
<td>0.20</td>
<td>0.402</td>
<td>0.09</td>
</tr>
<tr>
<td>5. Budget formulation</td>
<td>0.59</td>
<td>0.491</td>
<td>0.35</td>
</tr>
<tr>
<td>6. Budget allocation</td>
<td>0.73</td>
<td>0.444</td>
<td>0.45</td>
</tr>
<tr>
<td>7. Disciplinary policies</td>
<td>0.87</td>
<td>0.334</td>
<td>0.47</td>
</tr>
<tr>
<td>8. Assessment policies</td>
<td>0.74</td>
<td>0.437</td>
<td>0.48</td>
</tr>
<tr>
<td>9. Student admittance</td>
<td>0.67</td>
<td>0.472</td>
<td>0.30</td>
</tr>
<tr>
<td>10. Textbook selection</td>
<td>0.45</td>
<td>0.498</td>
<td>0.12</td>
</tr>
<tr>
<td>11. Course content</td>
<td>0.46</td>
<td>0.498</td>
<td>0.16</td>
</tr>
<tr>
<td>12. Courses offered</td>
<td>0.76</td>
<td>0.430</td>
<td>0.24</td>
</tr>
</tbody>
</table>

*Note. Significance level is set at \( \alpha = .05/12 = .00417 \).*
analysis given the categorical/binary outcomes (Beauducel and Herzberg 2006), and Theta parameterization was specified for the model identification purpose (Muthén and Asparouhov 2002).

FINDINGS

Single-group CFAs for Model Fit

As presented in Table 3, the fit indices overall suggested that the proposed one-factor model appeared to fit with principal data of both the United States and China. Although the chi-square value for the Chinese data was significant ($p < .01$), CFI, TLI, and RMSEA yielded acceptable model fit.

Multigroup CFA for Measurement Invariance

**Configural invariance**

In this model, we fixed the factor mean to 0 and the factor variance to 1, and constrained residual variance to 1 in both groups, and allowed factor loadings and thresholds to be estimated. As displayed in Table 4, the configural model suggested good model fit. Therefore, configural invariance held. We proceeded by testing more restrictive hypotheses, with the U.S. sample as the reference group.

**Metric invariance**

This model was used to investigate whether the same latent factor was measured across Chinese and U.S. principals. For identification, we fixed the factor means to 0 in both groups, set the factor variance to 1 in U.S. sample, but allowed the factor variance to be freely estimated for the Chinese sample. We constrained factor loadings to be equal across groups, and estimated all item thresholds. All residual variances were still constrained to 1 in both groups. As showed in Table 4, compared with the configural invariance model (functioned as the baseline model here), overall model fit of metric invariance model did not show signs of deterioration, Chi-square/df ratio <3, RMSEA <0.06, while CFI >0.96, and TLI >0.95, but testing of nested model fit with WLSMV estimator suggested that the model
fitted significantly worse than the configural invariance model, DIFFTEST (11) = 22.069, \( p = .024 \). Furthermore, the modification indices indicated the following items: item 2 (firing teachers), item 5 (budget formulation), item 12 (course offered), item 9 (student admittance), and item 6 (budget allocation) were the source of misfit after each time of modifications (MI = 5.364, 4.695, 4.239, 5.896, and 4.013, respectively). After relaxing each of above items one at a time (for detailed procedures, see Table 4), the final partial metric invariance model still did not fit significantly worse than the configural invariance model anymore, DIFFTEST (6) = 7.571, \( p = .271 \) while other fit indices did not show signs of deterioration.

Among those fitting indices, we inclined to regard DIFFTEST from Mplus as most important index, and chi-square and chi-square degrees of freedom as least important index, as in WLSMV estimation the degrees of freedom were estimated values (Schroeders and Wilhelm, 2011). Besides, Cheung and Rensvold (1999) concluded that invariance hypothesis should not be rejected when changes in CFI was smaller than \(-.01\). They did not report critical criteria for TLI and RMSEA (as cited in Vandenberg and Lance, 2000).

### TABLE 4
Model fit indices for multigroup CFAs

<table>
<thead>
<tr>
<th>Model</th>
<th># Free parms.</th>
<th>Chi-square value</th>
<th>Chi-square df</th>
<th>Chi-square p</th>
<th>CFI</th>
<th>TLI</th>
<th>RMSEA estimate</th>
<th>RMSEA CI</th>
<th>RMSEA CI</th>
<th>RMSEA p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Configural</td>
<td>48</td>
<td>177.035</td>
<td>108</td>
<td>&lt;.0001</td>
<td>0.981</td>
<td>0.977</td>
<td>0.054</td>
<td>0.039</td>
<td>0.068</td>
<td>.319</td>
</tr>
<tr>
<td>Metric</td>
<td>37</td>
<td>193.695</td>
<td>119</td>
<td>&lt;.0001</td>
<td>0.98</td>
<td>0.978</td>
<td>0.053</td>
<td>0.039</td>
<td>0.067</td>
<td>.333</td>
</tr>
<tr>
<td>Metric (free item 2)</td>
<td>38</td>
<td>190.758</td>
<td>118</td>
<td>&lt;.0001</td>
<td>0.98</td>
<td>0.978</td>
<td>0.053</td>
<td>0.039</td>
<td>0.066</td>
<td>.355</td>
</tr>
<tr>
<td>Metric (free items 2, 5)</td>
<td>39</td>
<td>188.735</td>
<td>117</td>
<td>&lt;.0001</td>
<td>0.98</td>
<td>0.978</td>
<td>0.053</td>
<td>0.038</td>
<td>0.066</td>
<td>.362</td>
</tr>
<tr>
<td>Metric (free items 2, 5, 12)</td>
<td>40</td>
<td>186.405</td>
<td>116</td>
<td>&lt;.0001</td>
<td>0.98</td>
<td>0.978</td>
<td>0.052</td>
<td>0.038</td>
<td>0.066</td>
<td>.374</td>
</tr>
<tr>
<td>Metric (free items 2, 5, 9)</td>
<td>41</td>
<td>182.04</td>
<td>115</td>
<td>&lt;.0001</td>
<td>0.98</td>
<td>0.979</td>
<td>0.051</td>
<td>0.037</td>
<td>0.065</td>
<td>.442</td>
</tr>
<tr>
<td>Metric (free items 2, 5, 12, 9)</td>
<td>42</td>
<td>180.401</td>
<td>114</td>
<td>&lt;.0001</td>
<td>0.98</td>
<td>0.979</td>
<td>0.051</td>
<td>0.037</td>
<td>0.065</td>
<td>.423</td>
</tr>
<tr>
<td>Scalar</td>
<td>31</td>
<td>197.074</td>
<td>125</td>
<td>&lt;.0001</td>
<td>0.98</td>
<td>0.979</td>
<td>0.051</td>
<td>0.037</td>
<td>0.064</td>
<td>.434</td>
</tr>
<tr>
<td>Scalar (free item 1)</td>
<td>32</td>
<td>194.235</td>
<td>124</td>
<td>&lt;.0001</td>
<td>0.98</td>
<td>0.978</td>
<td>0.051</td>
<td>0.036</td>
<td>0.064</td>
<td>.455</td>
</tr>
<tr>
<td>Scalar (free item 7)</td>
<td>33</td>
<td>191.076</td>
<td>123</td>
<td>&lt;.0001</td>
<td>0.98</td>
<td>0.978</td>
<td>0.05</td>
<td>0.036</td>
<td>0.063</td>
<td>.483</td>
</tr>
<tr>
<td>Residuals free</td>
<td>38</td>
<td>185.248</td>
<td>118</td>
<td>&lt;.0001</td>
<td>0.98</td>
<td>0.978</td>
<td>0.051</td>
<td>0.036</td>
<td>0.064</td>
<td>.448</td>
</tr>
<tr>
<td>Residuals fixed</td>
<td>33</td>
<td>191.076</td>
<td>123</td>
<td>&lt;.0001</td>
<td>0.98</td>
<td>0.978</td>
<td>0.05</td>
<td>0.036</td>
<td>0.063</td>
<td>.483</td>
</tr>
<tr>
<td>Residuals fixed (free item 3)</td>
<td>34</td>
<td>186.955</td>
<td>122</td>
<td>&lt;.0001</td>
<td>0.98</td>
<td>0.978</td>
<td>0.049</td>
<td>0.034</td>
<td>0.063</td>
<td>.528</td>
</tr>
<tr>
<td>Residuals fixed (free items 3, 4)</td>
<td>35</td>
<td>185.965</td>
<td>121</td>
<td>&lt;.0001</td>
<td>0.98</td>
<td>0.981</td>
<td>0.049</td>
<td>0.034</td>
<td>0.063</td>
<td>.519</td>
</tr>
</tbody>
</table>
indicating that, except for items 2, 5, 6, 9, and 12, the same latent factor was measured across Chinese and U.S. principals.

**Scalar invariance**

After establishing partial metric invariance, we proceeded to assess the equality of the unstandardized item thresholds between the Chinese and U.S. samples. We fixed the factor means to 0 and the factor variance to 1 for the U.S. sample, but allow them to be freely estimated for the Chinese sample. Factor loadings and all item thresholds were constrained to be equal across groups whereas residual variances were still constrained to 1 in both groups. As showed in Table 4, the overall model fit of full scalar invariance model did not show signs of deterioration, chi-square/df ratio was smaller than 3, TLI was the same, and RMSEA and CFI were slightly improved, but testing of nested model fit with WLSMV estimator suggested that the model fitted significantly worse than the partial metric invariance model (functioned as the baseline model here), DIFFTEST (11) = 24.104, \( p = .012 \). Furthermore, the modification indices indicated the misfit of item 1 (MI = 4.523) and item 7 (MI = 3.855). After relaxing factor loading of item 1 (hiring teachers) and threshold of item 7 (disciplinary policies), we improved model fit and the partial scalar invariance model did not fit worse than the partial metric invariance model, DIFFTEST (9) = 10.781, \( p = .291 \), and other fit indices did not show signs of deterioration. Establishment of partial scalar invariance implied that we could safely say that except for item 1 and 7, and items relaxed in the matrix invariance model, the factor loadings and intercepts were equal across the United States and China. As we still had a majority of factor loadings equal across group, the comparison of latent factor means could be considered meaningful (Byrne et al. 1989; Reise et al. 1993)—that Chinese principals appeared to have less decision-making power than U.S. principals (−0.74 vs. 0).

**Residual invariance**

Although many suggested that in many cases, the residual invariance was too strict to be practically meaningful (Vandenberg and Lance 2000), we were still interested in establishing residual invariance to avoid potential measurement bias (Schroeders and Wilhelm 2011). Based on the last scalar invariance model, we allowed all residual variance of scalar invariant items to be freely estimated in the Chinese sample (U.S. sample remained fixed at 1). Thus, we compared this model with the last partial scalar invariance model, within which residual variances of Chinese sample were constrained to 1 (U.S. sample remained fixed at 1). As showed in Table 4, compared with the invariance model with freed residual variances, the overall model fit of invariance model with constrained residual variances did not show signs of deterioration, chi-square/df ratio was smaller than 3, CFI and TLI were the same, and RMSEA was slightly improved, but testing of nested model fit with WLSMV estimator suggested that the model fitted significantly worse than the partial scalar invariance model (functioned as the baseline model here), DIFFTEST (5) = 8.573, \( p = .0127 \). Furthermore, the modification indices suggested the misfit of item 3 (Teacher start salaries) (MI = 4.124). After freeing item 3, item 4 (Salary increase) was also found to be the source of misfit (MI = 4.379). Thus, these
two items were both freed in the last model, which did not fit significantly worse than the partial scalar invariance model, DIFFTEST (3) = 1.226, \( p = .747 \), and other fit indices did not show signs of deterioration. No more modification indices were reported. Therefore, we concluded that the partial residual invariance could be held. The direct comparisons of raw scores might be meaningful for the remaining items, which were item 8 (assessment policies), item 10 (textbook selection), and item 11 (course content).

Table 5 presented unstandardized thresholds/factor loadings and standardized thresholds/factor loadings for Chinese and U.S. samples. The table suggested that at the average level (Theta = 0) of both groups, Chinese principals were less likely to be able to make decisions on item 7 (disciplinary policies) than U.S. principals were (−2.063 vs −1.209). Furthermore, item 1 (hiring teachers), item 2 (firing teachers), item 5 (budget formulation), item 6 (budget allocation), item 9 (student admittance), and item 12 (Courses offered) were more discriminating for Chinese principals than for U.S. principals on the primary latent factor of principal decision-making power.

### DISCUSSION

**Summary of Findings**

In this study, we sought to answer (a) whether the factorial structure of principal decision-making power is comparable between American and Chinese principals and (b) who has more decision-making power across the given areas. Although full measurement invariance across both groups was rejected in the initial tests, partial measurement invariance is still established. Since a majority of items (7 of 12) on the given latent construct—principal decision-making power—have factor loadings that are invariant across group, the construct of principal decision-making power can be meaningfully compared and interpreted (see Reise et al. 1993).
The initial comparison of principal power with chi-square tests suggested that U.S. principals appear to have significantly more power in all 12 decision-making areas. However, the initial findings should be interpreted with caution as chi-square test is sensitive to large sample size. Therefore, the comparison was further improved from the measurement perspective.

After the initial comparison with chi-square tests, we proceed with a series of measurement invariance tests between the two samples. To begin with, configural invariance suggested that the factor structure was same across Chinese and U.S. principals. The partial metric invariance indicated that except for the noninvariant items (e.g., firing teachers, budget formulation, budget allocation, student admittance, and courses offered), the factor structure of power was measured in the same way across the two countries. The partial scalar invariance indicated that mean differences of items were the results of factor mean difference except in the area of disciplinary policies. Therefore, item 3 (teacher start salaries), item 4 (teacher salary increase), item 8 (assessment policies), item 10 (textbook selection), and item 11 (course content) were invariant across the two countries in terms of item thresholds/means and factor loadings.

It can be safely concluded that U.S. principals had greater power across all these five domains than Chinese principals had. Further tests of residual variance indicated that the uniqueness of item 3 (teacher start salaries) and item 4 (teacher salary increase) were non-invariant between the two samples.

Findings of the measurement invariance tests have different, but progressive meaning for interpretations of cross-group comparisons. That is, although we are still able to compare factorial structure of principal decision-making power across the two countries with non-invariant items (at least scalar invariance), extra caution is needed given that some items held different factor loadings or threshold across the two samples. In other words, it is not rigorous enough to judge principal’s power with face values of these noninvariant items, as they are more discriminating and difficult for the other group of the principals that have the same true latent factor of decision-making power.

**IMPLICATION**

As the acceleration of globalization, educators around the world are becoming more and more aware of the fact that the growing complexity of the education system is the same challenge we all face, and that the issue is of interest and concern to the whole human race in general. Since their launch, large-scale, prominent international databases such as PISA, Trends in International Mathematics and Science Study, and Teaching and Learning International Survey, have carried a goal of allowing more researchers to collaboratively investigate various education issues from an international perspective. While those international databases have greatly facilitated the producing and verification of new knowledge for the international research communities, it is worth noting that many concepts and constructs surveyed may vary across nations. Therefore, using observed raw scores directly for cross-nation studies may raise issues, as the latent constructs/factors behind the observed raw scores may be different for various samples.
As an example, our present study shows how the construct of principal decision-making power are measured differently with U.S. and Chinese principals. Among the 12 decision-making domains, six items are found to be more discriminating for Chinese respondents than for American respondents, including item 1 (hiring teachers), item 2 (firing teachers), item 5 (budget formulation), item 6 (budget formulation), item 9 (student admittance), item 12 (courses offered). Item 7 (disciplinary policies) is found to be more difficult for American principals. Item 3 (teacher start salaries), item 4 (teacher salary increase), item 8 (assessment policies), item 10 (textbook selection), and item 11 (course content) show strong to strict invariant across the two countries, those items suggest that U.S. principals are more powerful than Chinese principals across these five decision-making domains.

Although it appeared that items selected in the present study were quite different from existing principal comparison studies, some of our findings are consistent with some existing studies. For example, our study echoes Chan and Du’s (2008) findings that U.S. and Chinese principals had great differences in administrative duties. Furthermore, our findings that Chinese principals had limited decision-making power in determining teacher salaries and assessment policies may help explain existing observations that Chinese principals tended to struggle with establishing orders in schools rather than improving student learning (Feng 2008).

Our study has the following implications: First and foremost, our study has implications for educational researchers who want to use international databases to study certain education issues. As we know, successful education and school leadership is highly context-dependent (see Hallinger 2003). One common practice to quantify and explore usefulness of educational theories in a different context is regression analysis within which raw scores from survey items are used. Such a practice may create potential issues in cross-nation studies, as survey respondents from different cultural, political, and professional background may interpreted survey items in conceptually different manners.

Second, our study also has some implication for policy makers. Bush (2009) called for reshaping of principal’s role to meet the trends of (a) the expansion of the role of school principal, (b) the increasing complexity of school contexts, (c) recognition that preparation is a moral obligation, and (d) recognition that effective preparation and development make a difference. Education reform in China already set an aim to distribute more power from central government to lower local levels. Given our findings that Chinese principals still have overall low power in school decision making than their U.S. counterparts, we suggested to give even more power to principals of local schools. It is vital to know that local government may not have as much knowledge about what is happening in schools as principals do, because principals are frontline information gatherers, and school contexts are becoming increasingly complex.

LIMITATION

Our study had the following limitations: First, we would like to caution our readers that the Chinese principal data are not nationally representative. As we mentioned earlier, only four provinces or cities (B-J-S-G) participated in the PISA 2015 survey. Therefore, findings of
this study cannot nationally be generalized. Second, the construct of power in the present study may be relatively simplified as power is a dialectic and dynamic concept by nature. No matter in China or the United States, power structures of the education system are very complex and complicated. Therefore, unless more stakeholders, including states, districts, schools, teachers, etc. are involved in a study, it is hard to tell exactly how power is shared. For example, Shen and Xia’s (2012) study discussed two mechanisms of the power sharing model (win-win vs. zero-sum) between principals and teachers. Gao (2016) revealed that schools’ decision-makings were tightly coupled with districts’ decision-makings, and he discussed top-down control and bottom-up reform models may work in different layers of the school system. Third, we are aware of the long-time debate about whether measurement invariance study should best be done under CFA framework or item response theory framework, which involved many unsettled discussions (see Meade and Lautenschlager 2004; Raju, Laffitte, and Byrne 2002). In the present study, since we analyzed the data under CFA framework, we intentionally avoid using item response theory parameters and its related technical terms to describe our findings. Future studies employing other analytic frameworks are recommended. Finally, in this comparative study, we mainly relied on quantitative data to compare principal power. We highly recommend experienced researchers use qualitative studies to further explore the in-depth reasons for the differences and similarities between Chinese and U.S. principals.

REFERENCES


